

Abdelaati Daouia
Anne Ruiz-Gazen *Editors*

Advances in Contemporary Statistics and Econometrics

Festschrift in Honor of
Christine Thomas-Agnan

 Springer

Advances in Contemporary Statistics and Econometrics



Published with permission from Clara and Didier Gazen

Abdelaati Daouia · Anne Ruiz-Gazen
Editors

Advances in Contemporary Statistics and Econometrics

Festschrift in Honor of
Christine Thomas-Agnan

 Springer

Editors

Abdelaati Daouia
Toulouse School of Economics
University of Toulouse Capitole
Toulouse, France

Anne Ruiz-Gazen
Toulouse School of Economics
University of Toulouse Capitole
Toulouse, France

ISBN 978-3-030-73248-6 ISBN 978-3-030-73249-3 (eBook)
<https://doi.org/10.1007/978-3-030-73249-3>

Mathematics Subject Classification: 62GXX, 62HXX, 62JXX, 62P20

© Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

I first met Christine in the spring of 2002 when I visited GREMAQ at Université Toulouse I, at her invitation to teach a short-course in spatial econometrics for graduate students. June 14 of that year was also the first *Workshop on Spatial Econometrics and Statistics*, organized by Christine, with Noel Cressie giving a keynote presentation. At that workshop I met Cem Ertur, Julie Le Gallo, and Catherine Baumont from Université de Bourgogne in Dijon who hosted the next workshop in May 2003. I also met Olivier Parent, a graduate student at the time, who drove from Strasbourg to Toulouse to attend the conference. In the years following these first two workshops, I had the good fortune to collaborate with many of the French researchers I met as well as their graduate students and to attend seven more of the workshops. Over the years, these workshops have been held in Avignon, Besançon, Dijon, Grenoble, Orleans, Paris, Strasbourg, and Toulon, and have attracted an international audience and several invited speakers from around the world. The success of this workshop has continued with the 19th workshop originally scheduled for May 2020 re-scheduled to May 2021 in Nantes due to the Covid-19 outbreak. This is one important legacy of Christine for those working in the areas of spatial statistics and econometrics.

This book contains five parts that reflect research areas in which Christine has worked over the years. These include nonparametric statistics and econometrics, quantiles and expectiles, spatial statistics and econometrics, compositional data analysis, and tools for empirical studies in economics and applications. So, spatial statistics and econometrics reflects only one area of Christine's past research efforts, but the continuing success of the workshop is a wonderful example of her ability to bring together researchers and promote collaboration.

In terms of collaboration, Christine has worked with nearly 50 co-authors during the last 25 years, on publications appearing in prestigious journals such as *Journal of the American Statistical Association*, *Annals of Statistics*, *Econometric Theory*, *Statistical Papers*, *Journal of Regional Science*, *Numerical Algorithms*, *Statistics & Probability Letters*, *Statistical Methodology*, *Journal of Nonparametric Statistics*, and *Computational Statistics*. Her 2004 book *Reproducing kernel Hilbert spaces in probability and statistics* co-authored with Alain Berlinet has received a great deal of attention in the literature, as has her *Econometric Theory* article: Nonparametric

frontier estimation: a conditional quantile-based approach, co-authored with Yves Aragon and Abdelaati Daouia.

This volume contains numerous contributions, some by those who have collaborated with Christine over the years, and we can all learn from these works. I have collaborated with Christine and found her to be a fabulous person to work with and have benefited from her scholarly expertise and insights. Christine is one of only a handful of scholars whose interests span very theoretical statistical issues as well as applied research that aims at tackling real-world problems. The contributions in this volume reflect that broad range of interests, so there is something for everyone to enjoy. Christine's past work and her ability to promote collaboration among researchers have been an inspiration to us all. Let us hope she continues her research and collaborative efforts for many years to come.

November 2020

James P. LeSage
Fields Endowed Chair
Texas State University
San Marcos, USA

Preface

Christine Thomas-Agnan became Senior Lecturer at Toulouse Capitole University (UTC) in 1988 and Professor in 1994, after completing a doctoral thesis at the University of California, Los Angeles, and a PRAG teaching position at Toulouse Jean Jaurès University. She founded and chaired the group STATISTIQUE—UT1 from 1994, when the faculty of Economics of UTC moved to the building of “Manufacture des Tabacs de Toulouse”. Currently, she heads the mathematics department at UTC.

Over her long and brilliant academic career, Christine Thomas-Agnan has worked on a variety of topics in mathematical and applied statistics, including nonparametric and semi-parametric inference, spatial statistics and econometrics, compositional data analysis, market share regression models, and political economics statistical models. Her work in these areas has found applications in a broad variety of fields, including efficiency measurements of French postal services and Spanish electricity distributors, optimal location of a new fire station in the surroundings of Toulouse, explaining the patterns of regional unemployment and doctors’ prescribing in the Midi-Pyrénées region, testing spatial dependence in air passenger flows, assessing the relations between socioeconomic factors and nutritional diet in Vietnam, and understanding the impact of the composition of media investments on automobile sales in the French automobile market, to cite a few. She has published 5 books and over 50 refereed works in top academic journals. She has been Chief Editor of the Journal CSBIGS (Case Studies in Business, Industry and Government) since 2015, and Member of the publications committee of the French Statistical Society. She has also supervised 12 Ph.D. students and 5 Habilitation degrees (HDR).

Christine Thomas-Agnan is not only a gifted and inspirational researcher and teacher but also a hard-working colleague with a fruitful and curious mind. She has boundless and communicative energy that she puts at the service of the University, her colleagues, co-authors, and students at all levels, especially her Ph.D. students. Her enthusiasm and open-mindedness are greatly appreciated by all. Working with her is an absolute pleasure for us, researchers and teachers in the statistics group, and more generally in the mathematics department, as she facilitated a high-level stimulating environment while maintaining a friendly and inviting demeanor that makes us feel like family.

The task of editing this volume was remarkably easy as the colleagues contacted were so enthusiastic about contributing to this Festschrift by writing and/or editing a research article in Christine's honor. She had dozens of collaborators on an extraordinary variety of research topics. As evidenced by the many tributes in this volume, all colleagues who have had the chance to work with Christine praise her human and scientific qualities.

The 35 articles in this volume are at the frontier of contemporary research in the fields of statistics and econometrics. They testify to Christine's numerous contributions in these fields at both theoretical and applied levels. Christine was first trained as a specialist in reproducing kernel Hilbert space theory and its use in statistical applications. The results she has established since her Ph.D. thesis were published in 2004 in a Springer book jointly with Alain Berlinet. In 1987, she started to explore nonparametric regression by elegantly using spline and kernel smoothing. Then, in 1993, she oriented her research toward functional estimation under form constraints. In the meantime, her intense work on nonparametric and semi-parametric modeling led her to the active fields of quantile/expectile regression and dimension reduction for multivariate response data. In 2002, she began to orient her research toward spatial statistics and econometrics through collaborations she initiated with James P. LeSage and Noel Cressie at the first spatial econometrics workshop she organized at UTC. By adopting the mathematical rigor of statistics and benefiting from the subtlety of econometrics, Christine has first generalized existing models to take into account spatial autocorrelation, and investigated Monte Carlo estimation of Markovian Gaussian fields, before moving to spatial point processes and their use to deal with spatial homogeneity tests, cluster detection, and optimal location-allocation problems. Her efforts have also focused on combining nonparametric methods with spatial statistics to estimate, for instance, autocovariance functions not only of processes but also of random fields, and to study the implications on kriging. Christine's attention was also directed toward the area of frontier and efficiency analysis in production econometrics, with her influential 2005 *Econometric Theory* paper in this literature. From 2011 to 2016, she has been the principal investigator of the interdisciplinary *ModULand* project on the modeling of land use, a prestigious research grant of the French National Research Agency. More recently, she has become interested in compositional data analysis and market share regression models with a particular attention to measuring the impact of covariates in spatial and compositional models. Her recent research allows her to investigate new areas while integrating various interdisciplinary components of her previous research.

Christine's impressive research record should not, however, hide her immense investment in education and services to the community and students. For more than 30 years, she has been heavily involved in the Master program of Econometrics and Statistics at the faculty of Economics of UTC and more recently at Toulouse School of Economics. Among other things, she created the statistical consultancy course of the Master 2 in Statistics and Econometrics more than 20 years ago. This course allows our students to develop their ability to confront concrete statistical problems, posed by companies, under reassuring university supervision. Students were given an invaluable opportunity to experience concrete and exciting projects.

As she nears retirement, one might think that Christine would have less energy or desire to invest in new areas or experience new things in her work, but nothing could be further from the truth. In addition to assuming responsibility for the Master 1 in Econometrics and Statistics at the Toulouse School of Economics, she has accepted the direction of the mathematics department, which she manages with great initiative and tact. She has also very recently accepted to supervise new doctoral students, namely Lukas Dargel and Thibault Laurent, in two stimulating research programs with applications in social sciences.

We would like to thank Christine Thomas-Agnan for being such an inspiring figure in our professional and personal lives. We join our colleague James P. LeSage in the hope that she will continue her excellent work for many years to come.

The five parts of this volume correspond to the topics that Christine has contributed much to. The contributions collected in each section answer important questions that reflect varied theoretical and/or applied interests of their authors. They provide nice examples of the new research ideas that are currently being developed. We expect that everyone will find something interesting in this rich collection of papers.

The first part contains seven papers related to the active area of nonparametric statistics and econometrics. Fadoua Balabdaoui and Piet Groeneboom elucidate the open question of whether a profile least squares estimator in the monotone single index model is \sqrt{n} convergent and asymptotically normal. Gérard Biau and Benoît Cadre present a general framework for studying two widespread gradient boosting algorithms from the perspective of functional optimization, and address the less-discussed problem of their convergence as the number of iterations tends to infinity. Sandrine Casanova and Eve Leconte introduce a novel nonparametric model-based estimator for the conditional distribution function of a right censored response, which is superior to its most known competitors in small domains. Eric Gautier suggests endogenous selection models, which allow for instrument nonmonotonicity and are based on nonparametric random coefficient indices. Camelia Goga gives a review of applications of B-spline regression in a survey sampling framework and design-based approach, including new properties of the (un)penalized estimators, and their improved consistency rates. Hadrien Lorenzo and Jérôme Saracco propose three computational devices to detect outliers in a single index regression model, when conducting sliced inverse regression along with kernel smoothing of the link function. Jan Meis and Enno Mammen revisit the uncoupled isotonic regression problem by improving the rate of convergence of the so-called minimum Wasserstein deconvolution estimator, for L_p -risks and for error distributions supported on a finite set of points.

The second part also contains seven contributions that are dedicated to the topic of (un)conditional quantiles and expectiles. The class of expectiles corresponds to a least squares analogue of quantiles. Cécile Adam and Irène Gijbels study multivariate partially linear expectile regression in which the nonlinear part is fitted using a local polynomial approach, along with an optimal choice of the bandwidth parameter. Delphine Blanke and Denis Bosq prove that, for estimating univariate quantiles, the reciprocal of the piecewise linear interpolation at the midpoints of a sample distribution function strictly improves the MISE of the usual sample quantile function. Axel

Bücher, Anouar El Ghouh, and Ingrid Van Keilegom propose a valid local linear smoothing approach to iteratively estimate a semi-parametric single-index model for conditional quantiles with right-censored data. Stéphane Girard, Gilles Stupfler, and Antoine Usseglio-Carleve construct kernel estimators of extreme regression L_p -quantiles, which encompass both families of expectiles and standard quantiles, and develop their asymptotic theory for heavy-tailed conditional distributions. Bao Hoang Nguyen and Valentin Zelenyuk perform a robust frontier and efficiency analysis of public hospitals in Queensland, Australia, by estimating both individual and aggregate quantile-based efficiency scores. Davy Paindaveine and Joni Virta unravel the behavior of extreme d -dimensional spatial quantiles under minimal conditions, in a general setup for both population and sample multivariate distributions. Fabian Otto-Sobotka, Radoslava Mirkov, Benjamin Hofner, and Thomas Kneib use shape-constrained expectile regression in conjunction with a geoadditive model to provide deeper insights into the behavior of gas flow within transmission networks.

The third part concerns spatial statistics and econometrics with eight contributions. François Bachoc provides a review of the asymptotic theory for maximum likelihood estimation of covariance parameters for Gaussian processes, under increasing and fixed-domain asymptotics. Florent Bonneu and Lionel Cucala adapt spatial scan methods, borrowed from local cluster detection, to test for global similarity between two spatial point patterns. Hervé Cardot and Antonio Musolesi rely on the use of additive models and conditional mixtures and on random forests to estimate the variation along time of the spillover effects of spatial policies. Raja Chakir and Julie Le Gallo review the current state of the literature on studies which account for spatial autocorrelation in econometric land use models or in the environmental impacts of land use. Noel Cressie and Christopher Wikle develop a modern hierarchical statistical approach to modeling spatio-temporal data on regular or irregular spatial lattices. Van Huyen Do, Thibault Laurent, and Anne Vanhems implement widely used methods in the areal interpolation problem using R software, and provide practical guidelines to concrete questions such as spatial scales, types of target variable, and border incompatibility. Thibault Laurent and Paula Margaretic apply prediction of spatial econometric models for areal data to model regional unemployment rates taking into account local interactions. Mary Lai Salvaña and Marc Genton propose a new estimation methodology for nonstationary covariance models of the Lagrangian type, by modeling the second-order nonstationarity parameters via thin plate splines and estimating all the parameters via two-step maximum likelihood estimation.

The fourth part contains six papers on the area of compositional data analysis that Christine has also contributed to over the last years. Peter Filzmoser, Karel Hron, and Alessandra Menafoglio present and discuss a log-ratio approach to distributional modeling in a unifying framework for the discrete and the continuous distributional data based on the theory of Bayes spaces. Built on ideas from the spatial Durbin model, Tingting Huang, Gilbert Saporta, and Huiwen Wang propose and estimate a new compositional linear model for areal data by employing the orthonormal log-ratio transformation and maximum likelihood method. Wilfredo Maldonado, Juan José Egozcue, and Vera Pawlowsky-Glahn contribute to the modeling and compositional analysis of exchange rate matrices and the corresponding no-arbitrage matrices, by

considering the Special Drawing Rights and by studying the relative exchange rate bubbles among the countries. Josep Antoni Martín-Fernández and Carles Barceló-Vidal revisit the basic concepts and properties of log ratios, log contrasts, and orthonormal coordinates for compositional data, and introduce a new approach that includes both the log-ratio orthonormal coordinates and an auxiliary variable carrying absolute information. Christoph Muehlmann, Kamila Fačevicová, Alžběta Gardlo, Hana Janečková, and Klaus Nordhausen review some basic methods of independent component analysis and show how to apply such analysis to compositional data. Michel Simioni, Huong Thi Trinh, Tuyen Thi Thanh Huynh, and Thao-Vy Vuong explore the association between food sources and diet quality in Vietnam by making use of recent advances in compositional data analysis.

The seven contributions collected in the last part provide useful tools for empirical studies in economics and applied work. Bastien Bernela, Liliame Bonnal, and Pascal Favard untangle the empirical reality of the phenomenon of geographical mobility among students and young graduates in France. Christophe Bontemps and Valérie Orozco show how the research process, from data collection to paper publication, could efficiently be reorganized to improve and promote reproducible research. Olivier de Mouzon, Thibault Laurent, and Michel Le Breton explore and estimate the departure from the “One Man, One Vote” principle in the context of political representation and its consequences for distributive politics. They also provide several applications of the Lorenz curve and the Gini and Dauer-Kelsay indices to the measurement of malapportionment and disproportionality. Jonathan Haughton and Dominique Haughton recommend and illustrate the use of cartograms as an effective complement to the more-traditional choropleth maps for conveying spatially distributed statistical data. Jérôme Mariette, Madalina Olteanu, and Nathalie Vialaneix present kernel and dissimilarity methods to perform exploratory analysis in the presence of multiple sources of data or of multiple kernels describing different features of the data. Finally, Alban Thomas develops and applies a generalized method of particle nonlinear filtering to estimate a system of structural equations for agricultural crop yield functions, when unobserved productivity depends on water availability that is only partially observed.

Toulouse, France
January 2021

Abdelaati Daouia
Anne Ruiz-Gazen

Acknowledgements

All contributions are peer-reviewed, and we would like to thank all the referees very warmly for their very careful work.

Yasser Abbas	Serge Garcia	Joanna Morais
Yves Aragon	Eric Gautier	Christoph Muhlmann
Cécile Adam	Marc G. Genton	Klaus Nordhausen
Denis Allard	Irène Gijbels	An Nguyen Huong
François Bachoc	David Ginsbourger	Madalina Olteanu
Fadoua Balabdaoui	Stéphane Girard	Jean Opsomer
Liliane Bel	Antoine Godichon-Baggioni	Fabian Otto-Sobotka
Philippe Besse	Michel Goulard	Davy Paindaveine
Roger Bivand	Dominique Haughton	Javier Palarea
Delphine Blanke	Xavier d'Haultfoeuille	Olivier Parent
Liliane Bonnal	David Haziza	Valentin Patilea
Florent Bonneu	Karel Hron	Vera Pawlowsky-Glahn
Christophe Bontemps	Thomas Kneib	Gilbert Saporta
Sandrine Casanova	Thibault Laurent	Jérôme Saracco
Hervé Cardot	Pascal Lavergne	Michel Simioni
Raja Chakir	Eve Leconte	Gilles Stupfler
Noel Cressie	Julie Le Gallo	Alban Thomas
Lionel Cucala	Erwan Le Pennec	Anne Vanhems
Lukas Dargel	Rik Lopuhaä	Karine Van der Straten
Philippe de Donder	Enno Mammen	Ingrid Van Keilegom
Juan José Egozcue	Paula Margaretic	Nathalie Vialaneix
Peter Filzmoser	Josep Antoni Martín-Fernández	Joni Virta
Edith Gabriel	Jorge Mateu	Valentin Zelenyuk
Sébastien Gadat		

Contents

Nonparametric Statistics and Econometrics

Profile Least Squares Estimators in the Monotone Single Index Model	3
Fadoua Balabdaoui and Piet Groeneboom	
Optimization by Gradient Boosting	23
G�rard Biau and Beno�t Cadre	
Nonparametric Model-Based Estimators for the Cumulative Distribution Function of a Right Censored Variable in a Small Area	45
Sandrine Casanova and Eve Leconte	
Relaxing Monotonicity in Endogenous Selection Models and Application to Surveys	59
Eric Gautier	
B-Spline Estimation in a Survey Sampling Framework	79
Camelia Goga	
Computational Outlier Detection Methods in Sliced Inverse Regression	101
Hadrien Lorenzo and J�r�me Saracco	
Uncoupled Isotonic Regression with Discrete Errors	123
Jan Meis and Enno Mammen	
Quantiles and Expectiles	
Partially Linear Expectile Regression Using Local Polynomial Fitting	139
C�cile Adam and Ir�ne Gijbels	
Piecewise Linear Continuous Estimators of the Quantile Function	161
Delphine Blanke and Denis Bosq	

Single-Index Quantile Regression Models for Censored Data	177
Axel Bücher, Anouar El Ghouch, and Ingrid Van Keilegom	
Extreme L^p-quantile Kernel Regression	197
Stéphane Girard, Gilles Stupfler, and Antoine Usseglio-Carleve	
Robust Efficiency Analysis of Public Hospitals in Queensland, Australia	221
Bao Hoang Nguyen and Valentin Zelenyuk	
On the Behavior of Extreme d-dimensional Spatial Quantiles Under Minimal Assumptions	243
Davy Paindaveine and Joni Virta	
Modelling Flow in Gas Transmission Networks Using Shape-Constrained Expectile Regression	261
Fabian Otto-Sobotka, Radoslava Mirkov, Benjamin Hofner, and Thomas Kneib	
Spatial Statistics and Econometrics	
Asymptotic Analysis of Maximum Likelihood Estimation of Covariance Parameters for Gaussian Processes: An Introduction with Proofs	283
François Bachoc	
Global Scan Methods for Comparing Two Spatial Point Processes	305
Florent Bonneu and Lionel Cucala	
Assessing Spillover Effects of Spatial Policies with Semiparametric Zero-Inflated Models and Random Forests	319
Hervé Cardot and Antonio Musolesi	
Spatial Autocorrelation in Econometric Land Use Models: An Overview	339
Raja Chakir and Julie Le Gallo	
Modeling Dependence in Spatio-Temporal Econometrics	363
Noel Cressie and Christopher K. Wikle	
Guidelines on Areal Interpolation Methods	385
Van Huyen Do, Thibault Laurent, and Anne Vanhems	
Predictions in Spatial Econometric Models: Application to Unemployment Data	409
Thibault Laurent and Paula Margaretic	
Lagrangian Spatio-Temporal Nonstationary Covariance Functions	427
Mary Lai O. Salvaña and Marc G. Genton	

Compositional Data Analysis

Logratio Approach to Distributional Modeling 451
Peter Filzmoser, Karel Hron, and Alessandra Menafoglio

A Spatial Durbin Model for Compositional Data 471
Tingting Huang, Gilbert Saporta, and Huiwen Wang

Compositional Analysis of Exchange Rates 489
Wilfredo L. Maldonado, Juan José Egozcue, and Vera Pawlowsky-Glahn

**Log-contrast and Orthonormal Log-ratio Coordinates
for Compositional Data with a Total** 509
Josep Antoni Martín-Fernández and Carles Barceló-Vidal

Independent Component Analysis for Compositional Data 525
Christoph Muehlmann, Kamila Fačevicová, Alžběta Gardlo,
Hana Janečková, and Klaus Nordhausen

**Diet Quality and Food Sources in Vietnam: First Evidence Using
Compositional Data Analysis** 547
Michel Simioni, Huong Thi Trinh, Tuyen Thi Thanh Huynh,
and Thao-Vy Vuong

Tools for Empirical Studies in Economics and Social Sciences

**Mobility for Study and Professional Integration: An Empirical
Overview of the Situation in France Based on the CÉREQ
generational surveys** 573
Bastien Bernela, Liliane Bonnal, and Pascal Favard

Toward a FAIR Reproducible Research 595
Christophe Bontemps and Valérie Orozco

**“One Man, One Vote” Part 2: Measurement of Malapportionment
and Disproportionality and the Lorenz Curve
A: Introduction and Measurement Tools** 615
Olivier de Mouzon, Thibault Laurent, and Michel Le Breton

**“One Man, One Vote” Part 2: Measurement of Malapportionment
and Disproportionality and the Lorenz Curve
B: Applications** 633
Olivier de Mouzon, Thibault Laurent, and Michel Le Breton

Visualizing France with Cartograms 653
Jonathan Haughton and Dominique Haughton

**Kernel and Dissimilarity Methods for Exploratory Analysis
in a Social Context** 669
Jérôme Mariette, Madalina Olteanu, and Nathalie Vialaneix

**Of Particles and Molecules: Application of Particle Filtering
to Irrigated Agriculture in Punjab, India 691**
Alban Thomas

Contributors

Cécile Adam Department of Mathematics, KU Leuven, Leuven (Heverlee), Belgium

François Bachoc Institut de Mathématiques, UMR; Université de Toulouse; CNRS, UPS IMT, Toulouse Cedex, France

Fadoua Balabdaoui Seminar für Statistik, ETH, Zürich, Switzerland

Carles Barceló-Vidal Department of IMAE, University of Girona, Girona, Spain

Bastien Bernela Université de Poitiers CRIEF, UFR de sciences économiques: rue Jean Carbonnier TSA, Poitiers cedex, France

Gérard Biau Sorbonne Université, CNRS, LPSM, Paris, France

Delphine Blanke Laboratoire de Mathématiques d'Avignon, LMA, Avignon Université, Avignon, France

Liliane Bonnal Université de Poitiers CRIEF and TSE, UFR de sciences économiques: rue Jean Carbonnier TSA, Poitiers cedex, France

Florent Bonneu Avignon Université, Avignon, France

Christophe Bontemps Toulouse School of Economics - INRAE, University of Toulouse Capitole, Toulouse, France

Denis Bosq Laboratoire de Probabilités, Statistique et Modélisation, LPSM, CNRS, Sorbonne Universités, Paris, France

Axel Bücher Mathematisches Institut, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany

Benoît Cadre Université Rennes, CNRS, IRMAR - UMR, Rennes, France

Hervé Cardot Université de Bourgogne, Institut de Mathématiques de Bourgogne, Dijon, France

Sandrine Casanova TSE-R, Université Toulouse Capitole, Toulouse cedex, France

Raja Chakir Université Paris-Saclay, INRAE, AgroParisTech, Economie Publique, France

Noel Cressie University of Wollongong, Wollongong, Australia

Lionel Cucala Université de Montpellier, Montpellier, France

Olivier de Mouzon Toulouse School of Economics, INRAE, University of Toulouse Capitole, Toulouse, France

Van Huyen Do Toulouse School of Economics, CNRS, University of Toulouse, Toulouse, France

Juan José Egozcue Technical University of Catalonia, Department of Civil and Environmental Engineering, Barcelona, Spain

Kamila Fačevicová Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, Olomouc, Czech Republic

Pascal Favard Université de Tours IRJI, Department of Economics, Tours Cedex, France

Peter Filzmoser Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Alžběta Gardlo Department of Clinical Biochemistry, University Hospital Olomouc and Palacký University Olomouc, Olomouc, Czech Republic

Eric Gautier TSE, Université Toulouse Capitole, Esplanade de l'Université, Toulouse, France

Marc G. Genton King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Anouar El Ghouch ISBA, UCLouvain, Louvain-la-Neuve, Belgium

Irène Gijbels Department of Mathematics and Leuven Statistics Research Center (LStat), Leuven (Heverlee), Belgium

Stéphane Girard University of Grenoble-Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France

Camelia Goga Laboratoire de Mathématiques de Besançon, Université de Bourgogne Franche-Comté, Besançon, France

Piet Groeneboom Delft University of Technology, DIAM, CD Delft, The Netherlands

Dominique Haughton Department of Mathematical Sciences, Bentley University, Waltham, MA, USA;
Université Paris I (SAMM), Paris, France;
Université Toulouse I (TSE-R), Toulouse, France

Jonathan Haughton Department of Economics, Suffolk University, Boston, MA, USA

Benjamin Hofner Section Biostatistics, Paul-Ehrlich-Institut, Langen, Germany

Karel Hron Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, Czech Republic

Tingting Huang School of Statistics, Capital University of Economics and Business, Beijing, China;

School of Economics and Management, Beihang University, Beijing, China;

Beijing Key Laboratory of Emergence Support Simulation Technologies for City Operations, Beijing, China

Tuyen Thi Thanh Huynh International Center for Tropical Agriculture (CIAT)—Asia Office, Hanoi, Vietnam

Hana Janečková Laboratory for Inherited Metabolic Disorders, Department of Clinical Biochemistry, University Hospital Olomouc and Palacký University Olomouc, Olomouc, Czech Republic

Thomas Kneib Department of Economics, Georg-August-Universität Göttingen, Göttingen, Germany

Thibault Laurent Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France

Michel Le Breton Institut Universitaire de France and Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France

Julie Le Gallo CESAER UMR, Agrosup Dijon, INRAE, Université de Bourgogne Franche-Comté, Dijon, France

Eve Leconte TSE-R, Université Toulouse Capitole, Toulouse cedex, France

Hadrien Lorenzo Inria BSO, Talence, France

Wilfredo L. Maldonado Faculty of Economics, Management and Accounting, University of São Paulo. Av. Professor Luciano Gualberto, São Paulo, Brazil

Enno Mammen Institute of Applied Mathematics, Heidelberg University, Heidelberg, Germany

Paula Margaretic University of San Andrés, Victoria, Argentina

Jérôme Mariette Université de Toulouse, INRAE, UR MIAT, Castanet-Tolosan, France

Josep Antoni Martín-Fernández Department of IMAE, University of Girona, Girona, Spain

Jan Meis Institute of Medical Biometry and Informatics, Heidelberg University, Heidelberg, Germany

Alessandra Menafoglio MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy

Radoslava Mirkov Department of Mathematics, Humboldt University Berlin, Berlin, Germany

Christoph Muehlmann Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Antonio Musolesi Department of Economics and Management and SEEDS, Ferrara University, Ferrara, Italy

Bao Hoang Nguyen School of Economics, University of Queensland, Brisbane, QLD, Australia

Klaus Nordhausen Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria

Madalina Olteanu SAMM, Université Paris, Paris, France

Valérie Orozco Toulouse School of Economics - INRAE, University of Toulouse Capitole, Toulouse, France

Fabian Otto-Sobotka Division of Epidemiology and Biometry, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

Davy Paindaveine Université libre de Bruxelles (ECARES and Department of Mathematics) and Université Toulouse Capitole (Toulouse School of Economics), Brussels, Belgium

Vera Pawlowsky-Glahn University of Girona, Department of Computer Science, Applied Mathematics and Statistics, Girona, Spain

Mary Lai O. Salvaña King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Gilbert Saporta CNAM, Center for Studies and Research in Computer Science and Communication, Paris, France

Jérôme Saracco Inria BSO & ENSC Bordeaux INP, Talence, France

Michel Simioni MOISA, INRAE, University of Montpellier, Montpellier, France

Gilles Stupfler University of Rennes, Ensai, CNRS, CREST - UMR, Rennes, France

Alban Thomas Toulouse School of Economics-Research, INRAE, University of Toulouse, Toulouse, France

Huong Thi Trinh Department of Mathematics and Statistics, Thuongmai University, Hanoi, Vietnam

Antoine Usseglio-Carleve University of Grenoble-Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France

Ingrid Van Keilegom ORSTAT, KU Leuven, Leuven, Belgium

Anne Vanhems TBS Business School, Toulouse, France

Nathalie Vialaneix Université de Toulouse, INRAE, UR MIAT, Castanet-Tolosan, France

Joni Virta University of Turku (Department of Mathematics and Statistics) and Aalto University School of Science (Department of Mathematics and Systems Analysis), Turun yliopisto, Finland

Thao-Vy Vuong College of Agriculture and Life Sciences, Cornell University, Ithaca, USA

Huiwen Wang School of Economics and Management, Beihang University, Beijing, China;
Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China

Christopher K. Wikle University of Missouri, Columbia, USA

Valentin Zelenyuk School of Economics and Centre for Efficiency and Productivity Analysis, University of Queensland, Brisbane, QLD, Australia

Nonparametric Statistics and Econometrics

Profile Least Squares Estimators in the Monotone Single Index Model



Fadoua Balabdaoui and Piet Groeneboom

Abstract We consider least squares estimators of the finite regression parameter α in the single index regression model $Y = \psi(\alpha^T X) + \varepsilon$, where X is a d -dimensional random vector, $\mathbb{E}(Y|X) = \psi(\alpha^T X)$, and ψ is a monotone. It has been suggested to estimate α by a profile least squares estimator, minimizing $\sum_{i=1}^n (Y_i - \psi(\alpha^T X_i))^2$ over monotone ψ and α on the boundary \mathcal{S}_{d-1} of the unit ball. Although this suggestion has been around for a long time, it is still unknown whether the estimate is \sqrt{n} -convergent. We show that a profile least squares estimator, using the same pointwise least squares estimator for fixed α , but using a different global sum of squares, is \sqrt{n} -convergent and asymptotically normal. The difference between the corresponding loss functions is studied and also a comparison with other methods is given.

1 Introduction

The monotone single index model tries to predict a response from the linear combination of a finite number of parameters and a function linking this linear combination to the response via a monotone *link function* ψ_0 which is unknown. So, more formally, we have the model

$$Y = \psi_0(\alpha_0^T X) + \varepsilon,$$

where Y is a one-dimensional random variable, $X = (X_1, \dots, X_d)^T$ is a d -dimensional random vector with distribution function G , ψ_0 is monotone, and ε is a one-dimensional random variable such that $\mathbb{E}[\varepsilon|X] = 0$ G almost surely. For identifiability, the regression parameter α_0 is a vector of norm $\|\alpha_0\|_2 = 1$, where

F. Balabdaoui (✉)

Seminar für Statistik ETH, Zürich, Switzerland
e-mail: fadoua.balabdaoui@stat.math.ethz.ch

P. Groeneboom

Delft University of Technology, DIAM, Mekelweg 4, 2628 CD Delft, The Netherlands
e-mail: P.Groeneboom@tudelft.nl

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_1

$\|\cdot\|_2$ denotes the Euclidean norm in \mathbb{R}^d , so $\alpha_0 \in \mathcal{S}_{d-1}$, the unit $(d-1)$ -dimensional sphere.

The ordinary profile least squares estimate of α_0 is an M -estimate in two senses: for fixed α , the least squares criterion

$$\psi \mapsto n^{-1} \sum_{i=1}^n \{Y_i - \psi(\alpha^T X_i)\}^2 \quad (1)$$

is minimized for all monotone functions ψ (either decreasing or increasing) which gives an α -dependent function $\hat{\psi}_{n,\alpha}$, and the function

$$\alpha \mapsto n^{-1} \sum_{i=1}^n \{Y_i - \hat{\psi}_{n,\alpha}(\alpha^T X_i)\}^2 \quad (2)$$

is then minimized over α . This gives a profile least squares estimator $\hat{\alpha}_n$ of α_0 , which we will call LSE in the sequel. Although this estimate of α_0 has been known now for a very long time (more than 30 years probably), it is not known whether it is \sqrt{n} -convergent (under appropriate regularity conditions), let alone that we know its asymptotic distribution. Also, simulation studies are rather inconclusive. For example, it is conjectured in Tanaka (2008) on the basis of simulations that the rate of convergence of $\hat{\alpha}_n$ is $n^{9/20}$. Other simulation studies, presented in Balabdaoui et al. (2019a), are also inconclusive. In that paper, it was also proved that an ordinary least squares estimator (which ignores that the link function could be non-linear) is \sqrt{n} -convergent and asymptotically normal under elliptic symmetry of the distribution of the covariate X . Another linear least squares estimator of this type, where the restriction on α is $\alpha^T S_n \alpha = 1$, S_n is the usual estimate of the covariance matrix of the covariates, and a renormalization at the end is not needed (as it is in the just mentioned linear least squares estimator) was studied in Balabdaoui et al. (2019b) and was shown to have similar behavior. If this suggests that the profile LSE should also be \sqrt{n} -consistent, the extended simulation study in Balabdaoui et al. (2019b) shows that it is possible to find other estimates which exhibit better performance in these circumstances.

An alternative way to estimate the regression vector is to minimize the criterion

$$\alpha \mapsto \left\| n^{-1} \sum_{i=1}^n \{Y_i - \hat{\psi}_{n,\alpha}(\alpha^T X_i)\} X_i \right\|^2 \quad (3)$$

over $\alpha \in \mathcal{S}_{d-1}$, where $\|\cdot\|$ is the Euclidean norm. Note that this is the sum of d squares. The rational behind minimizing (3) is the fact that the true index vector, α_0 , satisfies the (population) score equation

$$\mathbb{E} \{(Y - \psi_0(\alpha_0^T X)) X \theta(\alpha_0^T X)\} = \mathbf{0}, \quad (4)$$

where θ is any measurable and bounded function. This clearly follows from the iterative law of expectations and the fact that $\mathbb{E}\{Y|\alpha_0^T X\} = \psi_0(\alpha_0^T X)$. If the function θ is taken to be the constant 1, then the goal is to find the minimizer of the Euclidean norm of the empirical counterpart of the above score equation, after replacing the unknown link function, ψ_0 , by its estimator $\hat{\psi}_{n,\alpha}$.

We prove in Sect. 3 that this minimization procedure leads to a \sqrt{n} -consistent and asymptotically normal estimator, which is a more precise and informative result compared to what we know now about the LSE. Using the well-known properties of isotonic estimators, it is easily seen that the function (3) is piecewise constant as a function of α , with finitely many values, so the minimum exists and is equal to the infimum over $\alpha \in \mathcal{S}_{d-1}$. Notice that this estimator does not use any tuning parameters, just like the LSE.

In Balabdaoui et al. (2019b), a similar Simple Score Estimator (SSE) $\hat{\alpha}_n$ was defined as a point $\alpha \in \mathcal{S}_{d-1}$ where all components of the function

$$\alpha \mapsto n^{-1} \sum_{i=1}^n \left\{ Y_i - \hat{\psi}_{n,\alpha}(\alpha^T X_i) \right\} X_i$$

cross zero. If the criterion function were continuous in α , this estimator would have been the same as the least squares estimator, minimizing (3), with a minimum equal to zero, but in the present case we cannot assume this because of the discontinuities of the criterion function.

The definition of an estimator as a crossing of the d -dimensional vector $\mathbf{0}$ makes it necessary to prove the existence of such an estimator, which we found to be a rather non-trivial task. Defining our estimator directly as the minimizer of (3), so as a least squares estimator, relieves us from the duty to prove its existence. Since our estimator has the same limit distribution as the SSE, we refer to it here under the same name.

A fundamental function in our treatment is the function ψ_α , defined as follows.

Definition 1 Let \mathcal{S}_{d-1} denote again the boundary of the unit ball in \mathbb{R}^d . Then, for each $\alpha \in \mathcal{S}_{d-1}$, the function $\psi_\alpha : \mathbb{R} \rightarrow \mathbb{R}$ is defined as the nondecreasing function which minimizes

$$\psi \mapsto \mathbb{E}\{Y - \psi(\alpha^T X)\}^2$$

over all nondecreasing functions $\psi : \mathbb{R} \rightarrow \mathbb{R}$. The existence and uniqueness of the function ψ_α follows, for example, from the results in Landers and Rogge (1981).

The function ψ_α coincides in a neighborhood of α_0 with the ordinary conditional expectation function $\tilde{\psi}_\alpha$

$$\tilde{\psi}_\alpha(u) = \mathbb{E}\{\psi_0(\alpha_0^T X) | \alpha^T X = u\}, \quad u \in \mathbb{R}; \quad (5)$$

see Balabdaoui et al. (2019b), Proposition 1. The general definition of ψ_α uses conditioning on a σ -lattice, and ψ_α is also called a *conditional 2-mean* (see Landers and Rogge 1981).

The importance of the function ψ_α arises from the fact that we can differentiate this function w.r.t. α , in contrast with the least squares estimate $\hat{\psi}_{n,\alpha}$, and that ψ_α represents the least squares estimate of ψ_0 in the underlying model for fixed α , if we use $\alpha^T \mathbf{x}$ as the argument of the monotone link function.

It is also possible to introduce a tuning parameter and use an estimate of $\frac{d}{du} \psi_\alpha(u) \Big|_{u=\alpha^T X}$. This estimate is defined by

$$\tilde{\psi}'_{n,h,\alpha}(u) = \frac{1}{h} \int K\left(\frac{u-x}{h}\right) d\hat{\psi}_{n,\alpha}(x), \quad (6)$$

where K is one of the usual kernels, symmetric around zero and with support $[-1, 1]$, and h is a bandwidth of order $n^{-1/7}$ for sample size n . For fixed α , the least squares estimate $\hat{\psi}_{n,\alpha}$ is defined in the same way as above. Note that this estimate is rather different from the derivative of a Nadaraya-Watson estimate which is also used in this context and is in fact the derivative of a ratio of two kernel estimates. If we use the Nadaraya-Watson estimate, we need in principle two tuning parameters, one for the estimation of ψ_0 and another one for the estimation of the derivative ψ'_0 .

Using the estimate (6) of the derivative, we now minimize

$$\alpha \mapsto \left\| n^{-1} \sum_{i=1}^n \left\{ Y_i - \hat{\psi}_{n,\alpha}(\alpha^T X_i) \right\} X_i \tilde{\psi}'_{n,h,\alpha}(\alpha^T X_i) \right\|^2 \quad (7)$$

instead of (3), where $\|\cdot\|$ is again the Euclidean norm. The motivation for considering such a minimization problem is very similar to the one given above for the SSE. The only difference now is that the current approach allows us to take the function θ to be equal to the derivative of ψ'_0 , which is replaced in the empirical version of the population score in (4) by its estimator $\tilde{\psi}'_{n,h,\alpha}$. A variant of this estimator was defined in Balabdaoui et al. (2019b) and called the Efficient Score Estimator (ESE) there, since, if the conditional variance $\text{var}(Y|X = \mathbf{x}) = \sigma^2$, where σ^2 is independent of the covariate \mathbf{X} (the homoscedastic model), the estimate is efficient. As in the case of the simple score estimator (SSE), the estimate was defined as a crossing of zero estimate in Balabdaoui et al. (2019b) and not as a minimizer of (7). But the definition as a minimizer of (7) produces an estimator that has the same limit distribution.

The qualification “efficient” is somewhat dubious, since the estimator is no longer efficient if we do not have homoscedasticity. We give an example of that situation in Sect. 5, where, in fact, the SSE has a smaller asymptotic variance than the ESE. Nevertheless, to be consistent with our treatment in Balabdaoui et al. (2019b) we will call the estimate, $\hat{\alpha}_n$, minimizing (7), again the ESE.

Dropping the monotonicity constraint, we can also use as our estimator of the link function a cubic spline $\hat{\psi}_{n,\alpha}$, which is defined as the function minimizing

$$\sum_{i=1}^n \{\psi(\boldsymbol{\alpha}^T \mathbf{X}_i) - Y_i\}^2 + \mu \int_a^b \psi''(x)^2 dx, \quad (8)$$

over the class of functions $\mathcal{S}_2[a, b]$ of differentiable functions ψ with an absolutely continuous first derivative, where

$$a = \min_i \boldsymbol{\alpha}^T \mathbf{X}_i, \quad b = \max_i \boldsymbol{\alpha}^T \mathbf{X}_i,$$

see Green and Silverman (1994), pp. 18 and 19, where $\mu > 0$ is the penalty parameter. Using these estimators of the link function, the estimate $\hat{\boldsymbol{\alpha}}_n$ of $\boldsymbol{\alpha}_0$ is then found in Kuchibhotla and Patra (2020) by using a $(d - 1)$ -dimensional parameterization $\boldsymbol{\beta}$ and a transformation $S : \boldsymbol{\beta} \mapsto S(\boldsymbol{\beta}) = \boldsymbol{\alpha}$, where $S(\boldsymbol{\beta})$ belongs to the surface of the unit sphere in \mathbb{R}^d , and minimizing the criterion

$$\boldsymbol{\beta} \mapsto \sum_{i=1}^n \{Y_i - \hat{\psi}_{S(\boldsymbol{\beta}), \mu}(S(\boldsymbol{\beta})^T \mathbf{X}_i)\}^2,$$

over $\boldsymbol{\beta}$, where $\hat{\psi}_{S(\boldsymbol{\beta}), \mu}$ minimizes (8) for fixed $\boldsymbol{\alpha} = S(\boldsymbol{\beta})$.

Analogous to our approach above, we can skip the reparameterization and minimize instead

$$\left\| \frac{1}{n} \sum_{i=1}^n \{\hat{\psi}_{n, \boldsymbol{\alpha}, \mu}(\boldsymbol{\alpha}^T \mathbf{X}_i) - Y_i\} \mathbf{X}_i \tilde{\psi}'_{n, \boldsymbol{\alpha}, \mu}(u) \Big|_{u=\boldsymbol{\alpha}^T \mathbf{X}_i} \right\| \quad (9)$$

where $\tilde{\psi}_{n, \boldsymbol{\alpha}, \mu}$ minimizes (8) for fixed $\boldsymbol{\alpha}$ and $\tilde{\psi}'_{n, \boldsymbol{\alpha}, \mu}$ is its derivative. We call this estimator the spline estimator.

We finally give simulation results for these different methods in Sect. 5, where, apart from the comparison with the spline estimator, we make a comparison with other estimators of $\boldsymbol{\alpha}_0$ not using the monotonicity constraint: the Effective Dimension Reduction (EDR) method, proposed in Hristache et al. (2001) and implemented in the R package `edr`, the (refined) Mean Average conditional Variance Estimator (MAVE) method, discussed in Xia (2006), and implemented in the R package `MAVE`, and Estimation Function Method (EFM), discussed in Cui et al. (2011).

For reasons of space, the proofs of the statements of our paper are given in Balabdaoui and Groeneboom (2020).

2 General Conditions and the Functions $\hat{\psi}_{n, \hat{\boldsymbol{\alpha}}}$ and $\psi_{\hat{\boldsymbol{\alpha}}}$

We give general conditions that we assume to hold in the remainder of the paper here and give graphical comparisons of the functions $\hat{\psi}_{n, \boldsymbol{\alpha}}$ and $\psi_{\boldsymbol{\alpha}}$, where $\psi_{\boldsymbol{\alpha}}$ is defined in Definition 1.

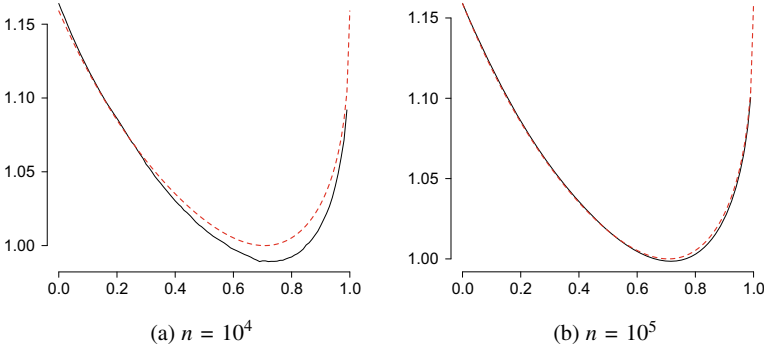


Fig. 1 The loss functions L^{LSE} (red, dashed) and $\widehat{L}_n^{\text{LSE}}$ (solid), where $n = 10^4$ and $n = 10^5$

Example 1 As an illustrative example, we take $d = 2$, $\psi_0(x) = x^3$, $\alpha_0 = (1/\sqrt{2}, 1/\sqrt{2})^T$, $Y_i = \psi_0(\alpha_0^T X_i) + \varepsilon_i$, where the ε_i are i.i.d. standard normal random variables, independent of the X_i , which are i.i.d. random vectors, consisting of two independent Uniform(0, 1) random variables. In this case, the conditional expectation function (5) is a rather complicated function of α which we shall not give here but can be computed by a computer package such as Mathematica or Maple. The loss functions:

$$L^{\text{LSE}} : \alpha_1 \mapsto \mathbb{E}\{Y - \psi_\alpha(\alpha^T X)\}^2 \quad \text{and} \quad \widehat{L}_n^{\text{LSE}} : \alpha_1 \mapsto n^{-1} \sum_{i=1}^n \{Y_i - \widehat{\psi}_{n,\alpha}(\alpha^T X_i)\}^2 \quad (10)$$

where the loss function $\widehat{L}_n^{\text{LSE}}$ is for sample sizes $n = 10,000$ and $n = 100,000$, and $\alpha = (\alpha_1, \alpha_2)^T$. For $\alpha_1 \in [0, 1]$ and α_2 equal to the positive root $\{1 - \alpha_1^2\}^{1/2}$, we get Fig. 1. The function L^{LSE} has a minimum equal to 1 at $\alpha_1 = 1/\sqrt{2}$, and $\widehat{L}_n^{\text{LSE}}$ has a minimum at a value very close to $1/\sqrt{2}$ (furnishing the profile LSE $\widehat{\alpha}_n$), which gives a visual evidence for consistency of the profile LSE.

In order to show the \sqrt{n} -consistency and asymptotic normality of the estimators in the next sections, we now introduce some conditions, which correspond to those in Balabdaoui et al. (2019b). We note that we do not need conditions on reparameterization.

- (A1) X has a density w.r.t. Lebesgue measure on its support \mathcal{X} , which is a convex set \mathcal{X} with a nonempty interior, and satisfies $\mathcal{X} \subset \{x \in \mathbb{R}^d : \|x\| \leq R\}$ for some $R > 0$.
- (A2) The function ψ_0 is bounded on the set $\{u \in \mathbb{R} : u = \alpha_0^T x, x \in \mathcal{X}\}$.
- (A3) There exists $\delta > 0$ such that the conditional expectation $\tilde{\psi}_\alpha$, defined by (5), is nondecreasing on $I_\alpha = \{u \in \mathbb{R} : u = \alpha^T x, x \in \mathcal{X}\}$ and satisfies $\tilde{\psi}_\alpha = \psi_\alpha$, so minimizes

$$\|\mathbb{E}\{Y - \psi(\boldsymbol{\alpha}^T \mathbf{X})\} \mathbf{X}\|^2,$$

over nondecreasing functions ψ , if $\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| \leq \delta$.

- (A4) Let a_0 and b_0 be the (finite) infimum and supremum of the interval $\{\boldsymbol{\alpha}_0^T \mathbf{x}, \mathbf{x} \in \mathcal{X}\}$. Then ψ_0 is continuously differentiable on $(a_0 - \delta R, a_0 + \delta R)$, where R and δ are as in Assumption A1 and A3.
- (A5) The density g of \mathbf{X} is differentiable and there exist strictly positive constants c_1 to c_4 such that $c_1 \leq g(\mathbf{x}) \leq c_2$ and $c_3 \leq \frac{\partial}{\partial x_i} g(\mathbf{x}) \leq c_4$ for \mathbf{x} in the interior of \mathcal{X} .
- (A6) There exists a $c_0 > 0$ and $M > 0$ such that $\mathbb{E}\{|Y|^m | \mathbf{X} = \mathbf{x}\} \leq m! M_0^{m-2} c_0$ for all integers $m \geq 2$ and $\mathbf{x} \in \mathcal{X}$ almost surely w.r.t. dG .

These conditions are rather natural, and are discussed in Balabdaoui et al. (2019b). The following lemma shows that, for the asymptotic distribution of $\hat{\boldsymbol{\alpha}}_n$, we can reduce the derivation to the analysis of $\psi_{\hat{\boldsymbol{\alpha}}_n}$. We have the following result (Proposition 4 in Balabdaoui et al. 2019b) on the distance between $\hat{\psi}_{n, \hat{\boldsymbol{\alpha}}}$ and $\psi_{\hat{\boldsymbol{\alpha}}}$.

Lemma 1 *Let conditions (A1)–(A6) be satisfied and let G be the distribution function of \mathbf{X} . Then we have, for $\boldsymbol{\alpha}$ in a neighborhood $\mathcal{B}(\boldsymbol{\alpha}_0, \delta)$ of $\boldsymbol{\alpha}_0$*

$$\sup_{\boldsymbol{\alpha} \in \mathcal{B}(\boldsymbol{\alpha}_0, \delta)} \int \left\{ \hat{\psi}_{n\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \mathbf{x}) - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \mathbf{x}) \right\}^2 dG(\mathbf{x}) = O_p((\log n)^2 n^{-2/3}).$$

3 The Limit Theory for the SSE

In this section, we derive the limit distribution of the SSE introduced above. In our derivation, the function $\psi_{\boldsymbol{\alpha}}$ of Definition 1 plays a crucial role. Below, we will use the following assumptions, additionally to (A1)–(A6).

- (A7) There exists a $\delta > 0$ such that for all $\boldsymbol{\alpha} \in (\mathcal{B}(\boldsymbol{\alpha}_0, \delta) \cap \mathcal{S}_{d-1}) \setminus \{\boldsymbol{\alpha}_0\}$, the random variable

$$\text{cov}((\boldsymbol{\alpha}_0 - \boldsymbol{\alpha})^T \mathbf{X}, \psi_0(\boldsymbol{\alpha}_0^T \mathbf{X}) \mid \boldsymbol{\alpha}^T \mathbf{X})$$

is not equal to 0 almost surely.

- (A8) The matrix

$$\mathbb{E}[\psi'_0(\boldsymbol{\alpha}_0^T \mathbf{X}) \text{cov}(\mathbf{X} \mid \boldsymbol{\alpha}_0^T \mathbf{X})]$$

has rank $d - 1$.

We start by comparing (3) with the function

$$\boldsymbol{\alpha} \mapsto \|\mathbb{E}\{Y - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \mathbf{X})\} \mathbf{X}\|^2. \quad (11)$$

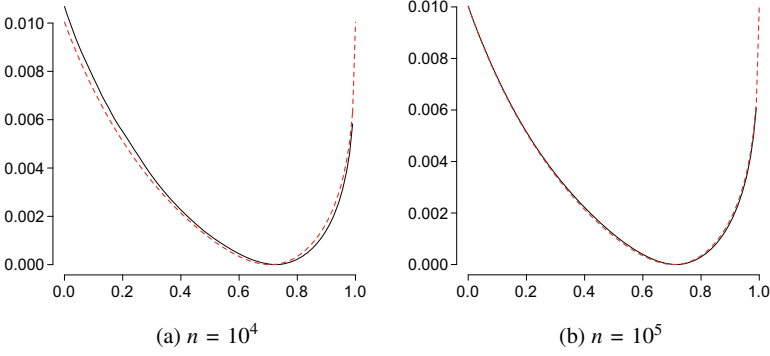


Fig. 2 The loss functions L^{SSE} (red, dashed) and $\widehat{L}_n^{\text{SSE}}$ (solid), where $n = 10^4$ and $n = 10^5$

As in Sect. 1, the function $\hat{\psi}_{n,\alpha}$ is just the (isotonic) least squares estimate for fixed α .

Example 2 (Continuation of Example 1) We consider the loss function given by

$$L^{\text{SSE}} : \alpha_1 \mapsto \left\| \mathbb{E} \{ Y - \psi_\alpha(\alpha^T X) \} X \right\|^2, \quad (12)$$

and compare this with the loss function

$$\widehat{L}_n^{\text{SSE}} : \alpha_1 \mapsto \left\| n^{-1} \sum_{i=1}^n \left\{ Y_i - \hat{\psi}_{n,\alpha}(\alpha^T X_i) \right\} X_i \right\|^2, \quad (13)$$

for the same data as in Example 1 in Sect. 2. If we plot the loss functions L^{SSE} and $\widehat{L}_n^{\text{SSE}}$ for the model of Example 1, where $\alpha = (\alpha_1, \alpha_2)^T$, for $\alpha_1 \in [0, 1]$ and α_2 the positive root $\sqrt{1 - \alpha_1^2}$, we get Fig. 2. The function L^{LSE} has a minimum equal to 0 at $\alpha_1 = 1/\sqrt{2}$ while $\widehat{L}_n^{\text{SSE}}$ attains its minimum at a value that is very close to $1/\sqrt{2}$.

In general, the curve $\widehat{L}_n^{\text{SSE}}$ will be smoother than the curve $\widehat{L}_n^{\text{LSE}}$. The rather striking difference in smoothness of the loss functions $\widehat{L}_n^{\text{LSE}}$ and $\widehat{L}_n^{\text{SSE}}$ can be seen in Fig. 3, where we zoom in on the interval $[0.65, 0.80]$ for $n = 10,000$ and the examples of Figs. 1 and 2. The question is whether this difference in smoothness explains why the SSE is \sqrt{n} -consistent while this might not be the case for the profile LSE.

In the computation of the SSE, we have to take a starting point. For this, we use the LSE, which is proved to be consistent in Balabdaoui et al. (2019a). The proof of the consistency of the SSE is a variation on the proof for corresponding crossing of the zero estimator in Balabdaoui et al. (2019b) in (D.2) of the supplementary material. We use the following lemma, which is a corollary to Proposition 2 in the material of Balabdaoui et al. (2019b).

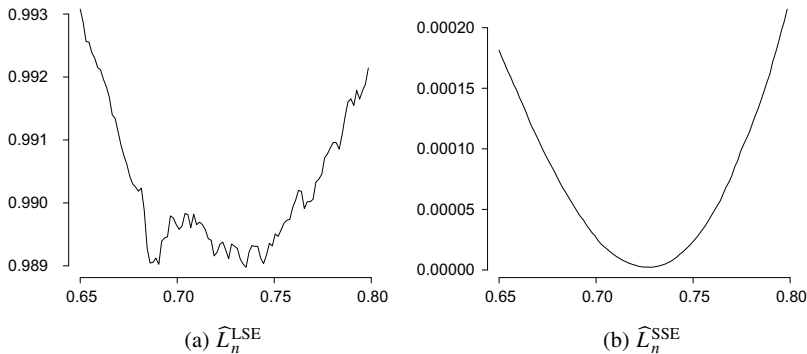


Fig. 3 The loss functions $\widehat{L}_n^{\text{LSE}}$ and $\widehat{L}_n^{\text{SSE}}$ on $[0.65, 0.80]$, for $n = 10^4$

Lemma 2 Let ϕ_n and ϕ be defined by

$$\phi_n(\boldsymbol{\alpha}) = \int \mathbf{x} \left\{ y - \widehat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \mathbf{x}) \right\} d\mathbb{P}_n(\mathbf{x}, y),$$

and

$$\phi(\boldsymbol{\alpha}) = \int \mathbf{x} \left\{ y - \psi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \mathbf{x}) \right\} dP(\mathbf{x}, y).$$

Then, uniformly for $\boldsymbol{\alpha}$ in a neighborhood $\mathcal{B}(\boldsymbol{\alpha}_0, \delta) \cap \mathcal{S}_{d-1}$ of $\boldsymbol{\alpha}_0$

$$\phi_n(\boldsymbol{\alpha}) = \phi(\boldsymbol{\alpha}) + o_p(1).$$

Remark 1 The proof in Balabdaoui et al. (2019b) used reparameterization, but this is actually not needed in the proof.

Theorem 1 (Consistency of the SSE) Let $\widehat{\boldsymbol{\alpha}}_n \in \mathcal{S}_{d-1}$ be the SSE of $\boldsymbol{\alpha}_0$ and let conditions (A1)–(A8) be satisfied. Then

$$\widehat{\boldsymbol{\alpha}}_n \xrightarrow{p} \boldsymbol{\alpha}_0.$$

Lemma 3 Let $\widehat{\boldsymbol{\alpha}}_n \in \mathcal{S}_{d-1}$ be a minimizer of

$$\left\| n^{-1} \sum_{i=1}^n \left\{ Y_i - \widehat{\psi}_{n,\boldsymbol{\alpha}}(\boldsymbol{\alpha}^T \mathbf{X}_i) \right\} \mathbf{X}_i \right\|^2, \quad (14)$$

for $\boldsymbol{\alpha} \in \mathcal{S}_{d-1}$, where $\|\cdot\|$ denotes the Euclidean norm. Then, under conditions (A1)–(A8) we have

$$n^{-1} \sum_{i=1}^n \left\{ Y_i - \hat{\psi}_{n, \hat{\alpha}_n}(\hat{\alpha}_n^T X_i) \right\} X_i = n^{-1} \sum_{i=1}^n \left\{ Y_i - \psi_{\hat{\alpha}_n}(\hat{\alpha}_n^T X_i) \right\} \left\{ X_i - \mathbb{E}(X | \hat{\alpha}_n^T X_i) \right\} + o_p(n^{-1/2}). \quad (15)$$

We now have the following limit result.

Theorem 2 (Asymptotic normality of the SSE) *Let $\hat{\alpha}_n$ be the minimizer of*

$$\left\| n^{-1} \sum_{i=1}^n \left\{ Y_i - \hat{\psi}_{n, \alpha}(\alpha^T X_i) \right\} X_i \right\|^2, \quad (16)$$

for $\alpha \in \mathcal{S}_{d-1}$, where $\|\cdot\|$ denotes the Euclidean norm. Let the matrices \mathbf{A} and $\mathbf{\Sigma}$ be defined by

$$\mathbf{A} = \mathbb{E} \left[\psi'_0(\alpha_0^T X) \text{Cov}(X | \alpha_0^T X) \right], \quad (17)$$

and

$$\mathbf{\Sigma} = \mathbb{E} \left[\left\{ Y - \psi_0(\alpha_0^T X) \right\}^2 \left\{ X - \mathbb{E}(X | \alpha_0^T X) \right\} \left\{ X - \mathbb{E}(X | \alpha_0^T X) \right\}^T \right]. \quad (18)$$

Then, under conditions (A1)–(A8), we have

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0) \rightarrow_d N(\mathbf{0}, \mathbf{A}^- \mathbf{\Sigma} \mathbf{A}^-),$$

where \mathbf{A}^- is the Moore-Penrose inverse of \mathbf{A} .

Example 3 (Continuation of Example 2) We compute the asymptotic covariance matrix for Example 2. In this case, we get for matrix \mathbf{A} in part (ii) of Theorem 2

$$\begin{aligned} \mathbf{A} &= \mathbb{E} \left[\psi'_0(\alpha_0^T X) \text{Cov}(X | \alpha_0^T X) \right] \\ &= \frac{3}{4} \mathbb{E} \left[\left(\frac{X_1 + X_2}{\sqrt{2}} \right)^2 (X - \mathbb{E}(X | \alpha_0^T X)) (X - \mathbb{E}(X | \alpha_0^T X))^T \right] \\ &= \begin{pmatrix} 1/15 & -1/15 \\ -1/15 & 1/15 \end{pmatrix}. \end{aligned}$$

The Moore-Penrose inverse of \mathbf{A} is given by

$$\mathbf{A}^- = \begin{pmatrix} 15/4 & -15/4 \\ -15/4 & 15/4 \end{pmatrix}.$$

Furthermore, we get

$$\begin{aligned}
\Sigma &= \mathbb{E} \left[\{Y - \psi_0(\alpha_0^T X)\}^2 \{X - \mathbb{E}(X|\alpha_0^T X)\} \{X - \mathbb{E}(X|\alpha_0^T X)\}^T \right] \\
&= \mathbb{E} \{X - \mathbb{E}(X|\alpha_0^T X)\} \{X - \mathbb{E}(X|\alpha_0^T X)\}^T \\
&= \begin{pmatrix} 1/24 & -1/24 \\ -1/24 & 1/24 \end{pmatrix}.
\end{aligned}$$

So the asymptotic covariance matrix is given by

$$A^- \Sigma A^- = \begin{pmatrix} 75/32 & -75/32 \\ -75/32 & 75/32 \end{pmatrix} \approx \begin{pmatrix} 2.34375 & -2.34375 \\ -2.34375 & 2.34375 \end{pmatrix}.$$

Remark 2 Theorem 2 corresponds to Theorem 3 in Balabdaoui et al. (2019b), but note that the estimator has a different definition. Reparameterization is also avoided.

4 The Limit Theory for ESE and Cubic Spline Estimator

The proofs of the consistency and asymptotic normality of the ESE and spline estimator are highly similar to the proofs of these facts for the SSE in the preceding section. The only extra ingredient is the occurrence of the estimate of the derivative of the link function. We only discuss the asymptotic normality.

In addition to the assumptions (A1)–(A7), we now assume the following:

(A8') ψ_α is twice differentiable on $(\inf_{x \in \mathcal{X}}(\alpha^T x), \sup_{x \in \mathcal{X}}(\alpha^T x))$.

(A9) The matrix

$$\mathbb{E} [\psi'_0(\alpha_0^T X)^2 \text{cov}(X|\alpha_0^T X)]$$

has rank $d - 1$.

An essential step is again to show that

$$\begin{aligned}
&\int \mathbf{x} \left\{ y - \hat{\psi}_{n, \hat{\alpha}_n}(\hat{\alpha}_n^T \mathbf{x}) \right\} \hat{\psi}'_{n, \hat{\alpha}_n}(\hat{\alpha}_n^T \mathbf{x}) d\mathbb{P}_n(\mathbf{x}, y) \\
&= \int \left\{ \mathbf{x} - \mathbb{E}(X|\hat{\alpha}_n^T X) \right\} \left\{ y - \hat{\psi}_{n, \hat{\alpha}_n}(\hat{\alpha}_n^T \mathbf{x}) \right\} \hat{\psi}'_{n, \hat{\alpha}_n}(\hat{\alpha}_n^T \mathbf{x}) d\mathbb{P}_n(\mathbf{x}, y) \\
&\quad + o_p(n^{-1/2}) + o_p(\hat{\alpha}_n - \alpha_0).
\end{aligned}$$

For the ESE, this is done by defining the piecewise constant function $\bar{\rho}_{n, \alpha}$ for u in the interval between successive jumps τ_i and τ_{i+1}) of $\hat{\psi}_{n, \alpha}$ by

$$\bar{\rho}_{n, \alpha}(u) = \begin{cases} \mathbb{E}[X|\alpha^T X = \tau_i] \psi'_\alpha(\tau_i) & \text{if } \psi_\alpha(u) > \hat{\psi}_{n, \alpha}(\tau_i) \text{ for all } u \in (\tau_i, \tau_{i+1}), \\ \mathbb{E}[X|\alpha^T X = s] \psi'_\alpha(s) & \text{if } \psi_\alpha(s) = \hat{\psi}_{n, \alpha}(s) \text{ for some } s \in (\tau_i, \tau_{i+1}), \\ \mathbb{E}[X|\alpha^T X = \tau_{i+1}] \psi'_\alpha(\tau_{i+1}) & \text{if } \psi_\alpha(u) < \hat{\psi}_{n, \alpha}(\tau_i) \text{ for all } u \in (\tau_i, \tau_{i+1}); \end{cases}$$

see Appendix E in the supplement of Balabdaoui et al. (2019b). The remaining part of the proof runs along the same lines as the proof for the SSE. For additional details, see Appendix E in the supplement of Balabdaoui et al. (2019b).

The corresponding step in the proof for the spline estimator is given by the following lemma.

Lemma 4 *Let the conditions of Theorem 5 in Kuchibhotla and Patra (2020) be satisfied. In particular, let the penalty parameter μ_n satisfy $\mu_n = o_p(n^{-1/2})$. Then we have for all α in a neighborhood of α_0 and for the corresponding natural cubic spline $\hat{\psi}_{n\alpha}$*

$$\int \mathbb{E}(X|\alpha^T X) \left\{ y - \hat{\psi}_{n\alpha}(\alpha^T x) \right\} \hat{\psi}'_{n\alpha}(\alpha^T x) d\mathbb{P}_n(x, y) = O_p(\mu_n) = o_p(n^{-1/2}).$$

Remark 3 The result shows that we have as our basic equation in α

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left\{ \hat{\psi}_{n\alpha}(\alpha^T X_i) - Y_i \right\} \hat{\psi}'_{n\alpha}(\alpha^T X_i) X_i \\ &= \frac{1}{n} \sum_{i=1}^n \left\{ \hat{\psi}_{n\alpha}(\alpha^T X_i) - Y_i \right\} \hat{\psi}'_{n\alpha}(\alpha^T X_i) \left\{ X_i - \mathbb{E}(X_i|\alpha^T X_i) \right\} + o_p(n^{-1/2}) \\ &= o_p(n^{-1/2}). \end{aligned}$$

The remaining part of the proof of the asymptotic normality can either run along the same lines as the proof for the corresponding fact for the SSE, using the function $u \mapsto \psi_\alpha(u) = \mathbb{E}\{\psi_0(\alpha^T x)|\alpha^T X = u\}$, or directly use the convergence of $\hat{\psi}_{n\hat{\alpha}_n}$ to ψ_0 and of $\hat{\psi}'_{n\hat{\alpha}_n}$ to ψ'_0 (see Theorem 3 in Kuchibhotla and Patra 2020). For the SSE and ESE, we were forced to introduce the intermediate function ψ_α to get to the derivatives, because for these estimators the derivative of $\hat{\psi}_{n\hat{\alpha}_n}$ did not exist.

We get the following result.

Theorem 3 *Let either $\hat{\alpha}_n$ be the ESE of α_0 and let Assumptions (A1)–(A7) and (A8') and (A9) of the present section be satisfied, or let $\hat{\alpha}_n$ be the spline estimator of α_0 and let Assumptions (A0)–(A6) and (B1)–(B3) of Kuchibhotla and Patra (2020) be satisfied. Moreover, let the bandwidth $h \asymp n^{-1/7}$ in the estimate of the derivative of ψ_α for the ESE. Define the matrices*

$$\tilde{\mathbf{A}} := \mathbb{E} \left[\psi'_0(\alpha_0^T X)^2 \text{Cov}(X|\alpha_0^T X) \right], \quad (19)$$

and

$$\tilde{\Sigma} := \mathbb{E} \left[\left\{ Y - \psi_0(\alpha_0^T X) \right\}^2 \psi'_0(\alpha_0^T X)^2 \left\{ X - \mathbb{E}(X|\alpha_0^T X) \right\} \left\{ X - \mathbb{E}(X|\alpha_0^T X) \right\}^T \right]. \quad (20)$$

Then

$$\sqrt{n}(\tilde{\alpha}_n - \alpha_0) \rightarrow_d N_d \left(\mathbf{0}, \tilde{\mathbf{A}}^{-} \tilde{\Sigma} \tilde{\mathbf{A}}^{-} \right),$$

where $\tilde{\mathbf{A}}^{-}$ is the Moore-Penrose inverse of $\tilde{\mathbf{A}}$.

This corresponds to Theorem 6 in Balabdaoui et al. (2019b) and Theorem 5 in Kuchibhotla and Patra (2020), but note that the formulation of Theorem 5 in Kuchibhotla and Patra (2020) still contains the Jacobian connected with the lower dimensional parameterization. Consequently, the ESE and the cubic spline estimator admit the same weak limit under the conditions stated above.

5 Simulation and Comparisons with Other Estimators

In this section, we compare the LSE with the Simple Score Estimator (SSE), the Efficient Score Estimator (ESE), the Effective Dimension Reduction (EDR) estimate, the spline estimate, the MAVE estimate, and the EFM estimate. We take part in the simulation settings in Balabdaoui et al. (2019a), which means that we take the dimension d equal to 2. Since the parameter belongs to the boundary of a circle in this case, we only have to determine a one-dimensional parameter. Using this fact, we use the parameterization $\alpha = (\alpha_1, \alpha_2) = (\cos(\beta), \sin(\beta))$ and determine the angle β by a golden section search for the SSE, ESE, and spline estimate. For EDR, we used the R package `edr`; the method is discussed in Hristache et al. (2001). The spline method is described in Kuchibhotla and Patra (2020), and there exists an R package `simest` for it, but we used our own implementation. For the MAVE method, we used the R package `MAVE`; for theory, see Xia (2006). For the EFM estimate (see Cui et al. 2011), we used an R script, due to Xia Cui and kindly provided to us by her and Rohit Patra. All runs of our simulations can be reproduced by running the R scripts in Groeneboom 2018.

In simulation model 1, we take $\alpha_0 = (1/\sqrt{2}, 1/\sqrt{2})^T$ and $\mathbf{X} = (X_1, X_2)^T$, where X_1 and X_2 are independent Uniform(0, 1) variables. The model is now

$$Y = \psi_0(\alpha_0^T \mathbf{X}) + \varepsilon,$$

where $\psi_0(u) = u^3$ and ε is a standard normal random variable, independent of \mathbf{X} .

In simulation model 2, we also take $\alpha_0 = (1/\sqrt{2}, 1/\sqrt{2})^T$ and $\mathbf{X} = (X_1, X_2)^T$, where X_1 and X_2 are independent Uniform(0, 1) variables. This time, however, the model is (Table 1)

$$Y = \text{Bin} \left(10, \exp(\alpha_0^T \mathbf{X}) / \{1 + \exp(\alpha_0^T \mathbf{X})\} \right);$$

see also Table 2 in Balabdaoui et al. (2019a). This means

Table 1 Simulation, model 1; ε_i is standard normal and independent of X_i , consisting of two independent Uniform(0, 1) random variables. The mean value $\hat{\mu}_i = \text{mean}(\hat{\alpha}_{in})$, $i = 1, 2$ and n times the variance-covariance $\hat{\sigma}_{ij} = n \cdot \text{cov}(\hat{\alpha}_{in}, \hat{\alpha}_{jn})$, $i, j = 1, 2$, of the Efficient Dimension Reduction Estimate (EDR), computed by the R package `edr`, the Least Squares Estimate (LSE), the Simple Score Estimate (SSE), the Efficient Score Estimate (ESE), the spline estimate, the MAVE estimate, and the EFM estimate for different sample sizes n . The line, preceded by ∞ , gives the asymptotic values (unknown for EDR and LSE). The values are based on 1000 replications

Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{12}$
EDR	100	0.621877	0.361894	11.409222	36.869184	9.152389
	500	0.701217	0.686094	7.334756	11.468453	-3.881349
	1000	0.701669	0.702244	6.437653	8.090771	-3.552562
	5000	0.706021	0.706798	7.344431	7.276717	-7.288047
	∞	0.707107	0.707107	?	?	?
LSE	100	0.672698	0.697350	3.148912	2.975246	-2.915427
	500	0.702163	0.701718	3.620960	3.665710	-3.588491
	1000	0.704706	0.704320	3.665561	3.664711	-3.637541
	5000	0.707262	0.705690	4.435842	4.485168	-4.453713
	∞	0.707107	0.707107	?	?	?
SSE	100	0.673997	0.6919403	3.338637	3.362656	-3.141408
	500	0.699986	0.706198	2.849647	2.807978	-2.793798
	1000	0.706477	0.704191	2.501106	2.510047	-2.494237
	5000	0.707090	0.706423	2.473765	2.485884	-2.477371
	∞	0.707107	0.707107	2.343750	2.343750	-2.343750
ESE	100	0.682781	0.687949	3.067802	2.991976	-2.855176
	500	0.702940	0.702462	3.100843	3.116337	-3.064151
	1000	0.704055	0.706387	2.676388	2.653164	-2.650667
	5000	0.707130	0.706444	2.257541	2.265547	-2.259443
	∞	0.707107	0.707107	1.885522	1.885522	-1.885522
Spline	100	0.690741	0.705485	1.801235	1.762567	-1.711552
	500	0.703670	0.702640	1.795384	1.778454	-1.773560
	1000	0.705684	0.706007	1.786589	1.781797	-1.777691
	5000	0.706404	0.707193	2.180466	2.181544	-2.179269
	∞	0.707107	0.707165	1.885522	1.885522	-1.885522
MAVE	100	0.686503	0.684887	2.423618	3.546768	-2.245708
	500	0.703333	0.705537	1.897806	1.876220	-2.040677
	1000	0.705840	0.705660	1.929966	1.907128	-1.911452
	5000	0.707328	0.706299	2.071168	2.082169	-2.074914
	∞	0.707107	0.707107	1.885522	1.885522	-1.885522
EFM	100	0.686292	0.684274	2.802308	3.280956	-2.312445
	500	0.703236	0.705133	2.082162	2.045150	-2.044960
	1000	0.705629	0.705950	1.866486	1.860184	-1.856340
	5000	0.707269	0.707251	1.953800	1.964081	-1.957351
	∞	0.707107	0.707107	1.885522	1.885522	-1.885522

Table 2 Simulation, model 2; $Y_i \sim \text{Bin}(10, \exp(\alpha_0^T X_i) / \{1 + \exp(\alpha_0^T X_i)\})$, where X_i consists of two independent Uniform(0, 1) random variables. The mean value $\hat{\mu}_i = \text{mean}(\hat{\alpha}_{in})$, $i = 1, 2$ and n times the variance-covariance $\text{ncov}(\hat{\alpha}_{in}, \hat{\alpha}_{jn})$, $i, j = 1, 2$, of the Efficient Dimension Reduction Estimate (EDR), computed by the R package `edr`, the Least Squares Estimate (LSE), the Simple Score Estimate (SSE), the Efficient Score Estimate (ESE), the spline estimate, the MAVE estimate, and the EFM estimate for different sample sizes n . The line, preceded by ∞ , gives the asymptotic values (unknown for EDR and LSE). The values are based on 1000 replications

Method	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_{11}$	$\hat{\sigma}_{22}$	$\hat{\sigma}_{12}$
EDR	100	0.587264	0.202005	13.33724	48.15572	11.87625
	500	0.670702	0.602469	26.76111	66.92737	14.09701
	1000	0.696075	0.666591	21.89080	49.31544	9.345753
	5000	0.704424	0.706604	11.39598	11.11493	-11.17376
	∞	0.707107	0.707107	?	?	?
LSE	100	0.658631	0.699725	4.069966	3.596783	-3.609490
	500	0.695541	0.703007	5.650618	5.362877	-5.358190
	1000	0.704497	0.701243	5.909494	6.043808	-5.911246
	5000	0.704805	0.707621	6.303320	6.321866	-6.298515
	∞	0.707107	0.707107	?	?	?
SSE	100	0.667908	0.694376	3.760921	3.420387	-3.356968
	500	0.698498	0.706423	3.358458	3.182044	-3.223734
	1000	0.707276	0.702390	3.179623	3.236283	-3.184724
	5000	0.706162	0.707286	2.718742	2.707549	-2.709870
	∞	0.707107	0.707107	2.727482	2.727482	-2.727482
ESE	100	0.684804	0.688063	2.892165	2.874755	-2.744223
	500	0.698078	0.706159	3.562625	3.457337	-3.446605
	1000	0.707879	0.701445	3.420159	3.470217	-3.418606
	5000	0.706321	0.707110	2.775092	2.760287	-2.764230
	∞	0.707107	0.707107	2.737200	2.737200	-2.737200
Spline	100	0.677287	0.695301	3.009781	2.779876	-2.714928
	500	0.699117	0.706946	2.952928	2.784383	-2.830415
	1000	0.707890	0.702001	3.027712	3.064772	-3.026082
	5000	0.706200	0.707312	2.764447	2.762986	-2.760530
	∞	0.707107	0.707232	2.737200	2.737200	-2.737200
MAVE	100	0.667849	0.654361	3.891510	8.700093	-2.325804
	500	0.699108	0.706377	3.155191	2.990569	-3.031249
	1000	0.707520	0.702341	3.040201	3.097965	-3.049075
	5000	0.707657	0.705827	2.572343	2.573418	-2.570275
	∞	0.707107	0.707107	2.737200	2.737200	-2.737200
EFM	100	0.663227	0.666070	5.681573	5.978194	-2.503058
	500	0.698920	0.706295	3.279110	3.055940	-3.118757
	1000	0.707878	0.706275	3.102414	3.157143	-3.108516
	5000	0.706043	0.701894	2.669352	2.650343	-2.656742
	∞	0.707107	0.707107	2.737200	2.737200	-2.737200

$$Y = \psi_0(\boldsymbol{\alpha}_0^T \mathbf{X}) + \varepsilon,$$

where

$$\psi_0(\boldsymbol{\alpha}_0^T \mathbf{X}) = 10 \exp(\boldsymbol{\alpha}_0^T \mathbf{X}) / \{1 + \exp(\boldsymbol{\alpha}_0^T \mathbf{X})\}, \quad \varepsilon = N_n - \psi_0(\boldsymbol{\alpha}_0^T \mathbf{X}),$$

and

$$N_n = \text{Bin} \left(10, \frac{\exp(\boldsymbol{\alpha}_0^T \mathbf{X})}{1 + \exp(\boldsymbol{\alpha}_0^T \mathbf{X})} \right).$$

Note that indeed $\mathbb{E}\{\varepsilon|\mathbf{X}\} = 0$, but that we do not have independence of ε and \mathbf{X} , as in the previous example.

It was noticed in Xia (2006), p. 1113, that, although it was shown in Hristache et al. (2001) that the \sqrt{n} rate of convergence for the estimation of $\boldsymbol{\alpha}_0$ can be achieved, the asymptotic distribution of the method proposed in Hristache et al. (2001) was not derived, which makes it difficult to compare the limiting efficiency of the estimation method with other methods. In Xia (2006), the asymptotic distribution of the rMAVE estimate is derived (see Theorem 4.2 of Xia 2006), which shows that this limit distribution is actually the same as that of the ESE and the spline estimate. Since Xia is one of the authors of the recent MAVER package, we assume that the rMAVE method has been implemented in this package, so we will identify MAVE with rMAVE in the sequel.

The proof of the asymptotic normality result for the MAVE method uses the fact that the iteration steps, described on p.1117 of Xia (2006), start in a neighborhood $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\| \leq Cn^{-1/2+c_0}\}$ of $\boldsymbol{\alpha}_0$, where $C > 0$ and $c_0 < 1/20$, and indeed our

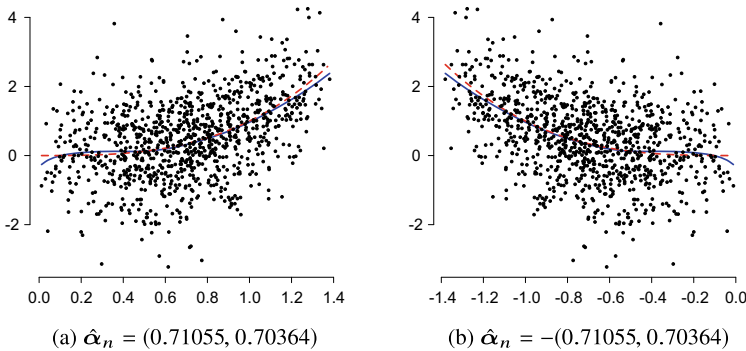


Fig. 4 Two MAVE estimates of $\boldsymbol{\alpha}_0 = 2^{-1/2}(1, 1)^T$ for model 1 with sample size $n = 1000$: **a** from starting the iterations at $\boldsymbol{\alpha}_0$, **b** from starting the iterations at $-\boldsymbol{\alpha}_0$; the blue solid curve is the estimate of the link function, based on $\hat{\boldsymbol{\alpha}}_n$; the blue dashed function is $t \mapsto t^3$ in **a** and $t \mapsto -t^3$ in **b**. Note that in **b** also the sign of the first coordinates of the points $(\hat{\boldsymbol{\alpha}}_n^T \mathbf{X}_i, Y_i)$ in the scatterplot is reversed. Under the restriction that the link function is nondecreasing **b** cannot be a solution

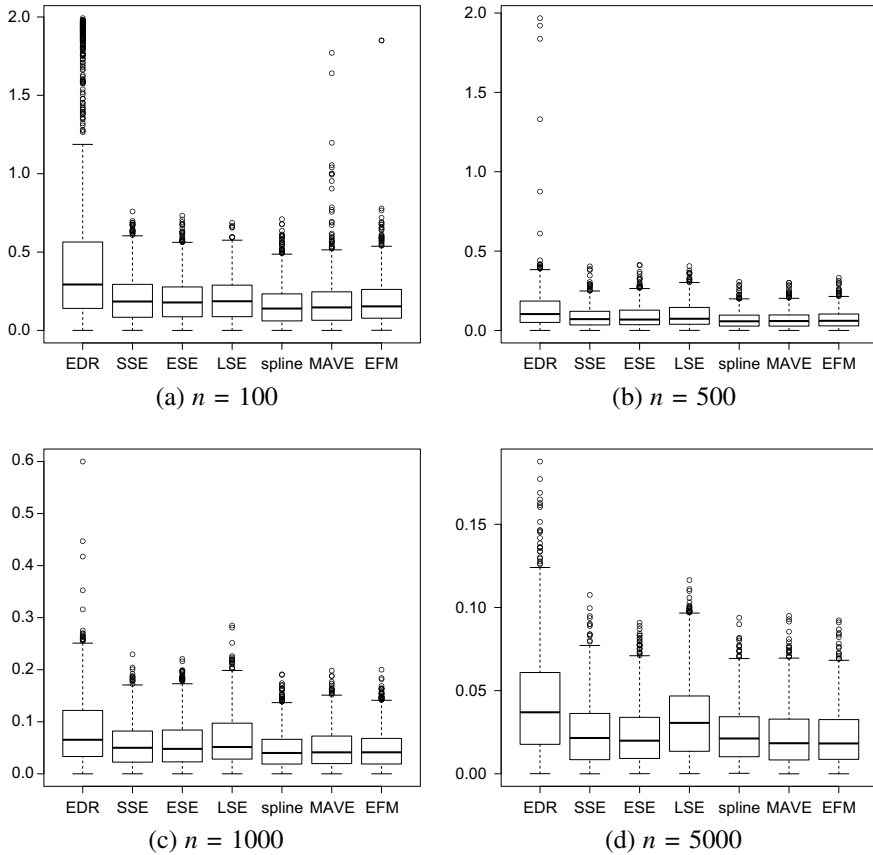


Fig. 5 Boxplots of $\sqrt{n/2} \|\hat{\alpha}_n - \alpha_0\|_2$ for model 1. In **b** and **c**, the values of EDR were truncated at 0.6 to show more clearly the differences between the other estimates

original experiments with the R package showed many outliers, probably due to starting values not sufficiently close to α_0 . A further investigation revealed that there were many solutions in the neighborhood of the points $-\alpha_0$. This phenomenon is illustrated in Fig. 4, generated by our own implementation of the algorithm in Xia (2006). The link function is constructed from the values $a_j^{\hat{\alpha}_n}$ in the algorithm in Xia (2006), p. 1117, where the ordered values of $\hat{\alpha}_n^T X_j$ are the first coordinates.

Because of the difficulty we just discussed, we reversed in the results of the MAVE R package the sign of the solutions in the neighborhood of $-\alpha_0$. By the parameterization with a positive first coordinate in Cui et al. (2011), situation (b) in Fig. 4 cannot occur for the EFM algorithm. We also tried a modification of the same type as our modification of the MAVE algorithm for the EDR algorithm, but this did not lead to a similar improvement of the results.

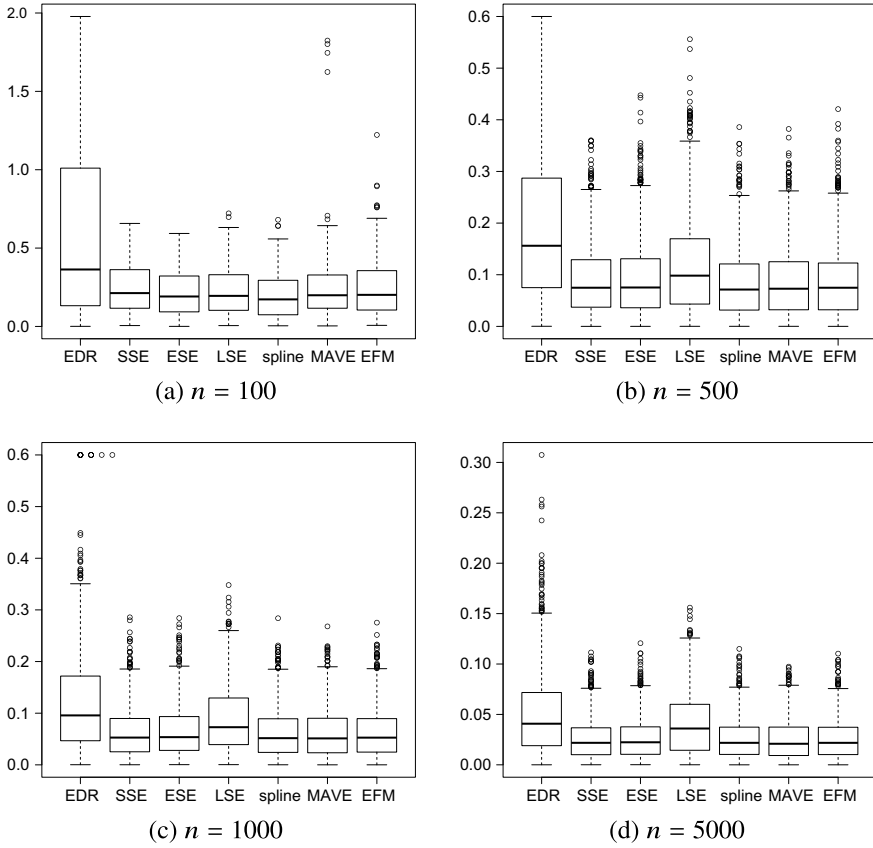


Fig. 6 Boxplots of $\sqrt{n/2} \|\hat{\alpha}_n - \alpha_0\|_2$ for model 2. In **b** and **c**, the values of EDR were truncated at 0.6 to show more clearly the differences between the other estimates

It follows from Theorem 2 that the variance of the asymptotic normal distribution for the SSE is equal to 2.727482 and from Theorem 3 that the variance of the asymptotic normal distribution for the ESE and spline estimator equals 2.737200. We already noticed in Sect. 4 that the present model is not homoscedastic. In this case, the asymptotic covariance matrix for the SSE of Theorem 2 is in fact given by $A^- = A^- \Sigma A^-$.

It is clear that the estimate EDR is inferior to the other methods for these models; even the LSE for which we do not know the rate of convergence has a better performance, see Figs. 5 and 6. In Hristache et al. (2001), not only it is assumed that the errors have a normal distribution, but also in model 1, where this condition is satisfied, the behavior is clearly inferior, in particular for the lower sample sizes.

6 Concluding Remarks

We replaced the “crossing of zero” estimators in Balabdaoui et al. (2019b) with profile least squares estimators. The asymptotic distribution of the estimators was determined and its behavior illustrated by a simulation study, using the same models as in Balabdaoui et al. (2019a).

In the first model, the error is independent of the covariate and homoscedastic and in this case, four of the estimators were efficient. In the other (binomial-logistic) model, the error was dependent on the covariates and not homoscedastic. It was shown that the Simple Score Estimate (SSE) had in fact a smaller asymptotic variance in this model than the other estimators for which the asymptotic variance is known, although the difference is very small and does not really show up in the simulations.

There is no uniformly best estimate in our simulation, but the EDR estimate is clearly inferior to the other estimates, including the LSE, in particular for the lower sample sizes. On the other hand, the LSE is inferior to the other estimators except for the EDR. All simulation results can be reproduced by running the R scripts in Groeneboom (2018).

Acknowledgements We thank Vladimir Spokoiny for helpful discussions during the Oberwolfach meeting “Statistics meets Machine Learning”, January 26–February 1, 2020. We also feel very honored to make this contribution to Christine’s Festschrift and wish her all the best in her future endeavors.

References

- Balabdaoui, F., & Groeneboom, P. (2020). Profile least squares estimators in the monotone single index model. Version with proofs. <https://arxiv.org/abs/2001.05454>.
- Balabdaoui, F., Durot, C., & Jankowski, H. (2019a). Least squares estimation in the monotone single index model. *Bernoulli*, 25(4), 3276–3310.
- Balabdaoui, F., Groeneboom, P., & Hendrickx, K. (2019b). Score estimation in the monotone single-index model. *Scandinavian Journal of Statistics*, 46(2), 517–544. ISSN 0303-6898.
- Cui, X., Härdle, W. K., & Zhu, L. (2011). The efm approach for single-index models. *Annals of Statistics*, 39(3), 1658–1688, 06. <https://doi.org/10.1214/10-AOS871>.
- Green, P. J., & Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. Monographs on statistics and applied probability (Vol. 58). London: Chapman & Hall. ISBN 0-412-30040-0. <https://doi.org/10.1007/978-1-4899-4473-3>. A roughness penalty approach.
- Groeneboom, P. (2018). Algorithms for computing estimates in the single index model. https://github.com/pietg/single_index.
- Hristache, M., Juditsky, A., & Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, 29(3), 595–623. ISSN 0090-5364. 10.1214/aos/1009210681. <https://doi.org/10.1214/aos/1009210681>.
- Kuchibhotla, A. K., & Patra, R. K. (2020). Efficient estimation in single index models through smoothing splines. *Bernoulli*, 26(2), 1587–1618. ISSN 1350-7265. <https://doi.org/10.3150/19-BEJ1183>.
- Landers, D., & Rogge, L. (1981). Isotonic approximation in L_s . *Journal of Approximation Theory*, 31(3), 199–223. ISSN 0021-9045. [https://doi.org/10.1016/0021-9045\(81\)90091-5](https://doi.org/10.1016/0021-9045(81)90091-5).

- Tanaka, H. (2008). Semiparametric least squares estimation of monotone single index models and its application to the iterative least squares estimation of binary choice models.
- Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, 22(6), 1112–1137. ISSN 0266-4666. <https://doi.org/10.1017/S0266466606060531>.

Optimization by Gradient Boosting



G rard Biau and Beno t Cadre

Abstract Gradient boosting is a state-of-the-art prediction technique that sequentially produces a model in the form of linear combinations of elementary predictors—typically decision trees—by solving an infinite-dimensional convex optimization problem. We provide in the present paper a thorough analysis of two widespread versions of gradient boosting, and introduce a general framework for studying these algorithms from the point of view of functional optimization. We prove their convergence as the number of iterations tends to infinity and highlight the importance of having a strongly convex risk functional to minimize. We also present a reasonable statistical context ensuring consistency properties of the boosting predictors as the sample size grows. In our approach, the optimization procedures are run forever (that is, without resorting to an early stopping strategy), and statistical regularization is basically achieved via an appropriate L^2 penalization of the loss and strong convexity arguments.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-73249-3_2) contains supplementary material, which is available to authorized users.

G. Biau (✉)

Sorbonne Universit , CNRS, LPSM, 4 place Jussieu, 75005 Paris, France
e-mail: gerard.biau@sorbonne-universite.fr

B. Cadre

Universit  Rennes, CNRS, IRMAR - UMR 6625, 35000 Rennes, France
e-mail: benoit.cadre@univ-rennes2.fr

  Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_2

1 Introduction

More than twenty years after the pioneering articles of Freund and Schapire (Schapire 1990; Freund 1995; Freund and Schapire 1996, 1997), boosting is still one of the most powerful ideas introduced in statistics and machine learning. Freund and Schapire's AdaBoost algorithm and its numerous descendants have proven to be competitive in a variety of applications, and are still able to provide state-of-the-art decisions in difficult real-life problems. In addition, boosting procedures are computationally fast and comfortable with both regression and classification problems. For surveys of various aspects of boosting algorithms, we refer to Meir and Rätsch (2003), Bühlmann and Hothorn (2007), and to the monographs by Hastie et al. (2009) and Bühlmann and van de Geer (2011). These references point in particular to approaches related to boosting, for example, Frank and Wolfe (1956) algorithm, Mallat and Zhang (1993) matching pursuit algorithm, and weak greedy algorithms of Temlyakov (2000).

In a nutshell, the basic idea of boosting is to combine the outputs of many "simple" predictors, in order to produce a powerful committee with performances improved over the single members. Historically, the first formulations of Freund and Schapire considered boosting as an iterative classification algorithm that is run for a fixed number of iterations, and, at each iteration, selects one of the base classifiers, assigns a weight to it, and outputs the weighted majority vote of the chosen classifiers. Later on, Breiman (1997, 1998, 1999, 2000, 2004) made in a series of papers and technical reports the breakthrough observation that AdaBoost is in fact a gradient-descent-type algorithm in a function space, thereby identifying boosting at the frontier of numerical optimization and statistical estimation. This connection was further emphasized by Friedman et al. (2000), who rederived AdaBoost as a method for fitting an additive model in a forward stagewise manner. Following this, Friedman (2001, 2002) developed a general statistical framework (both for regression and classification) that (i) yields a direct interpretation of boosting methods from the perspective of numerical optimization in a function space, and (ii) generalizes them by allowing optimization of an arbitrary loss function. The term "gradient boosting" was coined by the author, who paid special attention to the case where the individual additive components are decision trees. At the same time, Mason et al. (1999, 2000) embraced a more mathematical approach, revealing boosting as a principle to optimize a convex risk in a function space, by iteratively choosing a weak learner that approximately points in the negative gradient direction.

This functional view of boosting has led to the development of algorithms in many areas of machine learning and computational statistics, beyond regression and classification. The history of boosting goes on today with algorithms such as XGBoost (Extreme Gradient Boosting, Chen and Guestrin 2016), a tree boosting system widely recognized for its outstanding results in numerous data challenges. (An overview of its successes is given in the introductory section of the paper by Chen and Guestrin, 2016.) From a general point of view, XGBoost is but a scalable implementation of gradient boosting that contains various systems and algorithmic optimizations. Its

mathematical principle is to perform a functional gradient-type descent in a space of decision trees, while regularizing the objective to avoid overfitting.

However, despite a long list of successes, much work remains to be done to clarify the mathematical forces driving gradient boosting algorithms. Many influential articles regard boosting with a statistical eye and study the somewhat idealized problem of empirical risk minimization with a convex loss (e.g., Blanchard et al. 2003; Lugosi and Vayatis 2004). These papers essentially concentrate on the statistical properties of the approach (that is, consistency and rates of convergence as the sample size grows) and often ignore the underlying optimization aspects. Other important articles, such as Bühlmann and Yu (2003); Mannor et al. (2003); Zhang and Yu (2005); Bickel et al. (2006); Bartlett and Traskin (2007), take advantage of the iterative principle of boosting, but mainly focus on regularization via early stopping (that is, stopping the boosting iterations at some point), without paying too much attention to the optimization side. Notable exceptions are the pioneering notes of Breiman cited above, and the original paper by Mason et al. (2000), who envision gradient boosting as an infinite-dimensional numerical optimization problem and pave the way for a more abstract analysis. All in all, there is to date no sound theory of gradient boosting in terms of numerical optimization. This state of affairs is a bit paradoxical, since optimization is certainly the most natural mathematical environment for gradient-descent-type algorithms.

In line with the above, our main objective in this article is to provide a thorough analysis of two widespread models of gradient boosting, due to Friedman (2001) and Mason et al. (2000). We introduce in Sect. 2 a general framework for studying the algorithms from the point of view of functional optimization in an L^2 space, and prove in Sect. 3 their convergence as the number of iterations tends to infinity. Our results allow for a large choice of convex losses in the optimization problem (differentiable or not), while highlighting the importance of having a strongly convex risk functional to minimize. This point is interesting, since it provides some theoretical justification for adding a penalty term to the objective, as advocated, for example, in the XGBoost system of Chen and Guestrin (2016). Thus, the main message of Sect. 3 is that, under appropriate conditions, the sequence of boosted iterates converges toward the minimizer of the empirical risk functional over the set of linear combinations of weak learners. However, this does not guarantee that the output of the algorithms (i.e., the boosting predictor) enjoys good statistical properties, as overfitting may kick in. For this reason, we present in Sect. 4 a reasonable framework ensuring consistency properties of the boosting predictors as the sample size grows. In our approach, the optimization procedures are run forever (that is, without resorting to an early stopping strategy), and statistical regularization is basically achieved via an appropriate L^2 penalization of the loss and strong convexity arguments. For clarity, most proofs are gathered in the Supplementary Material Document.

Before embarking on the analysis, we would like to stress that the present paper is theoretical in nature and that its main goal is to clarify/solidify some of the optimization ideas that are behind gradient boosting. In particular, we do not report experimental results, and refer to the specialized literature on (extreme) gradient boosting for discussions on the computational aspects and experiments with real-world data.

2 Gradient Boosting

The purpose of this section is to describe the gradient boosting procedures that we analyze in the paper.

2.1 Mathematical Context

We assume to be given a sample $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ of i.i.d. observations, where each pair (X_i, Y_i) takes values in $\mathcal{X} \times \mathcal{Y}$. Throughout, \mathcal{X} is a Borel subset of \mathbb{R}^d , and $\mathcal{Y} \subset \mathbb{R}$ is either a finite set of labels (for classification) or a subset of \mathbb{R} (for regression). The vector space \mathbb{R}^d is endowed with the Euclidean norm $\|\cdot\|$.

Our goal is to construct a predictor $F : \mathcal{X} \rightarrow \mathbb{R}$ that assigns a response to each possible value of an independent random observation distributed as X_1 . In the context of gradient boosting, this general problem is addressed by considering a class \mathcal{F} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ (called the weak or base learners) and minimizing some empirical risk functional

$$C_n(F) = \frac{1}{n} \sum_{i=1}^n \psi(F(X_i), Y_i)$$

over the linear combinations of functions in \mathcal{F} . The function $\psi : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, called the loss, is convex in its first argument and measures the cost incurred by predicting $F(X_i)$ when the answer is Y_i . For example, in the least squares regression problem, $\psi(x, y) = (y - x)^2$ and

$$C_n(F) = \frac{1}{n} \sum_{i=1}^n (Y_i - F(X_i))^2.$$

However, many other examples are possible, as we will see below. Let δ_z denote the Dirac measure at z , and let $\mu_n = (1/n) \sum_{i=1}^n \delta_{(X_i, Y_i)}$ be the empirical measure associated with the sample \mathcal{D}_n . Clearly,

$$C_n(F) = \mathbb{E} \psi(F(X), Y),$$

where (X, Y) denotes a random pair with distribution μ_n and the symbol \mathbb{E} denotes the expectation with respect to μ_n . Naturally, the theoretical (i.e., population) version of C_n is

$$C(F) = \mathbb{E} \psi(F(X_1), Y_1),$$

where now the expectation is taken with respect to the distribution of (X_1, Y_1) . It turns out that most of our subsequent developments are independent of the context,

whether empirical or theoretical. Therefore, to unify the notation, we let throughout (X, Y) be a generic pair of random variables with distribution $\mu_{X,Y}$, keeping in mind that $\mu_{X,Y}$ may be the distribution of (X_1, Y_1) (theoretical risk), the standard empirical measure μ_n (empirical risk), or any smoothed version of μ_n .

We let μ_X be the distribution of X , $L^2(\mu_X)$ the vector space of all measurable functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that $\int |f|^2 d\mu_X < \infty$, and denote by $\langle \cdot, \cdot \rangle_{\mu_X}$ and $\| \cdot \|_{\mu_X}$ the corresponding norm and scalar product. Thus, for now, our problem is to minimize the quantity

$$C(F) = \mathbb{E}\psi(F(X), Y)$$

over the linear combinations of functions in a given subset \mathcal{F} of $L^2(\mu_X)$. A typical example for \mathcal{F} is the collection of all binary decision trees in \mathbb{R}^d using axis parallel cuts with k terminal nodes. In this case, each $f \in \mathcal{F}$ takes the form $f = \sum_{j=1}^k \beta_j \mathbb{1}_{A_j}$, where $(\beta_1, \dots, \beta_k) \in \mathbb{R}^k$ and A_1, \dots, A_k is a tree-structured partition of \mathbb{R}^d (Devroye et al. 1996, Chap. 20).

As noted earlier, we assume that, for each $y \in \mathcal{Y}$, the function $\psi(\cdot, y)$ is convex. In the framework we have in mind, the function ψ may take a variety of different forms, ranging from standard (regression or classification) losses to more involved penalized objectives. It may also be differentiable or not. Before discussing some examples in detail, we list assumptions that will be needed at some point. Throughout, we let $\xi(\cdot, y) = \partial_x^- \psi(\cdot, y)$ (left derivative) be a subgradient of the convex function $\psi(\cdot, y)$ (the choice of a specific subgradient $\xi(\cdot, y)$ has no impact on the results). In particular, for all $(x_1, x_2) \in \mathbb{R}^2$,

$$\psi(x_1, y) \geq \psi(x_2, y) + \xi(x_2, y)(x_1 - x_2). \tag{1}$$

Assumption A₁

A₁ One has $\mathbb{E}\psi(0, Y) < \infty$. In addition, for all $F \in L^2(\mu_X)$, there exists $\delta > 0$ such that

$$\sup_{G \in L^2(\mu_X) : \|G - F\|_{\mu_X} \leq \delta} \mathbb{E}|\partial_x^- \psi(G(X), Y)|^2 < \infty.$$

This assumption ensures that the convex functional C is locally bounded (in particular, $C(F) < \infty$ for all $F \in L^2(\mu_X)$, and C is continuous). Indeed, by inequality (1), for all $G \in L^2(\mu_X)$,

$$\psi(G(x), y) \leq \psi(0, y) + \xi(G(x), y)G(x).$$

Therefore, by Assumption **A₁** and the Cauchy-Schwarz inequality,

$$\mathbb{E}\psi(G(X), Y) \leq \mathbb{E}\psi(0, Y) + (\mathbb{E}\xi(G(X), Y)^2 \mathbb{E}G(X)^2)^{1/2},$$

so that C is locally bounded. Naturally, Assumption **A₁** is automatically satisfied for the choice $\mu_{X,Y} = \mu_n$.

Assumption \mathbf{A}_2

\mathbf{A}_2 There exists $\alpha > 0$ such that, for all $y \in \mathcal{Y}$, the function $\psi(\cdot, y)$ is α -strongly convex, i.e., for all $(x_1, x_2) \in \mathbb{R}^2$ and $t \in [0, 1]$,

$$\psi(tx_1 + (1-t)x_2, y) \leq t\psi(x_1, y) + (1-t)\psi(x_2, y) - \frac{\alpha}{2}t(1-t)(x_1 - x_2)^2.$$

This assumption will be used in most, but not all, results. Strong convexity will play an essential role in the statistical Sect. 4. We note that, under Assumption \mathbf{A}_2 , for all $(x_1, x_2) \in \mathbb{R}^2$,

$$\psi(x_1, y) \geq \psi(x_2, y) + \xi(x_2, y)(x_1 - x_2) + \frac{\alpha}{2}(x_1 - x_2)^2, \quad (2)$$

which is of course an inequality tighter than (1). Furthermore, the α -strong convexity of $\psi(\cdot, y)$ implies the α -strong convexity of the risk functional C over $L^2(\mu_X)$.

In addition to Assumptions \mathbf{A}_1 and \mathbf{A}_2 , we require the following:

\mathbf{A}_3 There exists a positive constant L such that, almost surely, for all $(x_1, x_2) \in \mathbb{R}^2$,

$$|\mathbb{E}(\xi(x_1, Y) - \xi(x_2, Y) | X)| \leq L|x_1 - x_2|.$$

(In the sequel, in order to lighten the text, we drop the ‘‘almost sure’’ wording wherever conditional expectations are involved.) However esoteric this assumption may seem, it is in fact mild and provides a framework that encompasses a large variety of familiar situations. In particular, Assumption \mathbf{A}_3 admits a stronger version \mathbf{A}'_3 , which is useful as soon as the function ψ is continuously differentiable with respect to its first variable:

\mathbf{A}'_3 For all $y \in \mathcal{Y}$, the function $\psi(\cdot, y)$ is continuously differentiable, and there exists a positive constant L such that, for all $(x_1, x_2, y) \in \mathbb{R}^2 \times \mathcal{Y}$,

$$|\partial_x \psi(x_1, y) - \partial_x \psi(x_2, y)| \leq L|x_1 - x_2|.$$

Assumption \mathbf{A}'_3 implies \mathbf{A}_3 , but the converse is not true. To see this, just note that, in the smooth situation \mathbf{A}'_3 , we have $\xi(x, y) = \partial_x \psi(x, y)$. Therefore,

$$\mathbb{E}(\xi(x_1, Y) | X) = \int \partial_x \psi(x_1, Y) \mu_{Y|X}(dy),$$

where $\mu_{Y|X}$ is the conditional distribution of Y given X . Assumption \mathbf{A}_3 (or \mathbf{A}'_3) plays a key role in controlling the decrease of the risk functional along the boosting iterations, as can be seen very clearly in Lemmas 1 and 2. This type of Lipschitz hypothesis is classical in the optimization literature (e.g., Bubeck 2015). We also note that, in the context of \mathbf{A}'_3 , the functional C is differentiable at any $F \in L^2(\mu_X)$ in the direction $G \in L^2(\mu_X)$, with differential

$$dC(F; G) = \langle \nabla C(F), G \rangle_{\mu_X},$$

where $\nabla C(F)(x) := \int \partial_x \psi(F(x), y) \mu_{Y|X=x}(dy)$ is the gradient of C at F . However, Assumption \mathbf{A}_3 allows to deal with a larger variety of losses, including non-differentiable ones, as shown in the examples below.

2.2 Some Examples

Each of the loss functions that we discuss in this subsection corresponds to a machine learning algorithm, as thoroughly explained in Bühlmann and Hothorn (2007), Sect. 3. We refer to this article for more properties of these losses and for issues regarding their practical implementation.

- A first canonical example, in the regression setting, is to let $\psi(x, y) = (y - x)^2$ (squared error loss), which is 2-strongly convex in its first argument (Assumption \mathbf{A}_2) and satisfies Assumption \mathbf{A}_1 as soon as $\mathbb{E}Y^2 < \infty$. It also satisfies \mathbf{A}'_3 , with $\partial_x \psi(x, y) = 2(x - y)$ and $L = 2$.
- Another example in regression is the loss $\psi(x, y) = |y - x|$ (absolute error loss), which is convex but not strongly convex in its first argument. Whenever strong convexity of the loss is required, a possible strategy is to regularize the objective via an L^2 -type penalty, and take

$$\psi(x, y) = |y - x| + \gamma x^2,$$

where γ is a positive parameter (possibly function of the sample size n in the empirical setting). This loss is (2γ) -strongly convex in x and satisfies \mathbf{A}_1 and \mathbf{A}_2 whenever $\mathbb{E}|Y| < \infty$, with $\xi(x, y) = \text{sgn}(x - y) + 2\gamma x$ (with $\text{sgn}(u) = 2\mathbb{1}_{\{u>0\}} - 1$ for $u \neq 0$ and $\text{sgn}(0) = 0$). On the other hand, the function $\psi(\cdot, y)$ is not differentiable at y , so that the smoothness Assumption \mathbf{A}'_3 is not satisfied. However,

$$\begin{aligned} \mathbb{E}(\xi(x_1, Y) - \xi(x_2, Y) | X) &= \int (\text{sgn}(x_1 - y) - \text{sgn}(x_2 - y)) \mu_{Y|X}(dy) + 2\gamma(x_1 - x_2) \\ &= \mu_{Y|X}((-\infty, x_1)) - \mu_{Y|X}((-\infty, x_2)) + 2\gamma(x_1 - x_2) \\ &\quad - \mu_{Y|X}((x_1, \infty)) + \mu_{Y|X}((x_2, \infty)). \end{aligned}$$

Thus, if we assume for example that $\mu_{Y|X}$ has a density (with respect to the Lebesgue measure) bounded by B , then

$$|\mathbb{E}(\xi(x_1, Y) - \xi(x_2, Y) | X)| \leq 2(B + \gamma)|x_1 - x_2|,$$

and Assumption \mathbf{A}_3 is therefore satisfied. Of course, in the empirical setting, assuming that $\mu_{Y|X}$ has a density precludes the use of the empirical measure μ_n for $\mu_{X,Y}$. A safe and simple alternative is to consider a smoothed version $\tilde{\mu}_n$ of μ_n (based,

for example, on a kernel estimate; see Devroye and Györfi [1985](#)), and to minimize the functional

$$C_n(F) = \int |y - F(x)| \tilde{\mu}_n(dx, dy) + \gamma \int F(x)^2 \tilde{\mu}_n(dx)$$

over the linear combinations of functions in \mathcal{F} .

- In the ± 1 -classification problem, the final classification rule is $+1$ if $F(x) > 0$ and -1 otherwise. Often, the function $\psi(x, y)$ has the form $\phi(yx)$, where $\phi : \mathbb{R} \rightarrow \mathbb{R}_+$ is convex. Classical losses include the choices $\phi(u) = \ln_2(1 + e^{-u})$ (logit loss), $\phi(u) = e^{-u}$ (exponential loss), and $\phi(u) = \max(1 - u, 0)$ (hinge loss). None of these losses is strongly convex, but here again, this can be repaired whenever needed by regularizing the problem via

$$\psi(x, y) = \phi(yx) + \gamma x^2, \tag{3}$$

where $\gamma > 0$. It is, for example, easy to see that $\psi(x, y) = \ln_2(1 + e^{-yx}) + \gamma x^2$ satisfies Assumptions \mathbf{A}_1 , \mathbf{A}_2 , and \mathbf{A}'_3 . This is also true for the penalized sigmoid loss $\psi(x, y) = (1 - \tanh(\beta yx)) + \gamma x^2$, where β is a positive parameter. In this case, $\psi(\cdot, y)$ is $2(\gamma - \beta^2)$ -strongly convex as soon as $\beta < \sqrt{\gamma}$. Another interesting example in the classification setting is the loss $\psi(x, y) = \phi(yx) + \gamma x^2$, where

$$\phi(u) = \begin{cases} -u + 1 & \text{if } u \leq 0 \\ e^{-u} & \text{if } u > 0. \end{cases}$$

We leave it as an easy exercise to prove that Assumptions \mathbf{A}_1 , \mathbf{A}_2 , and \mathbf{A}'_3 are satisfied. Examples could be multiplied endlessly, but the point we wish to make is that our assumptions are mild and allow considering a large variety of learning problems. We also emphasize that regularized objectives of the form [\(3\)](#) are typically in action in the Extreme Gradient Boosting system of Chen and Guestrin [\(2016\)](#).

2.3 Two Algorithms

Let $\text{lin}(\mathcal{F})$ be the set of all linear combinations of functions in \mathcal{F} , our collection of base predictors in $L^2(\mu_X)$. So, each $F \in \text{lin}(\mathcal{F})$ has the form $F = \sum_{j=1}^J \beta_j f_j$, where $(\beta_1, \dots, \beta_J) \in \mathbb{R}^J$ and f_1, \dots, f_J are elements of \mathcal{F} . Finding the infimum of the functional C over $\text{lin}(\mathcal{F})$ is a challenging infinite-dimensional optimization problem, which requires an algorithm. The core idea of the gradient boosting approach is to greedily locate the infimum by producing a combination of base predictors via a gradient-descent-type algorithm in $L^2(\mu_X)$. Focusing on the basics, this can be achieved by two related yet different strategies, which we examine in greater

mathematical details below. Algorithm 1 appears in Mason et al. (2000), whereas Algorithm 2 is essentially due to Friedman (2001).

It is implicitly assumed throughout this paragraph that Assumption \mathbf{A}_1 is satisfied. We recall that under this assumption, the convex functional C is locally bounded and therefore continuous. Thus, in particular,

$$\inf_{F \in \text{lin}(\mathcal{F})} C(F) = \inf_{F \in \overline{\text{lin}(\mathcal{F})}} C(F),$$

where $\overline{\text{lin}(\mathcal{F})}$ is the closure of $\text{lin}(\mathcal{F})$ in $L^2(\mu_X)$. Loosely speaking, looking for the infimum of C over $\overline{\text{lin}(\mathcal{F})}$ is the same as looking for the infimum of C over all (finite) linear combinations of base functions in \mathcal{F} . We note in addition that if Assumption \mathbf{A}_2 is satisfied, then there exists a unique function $\bar{F} \in \overline{\text{lin}(\mathcal{F})}$ (which we call the boosting predictor) such that

$$C(\bar{F}) = \inf_{F \in \overline{\text{lin}(\mathcal{F})}} C(F). \quad (4)$$

Algorithm 1. In this approach, we consider a class \mathcal{F} of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that $0 \in \mathcal{F}$, $f \in \mathcal{F} \Leftrightarrow -f \in \mathcal{F}$, and $\|f\|_{\mu_X} = 1$ for $f \neq 0$. An example is the collection \mathcal{F} of all ± 1 -binary trees in \mathbb{R}^d using axis parallel cuts with k terminal nodes (plus zero). Each nonzero $f \in \mathcal{F}$ takes the form $f = \sum_{j=1}^k \beta_j \mathbb{1}_{A_j}$, where $|\beta_j| = 1$ and A_1, \dots, A_k is a tree-structured partition of \mathbb{R}^d (Devroye et al. 1996, Chap. 20). The parameter k is a measure of the tree complexity. For example, trees with $k = d + 1$ are such that $\overline{\text{lin}(\mathcal{F})} = L^2(\mu_X)$ (Breiman 2000). Thus, in this case,

$$\inf_{F \in \overline{\text{lin}(\mathcal{F})}} C(F) = \inf_{F \in L^2(\mu_X)} C(F).$$

Although interesting from the point of view of numerical optimization, this situation is however of little interest for statistical learning, as we will see in Sect. 4.

Suppose now that we have a function $F \in \overline{\text{lin}(\mathcal{F})}$ and wish to find a new $f \in \mathcal{F}$ to add to F so that the risk $C(F + wf)$ decreases at most, for some small value of w . Viewed in function space terms, we are looking for the direction $f \in \mathcal{F}$ such that $C(F + wf)$ most rapidly decreases. Assume for the moment, to simplify, that ψ is continuously differentiable in its first argument. Then the knee-jerk reaction is to take the opposite of the gradient of C at F , but since we are restricted to choosing our new function in \mathcal{F} , this will in general not be a possible choice. Thus, instead, we start from the approximate identity

$$C(F) - C(F + wf) \approx -w \langle \nabla C(F), f \rangle_{\mu_X} \quad (5)$$

and choose $f \in \mathcal{F}$ that maximizes $-\langle \nabla C(F), f \rangle_{\mu_X}$. For an arbitrary (i.e., not necessarily differentiable) ψ , we simply replace the gradient by a subgradient and choose $f \in \mathcal{F}$ that maximizes $-\mathbb{E} \xi(F(X), Y) f(X)$. This motivates the following iterative algorithm:

Gradient Boosting Algorithm 1

1: **Require** $(w_t)_t$ a sequence of positive real numbers.

2: **Set** $t = 0$ and start with $F_0 \in \mathcal{F}$.

3: **Compute**

$$f_{t+1} \in \arg \max_{f \in \mathcal{F}} -\mathbb{E}\xi(F_t(X), Y)f(X) \quad (6)$$

and **let** $F_{t+1} = F_t + w_{t+1}f_{t+1}$.

4: **Take** $t \leftarrow t + 1$ and **go** to step 3.

(Throughout the article, it is assumed to simplify that maximizers as in (6) exist. This requirement can be avoided, for example, by working with approximate ε_t -maximizers, as long as the quality of the approximation ε_t is controlled. This essentially adds technical terms to the equations, without adding much to the general picture.) We emphasize that the method performs a gradient-type descent in the function space $L^2(\mu_X)$. At each iteration, it chooses a base predictor to include in the combination. This predictor is chosen so as to maximally reduce the value of the risk functional. However, the main difference with a standard gradient descent is that Algorithm 1 forces the descent direction to belong to \mathcal{F} . To understand the rationale behind this principle, assume that ψ is continuously differentiable in its first argument. As we have seen earlier, in this case,

$$-\mathbb{E}\xi(F_t(X), Y)f(X) = -\langle \nabla C(F_t), f \rangle_{\mu_X},$$

and, for $\nabla C(F_t) \neq 0$,

$$\frac{-\nabla C(F_t)}{\|\nabla C(F_t)\|_{\mu_X}} = \arg \max_{F \in L^2(\mu_X): \|F\|_{\mu_X}=1} -\langle \nabla C(F_t), F \rangle_{\mu_X}.$$

Thus, at each step, Algorithm 1 mimics the computation of the negative gradient by restricting the search of the supremum to the class \mathcal{F} , i.e., by taking

$$f_{t+1} \in \arg \max_{f \in \mathcal{F}} -\langle \nabla C(F_t), f \rangle_{\mu_X},$$

which is exactly (6). In the empirical case (i.e., $\mu_{X,Y} = \mu_n$), this descent step takes the form

$$f_{t+1} \in \arg \max_{f \in \mathcal{F}} -\frac{1}{n} \sum_{i=1}^n \nabla C(F_t)(X_i) \cdot f(X_i).$$

Finding this optimum is a non-trivial computational problem, which necessitates a strategy. For example, in the spirit of the CART algorithm of Breiman et al. (1984), Chen and Guestrin (2016) use in the XGBoost package a greedy approach that starts from a single leaf and iteratively adds branches to the tree.

The sequence $(w_t)_t$ is the sequence of step sizes, which are allowed to change at every iteration and should be carefully chosen for convergence guarantees. It is also stressed that the algorithm is assumed to be run forever, i.e., stopping or not the

iterations is not an issue at this stage of the analysis. As we will see in the next section, the algorithm is convergent under our assumptions (with an appropriate choice of the sequence $(w_t)_t$), in the sense that

$$\lim_{t \rightarrow \infty} C(F_t) = \inf_{F \in \text{lin}(\mathcal{F})} C(F).$$

Of course, in the empirical case, the statistical properties as $n \rightarrow \infty$ of the limit deserve a special treatment, connected with possible overfitting issues. This important discussion is postponed to Sect. 4.

Algorithm 2. The principle we used so far rests upon the simple Taylor-like identity (5), which encourages us to imitate the definition of the negative gradient in the class \mathcal{F} . Still starting from (5), there is however another strategy, maybe more natural, which consists in choosing f_{t+1} by a least squares approximation of $-\xi(F_t(X), Y)$. To follow this route, we modify a bit the collection of weak learners, and consider a class $\mathcal{P} \subset L^2(\mu_X)$ of functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that $f \in \mathcal{P} \Leftrightarrow -f \in \mathcal{P}$, and $af \in \mathcal{P}$ for all $(a, f) \in \mathbb{R} \times \mathcal{P}$ (in particular, $0 \in \mathcal{P}$, which is thus a cone of $L^2(\mu_X)$). Binary trees in \mathbb{R}^d using axis parallel cuts with k terminal nodes are a good example of a possible class \mathcal{P} . These base learners are of the form $f = \sum_{j=1}^k \beta_j \mathbb{1}_{A_j}$, where this time $(\beta_1, \dots, \beta_k) \in \mathbb{R}^k$, without any normative constraint.

Given F_t , the idea of Algorithm 2 is to choose $f_{t+1} \in \mathcal{P}$ that minimizes the squared norm between $-\xi(F_t(X), Y)$ and $f_{t+1}(X)$, i.e., to let

$$f_{t+1} \in \arg \min_{f \in \mathcal{P}} \mathbb{E}(-\xi(F_t(X), Y) - f(X))^2,$$

or, equivalently,

$$f_{t+1} \in \arg \min_{f \in \mathcal{P}} (2\mathbb{E}\xi(F_t(X), Y)f(X) + \|f\|_{\mu_X}^2).$$

A more algorithmic format is shown below.

Gradient Boosting Algorithm 2

1: **Require** ν a positive real number.

2: **Set** $t = 0$ and start with $F_0 \in \mathcal{P}$.

3: **Compute**

$$f_{t+1} \in \arg \min_{f \in \mathcal{P}} (2\mathbb{E}\xi(F_t(X), Y)f(X) + \|f\|_{\mu_X}^2) \tag{7}$$

and let $F_{t+1} = F_t + \nu f_{t+1}$.

4: **Take** $t \leftarrow t + 1$ and **go** to step 3.

We note that, contrary to Algorithm 1, the step size ν is kept fixed during the iterations. We will see in the next section that choosing a small enough ν (depending in particular on the Lipschitz constant of Assumption **A**₃) is sufficient to ensure the convergence of the algorithm. In the empirical setting, assuming that ψ is continuously differentiable in its first argument, the optimization step (7) reads

$$f_{t+1} \in \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (-\nabla C(F_t)(X_i) - f(X_i))^2.$$

Therefore, in this context, the gradient boosting algorithm fits f_{t+1} to the negative gradient instances $-\nabla C(F_t)(X_i)$ via a least squares minimization. When $\psi(x, y) = (y - x)^2/2$, then $-\nabla C(F_t)(X_i) = Y_i - F_t(X_i)$, and the algorithm simply fits f_{t+1} to the residuals $Y_i - F_t(X_i)$ at step t , in the spirit of original boosting procedures. This observation is at the source of gradient boosting, which Algorithm 2 generalizes to a much larger variety of loss functions and to more abstract measures.

3 Convergence of the Algorithms

This section is devoted to analyzing the convergence of the gradient boosting Algorithms 1 and 2 as the number of iterations t tends to infinity. Despite its importance, no results (or only partial answers) have been reported so far on this question.

3.1 Algorithm 1

The convergence of this algorithm rests upon the choice of the step size sequence $(w_t)_t$, which needs to be carefully specified. We take $w_0 > 0$ arbitrarily and set

$$w_{t+1} = \min(w_t, -(2L)^{-1} \mathbb{E} \xi(F_t(X), Y) f_{t+1}(X)), \quad t \geq 0, \quad (8)$$

where L is the Lipschitz constant of Assumption **A**₃. Clearly, the sequence $(w_t)_t$ is nonincreasing. It is also nonnegative. To see this, just note that, by definition,

$$f_{t+1} \in \arg \max_{f \in \mathcal{F}} -\mathbb{E} \xi(F_t(X), Y) f(X),$$

and thus, since $0 \in \mathcal{F}$, $-\mathbb{E} \xi(F_t(X), Y) f_{t+1}(X) \geq 0$. The main result of this section is encapsulated in the following theorem.

Theorem 1 *Assume that Assumptions **A**₁ and **A**₃ are satisfied, and let $(F_t)_t$ be defined by Algorithm 1 with $(w_t)_t$ as in (8). Then*

$$\lim_{t \rightarrow \infty} C(F_t) = \inf_{F \in \text{lin}(\mathcal{F})} C(F).$$

Proof See Supplementary Material Document. □

Observe that Theorem 1 holds without Assumption **A**₂, i.e., there is no need here to assume that the function $\psi(x, y)$ is strongly convex in x . However, whenever Assumption **A**₂ is satisfied, there exists as in (4) a unique boosting predictor

$\bar{F} \in \overline{\text{lin}(\mathcal{F})}$ such that $C(\bar{F}) = \inf_{F \in \text{lin}(\mathcal{F})} C(F)$, and the theorem guarantees that $\lim_{t \rightarrow \infty} C(F_t) = C(\bar{F})$.

The proof of the theorem relies on the following lemma, which states that the sequence $(C(F_t))_t$ is nonincreasing. Since $C(F)$ is nonnegative for all F , we conclude that $C(F_t) \downarrow \inf_k C(F_k)$ as $t \rightarrow \infty$. This is the key argument to prove the convergence of $C(F_t)$ toward $\inf_{F \in \text{lin}(\mathcal{F})} C(F)$.

Lemma 1 *Assume that Assumptions \mathbf{A}_1 and \mathbf{A}_3 are satisfied. Then, for each $t \geq 0$,*

$$C(F_t) - C(F_{t+1}) \geq Lw_{t+1}^2.$$

In particular, $C(F_t) \downarrow \inf_k C(F_k)$ as $t \rightarrow \infty$, $\sum_{t \geq 1} w_t^2 < \infty$, and $\lim_{t \rightarrow \infty} w_t = 0$.

Proof Let $t \geq 0$. Recall that $F_{t+1} = F_t + w_{t+1}f_{t+1}$. If $f_{t+1} = 0$, then $w_{t+1} = 0$ and $F_{t+1} = F_t$, so that there is nothing to prove. Thus, in the remainder of the proof, it is assumed that f_{t+1} is different from zero and, in turn, that $\|f_{t+1}\|_{\mu_X} = 1$. Applying technical Lemma 1 of the Supplementary Material Document, we may write

$$\begin{aligned} C(F_t) &\geq C(F_{t+1}) - w_{t+1}^2 L - w_{t+1} \mathbb{E} \xi(F_t(X), Y) f_{t+1}(X) \\ &\geq C(F_{t+1}) - w_{t+1}^2 L + 2Lw_{t+1} \min(w_t, -(2L)^{-1} \mathbb{E} \xi(F_t(X), Y) f_{t+1}(X)) \\ &= C(F_{t+1}) + Lw_{t+1}^2, \end{aligned}$$

by definition (8) of the sequence $(w_t)_t$. □

Theorem 1 ensures that the risk of the boosting iterates gets closer and closer to the minimal risk as the number of iterations grows. It turns out that, whenever $\overline{\text{lin}(\mathcal{F})} = L^2(\mu_X)$, under Assumption \mathbf{A}_2 and the smooth framework of Assumption \mathbf{A}'_3 , the sequence $(F_t)_t$ itself approaches $\bar{F} = \arg \min_{F \in L^2(\mu_X)} C(F)$, as shown in Corollary 1 below. This corollary is an easy consequence of Theorem 1 and the strong convexity of C .

Corollary 1 *Assume that $\overline{\text{lin}(\mathcal{F})} = L^2(\mu_X)$. Assume, in addition, that Assumptions \mathbf{A}_1 , \mathbf{A}_2 , and \mathbf{A}'_3 are satisfied, and let $(F_t)_t$ be defined by Algorithm 1 with $(w_t)_t$ as in (8). Then*

$$\lim_{t \rightarrow \infty} \|F_t - \bar{F}\|_{\mu_X} = 0,$$

where

$$\bar{F} = \arg \min_{F \in L^2(\mu_X)} C(F).$$

Proof By the α -strong convexity of C ,

$$C(F_t) \geq C(\bar{F}) + \mathbb{E} \xi(\bar{F}, Y)(F_t - \bar{F}) + \frac{\alpha}{2} \|F_t - \bar{F}\|_{\mu_X}^2,$$

which, under \mathbf{A}'_3 , takes the more familiar form

$$C(F_t) \geq C(\bar{F}) + \langle \nabla C(\bar{F}), F_t - \bar{F} \rangle_{\mu_X} + \frac{\alpha}{2} \|F_t - \bar{F}\|_{\mu_X}^2.$$

But, since $\bar{F} = \arg \min_{F \in L^2(\mu_X)} C(F)$, we know that $\langle \nabla C(\bar{F}), F_t - \bar{F} \rangle_{\mu_X} = 0$. Thus,

$$C(F_t) - C(\bar{F}) \geq \frac{\alpha}{2} \|F_t - \bar{F}\|_{\mu_X}^2,$$

and the conclusion follows from Theorem 1. \square

We would like to close this subsection by stressing that Theorem 1 is not quantitative, in the sense that nothing is known about the speed of convergence when increasing the number of iterations of the algorithm. This is an open question, which unfortunately cannot be dealt with in the present article. In line with the remarks of a referee, we believe that the existing analyses for L_2 Boosting (e.g., Bühlmann 2006) and weak greedy algorithms (e.g., Temlyakov 2000; Champion et al. 2014) could be a promising route to follow.

3.2 Algorithm 2

Recall that, in this context, each iteration picks an $f_{t+1} \in \mathcal{P}$ that satisfies

$$2\mathbb{E}\xi(F_t(X), Y) f_{t+1}(X) + \|f_{t+1}\|_{\mu_X}^2 \leq 2\mathbb{E}\xi(F_t(X), Y) f(X) + \|f\|_{\mu_X}^2 \quad \text{for all } f \in \mathcal{P}.$$

Theorem 2 *Assume that Assumptions \mathbf{A}_1 – \mathbf{A}_3 are satisfied, and let $(F_t)_t$ be defined by Algorithm 2 with $0 < \nu < 1/(2L)$. Then*

$$\lim_{t \rightarrow \infty} C(F_t) = \inf_{F \in \text{lin}(\mathcal{P})} C(F).$$

Proof See Supplementary Material Document. \square

The architecture of the proof is similar to that of Theorem 1. (Note however that this theorem requires the strong convexity Assumption \mathbf{A}_2). In particular, we need the following important lemma, which states that the risk of the iterates decreases at each step of the algorithm.

Lemma 2 *Assume that Assumptions \mathbf{A}_1 and \mathbf{A}_3 are satisfied, and let $0 < \nu < 1/(2L)$. Then, for each $t \geq 0$,*

$$C(F_t) - C(F_{t+1}) \geq \frac{\nu}{2} (1 - 2\nu L) \|f_{t+1}\|_{\mu_X}^2.$$

In particular, $C(F_t) \downarrow \inf_k C_k$ as $t \rightarrow \infty$, $\sum_{t \geq 1} \|f_t\|_{\mu_X}^2 < \infty$, and $\lim_{t \rightarrow \infty} \|f_t\|_{\mu_X} = 0$.

Proof Let $t \geq 0$. Applying technical Lemma 1 of the Supplementary Material Document, we may write

$$\begin{aligned} C(F_t) &\geq C(F_{t+1}) - \nu^2 L \|f_{t+1}\|_{\mu_X}^2 - \nu \mathbb{E} \xi(F_t(X), Y) f_{t+1}(X) \\ &= C(F_{t+1}) - \nu^2 L \|f_{t+1}\|_{\mu_X}^2 - \frac{\nu}{2} (2\mathbb{E} \xi(F_t(X), Y) f_{t+1}(X) + \|f_{t+1}\|_{\mu_X}^2) + \frac{\nu}{2} \|f_{t+1}\|_{\mu_X}^2. \end{aligned}$$

Upon noting that $2\mathbb{E} \xi(F_t(X), Y) f_{t+1}(X) + \|f_{t+1}\|_{\mu_X}^2 \leq 0$ (since $0 \in \mathcal{P}$), we conclude that

$$C(F_t) \geq C(F_{t+1}) + \frac{\nu}{2} (1 - 2\nu L) \|f_{t+1}\|_{\mu_X}^2.$$

□

Remark 1 The parameter ν can be regarded as controlling the learning rate of the boosting procedure. The lower bound of Lemma 2 suggests the optimal value $\nu^* = 1/(4L)$. In practice, ν is often chosen “small enough”, which leads to a larger number of iterations (and thus more computing time) for the same training risk. All in all, both ν and the number of iterations control prediction risk and these parameters do not operate independently.

As in Algorithm 1, the sequence $(F_t)_t$ approaches $\bar{F} = \arg \min_{F \in L^2(\mu_X)} C(F)$, provided $\overline{\text{lin}(\mathcal{P})} = L^2(\mu_X)$ and \mathbf{A}'_3 is satisfied in place of \mathbf{A}_3 . This is summarized in the following corollary. Its proof is similar to the proof of Corollary 1 and is therefore omitted.

Corollary 2 Assume that $\overline{\text{lin}(\mathcal{P})} = L^2(\mu_X)$. Assume, in addition, that Assumptions \mathbf{A}_1 , \mathbf{A}_2 , and \mathbf{A}'_3 are satisfied, and let $(F_t)_t$ be defined by Algorithm 2 with $0 < \nu < 1/(2L)$. Then

$$\lim_{t \rightarrow \infty} \|F_t - \bar{F}\|_{\mu_X} = 0,$$

where

$$\bar{F} = \arg \min_{F \in L^2(\mu_X)} C(F).$$

Theorem 1/Corollary 1 and Theorem 2/Corollary 2 guarantee that, under appropriate assumptions, Algorithms 1 and 2 converge toward the infimum of the risk functional. Given the unusual form of these algorithms, which have the flavor of gradient descents while being different, these results are all but obvious and cannot be deduced from general optimization principles. As far as we know, they are novel in the gradient boosting literature and extend our understanding of the approach.

Perhaps the most natural framework of Algorithms 1 and 2 is when $\mu_{X,Y} = \mu_n$, the empirical measure. In this statistical context, both algorithms track the infimum of the empirical risk functional $C_n(F) = \frac{1}{n} \sum_{i=1}^n \psi(F(X_i), Y_i)$ over the linear combinations of weak learners in \mathcal{F} (Algorithm 1) or in \mathcal{P} (Algorithm 2). This task is achieved by sequentially constructing linear combinations of base learners, of the

form $F_t = F_0 + \sum_{k=1}^t w_k f_k$ with $f_k \in \mathcal{F}$ for Algorithm 1, and $F_t = F_0 + \nu \sum_{k=1}^t f_k$ with $f_k \in \mathcal{P}$ for Algorithm 2. We stress that, in the empirical case, the boosted iterates F_t and their eventual limit \bar{F}_n are measurable functions of the data set \mathcal{D}_n . That being said, Theorem 1 and Theorem 2 are numerical-analysis-type results, which do not provide information on the statistical properties of the boosting predictor \bar{F}_n . From this point of view, more or less catastrophic situations can happen, depending on the “size” of $\text{lin}(\mathcal{F})$ (Algorithm 1) or $\text{lin}(\mathcal{P})$ (Algorithm 2), which should not be neither too small (to catch complex decisions) nor excessively large (to avoid overfitting).

To be convinced of this, consider for example Algorithm 1 with $\psi(x, y) = (y - x)^2$ (least squares regression problem) and $\mathcal{F} =$ all binary trees with $d + 1$ leaves. Denote by P_n the empirical measure based on the X_i only, $1 \leq i \leq n$. Then, by Theorem 1, $\lim_{t \rightarrow \infty} C_n(F_t) = C_n(\bar{F}_n)$, where

$$\bar{F}_n = \arg \min_{F \in L^2(P_n)} C_n(F).$$

Assume, to simplify, that all X_i are different. It is then easy to see that the boosting predictor \bar{F}_n takes the value Y_i at each X_i and is arbitrarily defined elsewhere. Of course, in general, such a function \bar{F}_n does not converge as $n \rightarrow \infty$ toward the regression function $F^*(x) = \mathbb{E}(Y|X = x)$, and this is a typical situation where the gradient boosting algorithms overfit. The overfitting issue of boosting procedures has been recognized for a long time, and various approaches have been proposed to combat it, in particular via early stopping (that is, stopping the iterations before convergence; see, e.g., Bühlmann and Yu 2003; Mannor et al. 2003; Zhang and Yu 2005; Bickel et al. 2006; Bartlett and Traskin 2007).

Nevertheless, the natural question we would like to answer is whether there exists a reasonable context in which the boosting predictors enjoy good statistical properties as the sample size grows, without resorting to any stopping strategy. The next section provides a positive response. The major constraint we face, imposed by the gradient-descent nature of the algorithms, is that we are required to perform a minimization over a vector space ($\text{lin}(\mathcal{F})$ for Algorithm 1 and $\text{lin}(\mathcal{P})$ for Algorithm 2). In particular, there is no question of imposing constraints on the coefficients of the linear combinations, which, for example, cannot reasonably be assumed to be bounded. As we will see, the trick is to carefully constrain the “complexity” of the vector spaces $\text{lin}(\mathcal{F})$ or $\text{lin}(\mathcal{P})$ in a manner compatible with the algorithms. The second message is the importance of having a strongly convex risk functional to minimize, which, in some way, restrict the norm of the sequence $(F_t)_{t \geq 0}$ of boosted iterates. As we have pointed out several times, if the loss function is not natively strongly convex in its first argument, then this type of regularization can be achieved by resorting to an L^2 -type penalty.

4 Large Sample Properties

We consider in this section a functional minimization problem whose solution can be computed by gradient boosting and enjoys non-trivial statistical properties. The context and notation are similar to that of the previous sections, but must be slightly adapted to fit this new framework.

For simplicity, it will be assumed throughout that \mathcal{X} is a compact subset of \mathbb{R}^d . We consider i.i.d. data $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ taking values in $\mathcal{X} \times \mathcal{Y}$, and let P_n be the empirical measure based on the X_i only, $1 \leq i \leq n$. We denote by P the common distribution of the X_i and assume that P has a density g with respect to the Lebesgue measure λ on \mathbb{R}^d , with

$$0 < \inf_{\mathcal{X}} g \leq \sup_{\mathcal{X}} g < \infty.$$

We concentrate on Algorithm 1 and take as weak learners a finite class \mathcal{F}_n of simple functions on \mathcal{X} with ± 1 values, which may possibly vary with the sample size n . It is actually easy to verify that all subsequent results are valid for Algorithm 2 by letting $\mathcal{P}_n = \{\lambda f : f \in \mathcal{F}_n, \lambda \in \mathbb{R}\}$.

The typical example we have in mind for \mathcal{F}_n is a finite class of binary trees using axis parallel cuts with k leaves. Of course, the parameter k has to be carefully chosen as a function of the sample size to guarantee consistency, as we will see below. The fact that the class \mathcal{F}_n is supposed to be finite should not be too disturbing, since in practice the optimization step (6) is typically performed over a finite family of functions. This is for example the case when a CART-style top-down recursive partitioning is used to compute the minimum at each iteration of the algorithm. In this approach, the optimal tree in (6) is greedily searched for by passing from one level of the node to the next one with cuts that are located between two data points. So, even though the collection \mathcal{F}_n may be very large, it is nevertheless fair to assume that its cardinal is finite.

As before, it is assumed that the identically zero function belongs to \mathcal{F}_n . So, in this framework, we see that there exists a (large) integer $N = N(n) \geq 1$ and a partition of \mathcal{X} into measurable subsets A_j^n , $1 \leq j \leq N$, such that any $F \in \text{lin}(\mathcal{F}_n)$ takes the form $F = \sum_{j=1}^N \alpha_j \mathbb{1}_{A_j^n}$, where $(\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N$. To avoid pathological situations, we assume that there exists a positive sequence $(v_n)_n$ such that $\min_{1 \leq j \leq N} \lambda(A_j^n) \geq v_n$. Of course, it is supposed that $N \rightarrow \infty$ as n tends to infinity.

We let $\phi : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ be a loss function, assumed to be convex in its first argument and to satisfy $\bar{\phi} := \sup_{y \in \mathcal{Y}} \phi(0, y) < \infty$. In line with the previous sections, we are interested in minimizing over $\text{lin}(\mathcal{F}_n)$ the empirical risk functional $C_n(F)$ defined by

$$C_n(F) = \frac{1}{n} \sum_{i=1}^n \psi(F(X_i), Y_i),$$

where $\psi(x, y) = \phi(x, y) + \gamma_n x^2$ and $(\gamma_n)_n$ is a sequence of positive parameters such that $\lim_{n \rightarrow \infty} \gamma_n = 0$. (Note that γ_n depends only on n and is therefore kept fixed during the iterations of the algorithm.) Put differently,

$$C_n(F) = A_n(F) + \gamma_n \|F\|_{P_n}^2, \quad (9)$$

where

$$A_n(F) = \frac{1}{n} \sum_{i=1}^n \phi(F(X_i), Y_i).$$

Assumption **A₁** is obviously satisfied (with $\mu_{X,Y} = \mu_n$, in the notation of Sect. 3), and the same is true for Assumption **A₂** by the α -strong convexity of the function $\psi(\cdot, y)$ for each fixed y , with α independent of y .

Remark 2 If the function $\phi(\cdot, y)$ is natively α -strongly convex with a parameter α independent of y , then we may consider the simpler problem of minimizing the functional $A_n(F)$. Indeed, in this case there is no need to resort to the $\gamma_n \|F\|_{P_n}^2$ penalty term since Lemma 3 of the Supplementary Material Document allows to bound $\|F\|_{P_n}^2$. As we have seen in Sect. 2, this is for example the case in the least squares problem, when $\phi(x, y) = (y - x)^2$. However, to keep a sufficient degree of generality, we will consider in the following the more general optimization problem (9).

Now, let

$$\bar{F}_n = \arg \min_{F \in \text{lin}(\mathcal{F}_n)} C_n(F).$$

We have learned in Theorem 1 that whenever Assumption **A₃** is satisfied, the boosted iterates $(F_t)_t$ of Algorithm 1 satisfy $\lim_{t \rightarrow \infty} C_n(F_t) = C_n(\bar{F}_n)$, i.e.,

$$\lim_{t \rightarrow \infty} (A_n(F_t) + \gamma_n \|F_t\|_{P_n}^2) = A_n(\bar{F}_n) + \gamma_n \|\bar{F}_n\|_{P_n}^2.$$

For $F \in L^2(P)$, the population counterpart of $A_n(F)$ is the convex functional $A(F) := \mathbb{E}\phi(F(X_1), Y_1)$, which is assumed to be locally bounded, and thus continuous. Throughout, we denote by F^* a minimizer of $A(F)$ over $L^2(P)$, i.e.,

$$F^* \in \arg \min_{F \in L^2(P)} A(F).$$

We have for example $F^*(x) = \mathbb{E}(Y|X = x)$ in the regression problem with $\phi(x, y) = (y - x)^2$ and $F^*(x) = \log(\frac{\eta(x)}{1-\eta(x)})$ in the classification problem with $\phi(x, y) = \log_2(1 + e^{-yx})$, where $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

Our goal in this section is to investigate the large sample properties of \bar{F}_n , i.e., to analyze the statistical behavior of the boosting predictor \bar{F}_n as $n \rightarrow \infty$. In particular, a sensible objective is to show that $A(\bar{F}_n)$ gets asymptotically close to the minimal risk $A(F^*)$ as the sample size grows. This necessitates a proof, since all we know for now is that

$$A_n(\bar{F}_n) + \gamma_n \|\bar{F}_n\|_{P_n}^2 - A(F^*) = \inf_{F \in \text{lin}(\mathcal{F}_n)} (A_n(F) + \gamma_n \|F\|_{P_n}^2 - A(F^*)),$$

which is our starting point. The following assumption on ϕ will be needed in the analysis:

A₄ For all $p \geq 0$, there exists a constant $\zeta(p) > 0$ such that, for all $(x_1, x_2, y) \in \mathbb{R}^2 \times \mathcal{Y}$ with $\max(|x_1|, |x_2|) \leq p$,

$$|\phi(x_1, y) - \phi(x_2, y)| \leq \zeta(p)|x_1 - x_2|.$$

It is readily seen that all classical convex losses in regression and classification satisfy this local Lipschitz assumption. Finally, we let $A^n(x) = A_j^n$ whenever $x \in A_j^n$, and, for $E \subset \mathbb{R}^d$,

$$\text{diam}(E) = \sup_{x, x' \in E} \|x - x'\|.$$

Recall that $\bar{\phi} := \sup_{y \in \mathcal{Y}} \phi(0, y) < \infty$.

Theorem 3 Assume that Assumptions **A₃** (with $\psi(x, y) = \phi(x, y) + \gamma_n x^2$) and **A₄** are satisfied, and that F^* is bounded. Assume, in addition, that $\text{diam}(A^n(X)) \rightarrow 0$ in probability as $n \rightarrow \infty$. Then, provided $\gamma_n \rightarrow 0$, $N \rightarrow \infty$, $\frac{\log N}{nv_n} \rightarrow 0$, and

$$\frac{1}{\sqrt{nv_n \gamma_n}} \zeta \left(\sqrt{\frac{2\bar{\phi}}{v_n \gamma_n \inf_{\mathcal{X}} g}} \right) \rightarrow 0,$$

we have $\lim_{n \rightarrow \infty} \mathbb{E}A(\bar{F}_n) = A(F^*)$.

Proof See Supplementary Material Document. □

The main message of this theorem is that, under appropriate conditions on the loss and provided the size of the weak learner classes are judiciously increased, gradient boosting does not overfit. In other words, in this framework, stopping the iterations is not necessary and the algorithms may be run indefinitely, without worrying about early stopping issues.

In line with Remark 2, we leave it as an exercise to prove that if the function $\phi(\cdot, y)$ is already α -strongly convex with a parameter α independent of y , then a similar result holds with the conditions $N \rightarrow \infty$, $\frac{\log N}{nv_n} \rightarrow 0$, and

$$\frac{1}{\sqrt{nv_n}} \zeta \left(\sqrt{\frac{a}{v_n \inf_{\mathcal{X}} g}} \right) \rightarrow 0,$$

where $a = \frac{2}{\alpha} \sup_{y \in \mathcal{Y}} |\xi(0, y)| + \sqrt{2\bar{\phi}/\alpha}$. In this case, we can take $\gamma_n = 0$ (i.e., no penalty) and resort to Lemma 3 of the Supplementary Material Document to bound the quantity $\|F\|_{P_n}^2$.

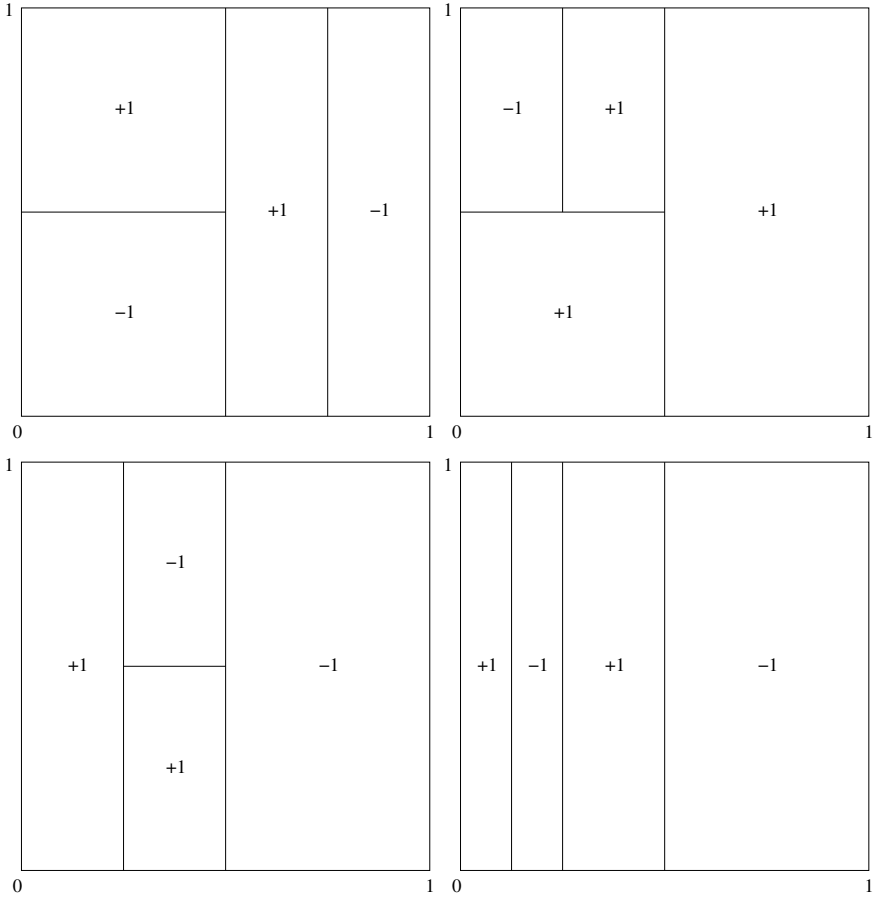


Fig. 1 Four examples of trees in the class \mathcal{F}_n , in dimension $d = 2$, with $k_n = 4$

Next, we point out that the conditions of Theorem 3 are mild and cover a wide variety of losses and possible classes of weak learners. As an example, let $\mathcal{X} = [0, 1]^d$ and take for \mathcal{F}_n the set of all binary trees on $[0, 1]^d$ with k_n leaves, where cuts are perpendicular to the axes and are located at the middle of the cells. Although combinatorially rich, this family of trees is finite (see Fig. 1 for an illustration in dimension $d = 2$).

It is easy to verify that any $F \in \text{lin}(\mathcal{F}_n)$ takes the form $F = \sum_{j=1}^N \alpha_j \mathbb{1}_{A_j^n}$, where $N \leq 2^{dk_n}$ and the A_j^n , $1 \leq j \leq N$, form a regular grid over $[0, 1]^d$. Thus, clearly, $v_n \geq 2^{-dk_n}$. In addition, considering for example the loss $\phi(x, y) = (y - x)^2$, we see that the conditions of Theorem 3 take the simple form

$$k_n \rightarrow \infty, \quad \frac{k_n 2^{dk_n}}{n} \rightarrow 0, \quad \text{and} \quad \frac{2^{dk_n}}{\sqrt{n}} \rightarrow 0.$$

Let us finally note that in the ± 1 -classification setting, each F defines a classifier g_F in a natural way, by

$$g_F(x) = \begin{cases} +1 & \text{if } F(x) > 0 \\ -1 & \text{otherwise,} \end{cases}$$

and the main concern is not the behavior of the theoretical risk $A(F)$ with respect to $A(F^*)$, but rather the proximity between the probability of error $L(g_F) := \mathbb{P}(g_F(X) \neq Y)$ and the Bayes risk $L^* := \inf_{g: \mathcal{X} \rightarrow \{-1,1\}} \mathbb{P}(g(X) \neq Y)$. For most classification losses (Zhang 2004; Bartlett et al. 2006), the difference $L(g_F) - L^*$ is small as long as $A(F) - A(F^*)$ is. In our framework, we conclude that for such well-behaved losses, under the assumptions of Theorem 3,

$$\lim_{n \rightarrow \infty} \mathbb{E}L(g_{\bar{F}_n}) = L^*.$$

Acknowledgements We greatly thank the Editors and two referees for valuable comments and insightful suggestions, which led to a substantial improvement of the paper.

References

- Bartlett, P. L., & Traskin, M. (2007). AdaBoost is consistent. *Journal of Machine Learning Research*, 8, 2347–2368.
- Bartlett, P. L., Jordan, M. I., & McAuliffe, J. D. (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101, 138–156.
- Bickel, P. J., Ritov, Y., & Zakai, A. (2006). Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7, 705–732.
- Blanchard, G., Lugosi, G., & Vayatis, N. (2003). On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4, 861–894.
- Breiman, L. (1997). *Arcing the edge*. Technical Report 486, Statistics Department, University of California, Berkeley.
- Breiman, L. (1998). Arcing classifiers (with discussion). *The Annals of Statistics*, 26, 801–849.
- Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11, 1493–1517.
- Breiman, L. (2000). *Some infinite theory for predictor ensembles*. Technical Report 577, Statistics Department, University of California, Berkeley.
- Breiman, L. (2004). Population theory for boosting ensembles. *The Annals of Statistics*, 32, 1–11.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Boca Raton: Chapman & Hall/CRC Press.
- Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8, 231–357.
- Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, 34, 559–583.
- Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22, 477–505.
- Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Berlin: Springer.
- Bühlmann, P., & Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98, 324–339.

- Champion, M., Cierco-Ayrolles, C., Gadat, S., & Vignes, M. (2014). Sparse regression and support recovery with L_2 -boosting algorithms. *Journal of Statistical Planning and Inference*, 155, 19–41.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York: ACM.
- Devroye, L., & Györfi, L. (1985). *Nonparametric density estimation: The L_1 view*. New York: Wiley.
- Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer.
- Frank, M., & Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3, 95–110.
- Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, 121, 256–285.
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In Lorenza, S. (Ed.) *Machine Learning: Proceedings of the Thirteenth International Conference on Machine Learning*, (pp 148–156). San Francisco: Morgan Kaufmann Publishers.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119–139.
- Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion). *The Annals of Statistics*, 28, 337–407.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29, 1189–1232.
- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38, 367–378.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.
- Lugosi, G., & Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32, 30–55.
- Mallat, S. G., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41, 3397–3415.
- Mannor, S., Meir, R., & Zhang, T. (2003). Greedy algorithms for classification – consistency, convergence rates, and adaptivity. *Journal of Machine Learning Research*, 4, 713–742.
- Mason, L., Baxter, L., Bartlett, P., & Frean, M. (1999). Boosting algorithms as gradient descent. In Solla, S. A., Leen, T. K., Müller, K. (Eds.) *Proceedings of the 12th International Conference on Neural Information Processing Systems* (pp. 512–518). Cambridge, MA: The MIT Press.
- Mason, L., Baxter, J., Bartlett, P., & Frean, M. (2000). Functional gradient techniques for combining hypotheses. In A. J. Smola, P. L. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 221–246). Cambridge, MA: The MIT Press.
- Meir, R., & Rätsch, G. (2003). An introduction to boosting and leveraging. In S. Mendelson & A. J. Smola (Eds.), *Advanced lectures on machine learning: Machine learning summer school 2002* (pp. 118–183). Berlin: Springer.
- Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5, 197–227.
- Temlyakov, V. N. (2000). Weak greedy algorithms. *Advances in Computational Mathematics*, 12, 213–227.
- Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32, 56–85.
- Zhang, T., & Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33, 1538–1579.

Nonparametric Model-Based Estimators for the Cumulative Distribution Function of a Right Censored Variable in a Small Area



Sandrine Casanova and Eve Leconte

Abstract In survey analysis, the estimation of the cumulative distribution function (cdf) is of great interest as it facilitates the derivation of mean/median estimators for both populations and sub-populations (i.e. domains). We focus on small domains and consider the case where the response variable is right censored. Under this framework, we propose a nonparametric model-based estimator that extends the cdf estimator of Casanova (2012) to the censored case: it uses auxiliary information in the form of a continuous covariate and utilizes nonparametric quantile regression. We then employ simulations to compare the constructed estimator with the model-based cdf estimator of Casanova and Leconte (2015) and the Kaplan–Meier estimator (Kaplan and Meier 1958), both of which use only information contained within the domain: the quantile-based estimator performs better than the former two for very small domain sample sizes. Access times to the first job for young female graduates in the Occitania region are used to illustrate the new methodology.

1 Introduction

In survey sampling, the classical literature studies estimation of totals or means, but in many applications, the parameters of interest are more complex: they can be quantiles (see e.g. Rueda et al. 2004) or other non-linear parameters derived from the cumulative distribution function (cdf) of the response variable. We consider the

Dedication *Thank you so much Christine for being there for us, for your kindness, your great energy and your friendship throughout the years as well as for allowing us to progress and fortunately, not to regress, especially nonparametrically!*

S. Casanova (✉) · E. Leconte
TSE-R, Université Toulouse 1 Capitole, 1, Esplanade de l'Université,
31080 Toulouse cedex 06, France
e-mail: sandrine.casanova@tse-fr.eu

E. Leconte
e-mail: eve.leconte@tse-fr.eu

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_3

estimation of the cdf in a possibly small domain when the variable of interest is right censored. This is the case when the response variable is a duration which is observed during a limited period of time, e.g. the access time to the first job for female graduates, which is utilized in the example section. This variable is censored for graduates who have not found a job at the end of the survey. The considered domains are defined by the different types and levels of education.

When the size of the domain is large enough, the estimation of the parameter of interest is based on the sampled individuals of the domain and the resulting direct estimators are sufficiently accurate. However, in many practical applications, sizes of the domain samples are not large enough. In this case, the estimation generally uses auxiliary information of a covariate and, in addition to it, some information is “borrowed” from the other domains.

In small area estimation, the classical technique which captures domain effects is the linear mixed model (see Rao 2003). When the errors are assumed to be Gaussian, estimates of the regression parameter, as well as predictions of random effects, are obtained by maximizing the likelihood function. This leads to the empirical best linear unbiased predictor (EBLUP) of the variable of interest. However, these models are very dependent on strong distributional assumptions such as the normality and homoscedasticity of the error terms. In order to explore a non-linear relationship between the variable of interest and the covariate, Salvati et al. (2010b) have proposed a nonparametric version of the EBLUP for the small area mean using penalized splines. Alternatively, Chambers and Tzavidis (2006) predict the mean in a small area using parametric conditional M-quantiles. A nonparametric estimator of the small area mean using M-quantiles via penalized splines can be found in Salvati et al. (2011).

As far as cdf estimation in a small area is concerned, an estimator based on conditional parametric M-quantiles can also be found in Chambers and Tzavidis (2006). Casanova (2012) extended their technique to the nonparametric setting using conditional M-quantiles via kernel regression. Alternatively, Salvati et al. (2010a) propose to estimate the cdf in a small area by a weighted sum of the sample data of the area of interest with calibrated sample weights based on the distribution of a covariate.

Casanova and Leconte (2015) proposed a nonparametric model-based estimator for the cdf of a right censored variable in a finite population, but to the best of our knowledge, there is no literature about the estimation of the cdf in small domains with right censored data. This can be due to the fact that methods that deal with censored data were originally developed in the medical field, where survey sampling is not usual.

In Sect. 2, we propose a nonparametric model-based estimator for the cdf in a small area when the variable of interest is right censored. Estimation is performed by adapting the technique proposed in Casanova (2012) to the censored case. In Sect. 3, model-based simulations compare the new estimator to the two following direct estimators: the naive Kaplan–Meier estimator (Kaplan and Meier 1958) and

the model-based estimator of Casanova and Leconte (2015) applied to a domain. Access times to the first job for young female graduates in the Occitania region are used to illustrate the new methodology in Sect. 4. Concluding remarks are given in Sect. 5.

2 Estimation of the Cdf of a Censored Variable in a Small Area

After presenting the framework, we review two direct estimators of the cdf in a small area in presence of censoring. We then develop a new small area indirect estimator.

2.1 Framework

Consider a finite population \mathcal{P} of size N which is partitioned into d sub-populations (i.e. domains) U_i of size N_i , $i = 1, \dots, d$. Let s be a sample of \mathcal{P} of size n and let $s_i = s \cap U_i$ be a sample of the domain U_i with size n_i . Let t_{ij} be the value of the variable of interest measured for the individual j of the domain U_i . The value of t_{ij} is supposed to be known only on s_i and possibly right censored by c_{ij} . So, for sample s_i , we observe $y_{ij} = \min(t_{ij}, c_{ij})$ and $\delta_{ij} = \mathbb{I}(t_{ij} \leq c_{ij})$. Let x_{ij} denote the value of a continuous covariate X measured for the individual j of the domain U_i . The cdf of the variable of interest T on the domain U_i is $F^i(t) = \frac{1}{N_i} \sum_{j \in U_i} \mathbb{I}(t_{ij} \leq t)$.

2.2 Direct Estimators

We recall that direct estimators use only domain information.

2.2.1 The Kaplan–Meier Domain Estimator

It is well known that the empirical cdf is not a consistent estimator of the cdf when the data are censored. On the other hand, we can naively but consistently estimate F^i using the Kaplan–Meier estimator (Kaplan and Meier 1958) computed on the sample s_i of the domain U_i . As the original Kaplan–Meier estimator is undetermined after the last observed time if this latter is censored, we use Efron’s version (Efron 1967) instead in order to obtain a distribution function as follows:

$$\hat{F}_{\text{KM}}^i(t) = \begin{cases} 1 - \prod_{j \in s_i} \left\{ 1 - \frac{1}{\sum_{r \in s_i} \mathbb{I}(y_{ir} \geq y_{ij})} \right\} \mathbb{I}(y_{ij} \leq t, \delta_{ij} = 1) & \text{if } t < y_{(in_i)}, \\ 1 & \text{otherwise,} \end{cases} \quad (1)$$

where $y_{(in_i)}$ denotes the last observed time of domain U_i .

2.2.2 The Casanova and Leconte (2015) Model-Based Domain Estimator

Since the cdf F^i of the domain U_i can be rewritten as

$$F^i(t) = \frac{1}{N_i} \left(\sum_{j \in s_i} \mathbb{I}(t_{ij} \leq t) + \sum_{j \in U_i \setminus s_i} \mathbb{I}(t_{ij} \leq t) \right), \quad (2)$$

we can probably improve its estimation by computing model-based estimators which use auxiliary information to predict the values of the variable of interest for the non-sampled individuals. In this context, an obvious estimator of the cdf in domain U_i can be derived from the estimator of the cdf in the population proposed by Casanova and Leconte (2015) by replacing s and \mathcal{P} by s_i and U_i , respectively, in all construction steps of the estimator.

This model-based approach requires defining a superpopulation model. In a non-parametric setting, we assume the following ξ model:

$$t_{ij} = m(x_{ij}) + e_{ij}, \quad i = 1, \dots, d, \quad j = 1, \dots, N_i,$$

where the e_{ij} are i.i.d. variables with cdf G^i and $m(x_{ij})$ is the conditional median of T given $X = x_{ij}$. Moreover, to obtain consistent and efficient estimators, we need to assume that the sampling design is not informative (or ignorable) which means that the same model holds for the sample and the population. This justifies the choice of a nonparametric model for which the risk of misspecification is reduced.

Since $\mathbb{E}_\xi(\mathbb{I}(t_{ij} \leq t)) = P(t_{ij} \leq t) = G^i(t - m(x_{ij}))$, a prediction of $\mathbb{I}(t_{ij} \leq t)$ can be obtained by estimating $G^i(t - m(x_{ij}))$. The conditional median $m(x_{ij})$ can be estimated by $\hat{m}(x_{ij})$, obtained by inverting the smoothed version \hat{F}_{SGKM}^i of the generalized Kaplan–Meier estimator of the conditional cdf proposed by Leconte et al. (2002). Let \hat{G}_{KM}^i denote the Kaplan–Meier estimator of the cdf G^i of the errors, computed with the residuals $\hat{e}_{ij} = y_{ij} - \hat{m}(x_{ij})$, $j \in s_i$. Note that the population

level bandwidths h_T and h_X have to be replaced by suitable domain level bandwidths h_T^i and h_X^i . The resulting estimator, based on formula (2), is

$$\hat{F}_M^i(t) = \frac{1}{N_i} \left(n_i \hat{F}_{KM}^i(t) + \sum_{j \in U_i \setminus s_i} \hat{G}_{KM}^i(t - \hat{m}(x_{ij})) \right). \tag{3}$$

2.3 The New Small Area Estimator

When the size of the domain sample is small, the previous estimators may have a large variance, and methods which use information from other domains are preferred in order to improve the precision of estimation. Therefore, we propose the following procedure: analogous to the Casanova and Leconte (2015) estimator, the first term between the parentheses of formula (2) is estimated using the Kaplan–Meier estimator on the sample s_i . In contrast, estimation of the second term of said equation will use information of the global sample s (and not only the sample s_i) in order to predict $\mathbb{I}(t_{ij} \leq t)$ for the non-sampled individuals of domain U_i . In this framework, we assume the superpopulation model ζ :

$$t_{ij} = m(q_i, x_{ij}) + \varepsilon_{ij}, \quad i = 1, \dots, d, \quad j = 1, \dots, N_i,$$

where the ε_{ij} are i.i.d. variables with cdf H^i , q_i is a coefficient in $(0, 1)$ characterizing the position of the domain U_i , and $m(q_i, x_{ij})$ is the conditional quantile of order q_i of T given $X = x_{ij}$. Each t_{ij} value can be considered as the conditional quantile of T given $X = x_{ij}$ for an order denoted $q(t_{ij}, x_{ij})$. Therefore, following Chambers and Tzavidis (2006), the coefficient q_i of the domain U_i can be defined by the mean or median of the conditional quantile orders $q(t_{ij}, x_{ij})$ of the units j in domain U_i .

Note that the conditional quantile orders are determined at the population level and we expect quantile orders of individuals of the same domain to have similar values if part of the data’s variability is explained by the domain.

Like the Casanova and Leconte (2015) estimator, the conditional quantile orders are estimated with the smoothed version of the generalized Kaplan–Meier estimator on the sample s as follows:

$$\hat{q}(t_{ij}, x_{ij}) = \hat{F}_{SGKM}(y_{ij} \mid x_{ij}).$$

Since the y_{ij} values can be right censored, so can be the $\hat{q}(t_{ij}, x_{ij})$. So, to estimate the global order q_i of domain U_i by the mean or the median of the $\hat{q}(t_{ij}, x_{ij})$, we first need to estimate their cdf while accounting for censoring. This can be easily performed by the Kaplan–Meier estimator. As the median is easier to compute than the mean in presence of censored data, we choose to estimate q_i using the median \hat{q}_i , obtained by inverting the Kaplan–Meier estimator.

Since $\mathbb{E}_\zeta (\mathbb{1}(t_{ij} \leq t)) = P(t_{ij} \leq t) = H^i(t - m(q_i, x_{ij}))$, $\mathbb{1}(t_{ij} \leq t)$ can be predicted by estimating $H^i(t - m(q_i, x_{ij}))$. A straightforward estimator $\hat{m}(\hat{q}_i, x_{ij})$ of $m(q_i, x_{ij})$ is the conditional quantile to x_{ij} of order \hat{q}_i , which is the solution in θ of $\hat{F}_{\text{SGKM}}(\theta | x_{ij}) = \hat{q}_i$ and is obtained by inversion of \hat{F}_{SGKM} . Notice that once again, similar to the estimation of the quantile order q_i , the whole sample is used to compute this estimator, allowing it to “borrow strength” from the other domains. Let us precise that the smoothed version of the Kaplan–Meier estimator on the sample s requires two suitable bandwidths h_T and h_X common to all domains. The cdf $H^i(t - m(q_i, x_{ij}))$ can then be estimated by the Kaplan–Meier estimator computed from the possibly right censored residuals $\hat{\varepsilon}_{ij} = y_{ij} - \hat{m}(\hat{q}_i, x_{ij})$, $j \in s_i$. We denote this estimator by \hat{H}_{KM}^i and derive the following estimator of the cdf of T in the domain U_i :

$$\hat{F}_Q^i(t) = \frac{1}{N_i} \left(n_i \hat{F}_{\text{KM}}^i(t) + \sum_{j \in U_i \setminus s_i} \hat{H}_{\text{KM}}^i(t - \hat{m}(\hat{q}_i, x_{ij})) \right). \quad (4)$$

It is obvious that the obtained estimator is a distribution function.

3 Model-Based Simulations

We present a simulation study to compare the performance of the three cdf estimators showcased in Sect. 2 where the population is partitioned into domains which may be small. We aim to estimate the cdf F^i in each domain. Therefore, we compute the new estimator \hat{F}_Q^i . Moreover, in order to measure the benefits of “borrowing strength” from neighbours, we also compute the estimator \hat{F}_M^i in each domain. The Kaplan–Meier estimator \hat{F}_{KM}^i of the cdf in each domain is also given as a naive estimator.

3.1 Description

We first generate 10 domain sizes N_i uniformly distributed over the interval (50, 150), leading to a population size $N = 901$. These sizes are kept fixed over iterations. Then, for each iteration, we generate the t_{ij} values for each domain U_i according to the accelerated failure time model $\log(t_{ij}) = 4 - 1.61x_{ij} + u_i + \varepsilon_{ij}$, where the covariates x_{ij} are uniformly distributed over the interval (1, 4). The error term ε_{ij} follows an extreme value distribution in order to obtain an exponential distribution for the t_{ij} . Note that for each domain, this model is a proportional hazard model with a hazard ratio (HR) equal to 5 (i.e. $\exp(1.61)$), which means that the ratio of the hazard rates of two individuals whose covariates x differ from one unit is constant over time and equal to 5. The domain effects u_i follow a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Note

that the variance of ε_{ij} is equal to 1.645, so that $\rho = \frac{\sigma^2}{\sigma^2 + 1.645}$ corresponds to the part of variability due to the domains ($\rho = 10\%$, 25% and 50% in the simulations). The times t_{ij} are censored by c_{ij} where c_{ij} is uniformly distributed on $(0, c)$, c being chosen in order to obtain population censoring rates τ of 10% , 25% and 50% .

For each domain, we then draw a simple random sample without replacement with a sampling fraction of 10% , leading to domain samples s_i of sizes n_i equal to 7, 9, 8, 7, 12, 9, 9, 8, 12 and 9 for the 10 domains. Samples with a sampling fraction equal to 5% are also drawn (with samples sizes n_i equal to 3, 5, 4, 4, 6, 5, 5, 4, 6 and 5). We perform $L = 1000$ iterations.

As far as smoothing is concerned, we choose the triweight kernel $K(x) = \frac{35}{32} (1 - x^2)^3 \mathbb{I}_{(-1,1)}(x)$. For each iteration and for each domain U_i , the bandwidths h_T^i and h_X^i for the estimator \hat{F}_M^i are chosen from a grid of bandwidths so that they minimize the averaged square error (ASE) criterion defined as

$$\text{ASE}(\hat{F}_M^i) = \frac{1}{5} \sum_{k=1}^5 \left(\hat{F}_M^i(tt_k) - F^i(tt_k) \right)^2 \tag{5}$$

where the evaluation times tt_k ($k = 1, \dots, 5$) are the 10th, 25th, 50th, 75th and 90th percentiles of the distribution of T (computed from a generated population of size 901 000). The cdf F^i of domain U_i is computed for each iteration using all the t_{ij} times of domain U_i (values generated before censoring the data).

As for the estimator \hat{F}_Q^i , the bandwidths h_T and h_X of the smoothed generalized Kaplan–Meier estimator do not depend on the domain and have been chosen in order to minimize the sum of the averaged square errors over the 10 domains $\sum_{i=1}^{10} \text{ASE}(\hat{F}_Q^i)$,

where $\text{ASE}(\hat{F}_Q^i)$ is defined in the same way as in formula (5). Notice that the estimators $\hat{m}(x_{ij})$ and $\hat{m}(\hat{q}_i, x_{ij})$ used in formulas (3) and (4) respectively are obtained by linear interpolation on a grid of 30 equally spaced values between the first and 75th percentiles of the T variable in samples s_i and s , respectively.

3.2 Results

Following Salvati et al. (2010a), we compare the performance of the three estimators of the cdf F^i of domain U_i in terms of absolute relative bias (ARB) and relative root mean squared error (RRMSE). Note that the cdf is different for each iteration as the population is generated at each iteration l and is, therefore, denoted by F_l^i . For each domain U_i and each estimator EST in the set $\{ \text{KM}, \text{M}, \text{Q} \}$, we compute the estimated ARB

$$\widehat{\text{ARB}}\left(\hat{F}_{\text{EST}}^i(t)\right) = \left(L^{-1} \sum_{l=1}^L F_l^i(t)\right)^{-1} L^{-1} \left| \sum_{l=1}^S \left(\hat{F}_{\text{EST},l}^i(t) - F_l^i(t)\right) \right|$$

and the estimated RRMSE

$$\widehat{\text{RRMSE}}\left(\hat{F}_{\text{EST}}^i(t)\right) = \left(L^{-1} \sum_{l=1}^L F_l^i(t)\right)^{-1} \sqrt{L^{-1} \sum_{l=1}^L \left(\hat{F}_{\text{EST},l}^i(t) - F_l^i(t)\right)^2}.$$

In practice, the above quantities were computed on the same grid of five time values as the one used for the ASE in formula (5) and averaged over the 10 domains for three censoring rates τ , two sampling fractions and three values of ρ , the part of the variability due to domains. As the results are very similar regardless of the different ρ values, we only present them for $\rho = 25\%$. Note that at least one uncensored event per domain sample is needed for the computation of the Kaplan–Meier estimator. Therefore, the number of iterations for which it was possible to compute the estimators depends on the censoring rate and the sampling fraction (only 489 iterations in the simulation study for 50% censoring and a sampling fraction of 5%).

Table 1 shows the area averages of the estimated MASE (mean of the estimated ASE over the iterations) of the estimators for two sampling fractions and three censoring rates. The model-based estimators \hat{F}_M^i and \hat{F}_Q^i have lower averaged MASE than the Kaplan–Meier estimator for all combinations of parameters. When the sampling fraction is small (i.e. 5%, leading to domain sample sizes n_i smaller than 6), the small area estimator \hat{F}_Q^i always behaves better than the domain based estimator \hat{F}_M^i . On the other hand, when the sampling fraction equals 10%, the two estimators perform very closely to one another.

As for the bias estimates, shown in Table 2, the direct Kaplan–Meier estimator enjoys smaller absolute bias compared to the model-based estimators \hat{F}_M^i and \hat{F}_Q^i for any censoring rate and any sampling fraction at almost all quantiles, which is expected. On the contrary, the estimators \hat{F}_M^i and \hat{F}_Q^i always record a lower RRMSE than the Kaplan–Meier estimator, with the difference decreasing for large quantiles (see Table 3). For a sampling fraction of 5%, the estimator \hat{F}_Q^i is more efficient

Table 1 Estimated area averages of MASE computed for the three cdf estimators for a domain effect measured by $\rho = 25\%$

Sampling fraction	$\tau^a = 10\%$			$\tau = 25\%$			$\tau = 50\%$		
	KM	M	Q	KM	M	Q	KM	M	Q
5%	32.90	20.25	16.68	34.39	23.13	18.05	44.26	35.57	28.36
10%	16.39	9.07	10.59	18.85	11.19	12.52	30.13	23.72	23.39

^a τ denotes the censoring rate

Table 2 Estimated area averages of absolute relative bias (ARB, %) computed for the three cdf estimators for a domain effect measured by $\rho = 25\%$

Sampling fraction: 5%

Quantile	$\tau^a = 10\%$			$\tau = 25\%$			$\tau = 50\%$		
	KM	M	Q	KM	M	Q	KM	M	Q
0.10	2.75	25.40	26.57	2.75	17.28	15.60	7.26	14.84	5.97
0.25	1.89	5.80	12.67	2.14	5.50	11.89	5.50	11.15	9.11
0.50	1.11	5.19	12.27	1.54	6.24	9.87	12.39	16.30	10.34
0.75	0.77	4.96	2.45	4.91	8.99	9.28	29.12	28.96	29.12
0.90	1.70	3.83	2.98	9.60	9.27	9.55	9.04	9.04	9.04

Sampling fraction: 10%

Quantile	$\tau = 10\%$			$\tau = 25\%$			$\tau = 50\%$		
	KM	M	Q	KM	M	Q	KM	M	Q
0.10	2.31	34.88	26.10	2.27	26.98	17.04	2.67	17.33	4.22
0.25	1.44	1.56	17.55	1.36	1.64	15.23	1.75	2.27	15.77
0.50	0.99	2.12	16.73	0.98	1.64	13.85	3.13	6.00	1.63
0.75	0.51	2.37	1.25	1.92	5.42	6.49	30.58	30.35	30.58
0.90	1.25	2.77	2.37	9.91	9.35	9.87	9.70	9.70	9.70

^a τ denotes the censoring rate

Table 3 Estimated area averages of relative root mean square errors (RRMSE, %) computed for the three cdf estimators for a domain effect measured by $\rho = 25\%$

Sampling fraction: 5%

Quantile	$\tau^a = 10\%$			$\tau = 25\%$			$\tau = 50\%$		
	KM	M	Q	KM	M	Q	KM	M	Q
0.10	129.64	101.45	75.21	129.91	99.95	70.41	133.13	110.16	72.47
0.25	73.13	53.97	40.04	73.31	55.93	40.82	75.83	61.20	45.83
0.50	42.80	33.52	28.48	44.08	35.46	29.72	53.25	45.12	39.12
0.75	25.67	20.86	22.05	28.40	23.88	24.42	32.81	32.65	32.81
0.90	15.24	12.02	13.55	12.88	12.40	12.82	12.18	12.18	12.18

Sampling fraction: 10%

Quantile	$\tau = 10\%$			$\tau = 25\%$			$\tau = 50\%$		
	KM	M	Q	KM	M	Q	KM	M	Q
0.10	90.70	65.95	61.33	90.84	60.93	57.38	91.60	58.40	55.55
0.25	52.23	29.10	34.18	52.39	29.47	33.96	54.01	33.70	36.90
0.50	30.05	22.35	25.02	30.83	23.58	24.53	37.12	29.87	27.71
0.75	17.88	15.37	15.99	21.06	17.57	18.72	34.45	34.22	34.45
0.90	11.52	9.49	10.50	13.25	12.42	13.20	12.90	12.90	12.90

^a τ denotes the censoring rate

than \hat{F}_M^i up to the median and presents comparable RRMSE values for larger quantiles. When the sampling fraction increases to 10%, the domain sample size becomes sufficient and the estimators \hat{F}_M^i and \hat{F}_Q^i perform similarly.

4 Example

We apply the new methods to data from the CEREQ (a French center of study and research on employment and skills). The CEREQ surveys young graduates about their professional careers three years after their diplomas (retrospective study about their monthly position for the previous three years). In our application, we only focus on the 10,135 girls from the Occitania region in France who left secondary education in 2010. The variable of interest T is the access time to the first job (in months), which is censored for the graduates who were still unemployed at the end of the survey (12.5% of the data). The aim of the CEREQ is to obtain statistics according to the level and the type of education, which partitions the population into 34 domains whose sizes vary from 6 to 1,443 girls. An unequal probability sampling led to a sample of 306 young female graduates and the domain sample sizes vary from 1 to 37. The auxiliary variable, known for the whole population, is the unemployment rate of the area where the graduate attended school. This variable is significantly and negatively correlated with the probability of finding a job (Wald test in the Cox model: $p = 0.013$). The bandwidths h_T^i and h_X^i of the \hat{F}_M^i estimator were selected by cross-validation techniques adapted to censoring (see formula (11) of Casanova and Leconte 2015). On the other hand, the \hat{F}_Q^i estimators require choosing a single pair of bandwidths (h_T, h_X) to accommodate all domains. To this aim, we select the pair that minimizes the following cross-validation criterion taking into account right censored data:

$$CV_Q = \sum_{i=1}^d \sum_{j \in s_u^i} \rho_{\hat{q}_i} (y_{ij} - \hat{m}_{-j}(\hat{q}_i, x_{ij}))$$

where s_u^i is the subset of uncensored individuals of s^i , and $\hat{m}_{-j}(\hat{q}_i, x_{ij})$ is the estimator of the conditional quantile of order \hat{q}_i based on the sample s excluding individual j of s_u^i . For $q \in (0, 1)$, ρ_q is the loss function associated with the quantile of order q , defined by

$$\rho_q(u) = \begin{cases} qu & \text{if } u \geq 0, \\ (q-1)u & \text{if } u < 0. \end{cases} \quad (6)$$

Figure 1 shows the curves of the three estimators \hat{F}_{KM}^i , \hat{F}_M^i and \hat{F}_Q^i for four domains of different sizes. The corresponding estimated quantiles of orders 0.25, 0.50 and 0.75 can be found in Table 4. As expected, acquiring a bachelor's degree or a certificate of professional competence with apprenticeship leads to shorter access times to the

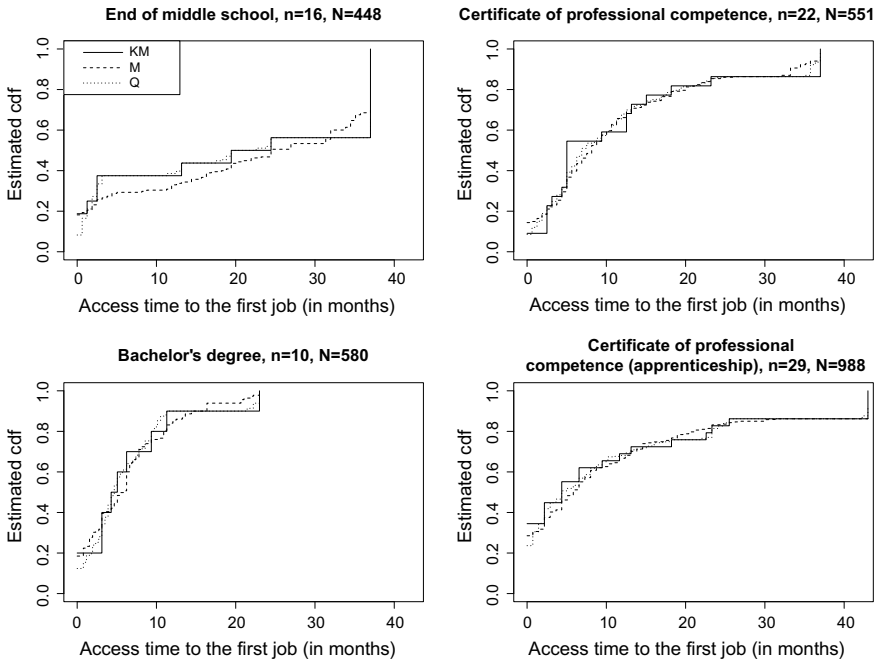


Fig. 1 Probability to access to the first job estimated by the three estimators for four different domains corresponding to a level and a type of education

Table 4 Estimated quantiles (in months) by the three methods for the four domains

Quantile order	End of middle school			Certificate of professional competence		
	KM	M	Q	KM	M	Q
0.25	1.57	2.34	1.65	2.98	3.65	3.17
0.50	21.64	25.12	20.69	6.40	7.77	6.74
0.75	33.24	40.93	49.81	15.05	16.45	15.66
Quantile order	Bachelor's degree			Certificate of professional competence (apprenticeship)		
	KM	M	Q	KM	M	Q
0.25	1.90	1.32	2.56	0	0	0.15
0.50	4.48	5.54	4.42	4.01	5.53	4.25
0.75	8.87	9.17	8.71	18.86	16.89	17.70

first job compared to the two other types of education presented above, with about 50% of female graduates with the two former qualifications finding a job in less than 5 months.

5 Concluding Remarks

Simulations show the gain in precision associated with predicting the variable of interest for non-sampled individuals, as reflected by the superiority of the estimators \hat{F}_M^i and \hat{F}_Q^i over the Kaplan–Meier estimator. For small area estimation, when the domain sample is very small (i.e. of size less than or equal to 6), it is preferable to borrow strength from neighbours and therefore, estimate the cdf in the domain by the estimator \hat{F}_Q^i . On the other hand, the median-based estimator \hat{F}_M^i is sufficient when the domain sample size is larger than 10.

The model-based approach is appropriate and will presumably lead to consistent estimators when the sampling is not informative. However, when a more complex sampling method is used or when the sampling is informative, a model-assisted approach which takes into account the sampling weights would be more suitable. For instance, following Chambers and Tzavidis (2006), computing the average order \hat{q}_i of domain U_i for the small area estimators could incorporate inclusion probabilities, which would lead to a model-assisted alternative to the estimator \hat{F}_Q^i .

The proposed estimators are based on the generalized Kaplan–Meier estimator of the conditional cdf, introduced by Beran (1981). Other estimators could have been used. In particular, Van Keilegom et al. (2001) defined an estimator of the conditional cdf which behaves better than the original Beran estimator in the right tail of the distribution even under heavy censoring. Alternatively, as proposed by Gannoun et al. (2005) in the censored case, the conditional quantiles could have been directly estimated by local linear polynomials.

Acknowledgements We thank H el ene Couprie for providing us the CEREQ data used in the example and the reviewers for their helpful comments. The authors acknowledge funding from ANR under grant ANR-17-EURE-0010 (Investissements d’Avenir program).

References

- Beran, R.: Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley (1981).
- Casanova, S. (2012). Using nonparametric conditional M-quantiles to estimate a cumulative distribution function in a domain. *Annales d’Economie et de Statistique*, 107–108, 287–297.
- Casanova, S., & Leconte, E. (2015). A nonparametric model-based estimator for the cumulative distribution function of a right censored variable in a finite population. *Journal of Surveys: Statistics and Methodology*, 3, 317–338.
- Chambers, R. L., & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255–268.

- Efron B.: The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium*4, 831–853 (1967).
- Gannoun, A., Saracco, J., Yuan, A., & Bonney, G. E. (2005). Non-parametric quantile regression with censored data. *Scandinavian Journal of Statistics*, 32(4), 527–550.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481.
- Leconte, E., Poiraud-Casanova, S., & Thomas-Agnan, C. (2002). Smooth conditional distribution function and quantiles under random censorship. *Lifetime Data Analysis*, 8, 229–246.
- Rao, J. N. K. (2003). *Small area estimation*. New-York: Wiley.
- Rueda, M. M., Arcos, A., Martínez-Miranda, M. D., & Román, Y. (2004). Some improved estimators of finite population quantile using auxiliary information in sample surveys. *Computational Statistics and Data Analysis*, 45, 825–848.
- Salvati, N., Chandra, H. & Chambers, R. (2010a). Model-based direct estimation of small area distributions. In *Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 20–10*.
- Salvati, N., Chandra, H., Ranalli, M. G., & Chambers, R. (2010b). Small area estimation using a nonparametric model-based direct estimator. *Computational Statistics and Data Analysis*, 54, 2159–2171.
- Salvati, N., Ranalli, M. G., & Pratesi, M. (2011). Small area estimation of the mean using nonparametric M-quantile regression: a comparison when a linear mixed model does not hold. *Journal of Statistical Computation and Simulation*, 81(8), 945–964.
- Van Keilegom, I., Akritas, M. G., & Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: A comparative study. *Computational Statistics and Data Analysis*, 35(4), 487–500.

Relaxing Monotonicity in Endogenous Selection Models and Application to Surveys



Eric Gautier

Abstract This paper considers endogenous selection models, in particular, nonparametric ones. Estimating the unconditional law of the outcomes is possible when one uses instrumental variables. Using a selection equation which is additively separable in a one dimensional unobservable has the sometimes undesirable property of instrument monotonicity. We present models which allow for nonmonotonicity and are based on nonparametric random coefficients indices. We discuss their nonparametric identification and apply these results to inference on nonlinear statistics such as the Gini index in surveys when the nonresponse is not missing at random.

1 Introduction

Empirical researchers often face a missing data problem. This is also called selection. Due to missing data, the observed data on an outcome variable corresponds to draws from the law of the outcome conditional on nonmissingness. Most of the time, the law of interest is the unconditional one. But the researcher can also be interested in the law of the outcome variable for the population that does not reveal the value of the outcome. For example, surveys rely on a sample drawn at random and the estimators require the observation of all sampled units. In practice, there is missing data and those estimators cannot be computed. A common practice is to rely on imputations. This means that the missing outcomes are replaced by artificial ones so that the estimator can eventually be computed. In the presence of endogenous selection, the law conditional on nonselection is the important one for imputation.

It is usual to assume that the data is Missing at Random (henceforth MAR, see Little and Rubin 2002) in which case there are perfectly observed variables such that the law of the outcome conditional on them and selection is the same as the law of outcome conditional on them and nonselection. Under such an assumption,

E. Gautier (✉)

TSE, Université Toulouse Capitole, 1 Esplanade de l'Université, 31000 Toulouse, France
e-mail: eric.gautier@tse-fr.eu

the estimable conditional law is the same as the one which is unconditional on selection. As a consequence, the researcher does not need a model for the joint law of the outcome and selection and the selection can be ignored. In survey sampling, the sampling frame can be based on variables available for the whole population, for example, if it involves stratification. In this case, those variables are natural candidates for conditioning variables for MAR to hold. In practice, there is noncompliance. It means that the researcher often does not have observations for all sampled units. This is called missing data in survey statistics. Though the original sampling law is known, the additional layer of missing data can be viewed as an additional selection mechanism conditional on the first one. The law of this second selection mechanism is unknown to the statistician. Oftentimes it can be suspected that units reveal the value of a variable partly depending on the value of that variable and the MAR assumption does not hold. This is a type of endogeneity issue commonly studied in econometrics. For example, wages are only observed for those who work and those who do not work might prefer not to work because their wage would be too low. Firms only carry out investment decisions if the net discounted value is nonnegative. An individual might be less willing to answer a question on his salary because it is not a typical one (either low or high). We expect a strong heterogeneity in the mechanism that drives individuals to not reveal the value of a variable.

When the MAR assumption no longer holds, the selection mechanism cannot be ignored. Identification of the distribution unconditional on selection or the distribution conditional on nonselection usually relies on the specification of a model for the vector formed by the outcome and a binary variable for selection. The alternative approach is to follow the partial identification route and recognize that the parameters of interest which are functionals of these distributions lie in sets. The Tobit and generalized Tobit models (also called Heckman selection model, see Heckman 1979) are classical parametric selection models to handle endogenous selection. The generalized Tobit model involves a system of two equations: one for the outcome and one for the selection. Each of these equations involves an error term and these errors are dependent, hence endogeneity. Identification in such systems relies on some variables which appear in the selection equation and are not measurable with respect to the sigma-field generated by the variables in the outcome equation and which do not have an effect on the errors. So these variables have an effect on the selection but not on the outcome. They are called instrumental variables or instruments.

This paper presents nonparametric models in Sects. 3 and 4. We explain in Sect. 4 that having a one-dimensional error term appearing in an additively separable form in the selection equation implies so-called instrument monotonicity. Instrument monotonicity has been introduced in Imbens and Angrist (1994). It has a strong identification power but at the same time leads to unrealistic selection equations as we detail in Sect. 4. To overcome this issue, we present in Sect. 5 selection equations where the error in the selection equation is multidimensional and appears in a non additively separable fashion. The baseline specification is a model where the selection equation involves an index with random coefficients. We show that we can rely on nonparametric models for these random coefficients. Finally, Sect. 6 presents a method to obtain a confidence interval around a nonlinear statistic like the Gini index

with survey data in the presence of non MAR missing data when we suspect that some instruments are nonmonotonic. These confidence intervals account for both the uncertainty due to survey sampling and the one due to missing data.

2 Preliminaries

2.1 Notations

Bold letters are used for vectors and matrices and capital letters for random elements. In the presence of an identically distributed sample, we add an index i for the marginal random element of index i . $1\{\cdot\}$ denotes the indicator function, ∂_p the derivative with respect to the variable p , $\langle \cdot, \star \rangle$ the inner product in the Euclidian space, $\|\cdot\|$ the euclidian norm, σ the spherical measure on the unit sphere in the Euclidian space. We denote it by \mathbb{S}^{d-1} when the Euclidian space is \mathbb{R}^d . $|\mathbb{S}^{d-1}|$ is its area. We write a.e. for almost everywhere. \mathbb{N}_0 is the set of nonnegative integers and \mathbb{N} are the positive integers. Quasi-analytic functions are functions which are infinitely differentiable and are characterized by the value of a function and all its derivatives at a point. A quasi-analytic class is defined via certain controls on the sup-norm of all the derivatives (Laplacians for functions defined on the sphere) as explained for example in Gaillac and Gautier (2019). Analytic functions are quasi-analytic.

All random elements are defined on the same probability space with probability \mathbb{P} and \mathbb{E} is the expectation. The support of a function or random vector is denoted by supp . We denote by $\text{supp}(U|X = \mathbf{x})$ the support of the conditional law of U given $X = \mathbf{x}$ when it makes sense. For a random vector Γ , f_Γ is its density with respect to a measure which will be clear in the text and d_Γ is its dimension. We use the notation $f_{\Gamma|X=\mathbf{x}}$ for a conditional density and $\mathbb{E}[U|X = \mathbf{x}]$ for the conditional expectation function evaluated at $\mathbf{x} \in \text{supp}(X)$. Below we usually write for all $\mathbf{x} \in \text{supp}(X)$ as if X were discrete. If X is continuous it should often be replaced by a.e. If X has both a discrete and a continuous component then the “for all” statement should hold for \mathbf{x} in the part of the support which is discrete. Equalities between random variables are understood almost surely. Random vectors appearing in models and which realizations are not in the observed data are called unobservable.

2.2 Baseline Setup

In this paper, the researcher is interested in features of the law of a variable Y given $X = \mathbf{x}$, where $\mathbf{x} \in \text{supp}(X)$. She has a selected sample of observations of Y , observations of a vector of which X is a subvector, for the selected and unselected samples, and R is a binary variable equal to 1 when Y is observed and else is 0. In this paper, the selection is often interpreted as a response and nonselection as nonresponse.

The law of Y given $X = \mathbf{x}$ is (nonparametrically) identified if, for a large class Φ of measurable functions ϕ , $\mathbb{E}[\phi(Y)|X = \mathbf{x}]$ can be characterized from the model equation, the restrictions on the primitives (such as conditional independence when using an instrumental variables strategy), and the distribution of the observed data. It is possible to take for Φ the bounded measurable functions, the bounded and continuous functions, the set of indicator functions $1\{\cdot \leq t\}$ for all $t \in \mathbb{R}$, the set of functions $\cos(t \cdot)$ and $\sin(t \cdot)$ for all $t \in \mathbb{R}$ (or certain countable subsets if Y is bounded). It is possible to add the assumption that Φ only contains functions which are nonnegative (a.e. if Y is continuous). For example, one can work with the functions $\cos(t \cdot) + 1$, and $\sin(t \cdot) + 1$ for all $t \in \mathbb{R}$. We call such class Φ an identifying class.

One can deduce, from the law of Y given $X = \mathbf{x}$, the law of a variable Y given $X = \mathbf{x}$, where $\mathbf{x} \in \text{supp}(X)$, and $R = 0$. This is the law of the outcome for the nonrespondants. It is the useful one for imputation. For all $\mathbf{x} \in \text{supp}(X)$, we have

$$\mathbb{E}[\phi(Y)|X = \mathbf{x}, R = 0] = \frac{\mathbb{E}[\phi(Y)|X = \mathbf{x}] - \mathbb{E}[\phi(Y)R|X = \mathbf{x}]}{\mathbb{P}(R = 0|X = \mathbf{x})}. \quad (1)$$

This paper presents identification results for a more fundamental object which is the joint distribution of the outcome and unobservable in the selection equation given $X = \mathbf{x}$, where $\mathbf{x} \in \text{supp}(X)$. One can clearly deduce from it by marginalization the distribution of Y given $X = \mathbf{x}$, where $\mathbf{x} \in \text{supp}(X)$.

2.3 *NMAR Missing Data*

Let \mathbf{W} be a vector, of which X is a subvector, which is observed for the selected and unselected samples. Inference on the conditional law of Y given X is possible if Y and R are independent given \mathbf{W} , namely if, for all bounded continuous function ϕ ,

$$\mathbb{E}[\phi(Y)R|\mathbf{W}] = \mathbb{E}[\phi(Y)|\mathbf{W}]\mathbb{E}[R|\mathbf{W}] \quad (2)$$

in which case

$$\mathbb{E}[\phi(Y)|\mathbf{W}] = \mathbb{E}[\phi(Y)|\mathbf{W}, R = 1] \quad (3)$$

and we conclude by the law of iterated expectations. Condition (2) is called Missing at Random (MAR, see Little and Rubin 2002). When it holds without the conditioning on \mathbf{W} , it is called Missing Completely at Random (MCAR). In econometrics \mathbf{W} such that (3) holds is called a control variable.

We consider cases where the researcher does not know that a specific vector \mathbf{W} is such that (2) holds. Then R is partly based on Y , even conditionally. This situation is

called Not Missing at Random (NMAR, see Little and Rubin 2002).¹ In the language of econometrics, this is called endogenous selection.

To handle NMAR missing data, it is usual to rely on a joint model for the determination of Y and R . This paper considers so-called triangular systems where a model for R , called a selection equation, is specified and does not involve Y but the dependence occurs via dependent latent (unobserved) variables. The identification arguments below rely on a vector \mathbf{Z} of so-called instrumental variables which are observed for the selected and unselected samples. It plays a completely different role as \mathbf{W} and \mathbf{X} .

3 Models with One Unobservable in the Endogenous Selection

Important parametric models rely on $Y = \mathbf{X}^\top \boldsymbol{\beta} + \sigma E_Y$ as a model equation for the variable of interest, $\boldsymbol{\beta}$ and σ are unknown parameters, \mathbf{X} and E_Y are independent, and E_Y is a standard normal random variable. In the Tobit model, $R = 1\{Y > y_L\}$ for a given threshold y_L . In the Heckman selection model (see Heckman 1979)

$$R = 1\{\mathbf{Z}^\top \boldsymbol{\gamma} - E_R > 0\}, \quad (4)$$

(E_Y, E_R) and $(\mathbf{X}^\top, \mathbf{Z}^\top)$ are independent,

$(E_Y, E_R)^\top$ is a mean zero gaussian vector with covariance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.

Equation (4) is the selection equation. The law of Y given \mathbf{X} and \mathbf{Z} , hence of Y given \mathbf{X} is identified and the model parameters can be estimated by maximum likelihood. Some functionals of the conditional law of Y given \mathbf{X} can be estimated for some semi-parametric extensions. For example, the conditional mean function can be obtained by estimating a regression model with an additional regressor which is a function of $\mathbf{Z}^\top \boldsymbol{\gamma}$. This leads to the interpretation that the endogeneity can be understood as a missing regressor problem.

A more general model is

$$R = 1\{\pi(\mathbf{Z}) > H\}, \quad (5)$$

$$\mathbf{Z} \text{ is independent of } (H, Y) \text{ given } \mathbf{X}, \quad (6)$$

$$\text{For all } \mathbf{x} \in \text{supp}(\mathbf{X}), \text{ the law of } H \text{ given } \mathbf{X} = \mathbf{x} \text{ is uniform on } (0, 1), \quad (7)$$

$$\text{For all } \mathbf{x} \in \text{supp}(\mathbf{X}), \text{ supp}(\pi(\mathbf{Z})|\mathbf{X} = \mathbf{x}) = [0, 1]. \quad (8)$$

¹The terminology nonignorable is also used but is defined for parametric models and requires parameter spaces to be rectangles. This is why we do not use this terminology in this paper.

Equation (5) is the selection equation. This model is quite general and clearly $\pi(\mathbf{Z}) = \mathbb{E}[R|\mathbf{X}, \mathbf{Z}] = \mathbb{E}[R|\mathbf{Z}]$. Equation (5) is as general as a selection equation that would be defined as $R = 1\{g(\mathbf{Z}) > E_R\}$, where g and the law of E_R are unknown. Indeed, one would obtain (5) from it by applying the nondecreasing CDF of E_R on both sides of the inequality $g(\mathbf{Z}) > E_R$. If we replace (6) by H and Y are independent given \mathbf{X}, \mathbf{Z} , assumption MAR holds by taking \mathbf{W} a vector which components are those of \mathbf{X} and \mathbf{Z} . Condition (6) allows for dependence between H and Y and R to be partly based on Y , even conditionally. The vector \mathbf{Z} thus plays a very different role from \mathbf{W} in the MAR assumption. By (6), \mathbf{Z} has a direct effect on R via $\pi(\mathbf{Z})$ which is non trivial but it does not have an effect on Y given \mathbf{X} . This type of properties for \mathbf{Z} is what makes it a vector of instrumental variables. It provides an alternative identification strategy. Equation (8) can be replaced by

$$\left. \begin{array}{l} \text{For all } \mathbf{x} \in \text{supp}(\mathbf{X}), \\ \text{supp}(\pi(\mathbf{Z})|\mathbf{X} = \mathbf{x}) \text{ contains a nonempty open set and,} \\ \text{for all } \phi \in \Phi_{\mathbf{x}}, \mathbb{E}[\phi(Y)1\{F(u) > H\}|\mathbf{X} = \mathbf{x}] \in C_{\mathbf{x}} \end{array} \right\}, \quad (9)$$

for a well chosen CDF F such that $F(x) = \int_{-\infty}^x f(t)dt$, where f is integrable and nonnegative a.e., and identifying classes $\Phi_{\mathbf{x}}$ and $C_{\mathbf{x}}$ a quasi-analytic classes of functions (see Gaillac and Gautier 2019).

Note that, by (13) below, for (9) to hold it is necessary that f is positive a.e. Also, for all $\mathbf{x} \in \text{supp}(\mathbf{X})$, the law of (Y, H) conditional on $\mathbf{X} = \mathbf{x}$ is identified if

$$\left. \begin{array}{l} \text{For all } \mathbf{x} \in \text{supp}(\mathbf{X}), \text{ there exists an identifying class } \Phi_{\mathbf{x}} \\ \text{and functions } F \text{ and } f \text{ as above such, that for all } \phi \in \bigoplus_{\mathbf{x}}, \\ \mathbb{E}[\phi(Y)|\mathbf{X} = \mathbf{x}, H = F(\cdot)]f(\cdot) \text{ is identified} \end{array} \right\} \quad (10)$$

We conclude this section with the following result:

Theorem 1 *If (5)–(7) and either (8) or (9) hold, then one has (28). Moreover, for all $\mathbf{x} \in \text{supp}(\mathbf{X})$,*

$$\mathbb{E}[\phi(Y)|\mathbf{X} = \mathbf{x}] = \int_{\mathbb{R}} \partial_u \mathbb{E}[\phi(Y)R|\mathbf{X} = \mathbf{x}, \pi(\mathbf{Z}) = F(u)]du \quad (11)$$

$$= \mathbb{E}[\phi(Y)R|\mathbf{X} = \mathbf{x}, \pi(\mathbf{Z}) = 1]. \quad (12)$$

Proof Based on (6), for all $\phi \in \Phi$, $\mathbf{x} \in \text{supp}(\mathbf{X})$, and u such that $F(u) \in \text{supp}(\pi(\mathbf{Z}))$,

$$\mathbb{E}[\phi(Y)R|\mathbf{X} = \mathbf{x}, \pi(\mathbf{Z}) = F(u)] = \mathbb{E}[\phi(Y)1\{F(u) > H\}|\mathbf{X} = \mathbf{x}]$$

so

$$\mathbb{E}[\phi(Y)|\mathbf{X} = \mathbf{x}, H = F(u)]f(u) = \partial_u \mathbb{E}[\phi(Y)R|\mathbf{X} = \mathbf{x}, \pi(\mathbf{Z}) = F(u)]. \quad (13)$$

The conclusion follows from either (8) or (9). By integration we obtain (11), hence

$$\mathbb{E}[\phi(Y)|X = \mathbf{x}] = \mathbb{E}[\phi(Y)R|X = \mathbf{x}, \pi(\mathbf{Z}) = 1] - \mathbb{E}[\phi(Y)R|X = \mathbf{x}, \pi(\mathbf{Z}) = 0],$$

and (12) follows because $\mathbb{E}[\phi(Y)R|X = \mathbf{x}, \pi(\mathbf{Z}) = 0] = 0$. \square

Remark 1 Similar formulas as (11) and (12) are given for a binary treatment effect model in Heckman and Vytlacil (2005) for effects that depend on an average (i.e. $\phi(y) = y$ for all $y \in \mathbb{R}$) rather than the whole law as above. There the integrand is called the local instrumental variable.

Condition (8) is strong. First, the support of \mathbf{Z} should be infinite so in practice, we think that at least a variable in \mathbf{Z} is continuous. Second, the variation of \mathbf{Z} should be large enough to move the selection probability $\pi(\mathbf{Z})$ from 0 to 1. This is a “large support” assumption. Using (12) for identification is called “identification at infinity”. Using it to construct an estimator does not make an efficient use of the data because it would make use of the subsample for which $\pi(\mathbf{Z}_i)$ is close to 1. In contrast, (11) can be used to form estimators which use all the data.

The techniques in Gaillac and Gautier (2019) allow for supports which are only countable in (8). It is possible to use the techniques in Gaillac and Gautier (2019), Gaillac and Gautier (2019) if we replace quasi-analytic with certain analytic classes. The advantage is to be able to use stable Fourier methods for extrapolation to build an estimator. Equation (8) were not required in the parametric Tobit and Heckman selection models. Condition (9) is a nonparametric middle ground between a parametric assumption made for convenience and a nonparametric one which is often too demanding for finding an instrument. Clearly, in this setup, building an estimator from (12) is simply impossible while building an estimator using (11) and the available data is possible.

4 Monotonicity

In this section, we show that the above nonparametric specification is not as general as we would think. From a modelling perspective, it is related (equivalent, see Vytlacil 2002) to the so-called instrument monotonicity introduced in Imbens and Angrist (1994).

For the sake of exposition, assume that \mathbf{Z} is discrete. For $\mathbf{z} \in \text{supp}(\mathbf{Z})$ and individuals that we index by $i \in I(\mathbf{z})$, such that $\mathbf{Z}_i = \mathbf{z}$, we have $R_i = 1\{\pi(\mathbf{z}) > H_i\}$. Suppose now that we could change exogenously (by experimental assignment) \mathbf{z} to \mathbf{z}' in $\text{supp}(\mathbf{Z})$ leaving unchanged the unobserved characteristics H_i for $i \in I(\mathbf{z})$. The corresponding R_i of those individuals are shifted monotonically. Indeed, we have either (1) $\pi(\mathbf{z}) \leq \pi(\mathbf{z}')$ or (2) $\pi(\mathbf{z}) > \pi(\mathbf{z}')$. In case (1),

$$\forall i \in I(\mathbf{z}), 1\{\pi(\mathbf{z}) > H_i\} \leq 1\{\pi(\mathbf{z}') > H_i\}$$

while in case (2),

$$\forall i \in \mathcal{I}(\mathbf{z}), 1\{\pi(\mathbf{z}) > H_i\} \geq 1\{\pi(\mathbf{z}') > H_i\}.$$

This instrument monotonicity condition has been formalized in Imbens and Angrist (1994).

Consider a missing data problem in a survey where $d_Z = 1$, $\mathbf{Z} = Z$ is the identity of a pollster, and $R = 1$ when the surveyed individual replies and else $R = 0$. The identity of the pollster could be Mr A ($z=0$) or Mrs B ($z=1$). This qualifies for an instrument because, usually, the identity of the pollster can have an effect on the response but not on the value of the surveyed variable. If the missing data model is any from Sect. 3 and pollster B has a higher response rate than pollster A, then in the hypothetic situation where all individuals surveyed by Mr A had been surveyed by Mrs B, then those who responded to Mr A respond to Mrs B and some who did not respond to Mr A respond to Mrs B, but no one who responded to Mr A would not respond to Mrs B. This last type of individuals corresponds to the so-called defiers in the terminology of Imbens and Angrist (1994): those for which $R_i = 1$ when $z = 1$ and $R_i = 0$ when $z = 0$. There, instrument monotonicity means that there are no defiers.

Remark 2 The terminology also calls compliers those who did not respond to Mr A but who would respond to Mrs B, never takers those who would respond to neither, and always takers those who would respond to both. \square

The absence of defiers can be unrealistic. For example, some surveyed individuals can answer a pollster because they feel confident with them. They can share the same traits which the statistician does not observe. For example, in the conversation, they could realize they share the same interest or went to the same school.

5 A Random Coefficients Model for the Selection Equation

Vytlacil (2002) showed that monotonicity is equivalent to modelling the selection equation as an additively separable latent index model with a single unobservable. In (5), the index is $\pi(\mathbf{Z}) - H$ and H is the unobservable. A nonadditively separable model takes the form $\pi(\mathbf{Z}, H)$. Heckman and Vytlacil (2005) proposes for a non-additively separable index with multiple unobservables a random coefficients binary choice model. They call it a benchmark. A random coefficients latent index model takes the form $A + \mathbf{B}^\top \mathbf{Z}$, where (A, \mathbf{B}^\top) and \mathbf{Z} are independent. This leads to

$$R = 1\{A + \mathbf{B}^\top \mathbf{Z} > 0\}. \quad (14)$$

The multiple unobservables are the coefficients (A, \mathbf{B}^\top) and play the role of H above. The model is nonadditively separable due to the products. The random intercept A

absorbs the usual mean zero error and deterministic intercept. The random slopes \mathbf{B} can be interpreted as the taste for the characteristic z . The components of (A, \mathbf{B}^\top) can be dependent.

To gain intuition, assume that \mathbf{Z} is discrete. For $\mathbf{z} \in \text{supp}(\mathbf{Z})$ and individuals $i \in \mathcal{I}(\mathbf{z})$ such that $\mathbf{Z}_i = \mathbf{z}$, we have

$$R_i = 1\{A_i + \mathbf{B}_i^\top \mathbf{z} > 0\}.$$

Suppose that the first component of \mathbf{B} takes positive and negative values with positive probability, that we change exogeneously \mathbf{z} to \mathbf{z}' in $\text{supp}(\mathbf{Z})$ by only changing the first component, and that we leave unchanged the unobserved characteristics (A_i, \mathbf{B}_i^\top) for $i \in \mathcal{I}(\mathbf{z})$. This model allows for populations of compliers (those for which the first component of \mathbf{B}_i is positive) and defiers (those for which the first component of \mathbf{B}_i is negative).

A parametric model for a selection equation specifies a parametric law for (A, \mathbf{B}^\top) . A parametric model for a selection model specifies a joint law of (A, \mathbf{B}^\top, Y) given \mathbf{X}, \mathbf{Z} . The model parameters can be estimated by maximum likelihood. The components of (A, \mathbf{B}^\top, Y) given \mathbf{X}, \mathbf{Z} could be modelled as dependent. (A, \mathbf{B}^\top) is a vector of latent variables and the likelihood involves integrals over \mathbb{R}^{dz+1} . As for the usual Logit or Probit models, a scale normalization is usually introduced for identification. Indeed $1\{A + \mathbf{B}^\top \mathbf{Z} > 0\} = 1\{c(A + \mathbf{B}^\top \mathbf{Z}) > 0\}$ for all $c > 0$. A nonparametric model allows the law of $(A, \mathbf{B}^\top, Y, \mathbf{Z})$ given $\mathbf{X} = \mathbf{x}$ to be a nonparametric class. Parametric and nonparametric models are particularly interesting when they allow for discrete mixtures to allow for different groups of individuals such as the compliers, defiers, always takers and never takers. But estimating a parametric model with latent variables which are drawn from multivariate mixtures can be a difficult exercise. In contrast, nonparametric estimators can be easy to compute.

The approach in this paper to relax monotonicity is based on Gautier and Hoderlein (2015). A few papers study in the treatment effects context which parameters can be identified without monotonicity (e.g. De Chaisemartin 2017).

5.1 Scaling to Handle Genuine Non Instrument Monotonicity

In this section, we rely on the approach used in the first version of Gautier and Hoderlein (2015) in the context of treatment effects models. This is based on the normalization in Gautier and Kitamura (2013), Gautier and Le Pennec (2018). Let $d - 1$ be the dimension of the vector of instrumental variables. For scale normalization, we define

$$\mathbf{\Gamma}^\top = \frac{(A, \mathbf{B}^\top)}{\|(A, \mathbf{B}^\top)\|} 1\{(A, \mathbf{B}^\top) \neq 0\}, \quad \mathbf{S}^\top = \frac{(1, \mathbf{Z}^\top)}{\|(1, \mathbf{Z}^\top)\|}$$

so that

$$R = 1\{\mathbf{\Gamma}^\top \mathbf{S} > 0\}.$$

We introduce some additional notations. When f is an integrable function on \mathbb{S}^{d-1} , we denote by \check{f} the function $\boldsymbol{\theta} \in \mathbb{S}^{d-1} \mapsto f(-\boldsymbol{\theta})$, by f^- the function $(f - \check{f})/2$. If $f \in L^2(\mathbb{S}^{d-1})$ (i.e. is square integrable) is nonnegative a.e. and $f\check{f} = 0$ a.e., then

$$f = 2f^- 1\{f^- > 0\} \text{ a.e.} \quad (15)$$

The hemispherical transform (see Rubin 1999) of an integrable function f on \mathbb{S}^{d-1} is defined as

$$\forall \mathbf{s} \in \mathbb{S}^{d-1}, \mathcal{H}[f](\mathbf{s}) = \int_{\boldsymbol{\theta} \in \mathbb{S}: (\mathbf{s}, \boldsymbol{\theta}) \geq 0} f(\boldsymbol{\theta}) d\sigma(\boldsymbol{\theta}).$$

This is a circular convolution in dimension $d = 2$

$$\forall \varphi \in [0, 2\pi), \mathcal{H}[f](\varphi) = \int_{\varphi \in [0, 2\pi): \cos(\varphi - \theta) \geq 0} f(\theta) d\theta.$$

We now recall a few useful properties of the hemispherical transform (see Gautier and Kitamura 2013 for more details). If $f \in L^2(\mathbb{S}^{d-1})$, then $\mathcal{H}[f]$ is a continuous function and $\mathcal{H}[f^-] = \mathcal{H}[f]^-$. The null space of \mathcal{H} consists of the integrable functions which are even (by a density argument) and integrate to 0 on \mathbb{S}^{d-1} . As a result

$$\mathcal{H}[f] = \int_{\boldsymbol{\theta} \in \mathbb{S}^{d-1}} f(\boldsymbol{\theta}) d\sigma(\boldsymbol{\theta})/2 + \mathcal{H}[f^-]. \quad (16)$$

\mathcal{H} is injective when acting on the cone of nonnegative almost everywhere functions in $L^2(\mathbb{S})$ such that $f\check{f} = 0$ a.e. (see Gautier and Kitamura 2013; Gautier and Le Pennec 2018). This means that f cannot be nonzero at two antipodal points of \mathbb{S} . We denote by \mathcal{H}^{-1} the unbounded inverse operator. We now present a formula for the inverse. If $g = \mathcal{H}(f)$, then

$$f^-(\boldsymbol{\gamma}) = \sum_{p \in \mathbb{N}_0} \frac{1}{\lambda_{2p+1, d}} \int_{\mathbb{S}^{d-1}} q_{2p+1, d}(\boldsymbol{\gamma}^\top \mathbf{s}) g(\mathbf{s}) d\sigma(\mathbf{s}) \quad (17)$$

$$= \sum_{p \in \mathbb{N}_0} \frac{1}{\lambda_{2p+1, d}} \int_{\mathbb{S}^{d-1}} q_{2p+1, d}(\boldsymbol{\gamma}^\top \mathbf{s}) g^-(\mathbf{s}) d\sigma(\mathbf{s}), \quad (18)$$

where

$$\lambda_{1,d} = \frac{|\mathbb{S}^{d-2}|}{d-1}, \forall p \in \mathbb{N}, \lambda_{2p+1,d} = \frac{(-1)^p |\mathbb{S}^{d-2}| 1 \cdot 3 \cdots (2p-1)}{(d-1)(d+1) \cdots (d+2p-1)},$$

$$L(k, d) = \frac{(2k+d-2)(k+d-2)!}{k!(d-2)!(k+d-2)}, q_{k,d}(t) := \frac{L(k, d)C_k^{(d-2)/2}(t)}{|\mathbb{S}^{d-1}|C_k^{(d-2)/2}(1)},$$

for all $\mu > -1/2$ and $k \in \mathbb{N}_0$, $C_k^\mu(t)$ are orthogonal polynomials on $[-1, 1]$ for the weight $(1-t^2)^{\mu-1/2}dt$. The Gegenbauer polynomials $C_k^\mu(t)$ can be obtained by the recursion $C_0^\mu(t) = 1$, $C_1^\mu(t) = 2\mu t$ for $\mu \neq 0$ while $C_1^0(t) = 2t$, and

$$(k+2)C_{k+2}^\mu(t) = 2(\mu+k+1)tC_{k+1}^\mu(t) - (2\mu+k)C_k^\mu(t).$$

Indeed, for all $p \in \mathbb{N}_0$, $s \rightarrow q_{2p+1,d}(\mathbf{y}^\top s)$ is odd.

Remark 3 Other inversion formulas when \mathcal{H} is restricted to odd functions or measures rather than the above cone are given in Rubin (1999).

We consider the following model restrictions, for all $\mathbf{x} \in \text{supp}(X)$,

$$\mathbb{P}(\mathbf{\Gamma} = 0 | X = \mathbf{x}) = 0, \tag{19}$$

$$S \text{ is independent of } (\mathbf{\Gamma}^\top, Y) \text{ given } X, \tag{20}$$

The conditional law of $\mathbf{\Gamma}$ given $X = \mathbf{x}$ is absolutely continuous

$$\text{with respect to } \sigma \text{ and the density belongs to } L^2(\mathbb{S}^{d-1}), \tag{21}$$

$$\text{For a.e. } \mathbf{y} \in \mathbb{S}^{d-1}, f_{\mathbf{\Gamma}|X=\mathbf{x}}(\mathbf{y}) \check{f}_{\mathbf{\Gamma}|X=\mathbf{x}}(\mathbf{y}) = 0, \tag{22}$$

$$\text{supp}(S|X = \mathbf{x}) = \{s \in \mathbb{S}^{d-1} : s_1 \geq 0\}. \tag{23}$$

Let $g_{\phi,\mathbf{x}} = \mathcal{H}[\mathbb{E}[\phi(Y)|X = \mathbf{x}, \mathbf{\Gamma} = \cdot] f_{\mathbf{\Gamma}|X=\mathbf{x}}(\cdot)]$, where, by a slight abuse of notations, the root $\mathbb{E}[\phi(Y)|X = \mathbf{x}, \mathbf{\Gamma} = \cdot] f_{\mathbf{\Gamma}|X=\mathbf{x}}(\cdot)$ is zero outside $\text{supp}(\mathbf{\Gamma}|X = \mathbf{x})$. By the above properties of \mathcal{H} , we have $g_{\phi,\mathbf{x}}^- = \mathcal{H}[(\mathbb{E}[\phi(Y)|X = \mathbf{x}, \mathbf{\Gamma} = \cdot] f_{\mathbf{\Gamma}|X=\mathbf{x}}(\cdot))^-]$.

Equation (23) can be replaced by, for identifying classes $\Phi_{\mathbf{x}}$ of functions which are nonnegative a.e. and a quasi-analytic classes $C_{\mathbf{x}}$,

$$\left. \begin{array}{l} \text{For all } \mathbf{x} \in \text{supp}(X), \\ \text{supp}(S|X = \mathbf{x}) \text{ has a nonempty interior} \\ \text{For all } \phi \in \Phi_{\mathbf{x}}, g_{\phi,\mathbf{x}} \in C_{\mathbf{x}}. \end{array} \right\} \tag{24}$$

This specification has the advantage that we do not assume that the researcher knows that one coefficient has a sign. Indeed it is easy to see that (22) contains such an assumption as a subcase. It allows for non instrument monotonicity for all instruments. Condition (23) is demanding because it means that $\text{supp}(Z|X = \mathbf{x})$ is the whole space for all $\mathbf{x} \in \text{supp}(X)$. Hence we provide (24) which allows for

an intermediate between nonparametric assumptions which are too demanding on the instruments and a parametric model. For further reference, we use the notation $H^+ = \{s \in \mathbb{S}^{d-1} : s_1 \geq 0\}$.

Remark 4 Proceeding like in Gaillac and Gautier (2019), Gaillac and Gautier (2019) allows an index of the form $\pi(\mathbf{Z}, \mathbf{H})$ where \mathbf{Z} are instrumental variables and \mathbf{H} is multidimensional of arbitrary dimension but has a sparse random series expansion on some classes of functions and the conditional law of \mathbf{Z} , given $\mathbf{X} = \mathbf{x}$, for all $\mathbf{x} \in \text{supp}(\mathbf{X})$, can have a support which is a subspace of the whole space. This means that a nonparametric random coefficients linear index already captures a large class of nonadditively separable models with multiple unobservables.

Using successively (20), the law of iterated expectations and (21), and (16), we obtain that, for all $(s^\top, \mathbf{x}^\top) \in \text{supp}(\mathbf{S}^\top, \mathbf{X}^\top)$,

$$\mathbb{E}[\phi(Y)R|\mathbf{X} = \mathbf{x}, \mathbf{S} = s] = \mathbb{E}[\phi(Y)1\{\Gamma^\top s > 0\}|\mathbf{X} = \mathbf{x}] \quad (25)$$

$$= g_{\phi, \mathbf{x}}(s), \quad (26)$$

$$= \frac{1}{2}\mathbb{E}[\phi(Y)|\mathbf{X} = \mathbf{x}] + g_{\phi, \mathbf{x}}^-(s). \quad (27)$$

We obtain the following theorem which states that $\mathbb{E}[\phi(Y)|\mathbf{X} = \mathbf{x}]$ can be identified at infinity under (23).

Theorem 2 Assume (19)–(23). For all \tilde{s} on the boundary of H^+ and $\mathbf{x} \in \text{supp}(\mathbf{X})$,

$$\mathbb{E}[\phi(Y)|\mathbf{X} = \mathbf{x}] = \lim_{s \rightarrow \tilde{s}, s \in H^+} \mathbb{E}[\phi(Y)R|\mathbf{X} = \mathbf{x}, \mathbf{S} = s] + \lim_{s \rightarrow -\tilde{s}, s \in H^+} \mathbb{E}[\phi(Y)R|\mathbf{X} = \mathbf{x}, \mathbf{S} = s].$$

Proof Let $\mathbf{x} \in \text{supp}(\mathbf{X})$. The result follows from (27) and the facts that $g_{\phi, \mathbf{x}}^-$ is odd and continuous as recalled at the beginning of the paragraph.

By (15), for all $\mathbf{x} \in \text{supp}(\mathbf{X})$, the law of (Y, Γ^\top) conditional on $\mathbf{X} = \mathbf{x}$ is identified if

$$\left. \begin{array}{l} \text{For all } \mathbf{x} \in \text{supp}(\mathbf{X}), \text{ there exists an identifying class } \Phi_{\mathbf{x}} \text{ of functions} \\ \text{which are nonnegative a.e. such that,} \\ \text{for all } \phi \in \bigoplus_{\mathbf{x}}, (\mathbb{E}[\phi(Y)|\mathbf{X} = \mathbf{x}, \Gamma = \cdot]f_{\Gamma}(\cdot))^- \text{ is identified.} \end{array} \right\} \quad (28)$$

The next theorem shows that, when ϕ is positive a.e., by integration, $\mathbb{E}[\phi(Y)|\mathbf{X} = \mathbf{x}]$ is nonparametrically identified with an alternative formula which does not involve taking limits.

Theorem 3 Assume (19)–(22) and either (23) or (24). Equation (28) holds. Moreover, under (23), for all $\mathbf{x} \in \text{supp}(\mathbf{X})$, $\phi \in \bigoplus_{\mathbf{x}}$, $\boldsymbol{\gamma} \in \mathbb{S}^{d-1}$, and $p \in \mathbb{N}_0$,

$$\begin{aligned} & \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(\boldsymbol{y}^\top \boldsymbol{s}) g_{\phi,x}(\boldsymbol{s}) d\sigma(\boldsymbol{s}) \\ &= 2\mathbb{E} \left[\frac{q_{2p+1,d}(\boldsymbol{y}^\top \boldsymbol{S})}{f_{S|X=x}(S)} \phi(Y) R \middle| X = \boldsymbol{x} \right] - \mathbb{E}[\phi(Y)|X = \boldsymbol{x}] \int_{H^+} q_{2p+1,d}(\boldsymbol{y}^\top \boldsymbol{s}) d\sigma(\boldsymbol{s}). \end{aligned} \quad (29)$$

Proof Let $\boldsymbol{x} \in \text{supp}(X)$ and $\phi \in \bigoplus_x$. Assuming (23), by Theorem 2, (27), and the fact that $g_{\phi,x}^-$ is odd, $g_{\phi,x}^-$ is identified. Hence, (28) holds. By the right-hand side of (25),

$$g_{\phi,x}(-\boldsymbol{s}) = \mathbb{E}[\phi(Y)|X = \boldsymbol{x}] - \mathbb{E}[\phi(Y)R|X = \boldsymbol{x}, S = \boldsymbol{s}],$$

which yields, for all $p \in \mathbb{N}_0$,

$$\begin{aligned} \int_{\mathbb{S}^{d-1}} q_{2p+1,d}(\boldsymbol{y}^\top \boldsymbol{s}) g_{\phi,x}(\boldsymbol{s}) d\sigma(\boldsymbol{s}) &= 2 \int_{H^+} q_{2p+1,d}(\boldsymbol{y}^\top \boldsymbol{s}) \mathbb{E}[\phi(Y)R|X = \boldsymbol{x}, S = \boldsymbol{s}] d\sigma(\boldsymbol{s}) \\ &\quad + \mathbb{E}[\phi(Y)|X = \boldsymbol{x}] \int_{H^+} q_{2p+1,d}(-\boldsymbol{y}^\top \boldsymbol{s}) d\sigma(\boldsymbol{s}), \end{aligned}$$

hence the moreover part.

Assuming (24), by (26) $g_{\phi,x}$, hence $g_{\phi,x}^-$, is identified. Hence, (28) holds. \square

Remark 1 By taking ϕ to 3 equal to 1, we obtain $f_{\Gamma|X=x}(\boldsymbol{y})$ for all \boldsymbol{x} and a.e. \boldsymbol{y} such that $(\boldsymbol{x}, \boldsymbol{y}) \in \text{supp}(X, \Gamma)$. \square

A simple estimator of $(\mathbb{E}[\phi(Y)|X = \boldsymbol{x}, \Gamma = \cdot] f_{\Gamma}(\cdot))^-$ on a grid of \boldsymbol{y} on \mathbb{S}^{d-1} under (23) takes the form

$$\sum_{p=1}^T \frac{\hat{c}_{2p+1}(\boldsymbol{y})}{\lambda_{2p+1,d}},$$

where $\hat{c}_{2p+1}(\boldsymbol{y})$ are estimators of the integrals in (29) and T is a smoothing parameter. This yields $\mathbb{E}[\phi(Y)|X = \boldsymbol{x}, \Gamma = \cdot] f_{\Gamma}(\cdot)$ using a plug-in and (15). A useful choice of ϕ for Algorithm 3 is $1\{\cdot \leq t\}$ for t on a grid on \mathbb{R} .

Algorithm 1 $\hat{c}_{2p+1}(\boldsymbol{y})$, for all $p = 1, \dots, T$, are obtained as follows:

1. Compute $\mathbb{E}[\phi(Y)|X = \boldsymbol{x}]$ using a local polynomial estimator of the right-hand side of the identity in Theorem 2 and compute numerically $\int_{H^+} q_{2p+1,d}(\boldsymbol{y}^\top \boldsymbol{s}) d\sigma(\boldsymbol{s})$,
2. Form, for the observations $i = 1, \dots, N$ in the sample,

$$\frac{q_{2p+1,d}(\boldsymbol{y}^\top \boldsymbol{S}_i)}{\hat{f}_{S|X=x}(S_i)} \phi(Y_i) R_i,$$

where $\hat{f}_{S|X=x}$ is a density estimator for directional data (see, e.g., Gautier and Kitamura 2013), and estimate $\mathbb{E} \left[\frac{q_{2p+1,d}(\boldsymbol{y}^\top \boldsymbol{S})}{f_{S|X=x}(S)} \phi(Y) R \middle| X = \boldsymbol{x} \right]$ using a local polynomial estimator.

In the approach in Gautier and Kitamura (2013), there is an additional damping of the high frequencies by an infinitely differentiable filter with compact support. The needlet estimator in Gautier and Le Pennec (2018) also builds on this idea. In the case of the estimation of $f_{\Gamma|X=x}$, Gautier and Le Pennec (2018) provides the minimax lower bounds for more general losses and an adaptive estimator based on thresholding the coefficients of a needlet expansion with a data driven level of hard thresholding.

Building an estimator based on (24), Hilbert space techniques, and assuming analyticity is an ongoing project.

To perform Algorithm 3, it is not useful to estimate the whole $\mathbb{E}[\phi(Y)|X = \mathbf{x}, \Gamma = \cdot]f_{\Gamma}(\cdot)$. Rather, the estimator $\mathbb{E}[\phi(Y)|X = \mathbf{x}]$ and local polynomial estimators are enough to obtain the elements in (1).

5.2 Alternative Scaling Under a Weak Version of Monotonicity

In this section, we still assume (14) and \mathbf{Z} is independent of $(A, \mathbf{B}^{\top}, Y)$ given X ((30) under the previous normalization). We maintain as well

$$\text{For all } \mathbf{x} \in \text{supp}(X), \exists \mathbf{P}_{\mathbf{x}} \in GL(d-1) : (\mathbf{P}_{\mathbf{x}}^{\top} \mathbf{B})_1 > 0 \text{ a.s.}, \quad (30)$$

where $GL(d-1)$ the general linear group over \mathbb{R}^{d-1} .

Under this assumption we can rewrite the model as follows. We denote by $V = (\mathbf{P}_{\mathbf{x}}^{-1} \mathbf{Z})_1$, $\bar{\mathbf{Z}} = (\mathbf{P}_{\mathbf{x}}^{-1} \mathbf{Z})_{2, \dots, d-1}$, $\Theta = -A / (\mathbf{P}_{\mathbf{x}}^{\top} \mathbf{B})_1$, and $\bar{\Gamma} = -(\mathbf{P}_{\mathbf{x}}^{\top} \mathbf{B})_{2, \dots, d-1} / (\mathbf{P}_{\mathbf{x}}^{\top} \mathbf{B})_1$. This yields

$$A + \mathbf{B}^{\top} \mathbf{Z} > 0 \Leftrightarrow V - \Theta - \bar{\Gamma}^{\top} \bar{\mathbf{Z}} > 0,$$

hence

$$R = 1\{V - \Theta - \bar{\Gamma}^{\top} \bar{\mathbf{Z}} > 0\} \quad (31)$$

and

$$(V, \bar{\mathbf{Z}}^{\top}) \text{ is independent of } (\Theta, \bar{\Gamma}^{\top}, Y) \text{ given } X. \quad (32)$$

By (32), (31) is equivalent to the fact that, for all $\mathbf{x} \in \text{supp}(X)$ and $\bar{\mathbf{z}} \in \text{supp}(\bar{\mathbf{Z}})$,

$$v \rightarrow \mathbb{P}(R = 1 | X = \mathbf{x}, \mathbf{Z} = \mathbf{P}_{\mathbf{x}}(v, \bar{\mathbf{z}}^{\top})^{\top}) = \mathbb{P}(\Theta + \bar{\Gamma}^{\top} \bar{\mathbf{z}} < v | X = \mathbf{x})$$

is a cumulative distribution, so the researcher can determine, from the distribution of the data, such an invertible matrix $\mathbf{P}_{\mathbf{x}}$.

The vector $(1 - \Theta - \bar{\Gamma}^\top)^\top$ of random coefficients in the linear index $V - \Theta - \bar{\Gamma}^\top \bar{Z}$ clearly satisfies (22). For this reason, the specification of the previous section is more general. There is instrument monotonicity in V , though not for \bar{Z} . This is a weak type of monotonicity because it is possible that there is instrument monotonicity for none of the instrumental variables in the original scale. This is the approach presented in the other versions of Gautier and Hoderlein (2015). It is shown in Gautier and Hoderlein (2015) that the equation

$$R = \mathbb{1} \left\{ V + f_0(\tilde{\mathbf{Z}}) - \Theta - \sum_{l=1}^{d-2} \bar{\Gamma}_l f_l(\tilde{\mathbf{Z}}_l) > 0 \right\},$$

where $d \geq 3$, f_0, \dots, f_{d-2} are unknown functions, can be transformed by reparametrization into (31) and the unknown functions are identified by similar arguments as for the additive model for a regression function.

We consider as well the following restrictions:

$$\text{For all } (\mathbf{x}^\top, \bar{\mathbf{z}}^\top) \in \text{supp}(\mathbf{X}^\top, \bar{\mathbf{Z}}^\top), f_{\Theta, \bar{\Gamma}|\mathbf{X}=\mathbf{x}} \text{ and } f_{\Theta+\bar{\Gamma}^\top \bar{\mathbf{z}}|\mathbf{X}=\mathbf{x}} \text{ exist,} \quad (33)$$

and either

$$\left. \begin{aligned} \forall (\mathbf{x}, \bar{\mathbf{z}}) \in \text{supp}(\mathbf{X}, \bar{\mathbf{Z}}), \text{supp}(V|\mathbf{X}=\mathbf{x}, \bar{\mathbf{Z}}=\bar{\mathbf{z}}) \supseteq \text{supp}(\Theta + \bar{\Gamma}^\top \bar{\mathbf{z}}|\mathbf{X}=\mathbf{x}), \\ \forall \mathbf{x} \in \text{supp}(\mathbf{X}), \text{supp}(\bar{\mathbf{Z}}|\mathbf{X}=\mathbf{x}) = \mathbb{R}^{d-2}, \end{aligned} \right\} \quad (34)$$

or, for identifying classes $\Phi_{\mathbf{x}}$ and quasi-analytic classes $C_{\mathbf{x}, \bar{\mathbf{z}}}^a$ and $C_{\mathbf{x}, \mathbf{x}}^b$, denoting by

$$\begin{aligned} a_{\phi, \mathbf{x}, \bar{\mathbf{z}}} &= \mathbb{E}[\phi(Y)|\Theta + \bar{\Gamma}^\top \bar{\mathbf{z}} = \cdot, \mathbf{X} = \mathbf{x}] f_{\Theta+\bar{\Gamma}^\top \bar{\mathbf{z}}|\mathbf{X}=\mathbf{x}}(\cdot), \\ b_{\phi, s, \mathbf{x}} &= \mathbb{E}[e^{i\Theta s} e^{i(\bar{\Gamma}^\top \cdot)^s} \phi(Y)|\mathbf{X} = \mathbf{x}], \end{aligned}$$

$$\left. \begin{aligned} \text{For all } (\mathbf{x}^\top, \bar{\mathbf{z}}^\top) \in \text{supp}(\mathbf{X}^\top, \bar{\mathbf{Z}}^\top), \text{supp}(V|\mathbf{X}=\mathbf{x}, \bar{\mathbf{Z}}=\bar{\mathbf{z}}) \\ \text{and } \text{supp}(\bar{\mathbf{Z}}|\mathbf{X}=\mathbf{x}) \text{ have nonempty interiors,} \\ \text{For all } \phi \in \bigoplus_{\mathbf{x}}, a_{\phi, \mathbf{x}, \bar{\mathbf{z}}} \in C_{\mathbf{x}, \bar{\mathbf{z}}}^a \text{ and for all } s \in \mathbb{R}, b_{\phi, s, \mathbf{x}} \in C_{\mathbf{x}, \mathbf{x}}^b. \end{aligned} \right\} \quad (35)$$

Clearly, for $a_{\phi, \mathbf{x}, \bar{\mathbf{z}}} \in C_{\mathbf{x}}$ to hold it is necessary that $f_{\Theta+\bar{\Gamma}^\top \bar{\mathbf{z}}|\mathbf{X}=\mathbf{x}}$ is positive a.e.. A simple sufficient condition for $b_{\phi, s, \mathbf{x}}$ to be analytic is

$$\text{For all } \mathbf{x} \in \text{supp}(\mathbf{X}), \exists R > 0 : \mathbb{E}[\exp(R \|\bar{\Gamma}\|)|\mathbf{X} = \mathbf{x}] < \infty,$$

This condition (which imply that $\bar{\Gamma}$ does not have heavy tails) and the support conditions in (24) are slightly stronger than necessary (see Gaillac and Gautier 2019).

Theorem 4 *Maintain (31)–(33) and either (34) or (35). For all $\mathbf{x} \in \text{supp}(\mathbf{X})$, the law of $(Y, \Theta, \bar{\Gamma}^\top)$ conditional on $\mathbf{X} = \mathbf{x}$ is identified.*

Proof Let $(\mathbf{x}, \bar{\mathbf{z}}) \in \text{supp}(X, \bar{Z})$ and $\phi \in \Phi_{\mathbf{x}}$. For all v in the interior of $\text{supp}(V|X = \mathbf{x}, \bar{Z} = \bar{\mathbf{z}})$, we have

$$\partial_v \mathbb{E}[\phi(Y)R|X = \mathbf{x}, V = v, \bar{Z} = \bar{\mathbf{z}}] = \mathbb{E}[\phi(Y)|X = \mathbf{x}, \Theta + \bar{\Gamma}^\top \bar{\mathbf{z}} = v] f_{\Theta + \bar{\Gamma}^\top \bar{\mathbf{z}}|X=\mathbf{x}}(v).$$

So, by the assumptions, the above right-hand side is identified for all $v \in \mathbb{R}$. Hence, for all $s \in \mathbb{R}$,

$$b_{\phi, s, \mathbf{x}} = \int_{\mathbb{R}} e^{isv} \mathbb{E}[\phi(Y)|X = \mathbf{x}, \Theta + \bar{\Gamma}^\top \bar{\mathbf{z}} = v] f_{\Theta + \bar{\Gamma}^\top \bar{\mathbf{z}}|X=\mathbf{x}}(v) dv \quad (36)$$

is identified on $\text{supp}(\bar{Z}|X = \mathbf{x})$. We conclude using either the large support assumption or the now usual argument involving quasi-analyticity. \square

Based on (36), it is not difficult to obtain an estimator of $b_{\phi, s, \mathbf{x}}$ under (34) and then the root $\mathbb{E}[\phi(Y)|X = \mathbf{x}, \Theta, \bar{\Gamma} = \cdot] f_{\Theta, \bar{\Gamma}|X=\mathbf{x}}(\cdot)$.

- Algorithm 2**
1. Compute a local polynomial estimator of $\partial_v \mathbb{E}[\phi(Y)R|X = \mathbf{x}, V = v, \bar{Z} = \bar{\mathbf{z}}]$,
 2. Take a smooth numerical approximation of the Fourier transform of it,
 3. Use a smoothed multivariate inverse Fourier transform and a change of variable $(s, s\bar{\mathbf{z}}) \rightarrow (s, \mathbf{z})$.

Alternatively, (Gautier and Hoderlein, 2015) uses a smooth regularized inverse of the Radon transform and an integration by part. It is also possible to turn the identification argument based on (35) into an estimation procedure as in (Gaillac and Gautier, 2019).

To perform Algorithm 3, it is not useful to estimate the whole $\mathbb{E}[\phi(Y)|X = \mathbf{x}, \Theta, \bar{\Gamma} = \cdot] f_{\Theta, \bar{\Gamma}|X=\mathbf{x}}(\cdot)$. Rather, one can estimate $\mathbb{E}[\phi(Y)|X = \mathbf{x}]$ by steps 1 and 2 (for $s = 0$) of Algorithm 2 and use local polynomial estimators of the remaining elements in (1).

Remark 5 Proceeding like in (Gaillac and Gautier, 2019), Gaillac and Gautier (2019) allows to work with an index of the form $\pi(\mathbf{Z}, \mathbf{H}) - V$ where \mathbf{H} is multidimensional of arbitrary dimension and $\pi(\mathbf{Z}, \mathbf{H})$ has a sparse random series expansion on some classes of functions and the conditional laws of \mathbf{Z} and V , given $X = \mathbf{x}$, for all $\mathbf{x} \in \text{supp}(X)$, can have a support which is a subspace of the whole space.

Remark 6 In a binary treatment effect model, the outcome can be written as $Y = (1 - R)Y_0 + RY_1$. Y_0 and Y_1 are the potential outcomes without and with treatment. They are unobservable. A selection model can be viewed as a degenerate case where $Y_0 = 0$ a.s. Quantities similar to the root in Theorem 3 have been introduced in Gautier and Hoderlein (2015). They are for the marginals of the potential outcomes $\mathbb{E}[\phi(Y_j)|X = \mathbf{x}, \Theta = \theta, \bar{\Gamma} = \bar{\gamma}]$ for $j \in \{0, 1\}$. An extension of the Marginal Treatment Effect in Heckman and Vytlačil (2005) to multiple unobservables and for laws is the Conditional on Unobservables Distribution

of Treatment Effects $\mathbb{E}[\phi(Y_1 - Y_0)|X = \mathbf{x}, \Theta = \theta, \bar{\Gamma} = \bar{\gamma}]$. Gautier and Hoderlein (2015) considers kernel estimators which rely on regularized inverses of the Radon transform. \square

6 Application to Missing Data in Surveys

When making inference with survey data, the researcher has available data on a vector of characteristics for units belonging to a random subset \mathcal{S} of a larger finite population \mathcal{U} . The law used to draw \mathcal{S} can depend on variables available for the whole population, for example, from a census. We assume that the researcher is interested in a parameter g which could be computed if we had the values of a variable y_i for all units of index $i \in \mathcal{U}$. This can be an inequality index, for example, the Gini index, and y_i the wealth of household i . In the absence of missing data, the statistician can produce a confidence interval for g , making use of the data for the units $i \in \mathcal{S}$ and his available knowledge on the law \mathcal{S} . We assume that the cardinality of \mathcal{S} is fixed and equal to n . When g is a total, it is usual to rely on an unbiased estimator, an estimator of its variance, and a Gaussian approximation. For more complex parameters, linearization is often used to approximate moments. The estimator usually rely on the survey weights $\pi_i = 1/\mathbb{P}(i \in \mathcal{S})$. For example an estimator of the Gini index is

$$\widehat{g}((y_i)_{i \in \mathcal{S}}) = \frac{\sum_{i=1}^n (2\hat{r}(i) - 1)\pi_i y_i}{\sum_{i=1}^n \pi_i \sum_{i=1}^n \pi_i y_i} - 1, \quad (37)$$

where $\hat{r}(i) = \sum_{j=1}^n w_j \mathbb{1}\{y_j \leq y_i\}$. The estimators of the variance of the estimators are more complex to obtain and we assume there is a numerical procedure to obtain them. Inference is based on the approximation

$$\widehat{g}((y_i)_{i \in \mathcal{S}}) \approx g + \sqrt{\widehat{\text{var}}(\widehat{g})((y_i)_{i \in \mathcal{S}})}\epsilon, \quad (38)$$

where ϵ is a standard normal random variable and $\widehat{\text{var}}(\widehat{g})((y_i)_{i \in \mathcal{S}})$ is an estimator of the variance of $\widehat{g}((y_i)_{i \in \mathcal{S}})$.

In practice, this is not possible when some of the y_i s are missing. There is a distinction between total nonresponse, where the researcher discards the data for some units $i \in \mathcal{S}$ or it is not available, and partial nonresponse. Let us ignore total nonresponse which is usually dealt with using reweighting and calibration and focus on partial nonresponse. We consider a case where y_i can be missing for some units $i \in \mathcal{S}$, while all other variables are available for all units $i \in \mathcal{S}$. We rely on a classical formalism where the vector of surveyed variables and of those used to draw $\mathcal{S} \subsetneq \mathcal{U}$, for each unit $i \in \mathcal{U}$, are random draws from a superpopulation. In this formalism, the parameter y_i for all indices i of households in the population and g are random and we shall now use capital letters for them. Let S_i and R_i be random variables,

where $S_i = 1$ if $i \in \mathcal{S}$ and $R_i = 1$ if unit i reveals the value of Y_i given $S_i = 1$, and \mathbf{X}_i and \mathbf{Z}_i be random vectors which will play a different role.

It is classical to rely on imputations to handle the missing data. This means that we replace missing data by artificial values obtained from a model forming predictions or simulating from a probability law and inject them in a formula like (37). In Gautier (2005), we discuss the use of the Heckman selection model when we suspect that the data is not missing at random. This relies on a parametric model for the partially missing outcome which is prone to criticism. Also, as this paper has shown, such a model relies on instrument monotonicity which is an assumption which is too strong to be realistic.

It is difficult to analyze theoretically the effect of such imputations. For example, when the statistic is nonlinear in the y_i s (e.g. (37)) then using predictions can lead to distorted statistics. It is also tricky to make proper inference when one relies on imputations. One way to proceed is to rely on a hierarchical model as in Gautier (2011). There the imputation model is parametric and we adopted the Bayesian paradigm for two reasons. The first is to account for parameter uncertainty and the second is to replace maximum likelihood with high dimensional integrals by a Monte Carlo Markov Chain Algorithm (a Gibbs sampler). The hierarchical approach also allows layers such as to model model uncertainty. The Markov chain produces sequences of values for each Y_i for $i \in \mathcal{S} \setminus \mathcal{R}$ in the posterior distribution given $(\mathbf{W}_i)_{i \in \mathcal{S}}$, the choice of which is discussed afterwards. Subsequently we get a path of

$$\tilde{G} = \hat{G}((Y_i)_{i \in \mathcal{S}}) + \sqrt{\widehat{\text{var}}(\hat{G})((Y_i)_{i \in \mathcal{S}})}\epsilon \quad (39)$$

where ϵ is a standard normal random variable independent from $(Y_i)_{i \in \mathcal{S}}$ given $(\mathbf{W}_i)_{i \in \mathcal{S}}$. Equation (39) is derived from (38). The variables $(\mathbf{W}_i)_{i \in \mathcal{S}}$ are those making the missing mechanism corresponding to R_i relative to Y_i MAR.² The last T values $(\tilde{G}_t)_{t=T_0+1}^{T_0+T}$ of the sample path for G allows to form credible sets C by adjusting the set so that the frequency that $\{\tilde{G}_t \in C\}$ exceeds $1 - \alpha$, where α is a confidence level. T_0 is the so-called burn-in. These confidence sets account for error due to survey sampling, parameter uncertainty, and nonresponse. They can be chosen from the quantiles of the distribution, to minimize the volume of the set, etc.

We now consider our nonparametric models of endogenous selection which allow for nonmonotonicity of the instrumental variables to handle a missing mechanism corresponding to R which is NMAR. For simplicity, we assume away parameter uncertainty, which would be taken into account more easily if we adopted a Bayesian framework and total nonresponse. The variables \mathbf{X}_i in Sect. 5 can be variables that are good predictors for Y_i . They are not needed to obtain valid inference but can be useful to make confidence intervals smaller. However, the selection corresponding to the binary variables R_i relative to the outcomes Y_i given $S_i = 1$ follow a NMAR mechanism. The (multiple) imputation approach becomes: for $t = 1, \dots, T$

²They can be those used by the survey statistician to draw \mathcal{S} if any (and usually made available) to handle a total nonresponse which is MAR via imputations.

1. Draw an i.i.d. sample of Y_i^t for $i \in \mathcal{S} \setminus \mathcal{R}$ from the law of Y given $X = x_i, S = 1$, and $R = 0$, an independent standard normal ϵ_t , and set $Y_i^t = y_i$ for $i \in \mathcal{R}$ where y_i are the observations in the selected sample
2. Compute

$$\tilde{G}_t = \widehat{G}((Y_i^t)_{i \in \mathcal{S}}) + \sqrt{\widehat{\text{var}}(\widehat{G})((Y_i^t)_{i \in \mathcal{S}})} \epsilon_t. \quad (40)$$

The confidence interval is formed from the sample $(\tilde{G}_t)_{t=1, \dots, T}$ for a given confidence level.

In practice, assuming away the conditioning on X , the draws from the law of Y given $S = 1$, and $R = 0$ can be obtained (approximately) by

- Algorithm 3**
1. Take $\phi = \{1, \dots, t\}$ for a grid of t ,
 2. Estimate the left-hand side of (1) using plug-in estimators of the elements on the right-hand side from the available data (corresponding to $S = 1$),
 3. Draw from a uniform random variable on $[0, 1]$,
 4. Apply a numerical approximation of the inverse CDF from step 2. □

Acknowledgements The author acknowledges financial support from the grant ERC POEMH 337665 and ANR-17-EURE-0010. The material of this paper was presented at the 7^{ème} Colloque Francophone sur les Sondages at ENSAI in 2012.

References

- De Chaisemartin, C. (2017). Tolerating defiance? local average treatment effects without monotonicity. *Quantitative Economics*, 8, 367–396.
- Gaillac, C., & Gautier, E. (2019). Identification in some random coefficients models when regressors have limited variation. Working paper.
- Gaillac, C., & Gautier, E. (2019). Adaptive estimation in the linear random coefficients model when regressors have limited variation. <https://arxiv.org/abs/1905.06584>.
- Gaillac, C., & Gautier, E. (2019). Estimates for the SVD of the truncated Fourier transform on $L^2(\exp(b|x|))$ and stable analytic continuation. <https://arxiv.org/abs/1905.11338>.
- Gautier, E. (2005). Eléments sur la sélection dans les enquêtes et sur la nonréponse non ignorable. Actes des Journées de Méthodologie Statistique.
- Gautier, E. (2011). Hierarchical Bayesian estimation of inequality measures with non-rectangular censored survey data with an application to wealth distribution of the French households. *Annals of Applied Statistics*, 5, 1632–1656.
- Gautier, E., & Hoderlein, C. (2015). A triangular treatment effect model with random coefficients in the selection equation. <https://arxiv.org/abs/1109.0362>.
- Gautier, E., & Kitamura, Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica*, 81, 581–607.
- Gautier, E., & Le Pennec, E. (2018). Adaptive estimation in the nonparametric random coefficients binary choice model by needlet thresholding. *Electronic Journal of Statistics*, 12, 277–320.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153–161.
- Heckman, J. J., & Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73, 669–738.

- Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*, 467–475.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.
- Rubin, B. (1999). Inversion and characterization of the hemispherical transform. *Journal of Analysis Mathematics*, *77*, 105–128.
- Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica*, *70*, 331–341.

B-Spline Estimation in a Survey Sampling Framework



Camelia Goga

Abstract Nonparametric regression models have been used more and more over the last years to model survey data and incorporate efficiently auxiliary information in order to improve the estimation of totals, means or other study parameters such as Gini index or poverty rate. *B*-spline nonparametric regression has the benefit of being very flexible in modeling nonlinear survey data while keeping many similarities and properties of the classical linear regression. This method proved to be efficient for deriving a unique system of weights which allowed to estimate in an efficient way and simultaneously many study parameters. Applications on real and simulated survey data showed its high efficiency. This paper aims at giving a review of applications of the *B*-spline nonparametric regression in a survey sampling framework and design-based approach. Handling item nonresponse by *B*-spline modeling is also considered. This review includes also new properties and improved consistency rates of the suggested penalized and unpenalized *B*-spline estimators.

1 Introduction

Consider a finite population $U = \{1, \dots, k, \dots, N\}$. We focus on the estimation of the total $t_y = \sum_{k \in U} y_k$ of a study variable \mathcal{Y} over the population U with y_k the non-random value of \mathcal{Y} for the k th unit. More general study parameters such as the Gini index or the functional median will be considered in Sect. 3. Let $s \subset U$ be a probability sample selected from U according to a sampling design $p(\cdot)$. More exactly, $p(\cdot)$ is a probability distribution defined on the set \mathcal{S} of all possible subsets of U and $p(s)$ is the probability of selecting the sample s . Given $p(\cdot)$, each unit k from the population has a known inclusion probability $\pi_k = \Pr(k \in s) = \sum_{k \ni s} p(s) > 0$ and a corresponding sampling design weight $d_k = 1/\pi_k$; here $k \ni s$ denotes that the sum is over those samples s that contain the given k .

C. Goga (✉)

Laboratoire de Mathématiques de Besançon, Université de Bourgogne Franche-Comté,
16 route de Gray, 25000 Besançon, France
e-mail: camelia.goga@univ-fcomte.fr

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_5

Without auxiliary information, the total t_y can be estimated by the well-known Horvitz–Thompson (HT) estimator (Horvitz and Thompson 1952)

$$\hat{t}_{yd} = \sum_{k \in s} d_k y_k = \sum_{k \in s} \frac{y_k}{\pi_k}. \quad (1)$$

If the first-order inclusion probabilities are all positive, $\pi_k > 0$, then the HT estimator is unbiased for t_y with respect to the sampling design $p(\cdot)$, namely $\mathbb{E}_p(\hat{t}_{yd}) = t_y$, where \mathbb{E}_p is the expectation with respect to the sampling design. Its variance with respect to $p(\cdot)$ is given by $\mathbb{V}_p(\hat{t}_{yd}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) d_k d_l y_k y_l$, where $\pi_{kl} = \Pr(k, l \in s)$ is the second-order inclusion probability. If $\pi_{kl} > 0$ for all $k \neq l$, then $\mathbb{V}_p(\hat{t}_{yd})$ can be estimated unbiasedly by the HT variance estimator $\hat{\mathbb{V}}_p(\hat{t}_{yd}) = \sum_{k \in s} \sum_{l \in s} (\pi_{kl} - \pi_k \pi_l) d_k d_l y_k y_l / \pi_{kl}$.

If auxiliary information $\{\mathbf{x}_k\}_{k \in U}$ is available, then it is of interest to improve the HT estimator of t_y by considering sampling designs based on this auxiliary information such as stratified or proportional to size sampling designs. The auxiliary information may also be used for building estimators more efficient than the HT estimator. Mainly, there are two ways to incorporate auxiliary information depending on whether or not a model is fitted to the data. In the model-assisted approach (Särndal et al. 1992), we assume that the $\{y_k\}_{k \in U}$ values are independent realizations from an infinite superpopulation and the working model ξ relating the auxiliary information \mathbf{x}_k to y_k as follows:

$$\xi : y_k = f(\mathbf{x}_k) + \varepsilon_k, \quad k \in U, \quad (2)$$

with $\mathbb{E}_\xi(\varepsilon_k) = 0$ and $\mathbb{V}_\xi(\varepsilon_k) = \sigma_k^2 > 0$. If $f(\mathbf{x}_k)$ was known for all $k \in U$, then t_y would be estimated by the generalized difference estimator (Cassel et al. 1976)

$$t_{y,\mathbf{x}}^{\text{diff}} = \hat{t}_{yd} - \left(\sum_{k \in s} d_k f(\mathbf{x}_k) - \sum_{k \in U} f(\mathbf{x}_k) \right), \quad (3)$$

which is the difference between the HT estimator \hat{t}_{yd} and the bias of $\hat{t}_{yd} - t_y$ under the model ξ . The estimator (3) can also be seen as the prediction of t_y under the model ξ plus a design-bias adjustment. In practice, we never know the true f , thus we have to build an estimator of it by using a two-step procedure: we first estimate f by \tilde{f} under the model ξ and using the data $\{(y_k, \mathbf{x}_k)\}_{k \in U}$, and next, we estimate \tilde{f} by \hat{f} under the sampling design and using the survey data $\{(y_k, \mathbf{x}_k)\}_{k \in s}$. Plugging \hat{f} in (3) yields the model-assisted estimator of t_y , $\hat{t}_{yw}^{\text{ma}} = \hat{t}_{yd} - (\sum_{k \in s} d_k \hat{f}(\mathbf{x}_k) - \sum_{k \in U} \hat{f}(\mathbf{x}_k))$. If $f(\mathbf{x}_k) = \mathbf{x}_k^T \boldsymbol{\beta}$, then \hat{t}_{yw}^{ma} is the well-known generalized regression estimator (GREG) extensively used in practice and described by Särndal et al. (1992)

$$\hat{t}_{yw}^{\text{greg}} = \hat{t}_{yd} - \left(\sum_{k \in s} d_k \mathbf{x}_k - \sum_{k \in U} \mathbf{x}_k \right)^T \hat{\boldsymbol{\beta}}, \quad (4)$$

where $\hat{\beta} = (\sum_{k \in S} d_k \mathbf{x}_k \mathbf{x}_k^T / \sigma_k^2)^{-1} \sum_{k \in S} d_k \mathbf{x}_k y_k / \sigma_k^2$ is the design-based least square estimator of β . The GREG estimator is efficient if the linear model fits the data well but, if the model is misspecified, the GREG estimator exhibits no improvement over the HT estimator and may even lead to a loss of efficiency. One way of guarding against model failure is to use nonparametric regression which does not require a predefined parametric mathematical expression for f . In a model-assisted approach, Breidt and Opsomer (2000) proposed local linear estimators and Breidt et al. (2005) considered penalized spline regression with a piecewise polynomial basis and fixed knots and McConville and Breidt (2013) extended the theoretical justification for that estimator, by allowing the number of knots to increase: Goga (2005) used B -spline nonparametric regression and Goga and Ruiz-Gazen (2014) used penalized B -spline regression.

We intend to give a review of the use of nonparametric B -spline estimation in a survey sampling framework and adopting a design-based inference point of view. This review includes the estimation of totals (Sect. 2) and nonlinear study parameters (Sect. 3) with model-assisted and calibration techniques. Some new properties and improved consistency rates of the suggested estimators are presented. When some sampled individuals do not respond, we deal with nonresponse, and Sect. 4 describes briefly how we can improve the estimation of finite population totals in the presence of item nonresponse by considering B -spline imputation models.

2 B-Spline Model-Assisted Estimator for Finite Population Totals

Consider the superpopulation model given in (2) with f an unknown function and a univariate x -variable. Without loss of generality, we suppose that $x_k \in [0, 1]$. We suppose also that x_k is known for all $k \in U$.

To estimate the unknown regression function f , we use spline approximation. For a fixed $m > 1$, the set $S_{K,m}$ of spline functions of order m with K equidistant interiors knots $0 = \xi_0 < \xi_1 < \dots < \xi_K < \xi_{K+1} = 1$ is the set of piecewise polynomials of degree $m - 1$ that are smoothly connected at the knots

$$S_{K,m} = \{t \in C^{m-2}[0, 1] : t(z) \text{ is a polynomial of degree } (m-1) \text{ on each interval } [\xi_i, \xi_{i+1}]\}.$$

For $m = 1$, $S_{K,m}$ is the set of step functions with jumps at knots. For each fixed set of knots, $S_{K,m}$ is a linear space of functions of dimension $q = K + m$. A basis for this linear space is provided by the B-spline functions $\{B_j(\cdot)\}_{j=1}^q$ defined by $B_j(x) = (\xi_j - \xi_{j-m}) \sum_{l=0}^m (\xi_{j-l} - x)_+^{m-1} / \prod_{r=0, r \neq l}^m (\xi_{j-l} - \xi_{j-r})$ with $(\xi_{j-l} - x)_+^{m-1} = (\xi_{j-l} - x)^{m-1}$ if $\xi_{j-l} \geq x$ and zero, otherwise (Schumaker 1981; Dierckx 1993). Each function $B_j(\cdot)$ has the knots ξ_{j-m}, \dots, ξ_j with $\xi_r = \xi_{\min(\max(r,0), K+1)}$ for $r = j - m, \dots, j$ (Zhou et al. 1998) which means that its support consists of a small, fixed, finite number of intervals between knots. Figure 1 exhibits the six B -

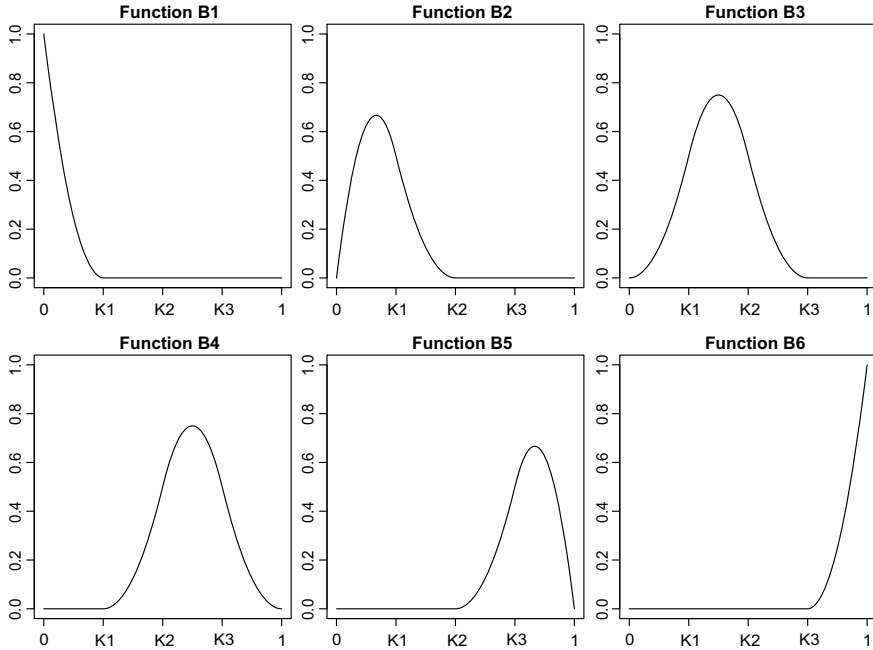


Fig. 1 B-spline basis functions for $K = 3$ interior knots and $m = 3$

spline basis functions for $K = 3$ interior knots and $m = 3$. Other important properties of B -splines are:

$$B_j(x) \geq 0 \text{ for all } x \in [0, 1]$$

and

$$\sum_{j=1}^q B_j(x) = 1, \quad x \in [0, 1]. \tag{5}$$

The above property proved particularly useful in a survey sampling framework. The B -spline estimation of f is given by

$$\tilde{f}(x_k) = \mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}, \quad k \in U, \tag{6}$$

where $\mathbf{b}^T(x_k) = (B_1(x_k), \dots, B_q(x_k))$ and $\tilde{\boldsymbol{\theta}}$ is the ordinary least square minimizer of

$$\tilde{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^q} \sum_{k \in U} (y_k - \mathbf{b}^T(x_k)\boldsymbol{\theta})^2 = (\mathbf{B}_U^T \mathbf{B}_U)^{-1} \mathbf{B}_U^T \mathbf{y}_U, \tag{7}$$

where \mathbf{B}_U is the $N \times q$ dimensional matrix with $\mathbf{b}^T(x_k)$ as rows, $\mathbf{B}_U = (\mathbf{b}^T(x_k))_{k \in U}$, and \mathbf{y}_U the $N \times 1$ dimensional vector of population y -values, $\mathbf{y}_U = (y_k)_{k \in U}$. There is no general rule for choosing the number of knots but it should be large enough to have enough points between knots. Ruppert et al. (2003) recommend that no more than 30–40 knots should be used and they give a simple rule for choosing K . As for the degree m , Ruppert et al. (2003) recommend $m = 3$ or $m = 4$. One way to overcome the issue of knot number, is to consider many knots and to constrain their influence by introducing a penalty. The penalized spline estimator of $f(x_k)$ is given by $\tilde{f}_\lambda(x_k) = \mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}_\lambda$ with $\tilde{\boldsymbol{\theta}}_\lambda$ the least square minimizer of

$$\tilde{\boldsymbol{\theta}}_\lambda = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^q} \sum_{k \in U} (y_k - \mathbf{b}^T(x_k)\boldsymbol{\theta})^2 + \lambda \int_0^1 [(\mathbf{b}^T(t)\boldsymbol{\theta})^{(\ell)}]^2 dt, \quad (8)$$

where (ℓ) represents the ℓ -th derivate with $\ell \leq m - 1$. The solution of (8) is a ridge-type estimator

$$\tilde{\boldsymbol{\theta}}_\lambda = \left(\sum_{k \in U} \mathbf{b}(x_k)\mathbf{b}^T(x_k) + \lambda \mathbf{D}_\ell \right)^{-1} \sum_{k \in U} \mathbf{b}(x_k)y_k = (\mathbf{B}_U^T \mathbf{B}_U + \lambda \mathbf{D}_\ell)^{-1} \mathbf{B}_U^T \mathbf{y}_U, \quad (9)$$

where \mathbf{D}_ℓ is the squared L^2 norm applied to the ℓ -th derivative of $\mathbf{b}^T \boldsymbol{\theta}$. Because the derivative of a B -spline function of order m may be written as a linear combination of B -spline functions of order $m - 1$, for equidistant knots, we obtain that $\mathbf{D}_\ell = K^{2\ell} \nabla_\ell^T \mathbf{R} \nabla_\ell$, where the matrix \mathbf{R} has elements $R_{ij} = \int_0^1 B_i^{(m-\ell)}(t) B_j^{(m-\ell)}(t) dt$ with $B_i^{(m-\ell)}$ as the B -spline function of order $m - \ell$ and ∇_ℓ as the matrix corresponding to the ℓ -th order difference operator (Claeskens et al. 2009).

The amount of smoothing is controlled by $\lambda > 0$. The case $\lambda = 0$ results in an unpenalized B-spline estimator whose properties have been extensively studied in the literature, see Agarwal and Studden (1980), Burman (1991), Zhou et al. (1998), among others. A review of spline use in statistics is given in Besse and Thomas-Agnan (1989). The case $\lambda \rightarrow \infty$ is equivalent to fitting a $(\ell - 1)$ -th degree polynomial. The theoretical properties of penalized splines with $\lambda > 0$, have been studied by Cardot (2002), Hall and Opsomer (2005), Kauermann et al. (2009), and Claeskens et al. (2009).

2.1 B-Spline Model-Assisted Estimation

In a survey sampling framework, the y_k 's values are available only for the sampled individuals, so $\tilde{f}(x_k)$ given in (6) cannot be used in practice. We estimate it by

$$\hat{f}(x_k) = \mathbf{b}^T(x_k)\hat{\boldsymbol{\theta}}, \quad k \in U, \quad (10)$$

where $\hat{\boldsymbol{\theta}}$ is the minimizer of the weighted least square sum

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg \min_{\boldsymbol{\theta} \in \mathbf{R}^q} \sum_{k \in S} d_k (y_k - \mathbf{b}^T(x_k) \boldsymbol{\theta})^2 \\ &= (\mathbf{B}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{B}_s)^{-1} \mathbf{B}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s = \left(\sum_{k \in S} d_k \mathbf{b}(x_k) \mathbf{b}^T(x_k) \right)^{-1} \sum_{k \in S} d_k \mathbf{b}(x_k) y_k, \quad (11)\end{aligned}$$

where $\mathbf{B}_s^T = (\mathbf{b}^T(x_k))_{k \in S}$, $\mathbf{y}_s = (y_k)_{k \in S}$ and $\boldsymbol{\Pi}_s = \text{diag}(\pi_k)_{k \in S}$ and provided that the matrix $\mathbf{B}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{B}_s$ is invertible. The sample-based estimator $\hat{\boldsymbol{\theta}}$ can be viewed as a substitution estimator of $\tilde{\boldsymbol{\theta}}$ given in (7) since every finite population total from the expression of $\tilde{\boldsymbol{\theta}}$ is substituted by its HT estimator.

The B -spline model-assisted estimator for estimating the total t_y has been suggested by Goga (2005) and obtained by plugging $\hat{f}(x_k)$ in (3)

$$\begin{aligned}\hat{t}_{bs} &= \sum_{k \in S} d_k (y_k - \hat{f}(x_k)) + \sum_{k \in U} \hat{f}(x_k) \\ &= \sum_{k \in S} d_k y_k - \left(\sum_{k \in S} d_k \mathbf{b}(x_k) - \sum_{k \in U} \mathbf{b}(x_k) \right)^T \hat{\boldsymbol{\theta}}. \quad (12)\end{aligned}$$

It is very important to note that, even if a nonparametric model (2) has been assumed, the B -spline model-assisted estimator \hat{t}_{bs} shares many properties of the GREG estimator. First of all, we can see from relation (12) that \hat{t}_{bs} may be written as a GREG estimator that uses the auxiliary information contained in vectors $\mathbf{b}(x_k) = (B_j(x_k))_{j=1}^q$. In practice, we choose the number of knots K and the degree m of splines and the estimator \hat{t}_{bs} can be seen as a GREG-type estimator with $q = K + m$ regressors; q should not be too large with respect to the sample size n . Note, however, that the B -spline model-assisted estimator, as well as any other nonparametric model-assisted estimator, requires that the values x_k 's to be known for all the population units. From an asymptotic point of view, the size of $\mathbf{b}(x_k)$ and $\hat{\boldsymbol{\theta}}$ is now $q \rightarrow \infty$, which is different from the linear GREG estimator, which considered that the number of regressors was fixed. However, these new regressors are uniformly bounded since $\|\mathbf{b}(x)\| \leq 1$ (Burman 1991) for all $x \in [0, 1]$; this property proved very useful in the technical proofs.

Another interesting property of \hat{t}_{bs} is the fact that the HT estimator of the residuals $y_k - \hat{f}(x_k)$ is zero

$$\sum_{k \in S} d_k \hat{f}(x_k) = \mathbf{1}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{B}_s \hat{\boldsymbol{\theta}} = \mathbf{1}_q^T \mathbf{B}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{B}_s \hat{\boldsymbol{\theta}} = \mathbf{1}_q^T \mathbf{B}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s = \mathbf{1}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s = \sum_{k \in S} d_k y_k, \quad (13)$$

where $\mathbf{1}_s, \mathbf{1}_q$ are vectors of ones of dimension n and respectively q . To obtain (13), we have used twice the fact that $\mathbf{1}_q^T \mathbf{B}_s^T = \mathbf{1}_s^T$ obtained from the property of B -spline

functions that $\sum_{j=1}^q B_j(x) = 1$. We then obtain that the estimator \hat{t}_{bs} is equal to the finite population total of the estimated predictions $\hat{f}(x_k)$, a form usually called the projection estimator

$$\hat{t}_{bs} = \sum_{k \in U} \hat{f}(x_k) = \sum_{k \in U} \mathbf{b}'(x_k) \hat{\boldsymbol{\theta}}. \quad (14)$$

In a linear modeling context, a similar property is shared by the usual GREG estimator given in (4) if the model variance of y satisfies the property $\sigma_k^2 = \boldsymbol{\lambda}^T \mathbf{x}_k$ for $\boldsymbol{\lambda}$ a vector of constants. This is true, for example, for $\sigma_k = \sigma$ for all $k \in U$ and if intercept is included in the model. With B -spline nonparametric modeling, the property of B -spline functions (5) is equivalent in a way to the presence of the intercept in the model.

Moreover, the B -spline model-assisted estimator from (12) may be written as a weighted sum of the y_k 's values as follows:

$$\hat{t}_{bs} = \sum_{k \in S} w_{ks}^{bs} y_k, \quad (15)$$

with nonparametric weights given by

$$w_{ks}^{bs} = d_k - d_k \mathbf{b}^T(x_k) \left(\sum_{l \in S} d_l \mathbf{b}(x_l) \mathbf{b}^T(x_l) \right)^{-1} \left(\sum_{l \in S} d_l \mathbf{b}(x_l) - \sum_{l \in U} \mathbf{b}(x_l) \right). \quad (16)$$

We have

$$d_k \mathbf{b}^T(x_k) \left(\sum_{l \in S} d_l \mathbf{b}(x_l) \mathbf{b}^T(x_l) \right)^{-1} \sum_{l \in S} d_l \mathbf{b}(x_l) \cdot \mathbf{1} = d_k \mathbf{b}^T(x_k) \mathbf{1}_q = d_k, \quad k \in S,$$

by the fact that $\mathbf{1} = \mathbf{b}^T(x_l) \mathbf{1}_q$ for all $l \in U$. So, the B -spline weights may be written as

$$w_{ks}^{bs} = d_k \mathbf{b}^T(x_k) \left(\sum_{l \in S} d_l \mathbf{b}(x_l) \mathbf{b}^T(x_l) \right)^{-1} \left(\sum_{l \in U} \mathbf{b}(x_l) \right), \quad k \in S, \quad (17)$$

The weights given in (16) are GREG-type weights based on the regressors $(\mathbf{b}(x_k))_{k \in U}$ and they are always equal to weights given in (17) which correspond to the projection form of \hat{t}_{bs} given in (14). The weights w_{ks}^{bs} do not depend on the study variable y , so they can be used to estimate totals of other variables than y or even more complicated study parameters as described in Sect. 3.

The B -spline model-assisted estimator is asymptotically design-unbiased and consistent for t_y (Goga 2005). However, as expected, the convergence rate of \hat{t}_{bs} depends on the number of knots $K \rightarrow \infty$. We propose here a different decomposition of $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}$

which allows getting a better convergence rate of $\hat{t}_{bs} - t_y$ than the one given in Goga (2005). This decomposition is the following:

$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = \hat{\mathbf{T}}^{-1} \left(\sum_{k \in S} d_k E_k - \sum_{k \in U} E_k \right), \quad (18)$$

where $\hat{\mathbf{T}} = \sum_{k \in S} d_k \mathbf{b}(x_k) \mathbf{b}^T(x_k)$ and $E_k = \mathbf{b}(x_k)(y_k - \mathbf{b}^T(x_k) \tilde{\boldsymbol{\theta}})$ with $\sum_{k \in U} E_k = 0$. Under the assumptions given in the Appendix, we obtain

$$\frac{1}{N^2} \mathbb{E}_p \left(\sum_{k \in S} d_k E_k - \sum_{k \in U} E_k \right)^2 \leq \left(\frac{1 - \tilde{c}}{N \tilde{c}} + \frac{n \max_{k \neq l \in U} |\pi_{kl} - \pi_k \pi_l|}{n \tilde{c}^2} \right) \frac{1}{N} \sum_{k \in U} \|E_k\|^2. \quad (19)$$

We have $\|\mathbf{b}(x_k)\| \leq 1$ for all $k \in U$, so

$$\frac{1}{N} \sum_{k \in U} \|E_k\|^2 \leq \frac{2}{N} \sum_{k \in U} y_k^2 + \frac{2}{N} \sum_{k \in U} |\mathbf{b}^T(x_k) \tilde{\boldsymbol{\theta}}|^2 \leq \frac{2}{N} \sum_{k \in U} y_k^2 + 2 \|\tilde{\boldsymbol{\theta}}\|^2 \|\frac{1}{N} \mathbf{B}^T \mathbf{B}\| \leq C,$$

with a constant $C > 0$ not depending on n and K , and consequently, $N^{-1}(\sum_{k \in S} d_k E_k - \sum_{k \in U} E_k) = O_p(n^{-1/2})$. We have used the fact that $N^{-1} \sum_{k \in U} |\mathbf{b}^T(x_k) \tilde{\boldsymbol{\theta}}|^2 = \tilde{\boldsymbol{\theta}}^T (N^{-1} \sum_{k \in U} \mathbf{b}(x_k) \mathbf{b}^T(x_k)) \tilde{\boldsymbol{\theta}} \leq \|\tilde{\boldsymbol{\theta}}\|^2 \|N^{-1} \mathbf{B}^T \mathbf{B}\| = O(1)$ by the fact that $\|\tilde{\boldsymbol{\theta}}\| = O(K^{1/2})$ (Goga 2005) and $\|N^{-1} \mathbf{B}^T \mathbf{B}\| = O(K^{-1})$ (Zhou et al. 1998). We have also $N \hat{\mathbf{T}}^{-1} = O_p(K)$ (Goga 2005) which gives

$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = O_p(K n^{-1/2}).$$

This consistency rate is $K^{1/2}$ smaller than the one obtained in Goga (2005). So, if $K = O(n^\alpha)$ with $0 < \alpha < 1/2$, then $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = o_p(1)$. Now, we can write

$$N^{-1}(\hat{t}_{bs} - t_y) = N^{-1}(\tilde{t}_{bs}^{\text{diff}} - t_y) + N^{-1} \left(\sum_{k \in S} d_k \mathbf{b}(x_k) - \sum_{k \in U} \mathbf{b}(x_k) \right)^T (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}), \quad (20)$$

where $\tilde{t}_{bs}^{\text{diff}} = \sum_{k \in S} d_k y_k - \left(\sum_{k \in S} d_k \tilde{f}(x_k) - \sum_{k \in U} \tilde{f}(x_k) \right)$ is the pseudo generalized B -spline difference estimator. Using the same arguments as in (19), we can prove that $N^{-1}(\hat{t}_{bs} - t_y) = O_p(n^{-1/2})$ as well as $N^{-1}(\tilde{t}_{bs}^{\text{diff}} - t_y) = O_p(n^{-1/2})$. The pseudo generalized B -spline difference estimator $\tilde{t}_{bs}^{\text{diff}}$ is design-unbiased for t_y and consistent with the parametric rate $n^{-1/2}$ and no longer $K n^{-1/2}$ as given in Goga (2005). This improved result was obtained by considering a similar decomposition as in (19) and that $N^{-1} \sum_{k \in U} \tilde{f}^2(x_k) = \tilde{\boldsymbol{\theta}}^T (N^{-1} \mathbf{B}^T \mathbf{B}) \tilde{\boldsymbol{\theta}} \leq \|\tilde{\boldsymbol{\theta}}\|^2 \|N^{-1} \mathbf{B}^T \mathbf{B}\| = O(1)$. We get

$$N^{-1}(\hat{t}_{bs} - t_y) = N^{-1}(\tilde{t}_{bs}^{\text{diff}} - t_y) + O_p(K n^{-1}).$$

If again $K = O(n^\alpha)$ with $0 < \alpha < 1/2$, then the B -spline model-assisted estimator is asymptotically consistent for t_y , namely $N^{-1}(\hat{t}_{bs} - t_y) = O_p(n^{-1/2})$ and asymptotically equivalent to the B -spline generalized difference estimator in the sense that $N^{-1}(\hat{t}_{bs} - t_y) = N^{-1}(\tilde{t}_{bs}^{\text{diff}} - t_y) + o_p(n^{-1/2})$. This means that the asymptotic design-based variance of \hat{t}_{bs} is the design-based variance of $\tilde{t}_{bs}^{\text{diff}}$,

$$A\mathbb{V}_p(\hat{t}_{bs}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) d_k d_l (y_k - \tilde{f}(x_k))(y_l - \tilde{f}(x_l)).$$

Goga and Ruiz-Gazen (2014) considered penalized B -spline model-assisted estimator given by

$$\begin{aligned} \hat{t}_{bs,\lambda} &= \sum_{k \in s} d_k (y_k - \hat{f}_\lambda(x_k)) + \sum_{k \in U} \hat{f}_\lambda(x_k) \\ &= \sum_{k \in s} d_k y_k - \left(\sum_{k \in s} d_k \mathbf{b}(x_k) - \sum_{k \in U} \mathbf{b}(x_k) \right)^T \hat{\boldsymbol{\theta}}_\lambda, \end{aligned}$$

where $\hat{\boldsymbol{\theta}}_\lambda = (\sum_{k \in s} d_k \mathbf{b}(x_k) \mathbf{b}^T(x_k) + \lambda \mathbf{D}_\ell)^{-1} (\sum_{k \in s} d_k \mathbf{b}(x_k) y_k)$ is obtained from (8) by estimating each total by its HT estimator. The penalized estimator $\hat{t}_{bs,\lambda}$ can be also written as a weighted sum of y_k 's values with weights $w_{ks}^{\text{bs}}(\lambda)$ similar to (16). However, it is very important to notice that the weights $w_{ks}^{\text{bs}}(\lambda)$ can be written in the projection form (17), namely $w_{ks}^{\text{bs}}(\lambda) = d_k \mathbf{b}^T(x_k) (\sum_{l \in s} d_l \mathbf{b}(x_l) \mathbf{b}^T(x_l) + \lambda \mathbf{D}_\ell)^{-1} (\sum_{l \in U} \mathbf{b}(x_l))$ which means that the penalized estimator $\hat{t}_{bs,\lambda}$ is also a projection estimator

$$\hat{t}_{bs,\lambda} = \sum_{k \in U} \hat{f}_\lambda(x_k) = \sum_{k \in U} \mathbf{b}^T(x_k) \hat{\boldsymbol{\theta}}_\lambda.$$

Again, this property is due to the fact that $\mathbf{b}^T(x_k) \mathbf{1}_q = 1$ for all $k \in U$ and $\mathbf{D}_\ell \mathbf{1}_q = \mathbf{0}_q$. We get $\sum_{l \in s} d_l \mathbf{b}(x_l) \cdot \mathbf{1} = (\sum_{l \in s} d_l \mathbf{b}(x_l) \mathbf{b}^T(x_l) + \lambda \mathbf{D}_\ell) \mathbf{1}_q$ and

$$d_k \mathbf{b}^T(x_k) \left(\sum_{l \in s} d_l \mathbf{b}(x_l) \mathbf{b}^T(x_l) + \lambda \mathbf{D}_\ell \right)^{-1} \sum_{l \in s} d_l \mathbf{b}(x_l) = d_k \mathbf{b}^T(x_k) \mathbf{1}_q = d_k, \quad k \in s.$$

An improved consistency rate is obtained again by using as for the unpenalized case the decomposition

$$\hat{\boldsymbol{\theta}}_\lambda - \tilde{\boldsymbol{\theta}}_\lambda = \hat{\mathbf{T}}_\lambda^{-1} \left(\sum_{k \in s} d_k E_{k,\lambda} - \sum_{k \in U} E_{k,\lambda} \right),$$

where $\hat{\mathbf{T}}_\lambda = \sum_{k \in s} d_k \mathbf{b}(x_k) \mathbf{b}^T(x_k) + \lambda \mathbf{D}_\ell$ and $E_{k,\lambda} = \mathbf{b}(x_k)(y_k - \mathbf{b}^T(x_k) \tilde{\boldsymbol{\theta}}_\lambda)$ with $\sum_{k \in U} E_{k,\lambda} = \lambda \mathbf{D}_\ell \tilde{\boldsymbol{\theta}}_\lambda$. Using the same lines as in (19) and assumptions from the Appendix, we obtain again that $N^{-1}(\sum_{k \in s} d_k E_{k,\lambda} - \sum_{k \in U} E_{k,\lambda}) = O_p(n^{-1/2})$ since

$N^{-1} \sum_{k \in U} \|E_{k,\lambda}\|^2 \leq 2 \sum_{k \in U} y_k^2 / N + 2 \|\tilde{\boldsymbol{\theta}}_\lambda\|^2 \|N^{-1} \mathbf{B}_U^T \mathbf{B}_U\| = O(1)$ by using the fact that $\|\tilde{\boldsymbol{\theta}}_\lambda\| = O(K^{1/2})$ (Goga and Ruiz-Gazen 2014). We also have $N \hat{\mathbf{T}}_\lambda^{-1} = O_p(K)$ (Goga and Ruiz-Gazen 2014), so

$$\hat{\boldsymbol{\theta}}_\lambda - \tilde{\boldsymbol{\theta}}_\lambda = O_p(Kn^{-1/2}).$$

Using the same decomposition as in (20), we get

$$N^{-1}(\hat{t}_{bs,\lambda} - t_y) = N^{-1}(\tilde{t}_{bs,\lambda}^{\text{diff}} - t_y) + O_p(Kn^{-1}),$$

where $\tilde{t}_{bs,\lambda}^{\text{diff}} = \sum_{k \in S} d_k y_k - \left(\sum_{k \in S} d_k \tilde{f}_\lambda(x_k) - \sum_{k \in U} \tilde{f}_\lambda(x_k) \right)$ is the penalized B -spline difference estimator which is again design-unbiased and $n^{-1/2}$ design-consistent, $N^{-1}(\tilde{t}_{bs,\lambda}^{\text{diff}} - t_y) = O_p(n^{-1/2})$ by using the same arguments as in the unpenalized case. If $K = O(n^\alpha)$ with $0 < \alpha < 1/2$, then the penalized B -spline model-assisted estimator $\hat{t}_{bs,\lambda}$ is asymptotically consistent for t_y and asymptotically equivalent to $\tilde{t}_{bs,\lambda}^{\text{diff}}$. So, the asymptotic variance of $\hat{t}_{bs,\lambda}$ is the variance of the HT estimator of residuals $y_k - \tilde{f}_\lambda(x_k)$ as in (21) for $\tilde{f}(x_k)$ replaced by $\tilde{f}_\lambda(x_k)$.

2.2 B -Spline Calibration Estimator

The calibration approach (Deville and Särndal 1992) is a method widely used in national statistical agencies. It consists of finding new weights $(w_{ks}^{\text{cal}})_{k \in S}$ that are as close as possible to the sampling weights $(d_k)_{k \in S}$ and such that $\sum_{k \in S} w_{ks}^{\text{cal}} \mathbf{x}_k$ perfectly estimates the known population total of auxiliary information $\sum_{k \in U} \mathbf{x}_k$. The calibrated estimator $\sum_{k \in S} w_{ks}^{\text{cal}} y_k$ is highly efficient for estimating t_y if the relationship f between y and x is close to a linear relationship but its efficiency may be worse than the HT estimator if f is nonlinear. In order to overcome this issue, Goga and Ruiz-Gazen (2019) suggest the B -spline calibration: they suggest finding the calibration weights $(w_{ks}^{\text{cal}})_{k \in S}$ that minimize a distance measure Υ_s to the sampling weights $(d_k)_{k \in S}$ and subject to the following calibration constraints:

$$\sum_{k \in S} w_{ks}^{\text{cal}} \mathbf{b}(x_k) = \sum_{k \in U} \mathbf{b}(x_k). \quad (21)$$

Constraints are now on the B -splines function values $\{B_j(x_k)\}_{j=1}^q$ and not directly on x_k as it is the case in the classical calibration as suggested by Deville and Särndal (1992), we need for that to know x_k for all $k \in U$. However, polynomials x^ℓ belong to the space spanned by $\{B_j(\cdot)\}_{j=1}^q$ for all $\ell = 0, \dots, q-1$. As a result, weights satisfying (21) will also satisfy

$$\sum_{k \in s} w_{ks}^{\text{cal}} = N, \quad \sum_{k \in s} w_{ks}^{\text{cal}} x_k = \sum_{k \in U} x_k,$$

$$\sum_{k \in s} w_{ks}^{\text{cal}} x_k^\ell = \sum_{k \in U} x_k^\ell, \quad \ell = 2, \dots, q - 1.$$

The calibration constraints on the first moments of x may be interpreted as a property of consistency with known totals, the population size N and that of the auxiliary information x , property mostly looked for in national statistical agencies. Considering calibration on higher powers of x is strongly advised by Särndal (2007) in order to cover a larger class of relationships such as higher order polynomials between y and x . We argue that considering calibration of B -spline functions as in (21) allows estimating totals of y 's efficiently when the relationship f between y and x is polynomial or even more general than polynomials since the calibration weights $(w_{ks}^{\text{cal}})_{k \in s}$ will also satisfy the calibration constraints

$$\sum_{k \in s} w_{ks}^{\text{cal}} \hat{f}(x_k) = \sum_{k \in U} \hat{f}(x_k),$$

where \hat{f} is the estimated prediction of the unknown f given in (10). The method we suggest is different from the model-calibration suggested by Montanari and Ranalli (2005), who considered calibration on estimated predictions and obtained weights depending on the study variable which does not happen in our case. The method we suggest can be extended easily to multivariate auxiliary information and additive models by adding additional constraints in (21).

We can consider different distance functions Υ_s as suggested in Deville and Särndal (1992). The resulting estimators are asymptotically equivalent to the estimator obtained by minimizing the chi-squared distance function $\Upsilon_s(\mathbf{w}) = \sum_{k \in s} (w_k - d_k)^2 / q_k d_k$ where the q_k 's are known positive constants used to control the variability of the observations and are unrelated to d_k . Most of the time, $q_k = 1$ for all $k \in U$ or $q_k = 1/\sigma_k^2$ with $\sigma_k^2 = \mathbb{V}_\xi(\varepsilon_k)$, where ε_k are the errors in model (2). With the chi-square distance Υ_s , the resulting calibration weights are given by

$$w_{ks}^{\text{cal}} = d_k - q_k d_k \mathbf{b}^T(x_k) \left(\sum_{l \in s} q_l d_l \mathbf{b}(x_l) \mathbf{b}^T(x_l) \right)^{-1} \left(\sum_{l \in s} d_l \mathbf{b}(x_l) - \sum_{l \in U} \mathbf{b}(x_l) \right), \quad k \in s$$

and the calibration estimator $\hat{t}_{yw}^{\text{cal}}$ is given by a GREG-type estimator

$$\hat{t}_{yw}^{\text{cal}} = \sum_{k \in s} d_k y_k - \left(\sum_{k \in s} d_k \mathbf{b}(x_k) - \sum_{k \in U} \mathbf{b}(x_k) \right)^T \hat{\boldsymbol{\theta}}(q)$$

where $\hat{\boldsymbol{\theta}}(q) = \left(\sum_{k \in s} d_k q_k \mathbf{b}(x_k) \mathbf{b}^T(x_k) \right)^{-1} \sum_{k \in s} d_k q_k \mathbf{b}(x_k) y_k$.

The B -spline calibration estimator $\hat{t}_{yw}^{\text{cal}}$ from (22) is similar to the B -spline model-assisted estimator \hat{t}_{bs} from (12) and they are equal if $q_k = 1$ for all $k \in U$.

It can be shown by using the same arguments as in Sect. 2.1 that the B -spline calibration estimator $\hat{t}_{yw}^{\text{cal}}$ is asymptotically equivalent to the generalized difference estimator $\sum_{k \in S} d_k y_k - \left(\sum_{k \in S} d_k \mathbf{b}(x_k) - \sum_{k \in U} \mathbf{b}(x_k) \right)^T \tilde{\boldsymbol{\theta}}(q)$ with $\tilde{\boldsymbol{\theta}}(q) = \left(\sum_{k \in U} q_k \mathbf{b}(x_k) \mathbf{b}^T(x_k) \right)^{-1} \sum_{k \in U} q_k \mathbf{b}(x_k) y_k$ and its asymptotic variance is similar to (21), namely it is equal to the variance of the HT estimator $\sum_{k \in S} (y_k - \mathbf{b}^T(x_k) \tilde{\boldsymbol{\theta}}(q))$.

3 B-Spline Model-Assisted Estimator for Complex Parameters

The estimation of nonlinear parameters Φ in finite populations has become a crucial problem in many recent surveys. For example, in the European Statistics on Income and Living Conditions (EU-SILC) survey, several indicators for studying social inequalities and poverty are considered; these include the Gini index, the at-risk-of-poverty rate, the quintile share ratio and the low-income proportion. Thus, deriving estimators and confidence intervals for such indicators is particularly useful.

Consider now a parameter Φ which is more complicated than a total or a mean. Broadly speaking, linearization techniques consist in obtaining an expansion of an estimator $\hat{\Phi}$ of Φ as follows:

$$\hat{\Phi} - \Phi \simeq \sum_{k \in S} d_k u_k - \sum_{k \in U} u_k = \hat{t}_{ud} - t_u, \quad (22)$$

where u_k is a kind of artificial variable called *the linearized variable* of Φ by Deville (1999b). The way it is derived depends on the type of linearization method used which could include Taylor series (Särndal et al. 1992), estimating equations (Binder 1983) or influence function (Deville 1999b) approaches. The right hand-side of (22) is the difference between the HT estimator and the corresponding population total of the variable u_k over the population U . Consequently, the variance of the right hand-side is easily obtained and given by

$$\sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) d_k d_l u_k u_l. \quad (23)$$

We can see from above that we will achieve a small approximate variance and good precision for $\hat{\Phi}$ if we estimate $t_u = \sum_{k \in U} u_k$ in an efficient way, namely the variance given in (23) is small. However, linearized variables may have complicated mathematical expressions and it is not obvious how to improve efficiently the estimation of t_u . In particular, fitting a linear model onto a linearized variable may not be the most appropriate choice.

An example of such a situation is the estimation from survey data of the Gini coefficient (Gini 1914) as considered in Goga and Ruiz-Gazen (2014). Gini coefficient is one of the most famous concentration measures often of interest in economical studies. In finite populations, the Gini index (Nygard and Sandström 1985) is given by (after neglecting the term $1/N$)

$$G = \frac{\sum_{k \in U} y_k (2F(y_k) - 1)}{t_y},$$

where $F(y) = \sum_{k \in U} \mathbf{1}_{\{y_k \leq y\}}/N$ is the finite population empirical distribution function. The expression of the linearized variable $u_{k,G}$ of the Gini index (Binder and Kovacevic 1995; Deville 1999a) is given by

$$u_{k,G} = 2F(y_k) \frac{y_k - \bar{y}_{k,<}}{t_y} - y_k \frac{1+G}{t_y} + \frac{1-G}{N}, \quad k \in U,$$

where $\bar{y}_{k,<}$ is the mean of y_j lower than y_k and t_y the total of the y_k on U . Goga and Ruiz-Gazen (2014) considered a dataset of size 1000, extracted from the French Labor Force Survey; y_k (the wages of person k in 2000) was the study variable and x_k (the wages of person k in 1999) the auxiliary variable. In the left (resp. right) graphic of Fig. 2, the study variable y_k is plotted (resp. the linearized variable u_k) on the y -axis and the auxiliary variable x_k is plotted on the x -axis. The relationship between y and x is almost linear; however, the relationship between the linearized variable u_G and x is no longer linear. Therefore, nonparametric models should be preferred to parametric models to estimate the Gini index.

Let the superpopulation model ξ' relating the auxiliary information x_k to the linearized variable u_k given by

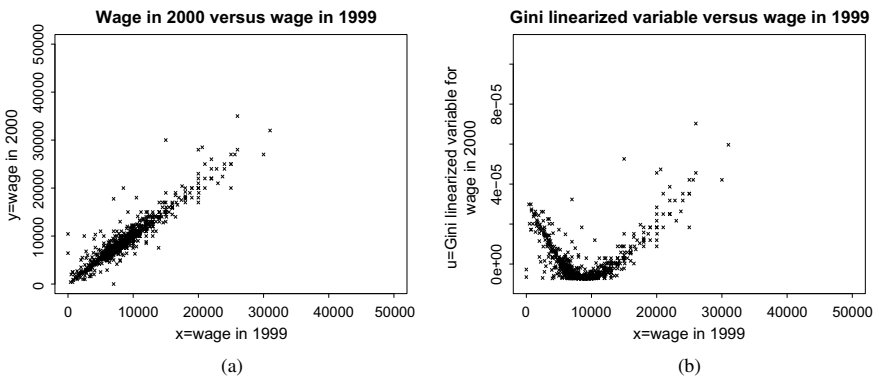


Fig. 2 Left plot: y_k : the wages of person k in 2000 against z_k : the wages of person k in 1999. Right plot: u_k : linearized variable of the Gini index for the wages in 2000 for person k against z_k : the wages of person k in 1999

$$\xi' : u_k = g(x_k) + \eta_k, \quad k \in U, \tag{24}$$

where g is unknown and η_k are centered and uncorrelated. Note that it is not really a model since we do not observe the linearized variables u_k . It can be viewed as a tool used to construct new weights for estimating $t_u = \sum_{k \in U} u_k$ efficiently and so by (22), for estimating efficiently G . From Sect. 2.1, we obtain that the nonparametric weights $(w_{ks}^{bs})_{k \in s}$ are given by the same relation (16). This fact is not really surprising since these weights have been obtained in the case of the estimation of the finite population total of y but they do not depend on y , so they can be used to estimate nonlinear parameters such as the Gini index

$$\widehat{G}^{bs} = \frac{\sum_{k \in s} w_{ks}^{bs} y_k (2\widehat{F}^{bs}(y_k) - 1)}{\sum_{k \in s} w_{ks}^{bs} y_k},$$

where $\widehat{F}^{bs}(y_k) = \sum_{l \in s} w_{ls}^{bs} \mathbf{1}_{\{y_l \leq y_k\}} / \sum_{l \in s} w_{ls}^{bs}$ is the nonparametric estimator of F . The asymptotic variance of \widehat{G}^{np} is given by Goga and Ruiz-Gazen (2014)

$$A\mathbb{V}_p(\widehat{G}^{np}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) d_k d_l (u_k - \tilde{g}(x_k))(u_l - \tilde{g}(x_l)),$$

where $\tilde{g}(x_k)$ is similar to (6) but computed from the data $\{(x_k, u_k)\}_{k \in U}$. The asymptotic variance is, in fact, the HT variance for the residuals $u_k - \tilde{g}(x_k)$ of the linearized variable u_k under the model ξ' given in (24). The smaller the residuals $u_k - \tilde{g}(x_k)$, $k \in U$ are, the better the estimator \widehat{G}^{np} for G is. Considering nonparametric models ξ' as in (24) and B -spline regression provide good prediction for rather complicated u_k and lead to low residuals $u_k - \tilde{g}(x_k)$. Nevertheless, unlike the GREG estimators derived under a linear model, nonparametric model-assisted estimators need x_k to be known for all the individuals from the population. Goga and Ruiz-Gazen (2014) suggested a variance estimator and gave assumptions under which the suggested variance estimator is consistent. They also conducted a large simulation study on data extracted from the French Labour Force which showed that the suggested estimator had a large efficiency gain for estimating nonlinear parameters such as the Gini index or the low-income proportion compared to usual estimators such as the HT estimator, the linear GREG or the poststratified estimator. They noticed that $m = 3$ was the best choice, especially for sample sizes smaller than 1000. Moreover, for $m = 3$ the coverage rates were good and results do not depend heavily on the number of knots and are similar for K between 2 and 4.

We consider here the estimation with B -splines of another nonlinear parameter, the functional median. Consider that the study variable \mathcal{Y} is now a curve or functional belonging to $L^2[0, \mathcal{T}]$, $Y_k(t)$ is the value of \mathcal{Y} recorded for the k -th individual at the instant t . For example, national companies of electricity such as the French company EDF (Electricité de France) or the Irish company have installed smart meters in households and companies over the past years. These meters are capable to record and send the electricity consumption at a very fine scale. The electricity consumption

is considered in this situation as a functional variable. With high dimensional data, it is not uncommon to have outlying curves, such as consumers with very high levels of electricity consumption. In such a situation, it is advisable to consider indicators which are more robust to outlying data than the mean profile, and the median is one of them. With a finite population point of view and functional data, the median curve calculated from the elements $\{Y_k\}_{k \in U}$ belonging to $L^2[0, \mathcal{T}]$ is defined by Gervini (2008)

$$m_N = \operatorname{argmin}_{y \in L^2[0, \mathcal{T}]} \sum_{k \in U} \|Y_k - y\|$$

and estimated from survey data by Chaouch and Goga (2012)

$$\hat{m}_n = \operatorname{argmin}_{y \in L^2[0, \mathcal{T}]} \sum_{k \in s} d_k \|Y_k - y\|.$$

If $Y_k, k \in s$ are not on a straight line and $\hat{m}_n \neq Y_k$, then the design-based estimator \hat{m}_n is the unique solution of the estimating equation

$$\sum_{k \in s} d_k \frac{Y_k - \hat{m}_n}{\|Y_k - \hat{m}_n\|} = 0. \tag{25}$$

Under broad assumptions, we linearize \hat{m}_n (Deville 1999b; Chaouch and Goga 2012) as follows:

$$\hat{m}_n = m_N + \sum_{k \in s} d_k u_{k, m_N} - \sum_{k \in U} u_{k, m_N} + o_p(n^{-1/2}), \tag{26}$$

where $u_{k, m_N} = \Gamma^{-1}((Y_k - m_N)/\|Y_k - m_N\|)$ is the linearized variable of m_N and $\Gamma = \sum_{k \in U} \|Y_k - m_N\|^{-1} [\mathbf{I} - (Y_k - m_N) \otimes (Y_k - m_N) \|Y_k - m_N\|^{-2}]$ is the Jacobian operator of $\sum_{k \in U} (Y_k - m_N)/\|Y_k - m_N\|$. Chaouch and Goga (2012) considered the estimation of the functional median with various sampling designs such as the simple random sampling without replacement (SRSWOR), the stratified sampling (STRAT) and proportional to size sampling designs on a population test of load electricity curves recorded every thirty minutes during two consecutive weeks. We consider here a set of about $N = 15000$ electricity curves. We plot in Fig. 3a the electricity load curves recorded during one week ($D = 336$ discretized points) for a sample of ten firms, the median curve computed from the population is plotted in red. The electricity consumption recorded during the first week was used as auxiliary information to improve the estimation of m_N at the sampling stage by stratifying the population or by selecting firms with a probability proportional to their past consumption. STRAT performed better than the SRSWOR for estimating m_N . However, even if the relationship between Y_k and the considered x_k was linear and going through the origin, proportional to size sampling design, performed very poorly for the estimation of the median m_N , in fact, it was even worse than the SRSWOR, while

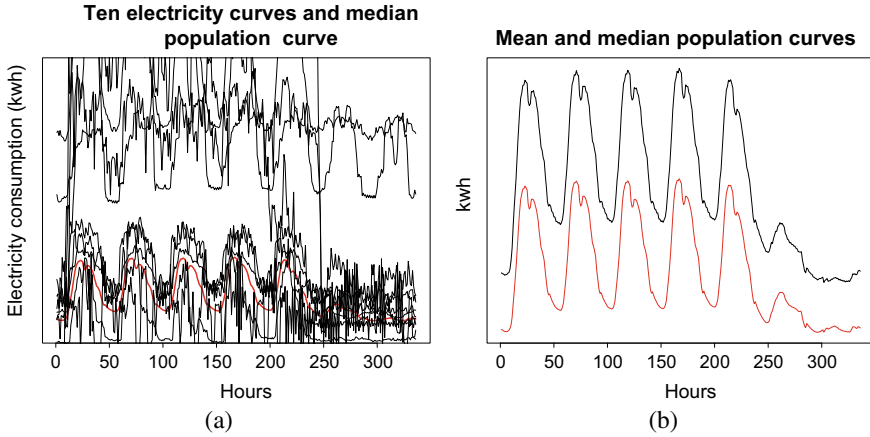


Fig. 3 a Electricity consumption curves for ten individuals, population median curve is plotted in red. b Mean curve of the test population in black and median curve of the test population in red

this design performed very good for the estimation of the total or mean consumption curve (Cardot et al. 2012). We suggest in the next *B*-spline model-assisted estimators to ameliorate the estimation of the median m_N obtained with the SRSWOR design or with proportional to size and without replacement sampling design (π ps).

Consider that a sample of size n is selected according to a SRSWOR design from the population of size N , so the inclusion probabilities are $\pi_k = n/N$ for all $k \in U$. The estimator \hat{m}_n without auxiliary information is obtained from (25) for $d_k = N/n$. Let x_k be the mean consumption during the first week, $x_k = \sum_{t=1}^D X_k(t_d)/D$, $k \in U$ which will be used as auxiliary information for improving the estimation of m_N when firms are selected according to SRSWOR. In order to do that, we use the nonparametric weights $(w_{ks}^{bs})_{k \in s}$ given in (17) to obtain the *B*-spline model-assisted estimator \hat{m}_n^{bs}

$$\sum_{k \in s} w_{ks}^{bs} \frac{Y_k - \hat{m}_n^{bs}}{\|Y_k - \hat{m}_n^{bs}\|} = 0.$$

To check the performance of \hat{m}_n^{bs} and compare it to \hat{m}_n , we select $I = 1000$ samples of size $n = 2000$ according to SRSWOR. In each sample, we compute \hat{m}_n by using the sampling weights d_k and \hat{m}_n^{bs} with the *B*-spline weights w_{ks}^{bs} . We compute for each sample the absolute errors, $R(\hat{m}_n) = \int_0^T |\hat{m}_n - m_N| \simeq \sum_{d=1}^D |\hat{m}_n(t_d) - m_N(t_d)|/D$ where $D = 336$ the number of discretized points. Figure 4a gives the distribution of these I absolute errors computed for both strategies. We can notice that considering *B*-spline estimation ($m = 3$ and $K = 8$ interior knots) of m_N leads to a substantial improvement of the median estimation.

Consider now the estimation of m_N with a π ps design. This design consists of selecting firms without replacement with probabilities of inclusion π_k proportional

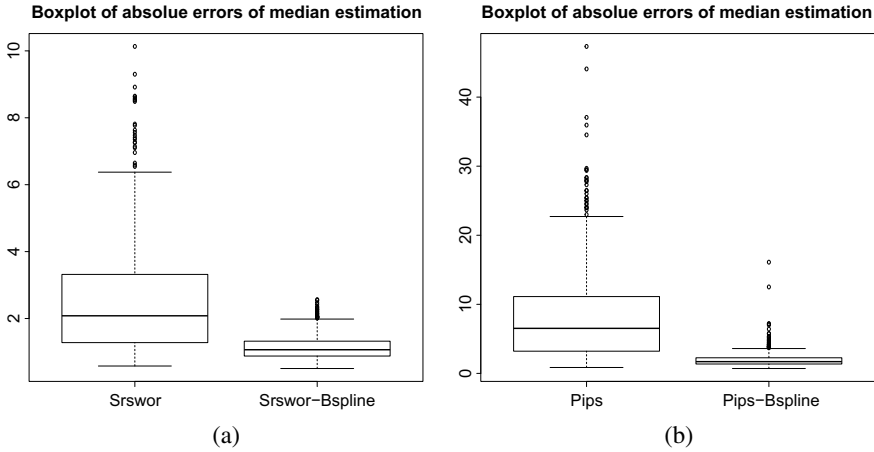


Fig. 4 **a:** Boxplots of absolute errors of I estimations of the median for the SRSWOR design and sampling weights (left boxplot) and B -spline weights (right boxplot); **b:** Boxplots of absolute errors of I estimations of the median for the π ps design and sampling weights (left boxplot) and B -spline weights (right boxplot)

to x_k , i.e., $\pi_k = nx_k / \sum_{k \in U} x_k$ for all $k \in U$. The estimator \hat{m}_n with a π ps sampling is obtained by solving (25) for $d_k = 1/\pi_k$.

This design performs very poorly for the estimation of the median curve, usually the estimation of the median curve with π ps fails for high and low values of the median. In order to understand the reason of it, consider the linearization of \hat{m}_n given in (26) with asymptotic HT variance given in the general formula (23) which, for π ps designs, is equal to Yates and Grundy (1953)

$$A\mathbb{V}_p(\hat{m}_n)(t) = -2^{-1} \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) (d_k u_{k,m_N}(t) - d_l u_{l,m_N}(t))^2, \quad t \in [0, T].$$

This means that the π ps sampling is efficient for estimating m_N if $u_k(t)$ is approximately proportional to x_k which is not the case here since the relationship between the linearized variable u_{k,m_N} and π_k is not linear. However, the π ps design is highly efficient for estimating the total consumption curve during the second week $t_Y = \sum_{k \in U} Y_k$ because, for all instants t from the second week, $Y_k(t)$ is approximately proportional to the consumption from the previous week, so $Y_k(t)$ is approximately proportional to π_k (Cardot et al. 2012). In order to improve the estimation of the median with a π ps design, we suggest an estimator of m_N which consists of modifying the sampling weights $d_k = 1/\pi_k$ by using a superpopulation model explaining the relationship between u_k and π_k as follows:

$$u_{k,m_N}(t) = g(\pi_k, t) + \eta_{kt}, \quad k \in U$$

where g is unknown and the errors η_{kt} are centered. The function g can be estimated by using the B -spline regression as proposed by Goga and Ruiz-Gazen (2014) and described before. This leads to consider relation (17) for the auxiliary information given now by π_k leading to the following smoothed weights: $w_{ks}^{bs}(\pi) = d_k \mathbf{b}^T(\pi_k) \left(\sum_{k \in s} d_k \mathbf{b}(\pi_k) \mathbf{b}^T(\pi_k) \right)^{-1} \sum_{k \in U} \mathbf{b}(\pi_k)$, $k \in s$. The improved estimator of the median is obtained from (25) by replacing d_k with the weights $w_{ks}^{bs}(\pi)$. In a model-based setting, Zheng and Little (2003, 2005) used a similar idea and penalized spline in order to estimate finite population totals with π ps sampling designs.

Consider again the same population test and draw now $I = 1000$ π ps samples. For each sample, we compute \hat{m}_n by using the sampling weights d_k and the B -spline weights $w_{ks}^{bs}(\pi)$ and compute again the absolute errors, $R(\hat{m}_n) \simeq \sum_{d=1}^D |\hat{m}_n(t_d) - m_N(t_d)|/D$. Figure 4b gives the distribution of these I absolute errors computed for both strategies. We note again a marked improvement of the estimation of the median for the π ps sampling design by using B -spline model-assisted estimator. This can be explained by the fact that the nonparametric weights $w_{ks}^{bs}(\pi)$ are more adapted to estimate m_N than the π ps weights.

4 B-Spline Imputation for Handling Item Nonresponse

The theory presented in the above sections supposed that all the sampled individuals respond, so we have full sample data $\{y_k\}_{k \in s}$. In practice however, due to various reasons, some individuals do not respond to the survey questionnaire (*unit nonresponse*) or respond only partially (*item nonresponse*). Unit nonresponse is treated by weighting methods while item nonresponse is treated by imputation. We focus here on item nonresponse and the estimation of finite population total t_y .

Let s_r be the subset of the original sample s containing the individuals that responded to item y and $s_m = s - s_r$, the subset of s containing the nonrespondents. To estimate t_y , we use an imputed estimator \hat{t}_I which is obtained from the HT estimator given in (1) by replacing or imputing the missing values y_k , $k \in s_m$ by values \hat{y}_k

$$\hat{t}_I = \sum_{k \in s_r} d_k y_k + \sum_{k \in s_m} d_k \hat{y}_k$$

The imputed values are obtained by fitting an imputation model. It is usually assumed that the response mechanism is MAR (*missing at random*), namely the distribution of \mathcal{Y} is the same within respondents and nonrespondents given fully observed covariates. Under the MAR assumption and provided that the auxiliary information x_k is available for all $k \in s$, the respondent data $\{(y_k, x_k)\}_{k \in r}$ may be used to build imputation models and to predict y_k for the nonrespondents. Goga et al. (2020) suggested a B -spline imputation procedure. We consider the model (2) as the imputation model and we estimate f by B -splines from the respondent data, the imputed value \hat{y}_k is given by

$$\hat{y}_k = \mathbf{b}^T \hat{\boldsymbol{\theta}}_r, \quad k \in s_m$$

where $\hat{\boldsymbol{\theta}}_r = (\sum_{k \in s_r} d_k \mathbf{b}(x_k) \mathbf{b}^T(x_k))^{-1} \sum_{k \in s_r} d_k \mathbf{b}(x_k) y_k$. Using the same techniques as in Sect. 2, we can show that $\sum_{k \in s_r} d_k (y_k - \hat{y}_k) = 0$ so the imputed estimator can be also written in a projection form, namely $\hat{t}_I = \sum_{k \in s} d_k \hat{y}_k$. Under the assumptions described in the Appendix and assuming that the response probabilities are all bounded away from 0, the imputed estimator is consistent for t_y but with a consistency rate which is slower than in the full response case. The imputed estimator can be written as a weighted sum of y_k 's values with weights not depending on \mathcal{Y} , so the approach supposed by Beaumont and Bissonnette (2011) can be used to compute and estimate the variance of \hat{t}_I . Goga et al. (2020) also suggest random B-spline imputation which consists in replacing the missing y_k by

$$\hat{y}_k = \hat{f}(x_k) + \epsilon_k^*, \quad k \in s_m$$

where ϵ_k^* is a residual selected at random from the set of standardized residuals observed from the responding units, $\{\tilde{e}_k; k \in s_r\}$, with probability $P(\epsilon_k^* = \tilde{e}_k) = \pi_k^{-1} / \sum_{l \in s_r} \pi_l^{-1}$, where $\tilde{e}_k = e_k - \bar{e}_r$ and $e_k = y_k - \hat{f}(x_k)$ with $\bar{e}_r = \sum_{k \in s_r} d_k e_k / \sum_{k \in s_r} d_k$. Simulation studies conducted by Goga et al. (2020) show that the deterministic and random B-spline imputation estimators perform much better than those obtained through regression imputation and nonparametric nearest neighbor imputation. In particular, the imputation estimator based on random B-spline imputation performs very well for the quantile estimation. The suggested method can be easily generalized to multiple auxiliary information by considering additive models.

Appendix

Assumptions on the Sampling Design and the Study Variable

We assume the following assumptions classical in survey sampling theory (Breidt and Opsomer 2000; Goga 2005).

- Assume that $\lim_{N \rightarrow \infty} N^{-1}n = \pi \in (0, 1)$.
- Assume that $\min_{k \in U} \pi_k \geq \tilde{c}$ and $\min_{k,l \in U} \pi_{kl} \geq c^*$ with \tilde{c} and c^* some positive constants and $\bar{\lim}_{N \rightarrow \infty} n \max_{k \neq l \in U} |\pi_{kl} - \pi_k \pi_l| < C_1 < \infty$, with C_1 a positive constant.
- Assume that $\lim_{N \rightarrow \infty} N^{-1} \sum_{k \in U} y_k^2 < \infty$.

We assume the following assumptions classical in nonparametric regression (Agarwal and Studden 1980; Burman 1991; Zhou et al. 1998; Claeskens et al. 2009).

Assumptions on B-Splines

- Assume that there exists a distribution function $Q(x)$ with strictly positive density on $[0, 1]$ such that $\sup_{x \in [0,1]} |Q_N(x) - Q(x)| = o(K^{-1})$, with $Q_N(x)$ the empirical distribution of $(x_i)_{i=1}^N$.
- Assume that the number of interior knots K satisfies $K = o(N)$.
- Assume that $K_\ell = (K + m - \ell)(\lambda\tilde{C})^{1/(2\ell)}N^{-1/(2\ell)} < 1$ where $\tilde{C} = C\{1 + o(1)\}$ with C a constant depending only on ℓ and the design density.

References

- Agarwal, G. G., & Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Annals of Statistics*, 8(6), 1307–1325.
- Beaumont, J.-F., & Bissonnette, J. (2011). Variance estimation under composite imputation: The methodology behind sevani. *Survey Methodology*, 37(2), 171–179.
- Besse, P. & Thomas-Agnan, C. (1989). Le lissage par fonctions splines en statistique, revue bibliographique. *Statistique et analyse des données*, 14(1), 55–84.
- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279–292.
- Binder, D. A., & Kovacevic, M. S. (1995). Estimating some measures of income inequality from survey data: An application of the estimating equations approach. *Survey Methodology*, 21, 137–145.
- Breidt, F., Claeskens, G., & Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831–846.
- Breidt, F.-J., & Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4), 1023–1053.
- Burman, P. (1991). Regression function estimation from dependent observations. *Journal of Multivariate Analysis*, 36, 263–279.
- Cardot, H. (2002). Local roughness penalties for regression splines. *Computational Statistics*, 17(1), 89–102.
- Cardot, H., Goga, C., & Lardin, P. (2012). Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. *Scandinavian Journal of Statistics*, 41, 516–534.
- Cassel, C., Särndal, C., & Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615–620.
- Chaouch, M., & Goga, C. (2012). Using complex surveys to estimate the L_1 -median of a functional variable: Application to electricity load curves. *International Statistical Review*, 80(1), 40–59.
- Claeskens, G., Krivobokova, T., & Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3), 529–544.
- Deville, J.-C. (1999a). Simultaneous calibration of several surveys. In *Proceedings of Statistics Canada Symposium 99 of Statistics Canada* (pp. 207–212).
- Deville, J.-C. (1999b). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193–203.
- Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.
- Dierckx, P. (1993). *Curves and surface fitting with splines*. Oxford: Clarendon.
- Gervini, D. (2008). Robust functional estimation using the spatial median and spherical principal components. *Biometrika*, 95, 587–600.

- Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del R. Istituto Veneto di Scienze Lettere ed Arti*.
- Goga, C., & Ruiz-Gazen, A. (2019). Nonparametric B-spline calibration. in work.
- Goga, C., Haziza, D., & Dagdoug, M. (2020). B-spline based imputation procedures for the treatment of item nonresponse in surveys. in work.
- Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire: une approche non paramétrique par splines de régression. *Canadian Journal of Statistics*, 33(2), 163–180.
- Goga, C., & Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society, B*, 76, 113–140.
- Hall, P., & Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, 92(1), 105–118.
- Horvitz, D., & Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.
- Kauermann, G., Krivobokova, T., & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society Series B Statistical Methodology*, 71(2), 487–503.
- McConville, K., & Breidt, F. J. (2013). Survey design asymptotics for the model-assisted penalized spline regression estimator. *Journal of Nonparametrics Statistics*, 25, 745–763.
- Montanari, G. E., & Ranalli, M. G. (2005). Nonparametric model calibration in survey sampling. *Journal of the American Statistical Association*, 100, 1429–1442.
- Nygard, F., & Sandström, A. (1985). The estimation of Gini and the entropy inequality parameters in finite population. *Journal of Official Statistics*, 4, 399–412.
- Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*, volume 12 of Cambridge series in statistical and probabilistic mathematics. Cambridge: Cambridge University Press.
- Särndal, C.-E., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. New York: Springer.
- Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99–119.
- Schumaker, L. L. (1981). *Spline functions: Basic theory*. New York: Wiley.
- Yates, F., & Grundy, P.-M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 235–261.
- Zheng, H., & Little, R. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99–117.
- Zheng, H., & Little, R. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1–20.
- Zhou, S., Shen, X., & Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, 26(5), 1760–1782.

Computational Outlier Detection Methods in Sliced Inverse Regression



Hadrien Lorenzo and Jérôme Saracco

Abstract Sliced inverse regression (SIR) focuses on the relationship between a dependent variable y and a p -dimensional explanatory variable x in a semiparametric regression model, in which, the link relies on an index $x'\beta$ and link function f . SIR allows estimating the direction of β that forms the effective dimension reduction (EDR) space. Based on the estimated index, the link function f can then be nonparametrically estimated using kernel estimator. This two-step approach is sensitive to the presence of outliers in the data. The aim of this paper is to propose computational methods to detect outliers in that kind of single-index regression model. Three outlier detection methods are proposed and their numerical behaviors are illustrated on a simulated sample. To discriminate outliers from “normal” observations, they use IB (in-bags) or OOB (out-of-bags) prediction errors from subsampling or resampling approaches. These methods, implemented in R, are compared with each other in a simulation study. An application on a real data is also provided.

1 Introduction

On one hand, classical linear regression or more generally parametric regression have achieved resounding success in many real problems whose goal is to investigate the relationship between a response variable $y \in \mathbb{R}$ and a covariate $x \in \mathbb{R}^p$. However, it can be argued that assuming specific structural constraints on the link function of y on x is too stringent. On the other hand, nonparametric regression is clearly a more flexible approach, but it is well-known that it typically suffers from the curse of dimensionality, i.e., a poor rate of convergence when the dimension p of x increases.

H. Lorenzo
Inria BSO, 33400 Talence, France
e-mail: hadrien.lorenzo@inria.fr

J. Saracco (✉)
Inria BSO & ENSC Bordeaux INP, 33400 Talence, France
e-mail: jerome.saracco@ensc.fr

To address these problems from purely parametric or nonparametric approaches, several authors studied single-index or multiple-index models. This kind of regression model can be viewed as an alternative semiparametric approach based on sufficient dimension reduction. So, in a dimension reduction setting, many authors suppose that x can be replaced by a linear combination of its components, $\beta'x$, without losing information on the conditional distribution of y given x . One way to express this assumption is

$$y \perp x \mid \beta'x \quad (1)$$

where the notation $v_1 \perp v_2 \mid v_3$ means that the random variable v_1 is independent of the random variable v_2 given any values for the random variable v_3 . One can write (1) as, for instance, the following single-index model with an additive error:

$$y = f(\beta'x) + \varepsilon, \quad (2)$$

where f is an unknown real-valued function, the distribution of ε is arbitrary and unknown, and $\varepsilon \perp x$. Since f is unknown, the p -dimensional parameter β is not totally identifiable, but the subspace spanned by β is identifiable. This subspace is referred to as the effective dimension reduction (EDR) subspace following Duan and Li (1991) in their original presentation of sliced inverse regression (SIR). Li (1991) consider a multiple-index regression model. The Euclidean parameter β is now a $p \times K$ matrix: $\beta = [\beta_1, \dots, \beta_K]$ where the vectors β_k are assumed to be linearly independent. The EDR subspace is then the K -dimensional linear subspace of \mathbb{R}^p spanned by the β_k 's.

Note that the dimension reduction is very useful in an exploratory stage of data analysis since model (1) relies on very few structural assumptions. For instance, it is not assumed that the indices act additively as often assumed in multiple-index models. It is likewise not necessary to assume that the error term is additive on the mean (as for the model (2)), thus heteroscedastic models are potentially included in this modeling. Note also that sufficient dimension reduction of the regression leads to a summary plot of y versus estimated indices which provides useful graphical modeling information.

In a second step, to study the relationship between the response variable and the few estimated indices, standard nonparametric approaches (such as kernel or spline smoothing) can be used. This stage usually involves additional assumptions such as an additive error term (as in model (2)) to get consistent properties of the corresponding estimate of the link function f .

In the statistical literature, different methods have been developed with the aim of estimating the EDR subspace. SIR, SIR-II, SIR_α , SAVE (sliced average variance estimation), and pHd (principal Hessian directions) approaches are the most popular, see Azais et al. (2012), Cai et al. (2020), Chavent et al. (2014), Chen and Li (1998), Cook (2000), Gannoun and Saracco (2003), Hsing (1999), Jlassi and Saracco (2019), Li (2018, 1992), Li et al. (2003), Li and Zhu (2007), Saracco (1997, 2005), Yin and Seymour (2005), Zhu et al. (2006), Zhu and Zhu (2007) among others. The important

question of the determination of the EDR space dimension in SIR and related methods has also been much studied, see for example Ferré (1998), Liquet and Saracco (2008).

SIR is known to be a relevant technique for the purpose of dimension reduction. Several properties of SIR have been extensively studied and numerous extensions have been already proposed. However, little attention has been paid to the sensitivity of SIR to outliers or to robustness aspects. Since SIR theory is based on conditional expectation and covariance matrix properties (see Sect. 2 for details), it is obvious that SIR can be severely influenced by outliers in the data, see Gather et al. (2002) or Cook and Critchley (2000) for instance. In Prendergast (2006, 2007), the detection of influential observations on the estimation of the dimension reduction subspaces returned by SIR, SIR-II, and SAVE have been studied using the notion of influence functions of single observations. However, the proposed empirical influence values are very sensitive to the choice of the number H of slices (introduced in the next section) in detecting influential observations, which makes this approach complicated to use in practice. Robust SIR methods were then developed and only focused on the estimation of the EDR space (regardless of the estimation of the link function f). For example, in Chiancone et al. (2017), the inverse regression formulation of SIR is, therefore, extended to non-Gaussian errors with heavy-tailed distributions (Student). The underlying Expectation-Maximization algorithm was tested in presence of outliers and provided good numerical results. Dong et al. (2015) also mentioned that classical sufficient dimension reduction methods are sensitive to outliers present in predictors, and may not perform well when the distribution of the predictors is heavy-tailed. Two robust inverse regression methods which are insensitive to data contamination (weighted inverse regression estimation and sliced inverse median estimation) were then introduced and they demonstrated very interesting numerical performances in the presence of potential outliers. In the same spirit, Babos and Artemiou (2020) proposed sliced inverse median difference regression to robustify SIR methodology at the presence of outliers. In Dikheel (2014), robust SIR extensions were presented through robust estimates of the covariance matrix.

The goal of this paper is to propose computational methods to detect outliers in a single-index regression model, comprising EDR space estimation using SIR and link function estimation based on kernel smoothing. In practice, it is always interesting to detect outliers (rather than only developing robust methods), to isolate them, and to understand why these observations are aberrant (wrong numerical values, unusual individuals, ...). Once the dataset has been cleaned, it is then possible to implement the usual methodology, SIR followed by a nonparametric estimation of f .

In Sect. 2, a brief overview on usual SIR is given. Three outlier detection methods, named MONO, TTR, and BOOT hereafter, are presented in Sect. 3. They use IB (in-bags) or OOB (out-of-bags) prediction errors from subsampling or resampling approaches in order to discriminate outliers from “normal” observations. These methods have been implemented in R. How these methodologies work is described on a simulated example in Sect. 4. Section 5 provides a more extensive simulation study that compares the numerical performances of the proposed methods. A real dataset is also used to illustrate these approaches in Sect. 6. Finally, concluding remarks are given in Sect. 7.

2 A Brief Review on Usual SIR

In order to estimate the EDR space, various methods based on the use of inverse regression are widely available in the literature. In order for inverse regression to be useful in estimating the EDR space, some of them, like SIR or SAVE or principal Hessian direction, need additional conditions on the marginal distribution of the covariate x . In this paper, we focus the usual SIR approach which relies on the following linearity condition (LC) on x :

$$\text{For all } b \in \mathbb{R}^p, \mathbb{E}[b'x \mid \beta'x] \text{ is linear in } x'\beta. \quad (3)$$

Note that the LC is required to hold only for the true Euclidean parameter β . Since β is unknown, it is not possible, in practice, to verify a priori this assumption. Therefore, we can assume that LC holds for all possible values of β , this is equivalent to assume an elliptical symmetry of the distribution of x : for instance, the well-known multivariate normal distribution satisfies this condition. Finally, following Hall and Li (1993), the LC is not a severe restriction because this LC holds to a good approximation in many problems as the dimension p of the predictors increases. Interesting discussions on the LC can also be found in Chen and Li (1998), Li (2018) for instance.

Let us now consider a monotone transformation T . Under model (1) and LC, Duan and Li (1991) showed that the centered inverse regression curve satisfies

$$\mathbb{E}[x \mid T(y)] - \mu \in \text{Span}(\Sigma\beta), \quad (4)$$

where $\mu := \mathbb{E}[x]$ and $\Sigma := \mathbb{V}(x)$. Therefore, the space spanned by the centered inverse curve, $\{\mathbb{E}[x \mid T(y)] - \mathbb{E}[x] : y \in \mathcal{Y}\}$ where \mathcal{Y} is the support of response variable y , is a subspace of the EDR space, but it does not guarantee equality. A pathological model, often called symmetric dependent model, has been identified in the literature, and this is the model for which the centered inverse regression curve is degenerated. To solve this problem, specific methods (based on higher order inverse moments), such as SIR-II, SIR_α or SAVE, have been developed.

When the model is not pathological (which is often the case in practice), the centered inverse regression curve can be used to recover the EDR space from (4). Indeed, a direct consequence of this result is that the covariance matrix of this curve,

$$\Gamma := \mathbb{V}(\mathbb{E}[x \mid T(y)]),$$

is degenerate in any direction Σ -orthogonal to β (i.e., to the β_k 's for a multiple-index model). Therefore, the eigenvectors associated with the nonnull eigenvalues of $\Sigma^{-1}\Gamma$ are EDR directions, which means that they span the EDR space E .

In the slicing step of SIR, the range of y is partitioned into H nonoverlapping slices $\{s_1, \dots, s_H\}$. With such slicing, the covariance matrix Γ can be straightforwardly written as

$$\Gamma := \sum_{h=1}^H p_h(m_h - \mu)(m_h - \mu)'$$

where $p_h = P(y \in s_h)$ and $m_h = \mathbb{E}[x \mid y \in s_h]$.

Let us now consider a random sample $\{(x_i, y_i), i = 1, \dots, n\}$ generated from the single-index regression model (2). By substituting the empirical versions of μ, Σ, p_h and m_h for their theoretical counterparts, we obtain an estimated basis of E spanned by the eigenvector \hat{b}_{SIR} associated with the largest eigenvalue of the estimate $\hat{\Sigma}_n^{-1} \hat{\Gamma}_n$ of $\Sigma^{-1} \Gamma$ where

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(x_i - \bar{x}_n)' \quad \text{and} \quad \hat{\Gamma}_n = \sum_{h=1}^H \hat{p}_{h,n}(\hat{m}_{h,n} - \bar{x}_n)(\hat{m}_{h,n} - \bar{x}_n)',$$

with $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$, $\hat{n}_{h,n} = \sum_{i=1}^n \mathbb{I}_{[y_i \in s_h]}$, $\hat{p}_{h,n} = \frac{\hat{n}_{h,n}}{n}$, $\hat{m}_{h,n} = \frac{1}{\hat{n}_{h,n}} \sum_{i \in s_h} x_i$, the notation $\mathbb{I}_{[\cdot]}$ standing for indicator function. This approach is the one proposed by Duan and Li (1991), Li (1991) when they initially introduced the SIR approach. Since the early 1990s, the SIR method has been extensively studied by many authors, see, for instance, all the references mentioned in the introduction.

The link function f of model (2) can then be estimated by the usual kernel estimator (see for example Schimek 2013) based on the sample $\{(x'_i \hat{b}_{\text{SIR}}, y_i), i = 1, \dots, n\}$ where the $x'_i \hat{b}_{\text{SIR}}$'s are the values of the estimated index. For a given value x_0 of x , the kernel estimation of $f(\beta' x_0)$ is given by

$$\hat{f}_n(\hat{b}'_{\text{SIR}} x_0) = \frac{\sum_{i=1}^n K\left(\frac{x'_i \hat{b}_{\text{SIR}} - x'_0 \hat{b}_{\text{SIR}}}{h_n}\right) y_i}{\sum_{i=1}^n K\left(\frac{x'_i \hat{b}_{\text{SIR}} - x'_0 \hat{b}_{\text{SIR}}}{h_n}\right)},$$

where K is the kernel and h_n is the bandwidth. The kernel is usually a positive symmetric weighting function with an integral equal to 1. In the rest of the paper, the chosen kernel is the density of the normal distribution $\mathcal{N}(0, 1)$, called the Gaussian kernel. The bandwidth $h_n > 0$ is called the smoothing parameter in kernel regression because it controls variance and bias of the estimator. This parameter must, therefore, be correctly tuned using cross-validation for instance:

$$h_n^{\text{opt}} = \arg \min_{h_n > 0} \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}_n^{(-i)}(\hat{b}'_{\text{SIR}} x_0) \right)^2,$$

where $\hat{f}_n^{(-i)}(\hat{b}'_{\text{SIR}} x_0)$ stands for the estimation of $f(\beta' x_0)$ based on the sample $\{(x'_j \hat{b}_{\text{SIR}}, y_j), j \neq i\}$.

When the underlying regression model is a multiple-index model, the estimated EDR space is spanned by the eigenvectors associated with the largest K eigenvalues of the estimate $\hat{\Sigma}_n^{-1} \hat{\Gamma}_n$. Let \hat{B}_{SIR} be the $p \times K$ matrix of these K eigenvectors.

The estimated indices $x'_i \hat{B}'_{\text{SIR}}$'s are now K -dimensional and the kernel estimation of $f(\beta'x_0)$ is then based on multivariate kernel. For example, K can be the density of the multivariate normal distribution $\mathcal{N}(0_K, I_K)$ where 0_K (resp. I_K) stands for the null vector of dimension K (resp. the identity matrix of order K), and the associated smoothing parameter h_n can be K -dimensional. Another way is to consider the following kernel estimator:

$$\hat{f}_n(\hat{B}'_{\text{SIR}}x_0) = \frac{\sum_{i=1}^n K\left(\frac{\|x'_i \hat{B}'_{\text{SIR}} - x'_0 \hat{B}'_{\text{SIR}}\|}{h_n}\right) y_i}{\sum_{i=1}^n K\left(\frac{\|x'_i \hat{B}'_{\text{SIR}} - x'_0 \hat{B}'_{\text{SIR}}\|}{h_n}\right)},$$

where $\|\cdot\|$ stands for a chosen norm in \mathbb{R}^K and the bandwidth h_n is unidimensional.

3 Outlier Detection Methods in SIR

Three outlier detection methods for single-index regression model (2) are presented. Let us consider a sample $S = \{(x_i, y_i), i = 1, \dots, n\}$ of n individuals among which some may be outliers.

For each of the three methods, the parameter β (more properly, the EDR direction b) is estimated by the usual SIR method (with the number of slices $H = 10$) and the link function f is estimated using the kernel estimator with the Gaussian kernel and the bandwidth tuned via cross-validation.

3.1 A Naive Method

This naive method relies on the following three steps:

STEP 1. Estimation the EDR direction from the sample S .

The usual SIR provides the estimate \hat{b}'_{SIR} of b . The corresponding indices $\{\hat{b}'_{\text{SIR}}x_i, i = 1, \dots, n\}$ are then calculated.

STEP 2. Estimation of the adjusted value $f(\beta'x_i)$'s.

From the sample $\{(\hat{b}'_{\text{SIR}}x_i, y_i), i = 1, \dots, n\}$, the adjusted values are obtained via the kernel estimator based on the Gaussian kernel and the bandwidth tuned via cross-validation. Let $\hat{y}_i = \hat{f}_n(\hat{b}'_{\text{SIR}}x_i)$ for $i = 1, \dots, n$.

STEP 3. Evaluation of the error associated with the model estimation and outlier detection.

The errors considered are naturally the residuals: for $i = 1, \dots, n$, $\hat{e}_i = y_i - \hat{y}_i$. The detection of potential outliers is simply based on the definition of outliers in the boxplot of the absolute error $|\hat{e}_i|$'s, i.e., the outliers correspond to individuals whose values are greater than the value of the 3rd quartile plus 1.5 times the interquartile interval.

Note that, in the same spirit, the bootstrap histogram of “mean—trimmed mean” for a suitable trimming number was proposed by Singh and Xie (2003) as a nonparametric graphical tool for detecting outlier(s) in a dataset. The bootlier plot was introduced and it is shown that the multimodality in the bootlier plot is caused by outlier(s) in the sample.

This naive method is called MONO hereafter. The name MONO stands for a single use of the initial sample S and a single estimate of the underlying single-index model. In the numerical example of Sect. 4, Fig. 1 allows to visualize the position of the outliers in the corresponding boxplot.

3.2 TTR Method

This method relies on training sample and test sample replications for evaluating the “stability” of the estimated model, hence the name TTR of the method for Training Test Replications.

The TTR approach works in two major steps. Let R be the number of replications chosen by the user. In practice, $R = 2000$ is more than enough for reasonable sample sizes, i.e., $n \leq 500$. Let $\alpha \in [0, 1]$ be the proportion of the sample which will constitute the test sample. In the rest of the paper, the parameter is fixed to $\alpha = 0.1$, thus 90% of the sample S is used as the training sample S_{train} and the remaining 10% of the sample S constitutes the test sample S_{test} . Note that individuals are drawn with equal weight and without replacement.

STEP 1. For each replication r (with $r = 1, \dots, R$)

- 1.a. Split the initial sample S into a training sample $S_{\text{train}}^{(r)}$ and a test sample $S_{\text{test}}^{(r)}$ containing, respectively, $(1 - \alpha)\%$ and $\alpha\%$ of the individuals.
- 1.b. Using $S_{\text{train}}^{(r)}$, calculate the estimated EDR direction $\hat{b}_{\text{SIR}}^{(r)}$ and the associated indices $\{(\hat{b}_{\text{SIR}}^{(r)})'x_i, i \in S_{\text{train}}^{(r)}\}$.
- 1.c. For all the individuals $i^* \in S_{\text{test}}^{(r)}$, calculate the error of prediction of the response variable y as follows:

$$\hat{e}_{i^*}^{(r)} = y_{i^*} - \hat{f}_n^{(r)} \left((\hat{b}_{\text{SIR}}^{(r)})'x_{i^*} \right),$$

where the estimate $\hat{f}_n^{(r)}(\cdot)$ is based on the sample $\{((\hat{b}_{\text{SIR}}^{(r)})'x_i, y_i), i \in S_{\text{train}}^{(r)}\}$.

STEP 2. Evaluation of the error means.

For each $i^* = 1, \dots, n$, calculate the associated error mean over the R replications (when the individual i^* is present in the corresponding test sample):

$$\bar{e}_{i^*} = \frac{\sum_{r=1}^R \mathbb{I}_{[i^* \in S_{\text{test}}^{(r)}]} \hat{e}_{i^*}^{(r)}}{\sum_{r=1}^R \mathbb{I}_{[i^* \in S_{\text{test}}^{(r)}]}}.$$

STEP 3. Detection of the outliers via a change point detection.

The idea is to find a single change point position in the sequence of the errors' means $\{\bar{e}_{(i^*)}, i^* = 1, \dots, n\}$ ordered by decreasing values (where the subscript (i^*) enclosed in parentheses indicates the i^* th order statistic of the sample). Indeed, if there are no outliers in the data, no change points should clearly appear in this sequence of ordered absolute mean errors. On the other hand, in the presence of outliers, the corresponding mean absolute errors should naturally be significantly larger than the errors associated with other individuals. Thus, looking for a single change point in mean and variance in this sequence should intuitively allow us to separate outliers from other observations.

Many authors have proposed a single search method to detect change points. Recently, Killick and Eckley (2014) have developed the R package `changePoint` that helps to detect the location of different change points. For single or multiple change point detection, the approach allows estimating the points at which the statistical properties of a sequence of observations change. Within this package, several changes in mean methods are available as well as methods focusing on detection of change in variance and methods searching a change in both mean and variance. Briefly, let us give an overview of the underlying approach. Let $z_{1:n} = (z_1, \dots, z_n)$ be the ordered sequence of the errors' means and $\tau_{i:m} = (\tau_1, \dots, \tau_m)$ the positions of the m change points (each change point position is between 1 and $n - 1$, $\tau_0 = 0$ and $\tau_{m+1} = n$). The idea is to minimize

$$\sum_{i=1}^{m+1} [C(z_{(\tau_{i-1}+1):\tau_i})] + \gamma g(m) \quad (5)$$

where C is a cost function (for instance negative log-likelihood ratio test statistic) and $\gamma g(m)$ is a penalty to guard against over fitting. This package implements several algorithms to minimize (5): binary segmentation (Edwards and Cavall-Sforza 1965), segment neighborhood (Auger and Lawrence 1989) and pruned exact linear time (PELT) (Killick et al. 2012). Here the `changePoint` package is used to detect only one change point ($m = 1$) in mean and variance with the binary segmentation algorithm in the ordered sequence of means $\{\bar{e}_{(i^*)}, i^* = 1, \dots, n\}$. In the numerical example of Sect. 4, Fig. 2 (top left) visualizes the position of the estimated single change point.

An individual associated with an ordered error's mean before the single change point position is then considered as an outlier.

Remark In the associated R code, the bandwidth is tuned only once in step 1.c for the kernel estimation of each iteration. This “optimal” bandwidth is obtained via cross-validation using the whole sample of the y_i 's versus the estimated indices $x'_i \hat{b}_{\text{SIR}}$. This is a reasonable choice if one assumes that there are no outliers in the x_i 's and thus in the $x'_i \hat{b}$'s or in the $x'_i \hat{b}^{(r)}$'s. Note that the presence of visible outliers in the x_i 's would have been detected in a preliminary step and the dataset would then have been cleaned. This choice of only one tuned bandwidth clearly saves calculation time for the TTR method. Finally, note also that, in each iteration of step 3 for the

TTR method, it is easy to integrate an automatic optimal bandwidth selection in the R code. The same strategy is used for the BOOT method in step 1.c presented in the following section:

3.3 *BOOT Method*

The MONO method deals with in-bag (IB) errors and the TTR method with out-of-bag (OOB) errors. While MONO risks overfitting, TTR risks significance loss of statistical power (since the training sample is a subsample of the initial sample) but cannot quantify the impact of IB individuals. The current BOOT method uses IB errors in that objective.

Isolated individuals that are not outliers, in the plot of the estimated index versus the response y are usually hard to predict especially if those individuals are not in the training dataset. However, if any of those individuals are included in the training dataset, it has a beneficial effect on the built model. Indeed, those individuals are, therefore, better predicted while the nonisolated individuals are still well predicted since those isolated individuals are in line with the regression model. For those isolated individuals, the OOB error is then high while the IB error is potentially low. They are denoted as “borderline” observations in the following: On the other hand, the “outliers” are always badly predicted with high IB and OOB errors and the “normal” individuals are always well predicted with low IB and OOB errors (see an illustration of these comments in Fig. 4 that gives examples of those three types of observations).

The BOOT method is based on two simple decision rules to discriminate between these three types of individuals (“normal” observation, “borderline” observation, “outlier”) using the IB error and its logarithmic transformation. This method relies on bootstrap samples of S . Let B be the number of bootstraps chosen by the user. In practice $B = 2000$ is more than enough for reasonable sample sizes, i.e., $n \leq 500$. Note that individuals are drawn with equal weight and with replacement.

STEP 1. For $b = 1, \dots, B$,

- 1.a. Draw a bootstrap sample $S^{(b)}$ from the initial sample S . Let $n_i^{(b)}$ denote the number of times the observation i is present in the bootstrap sample $S^{(b)}$.
- 1.b. Using $S^{(b)}$, calculate the corresponding estimated EDR direction $\hat{b}_{\text{SIR}}^{(b)}$ and the associated indices $\{(\hat{b}_{\text{SIR}}^{(b)})'x_i, i \in S^{(b)}\}$.
- 1.c. For all the individuals $i \in S^{(b)}$, calculate the IB error of prediction of the response variable y as follows:

$$\hat{e}_i^{(b)} = y_i - \hat{f}_n^{(b)}\left((\hat{b}_{\text{SIR}}^{(b)})'x_i\right),$$

where the estimate $\hat{f}_n^{(b)}(\cdot)$ is based on the sample $\{((\hat{b}_{\text{SIR}}^{(b)})'x_i, y_i), i \in S^{(b)}\}$.

Note that, even if they are not used in Steps 2 and 3, the OBB errors (for all the individuals $i \notin S^{(b)}$) have also been calculated since they are used in graphical representations (see Fig. 4).

STEP 2. Evaluation of the error means.

For each $i = 1, \dots, n$, calculate the associated error mean over the B replications (when the individual i is present at least once in the corresponding bootstrap sample)

$$\bar{e}_{(i)} = \frac{\sum_{b=1}^B \left| \hat{e}_i^{(b)} \right| \mathbb{I}_{[i \text{ such that } n_i^{(b)} \geq 1]}}{\sum_{b=1}^B \mathbb{I}_{[i \text{ such that } n_i^{(b)} \geq 1]}}.$$

STEP 3. Detection of outliers and “borderline” observations

The idea here is to first identify among the errors $\{\bar{e}_{(i)}, i = 1, \dots, n\}$ those which are particularly high and which will naturally correspond to these “big” outliers. For this purpose, the log scale was used to detect these outliers. Then, in a second step, the usual scale is used in order to identify other possible “small” remaining outliers which are then called “borderline” observations.

- 3.a. The detection of potential outliers is based on the definition of outliers in the boxplot¹ of the log $(\bar{e}_{(i)})$'s. The corresponding observations are plotted in blue in Fig. 3 (on the left).
- 3.b. The detection of potential “borderline” observations is based on the definition of outliers in the boxplot of the $\bar{e}_{(i)}$'s, these “current outliers” are plotted with orange triangle in Fig. 3 (in the middle). The “borderline” observations are thus defined as the current detected outliers not identified as outliers in the previous step 3.a. (plotted with blue circle behind the orange triangle in this graphic). The corresponding “borderline” observations are, therefore, those represented only in orange on the graphic in Fig. 3 (on the right).

Remark In Step 3.a., the log transformation is used by default to detect the potential outliers. However, the relevant transformation of the considered errors, $\bar{e}_{(i)}$, $i = 1, \dots, n$, is probably not always log but that it may depend on the link function f itself and on the distribution of ϵ in the regression model (2).

4 A Numerical Example

Let us consider a simulated sample to clearly illustrate how the previous three outlier detection methods (MONO, TTR, and BOOT) work. Note that steps 3 of the different methods (MONO, TTR, and BOOT) are interchangeable with each other, and thus they can be used after any of the error calculation steps (steps 1 and 2). In this

¹Already described in the presentation of the MONO method.

section, and Sect. 5, only the MONO, TTR, and BOOT methods are compared with each other, not ideally all possible combinations. Thus, we are well aware that this will make it difficult to identify whether the success of the method is due mainly to the different error calculation processes (steps 1 and 2) or to the technique of detecting “abnormally large” errors (step 3).

4.1 Description of the Simulated Dataset

The following single-index regression model is used in all numerical studies in Sects. 4 and 5:

$$y = \frac{(x'\beta)^3}{100} + \epsilon, \quad (6)$$

where $\beta = (2, 2, 1, -2, -3, 0, \dots, 0)' \in \mathbb{R}^p$, x follows the p -dimensional uniform distribution on $[-2; 2]^p$, and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 0.25)$ is independent of x . In a first step, $\tilde{n} = 200$ observations $\{(x_i, y_i), i = 1, \dots, \tilde{n}\}$ are generated from model (6) with $p = 5$. Then in a second step, $\tilde{\tilde{n}} = 10$ new individuals are generated as follows: for $i = \tilde{n} + 1, \dots, \tilde{n} + \tilde{\tilde{n}}$

- x_i is drawn from the uniform distribution on $[-2; 2]^p$,
- y_i is drawn (independently from x_i) from the uniform distribution on the support of the first \tilde{n} values of y .

These $\tilde{\tilde{n}}$ new observations are then “potential” outliers for the model (6) since their y_i ’s are not linked to the x_i ’s via this model. Note that these observations are not outliers regarding the distribution of the x_i ’s (resp. of the y_i ’s). The term “potential” refers to the fact that an observation (x_i, y_i) (for an $i \in \{\tilde{n} + 1, \dots, \tilde{n} + \tilde{\tilde{n}}\}$) may be close, just by chance, to the “true” structure of the data (based on the underlying model (6)). The objective is to detect these potential $\tilde{\tilde{n}}$ outliers in the sample $S = \{(x_i, y_i), i = 1, \dots, n\}$ where $n = \tilde{n} + \tilde{\tilde{n}}$ and then to estimate as best as possible the relationship between y and x through the single-index $x'\beta$.

4.2 Numerical Results

In a first step, based on the available sample $S = \{(x_i, y_i), i = 1, \dots, n\}$, the EDR direction b is estimated by \hat{b}_{SIR} using the usual SIR method (with the number of slices $H = 10$) and the link function f is estimated by $\hat{f}_n(\cdot)$ using the kernel estimator with the Gaussian kernel and the bandwidth tuned via cross-validation. The distance between the true EDR space and the estimated one is defined as

$$d^2(E, \hat{E}) = 1 - \frac{\text{Trace}(P_E P_{\hat{E}})}{K} \in [0, 1],$$

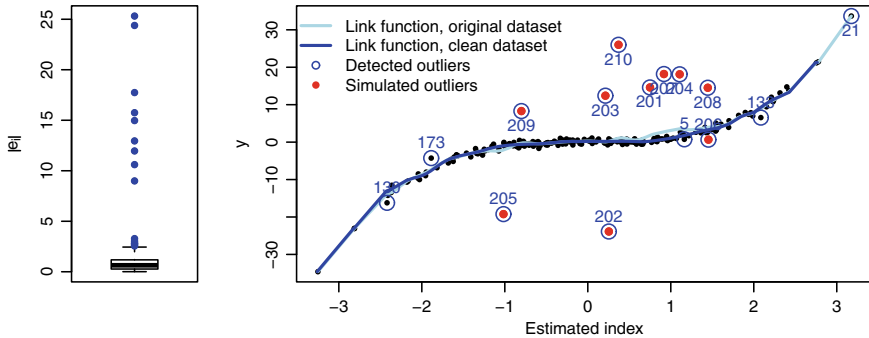


Fig. 1 MONO method description. Left graphic provides the boxplot of the absolute errors, the detected outliers are in blue. On the plot of the estimated indices $x_i' \hat{b}_{SIR}$ versus the y_i 's (right graphic), these outliers are also plotted in blue, red points correspond to the \tilde{n} (true) potential outliers. The kernel estimations of the link function are superimposed for both the original dataset (in light blue) and the dataset without the detected outliers (in dark blue)

where $P_E = \beta(\beta' \beta)^{-1} \beta'$ (resp. $P_{\hat{E}}$) is the orthogonal projector onto E (resp. \hat{E}) with K the dimension of the EDR space (here $K = 1$ for a single-index model). The closer this distance is to zero, the better the estimation \hat{E} of E . On the simulated sample, we have $d^2(E, \hat{E}) = 0.0093$. The corresponding MSE (mean squared error) defined as

$$MSE = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{f}_n(x_i' \hat{b}_{SIR}) \right)^2$$

is equal to $MSE = 12.99$.

Using the naive MONO method, 15 outliers have been detected, see Fig. 1 (left) for the boxplot of the absolute residual errors (with outliers in blue) and Fig. 1 (right) for the visualization of these outliers on the plot of the estimated indices $x_i' \hat{b}_{SIR}$ versus the y_i 's. Note that all the $\tilde{n} = 10$ generated outliers have been identified. Among these 15 detected outliers, 5 are false positive, however, the individual 21 (at the top right of the plot of Fig. 1 (right)) can be considered as an “extreme” observation. An “extreme” observation may obviously be detected as an outlier by the method because the nonparametric estimation of f by the kernel method is based on local smoothing. Thus, since an “extreme” observation is too isolated in the plot of the estimated indices ($x_i' \hat{b}_{SIR}$, $i = 1, \dots, n$) versus the y_i 's, its kernel prediction is difficult due to the lack of observations around it (this is the problem of data sparsity in nonparametric regression). Using the initial sample without these 15 outliers, the associated MSE is now equal to 0.24, and we have $d^2(E, \hat{E}) = 0.00361$. These two quantities clearly show the benefits of removing the detected outliers.

Using the TTR method with $R = 3000$, 11 outliers have been detected, see Fig. 2 (top left) for the detection of the unique change point position in the sequence of the ordered errors' means, $\{\bar{e}_{(i^*)}, i^* = 1, \dots, n\}$. Figure 2 (top right) provides the

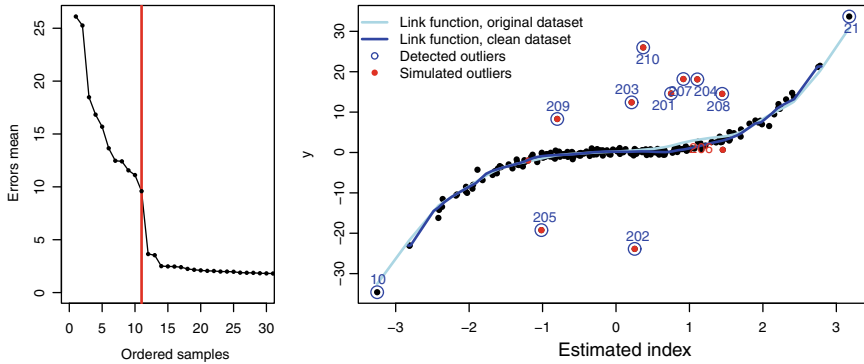


Fig. 2 TTR method description. Top-left graphic shows the ordered means errors with the red vertical line providing the estimated single change point position. On the plot of the estimated indices $x'_i \hat{b}_{SIR}$ versus the y_i 's (top-right graphic), these outliers are also plotted in blue, red points correspond to the \tilde{n} (true) potential outliers. The kernel estimations of the link function are superimposed for both the original dataset (in light blue) and the dataset without the detected outliers (in dark blue)

visualization of these outliers on the plot of the estimated indices $x'_i \hat{b}_{SIR}$ versus the y_i 's. Among these 11 outliers, only 2 are false positive: observations 10 and 21 (at the bottom left and at the top right of the plot of Fig. 2 (top right)) can naturally be considered as “extreme” observations but they are still selected as outliers for the same reasons of nonparametric kernel estimation as those mentioned for the MONO method. Note also that observation 206 (in red) has not been detected as an outlier by TTR method, but its projection is very close to the “true data” (in black, i.e., that is those generated by the underlying model), and thus this observation is not really a significant outlier. Using the initial sample without these 11 outliers, the associated MSE is now equal to 0.29, and we have $d^2(E, \hat{E}) = 0.00367$. The benefits of removing these detected outliers are again very clear. Figure 2 (bottom) provides the plot of the estimated indices $x'_i \hat{b}_{SIR}$ versus the y_i 's considering the dataset without the detected outliers. The kernel estimation of the link function (in blue) is superimposed on the plot. One can observe the very good fit of the data to the underlying model.

Using the BOOT method with $B = 3000$, 9 out of the $\tilde{n} = 10$ outliers were detected, and 4 “borderline” observations have been identified. Figure 3 (right) provides the visualization of the outliers (in blue) and of the “borderline” observations (in orange) on the plot of the y_i 's versus the estimated indices $x'_i \hat{b}_{SIR}$. The boxplot on the right allows to detect the outliers, while the boxplot in the middle identifies the “borderline” observations. The individual 206 (simulated as an outlier) is here detected as a “borderline” observation. Note that there is no false positive. Graphics in Fig. 4 provide the plot of the $n_i^{(b)}$'s versus the $|e_i^{(b)}|$'s (for $b = 1, \dots, B$) for three individuals. The horizontal line on each plot represents the corresponding error mean $\overline{|e_{(i)}|}$ over the B replications (when the individual i was present at least once in the corresponding bootstrap sample). One can observe that

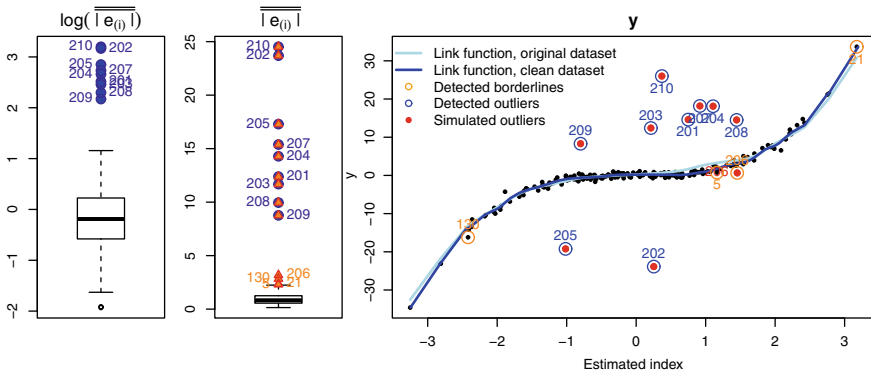


Fig. 3 BOOT method description. The two graphics on the left correspond to the boxplots of the \log mean absolute errors (defining outliers, in blue) and of the mean absolute errors (defining “borderline” observations, in orange). Selected outliers (in blue) and selected “borderline” observations (in orange) are shown on the plot of the estimated indices $x_i^{\hat{b}} \hat{b}_{SIR}$ versus the y_i ’s (right graphic), the red points correspond to the \tilde{n} (true) potential outliers. The kernel estimations of the link function are superimposed for both the original dataset (in light blue) and the dataset without the detected outliers and “borderline” observations in dark blue)

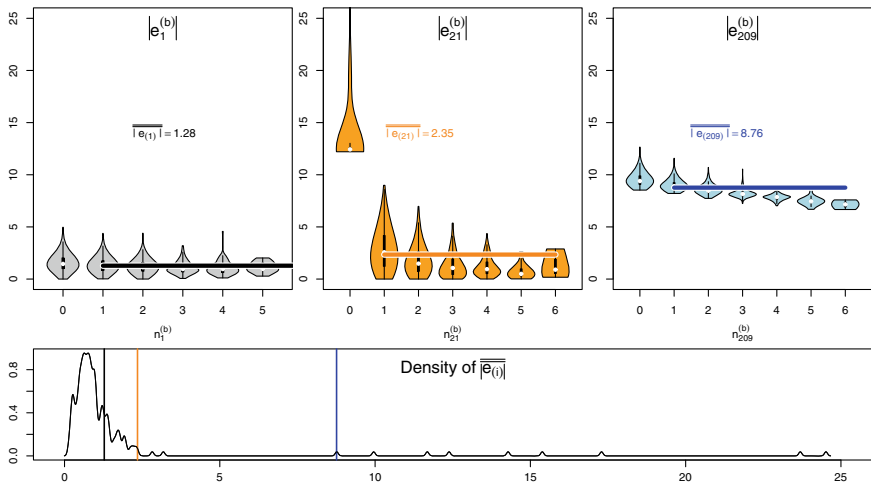


Fig. 4 For the BOOT method, plots of the $|e_{(i)}^{(b)}|$ ’s versus the $n_i^{(b)}$ ’s (for $b = 1, \dots, B$) for three individuals: a “normal” individual ($i = 1$) on the left, a “borderline” individual ($i = 21$) in the middle, and an outlier ($i = 209$) on the right. Colored (resp. black, orange and blue) segments show their corresponding computed IB error means $\overline{|e_{(i)}|}$ (for $n_i^{(b)} \geq 1$). The last plot at the bottom provides a density estimation of the $\overline{|e_{(i)}|}$ ’s with these three individuals showed through colored vertical lines. Since the OOB errors (for $n_i^{(b)} = 0$) are not used, individual $i = 21$ (in orange) is not considered as outlier but as “borderline” observation

- for observation 1 (which is a “normal” observation), the corresponding mean $\overline{|e_{(1)}|}$ is low,
- for observation 21 (which is characterized as a “borderline” observation), the corresponding mean $\overline{|e_{(21)}|}$ is intermediate. The model learns its position and modifies its tail, which explains the fall in error between $n_{(21)}^{(b)} = 0$ and $n_{(21)}^{(b)} = 1$,
- for observation 209 (which was detected as an outlier), the corresponding mean $\overline{|e_{(209)}|}$ is clearly higher than the previous ones, no matter the number of times that observation is present in the bootstrap sample.

Using the initial sample without these 9 outliers and 4 “borderline” observations, the associated MSE is now equal to 0.32 and we have $d^2(E, \widehat{E}) = 0.00172$, which highlights the high effectiveness of the BOOT method. Finally, let us remark that, in the computation of the mean descriptors, we chose to consider only the (absolute) error values for which each individual is represented at least once in the bootstrap sample as to prevail from selecting “extreme” observations, as discussed in the comments of the previous two methods.

5 Simulation Results

In this simulation study, $N = 100$ replications of samples from model (6) have been generated with various values of the sample size \tilde{n} ($= 100, 200, 300$), various values of the dimension p ($= 5, 20$) of the covariate x , and two numbers of potential outliers $\tilde{\tilde{n}}$ ($= 3, 10$). For each generated sample and each outlier detection method (MONO, TTR with $R = 2000$ and BOOT with $B = 2000$), the following quantities were calculated:

- the quality of the estimated EDR direction $d^2(E, \widehat{E})$ where \widehat{E} is the estimated EDR space based on the complete sample (unique for all the three methods),
- the MSE evaluated on the complete sample (unique for all the three methods),
- the number of detected outliers (and the number of “borderline” observations for the BOOT method),
- the number of false positives,
- the quality of the estimated EDR direction $d^2(E, \widehat{E}_\star)$ where \widehat{E}_\star is the estimated EDR space based on the sample without the outliers (and the “borderline” observations) detected by the method \star ,
- the MSE evaluated on the sample without the detected outliers (and the “borderline” observations for BOOT method).

To visualize and easily compare all these indicators, boxplots were used. According to results available in Fig. 5, all the three methods allow to reduce the distance to the true model. All methods, and even the model based on the complete dataset (in yellow), naturally perform better if the size of the available sample ($\tilde{n} + \tilde{\tilde{n}}$) increases. If the number of outliers is $\tilde{\tilde{n}} = 10$, the model based on the complete dataset shows

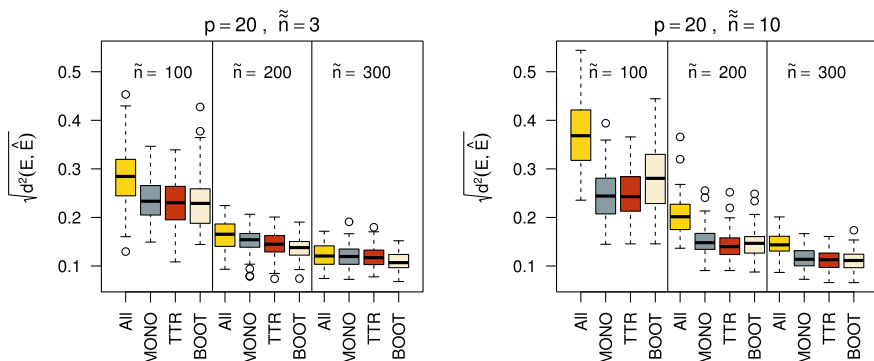


Fig. 5 Boxplots the quality measures of the estimated EDR space based on the $\sqrt{d^2(E, \hat{E}_*)}$ values, for simulated datasets with $p = 20$ and $\tilde{n} \in \{3, 10\}$. “All” stands for the estimation of the EDR using the complete sample

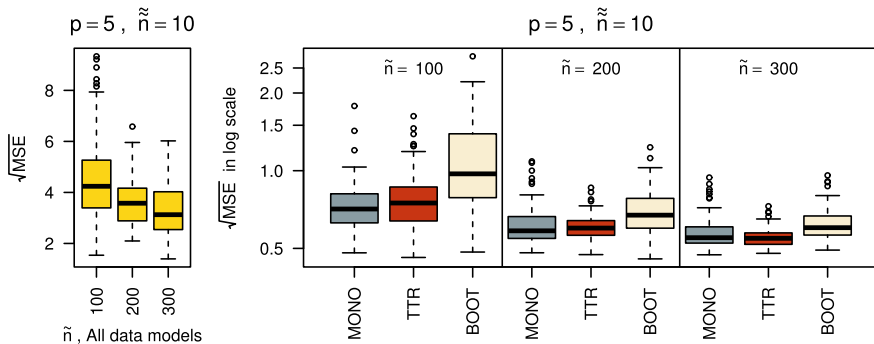


Fig. 6 Root MSE for the different methods. Visualizations of the errors for the complete dataset model (in yellow) and for the three outlier detection methods have been split into two graphics since scales are different. Here $p = 5$ and $\tilde{n} = 10$ have been detailed

poorer results whatever the number \tilde{n} . Note that, for a given number \tilde{n} of outliers, the proportion of outliers naturally decreases as the sample size increases. BOOT seems to suffer from a large proportion of outliers only when the sample size is small.

Figure 6 shows the MSE’s for all the proposed methods for $p = 5$ and $\tilde{n} = 10$. Other simulations have been conducted and results are not provided because of redundancy in the associated comments. Errors are larger for the complete dataset model (in yellow) than for any of the three methods but tend to decrease as \tilde{n} increases, and thus the proportion of outliers decreases. TTR seems to provide the best results for large sample sizes (and thus for low proportions of outliers), while BOOT shows larger errors, especially when \tilde{n} is small (and thus when the proportion of outliers is high). An explanation of the phenomenon is that MSE is computed on the sample without the outliers. In that context, the MONO and TTR methods that select extreme (or “borderline”) observations as outliers tend to get smaller MSE. On the contrary,

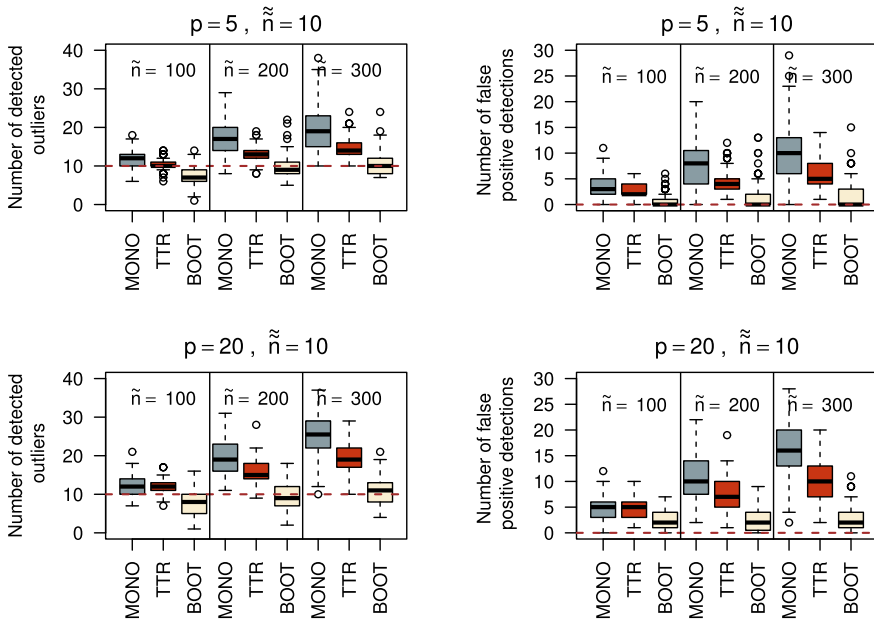


Fig. 7 Number of detected outliers (left column) and number of false positive detections (right column) for different values of p , \tilde{n} and $\tilde{n} = 10$

the BOOT approach does not exclude these “borderline” observations which are more difficult to predict correctly, leading to a larger MSE. The MSE descriptor must be interpreted with this remark in mind, as well as by taking into account the number of false positives of each method, which is done, thanks to Fig. 7.

Whatever the sets of parameters in Fig. 7, BOOT is the only method that seems to be able to select the true outliers without selecting too many false positives (i.e., individuals detected as outliers when they are not). BOOT seems to be the most efficient method by showing the lowest number of false positives for all \tilde{n} . The number of false positives stays somewhat constant over the sample size \tilde{n} for BOOT but increases with \tilde{n} for the other two methods. MONO and TTR methods seem to have a sensibility to \tilde{n} with an increase of the numbers of detected outliers and false positives as the sample size increases (and thus as the proportion of outliers decreases since their number \tilde{n} is fixed at 10).

6 A Real Data Application

Daily measurements of meteorological variables and ozone concentration are available in the dataset “ozone” (Source: Cornillon et al. 2012). More precisely, this dataset contains $n = 112$ daily measurements of meteorological variables (wind speed, tem-

perature, rainfall, cloudiness) and ozone concentration recorded in Rennes (France) in summer 2001. In this study, an individual is a day. Eleven numerical variables are measured with no missing values:

- $\max\text{O3}$: maximum of daily ozone concentration measured in gr/m^3 ,
- T_9, T_{12}, T_{15} : daily temperatures measured in degree Celsius at 9, 12, and 15 h (called “temperature” variables hereafter),
- $\text{Ne}_9, \text{Ne}_{12}, \text{Ne}_{15}$: cloudiness measured at 9, 12, and 15 h (called “cloudiness” variables hereafter),
- $V_{\times 9}, V_{\times 12}, V_{\times 15}$: wind speed (E-W component) measured at 9, 12, and 15 h (called “wind” variables hereafter),
- $\max\text{O3v}$: maximum concentration of ozone measured the day before.

The initial objective is to explain the maximum of daily ozone concentration (the response variable y is thus $\max\text{O3}$) by the $p = 10$ variables available ($T_9, T_{12}, T_{15}, \text{Ne}_9, \text{Ne}_{12}, \text{Ne}_{15}, V_{\times 9}, V_{\times 12}, V_{\times 15}, \max\text{O3v}$). Hereafter, let x be the vector of these ten covariates. To do this, the semiparametric regression model (2) is used and the EDR space $E = \text{Span}(\beta)$ is estimated by the usual SIR method (with the number of slices $H = 10$), while the link function f is estimated using the kernel estimator with the Gaussian kernel and the bandwidth tuned via cross-validation. Our aim is here to detect the presence or absence of outliers in this dataset. The proposed three outlier detection methods (MONO, TTR with $R = 1000$, and BOOT with $B = 1000$) are compared.

The naive MONO method does not detect outliers. The TTR method provides 9 outliers and the BOOT method identifies 4 “borderline” observations and no outlier (see the corresponding plots in Fig. 8, respectively, at the top left and at the top right). Among the 9 TTR’s outliers, 4 of them are the BOOT’s “borderline” observations. These 4 observations correspond to specific days in terms of road traffic, since these are days of major departures or returns from summer holidays in France. It is known that ozone pollution is also due to car traffic, but the built model is based only on weather data and does not take into account this important source of pollution. It is, therefore, quite natural that these 4 days correspond to individuals outside the model’s standards. The 5 other specific TTR’s outlier observations are closer to the scatterplot structure and they correspond to the days of early June, mid-June (music festival on the first day of summer), late July (end of a week), and mid-September.

In order to improve the final model, the method introduced by Jlassi and Saracco (2019) for selecting the relevant variables based on variable importance is now applied on the sample without the outliers (TTR method) or the “borderline” observations (BOOT method). Only the following $p^* = 4$ covariates are then selected: a temperature variable, T_{12} , a cloudiness variable, Ne_9 , a wind variable, $V_{\times 9}$, and the maximum concentration of ozone measured the day before, $\max\text{O3v}$. This is not surprising since the 3 variables of temperature (resp. cloudiness, wind speed) are strongly correlated with each other. The corresponding EDR directions are very close: $\hat{b}_{\text{SIR}}^{\text{TTR}} = (0.778, -0.565, 0.258, 0.094)'$ and $\hat{b}_{\text{SIR}}^{\text{BOOT}} = (0.660, -0.724, 0.175, 0.094)'$.

Finally, for the outlier detection method TTR (resp. BOOT), the plot of the estimated indices $x_i' \hat{b}_{\text{SIR}}^{\text{TTR}}$ (resp. $x_i' \hat{b}_{\text{SIR}}^{\text{BOOT}}$) versus the y_i ’s for the corresponding samples

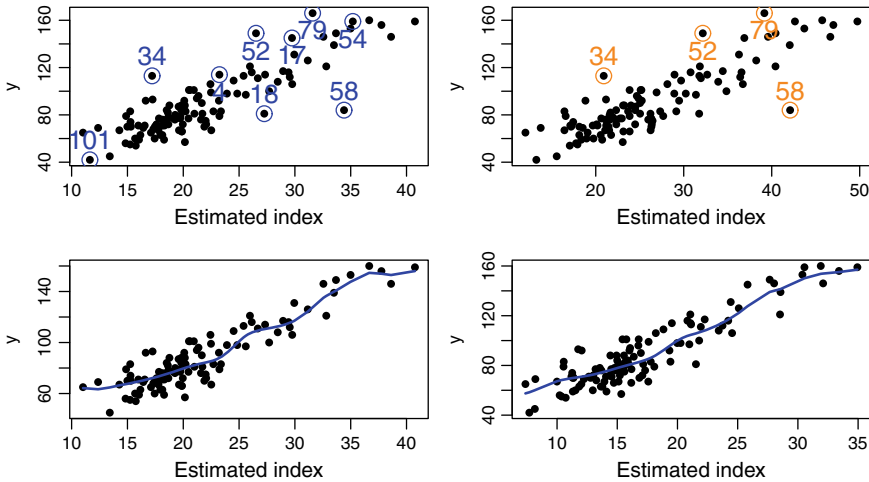


Fig. 8 Study of Ozone dataset. Top Left: Selected outliers with TTR method. Top Right: Selected “borderline” observations with BOOT method. Bottom Left: plot of the y_i ’s (values of $\max O_3$) versus the final estimated indices based on the $n_{TTR}^* = n - 9$ observations, i.e., removing the 9 selected outliers. Bottom Right: plot of the y_i ’s (values of $\max O_3$) versus the final estimated indices based on the $n_{BOOT}^* = n - 4$, removing the 4 selected “borderline” observations. The corresponding estimated link functions (solid blue curve) are superimposed on the last two plots

without outliers (resp. “borderline” observations) and the associated estimated link function (solid blue curve) are provided in Fig. 8 (at the bottom, on the right, resp. on the right). These two graphics are very similar and show an increasing link between the estimated index and the response variable $\max O_3$. Then, it is possible to interpret the coefficients of the estimated EDR direction \hat{b}_{SIR}^{TTR} (or similarly \hat{b}_{SIR}^{BOOT}) using their signs. The variable T_{12} (resp. $V_{\times 9}$ and $\max O_3 \hat{v}$) has a positive coefficient which means that an increase in daily temperatures at 12 h (resp. of the wind speed at 9h, or of maximum concentration of ozone measured the day before) implies an increase of the estimated index and this then implies (not surprisingly) an increase of a maximum of daily ozone concentration. On the contrary, the variable N_{e09} has a negative coefficient and then an increase of its values leads to a decrease in the maximum of daily ozone concentration, which is relevant from an air pollution point of view.

7 Concluding Remarks and Extensions

Three computational outlier detection approaches for sliced inverse regression have been presented. In this work, the original idea is to consider potential outliers that are outliers only in the SIR model and that are not detectable outliers by studying only their distribution in x or y . Thus, considering the plot of the estimated indices

versus the dependent variable, only outliers can appear in y . The case of outliers in x or in y is not considered here since the corresponding observations should be detectable as outliers in an early stage before the SIR modeling step, and the dataset should then be cleaned up accordingly. The MONO, TTR, and BOOT approaches were implemented in R and the code is available on <https://github.com/hlorenzo/outlierSIR>.

The philosophy of these approaches relies neither on the SIR method used in the first estimation step nor on the nonparametric regression used in the second estimation step. For example, instead of the usual SIR method, it is possible to use the SIR-II, SIR_α or SAVE methods among others. Moreover, the proposed approaches are also easily generalizable to the multiple-index model framework, i.e., when the dimension of the EDR space is equal to $K > 1$. All SIR-related methods, as well as nonparametric regression methods (like multivariate kernels), work well in this framework. However, the nonparametric regression methods might suffer from the well-known curse of dimensionality. Note that the choice of the dimension K of this EDR subspace should be then discussed. Finally, these outlier detection approaches can also be extended to a q -dimensional response variable y . Several authors developed SIR-based methods to estimate the EDR space that is common to the q components of the multivariate response variable, see, for instance, Barreda et al. (2007), Coudret et al. (2014), Li et al. (2003), Lue (2009), Saracco (2005) among others. However, the concept of an outlier in this multivariate framework must be first clarified since it is not entirely natural.

Acknowledgements Jérôme Saracco would like to sincerely thank Prof. Christine Thomas-Agnan for having “brought back in her luggage” the SIR method in Toulouse in the early 90s. His first research focused on contributions to SIR (master’s thesis, doctoral thesis, first articles in international journals). Thank you for all the discussions I had with you Christine throughout my career on many scientific subjects (nonparametric estimation, conditional quantiles, etc.) and many other subjects.

References

- Auger, I., & Lawrence, C. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1), 39–54.
- Azais, R., Gegout-Petit, A., & Saracco, J. (2012). Optimal quantization applied to sliced inverse regression. *Journal of Statistical Planning and Inference*, 142, 481–492.
- Babos, S., & Artemiou, A. (2020). Sliced inverse median difference regression. *Stat Methods & Applications*.
- Barreda, L., Gannoun, A., & Saracco, J. (2007). Some extensions of multivariate sliced inverse regression. *Journal of Statistical Computation and Simulation*, 77(1), 1–17.
- Cai, Z., Li, R., & Zhu, L. (2020). Online sufficient dimension reduction through sliced inverse regression. *Journal of Machine Learning Research*, 21(10), 1–25.
- Chavent, M., Girard, S., Kuentz-Simonet, V., Liquet, B., Nguyen, T. M. N., & Saracco, J. (2014). A sliced inverse regression approach for data stream. *Computational Statistics*, 29, 1129–1152.
- Chen, C.-H., & Li, K.-C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8(2), 289–316.

- Chiancone, A., Forbes, F., & Girard, S. (2017). Student sliced inverse regression. *Computational Statistics & Data Analysis*, *113*, 441–456.
- Cook, R. D. (2000). SAVE: A method for dimension reduction and graphics in regression. *Communications in Statistics - Theory Methods*, *29*, 2109–2121.
- Cook, R., & Critchley, F. (2000). Identifying regression outliers and mixtures graphically. *Journal of the American Statistical Association*, *95*(451), 781–794.
- Cornillon, P.-A., Guyader, A., Husson, F., Jégou, N., Josse, J., Kloareg, M., et al. (2012). *R for statistics*. Rennes: Chapman & Hall/CRC.
- Coudret, R., Girard, S., & Saracco, J. (2014). A new sliced inverse regression method for multivariate response. *Computational Statistics and Data Analysis*, *77*, 285–299.
- Dikheel, T. R. (2014). Robust sliced inverse regression. *Journal of Administrative and Economic Sciences*, *15*(1), 227–242.
- Dong, Y., Yu, Z., & Zhu, L. (2015). Robust inverse regression for dimension reduction. *Journal of Multivariate Analysis*, *134*, 71–81.
- Duan, N., & Li, K. C. (1991). Slicing regression: A link-free regression method. *Annals of Statistics*, *19*, 505–530.
- Edwards, A., & Cavalli-Sforza, L. (1965). Method for cluster analysis. *Biometrics*, *21*(2), 362–375.
- Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, *93*, 132–140.
- Gannoun, A., & Saracco, J. (2003). An asymptotic theory for SIR_α method. *Statistica Sinica*, *13*, 297–310.
- Gather, U., Hilker, T., & Becker, C. (2002). A note on outlier sensitivity of sliced inverse regression. *Statistics*, *36*(4), 271–281.
- Hall, P., & Li, K.-C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Annals of Statistics*, *21*(2), 867–889.
- Hsing, T. (1999). Nearest neighbor inverse regression. *Annals of Statistics*, *27*(2), 697–731.
- Jlassi, I., & Saracco, J. (2019). Variable importance assessment in sliced inverse regression for variable selection. *Communications in Statistics - Simulation and Computation*, *48*(1), 169–199.
- Killick, R., & Eckley, I. A. (2014). Changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, *58*(3)
- Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590–1598.
- Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. New York: Chapman and Hall/CRC.
- Li, K.-C. (1991). Sliced inverse regression for dimension reduction, with discussion. *Journal of the American Statistical Association*, *86*, 316–342.
- Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, *87*, 1025–1039.
- Li, K.-C., Aragon, Y., Shedden, K., & Thomas Agnan, C. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, *98*(461), 99–109.
- Li, Y., & Zhu, L. (2007). Asymptotics for sliced average variance estimation. *Annals of Statistics*, *35*, 41–69.
- Liquet, B., & Saracco, J. (2008). Application of the bootstrap approach to the choice of dimension and the α parameter in the SIR_α method. *Communications in Statistics - Simulation and Computation*, *37*(6), 1198–1218.
- Lue, H. (2009). Sliced inverse regression for multivariate response regression. *Journal of Statistical Planning and Inference*, *139*(8), 2656–2664.
- Prendergast, L. A. (2006). Detecting influential observations on the SIR e.d.r. space. *Australian & New Zealand Journal of Statistics*, *48*, 285–304.
- Prendergast, L. A. (2007). Implications of influence function analysis for sliced inverse regression and sliced average variance estimation. *Biometrika*, *94*(3), 585–601.

- Saracco, J. (1997). An asymptotic theory for sliced inverse regression. *Communications in Statistics - Theory Methods*, 26, 2141–2717.
- Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis*, 96, 117–135.
- Schimek, M. G. (Ed.). (2013). *Smoothing and regression: approaches, computation, and application*. New York: Wiley.
- Singh, K., & Xie, M. (2003). Bootlier-Plot: Bootstrap based outlier detection plot. *Sankhya, Series A*, 65(3), 532–559.
- Yin, X., & Seymour, L. (2005). Asymptotic distributions for dimension reduction in the SIR-II method. *Statistica Sinica*, 15(4), 1069–1079.
- Zhu, L. X., Miao, B., & Peng, H. (2006). On sliced inverse regression with large dimensional covariates. *Journal of the American Statistical Association*, 101, 630–643.
- Zhu, L., & Zhu, L. (2007). On kernel method for sliced average variance estimation. *Journal of Multivariate Analysis*, 98, 970–991.

Uncoupled Isotonic Regression with Discrete Errors



Jan Meis and Enno Mammen

Abstract In Rigollet and Weed (2019), an estimator was proposed for the uncoupled isotonic regression problem. It was shown that a so-called minimum Wasserstein deconvolution estimator achieves the rate $\log \log n / \log n$. Furthermore, it was shown that for normally distributed errors, this rate is optimal. In this note, we will show that for error distributions supported on a finite set of points, this rate can be improved to the order of $n^{-1/(2p)}$ for L_p -risks. We also show that this rate is optimal and cannot be improved for Bernoulli errors.

1 Introduction

In this note, we consider the nonparametric uncoupled isotonic regression problem. This estimation is related to nonparametric isotonic regression, where one observes Y_1, \dots, Y_n and x_1, \dots, x_n in the model

$$Y_i = m(x_i) + \varepsilon_i, \quad (1)$$

with x_1, \dots, x_n deterministic points in $[0, 1]$ and independent zero mean error variables $\varepsilon_1, \dots, \varepsilon_n$. In uncoupled isotonic regression, one does not observe the link between x_i and Y_i . That means, instead of Y_1, \dots, Y_n and x_1, \dots, x_n , one observes $Y_{\tau(1)}, \dots, Y_{\tau(n)}$ and x_1, \dots, x_n where $(\tau(1), \dots, \tau(n))$ is an unobserved permutation of $\{1, \dots, n\}$. In this model, identification of the function m is guaranteed if the distribution \mathcal{D} of the error variables $\varepsilon_1, \dots, \varepsilon_n$ is known and fulfils certain regularity

J. Meis

Institute of Medical Biometry and Informatics, Heidelberg University, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany
e-mail: meis@imbi.uni-heidelberg.de

E. Mammen (✉)

Institute of Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
e-mail: mammen@math.uni-heidelberg.de

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_7

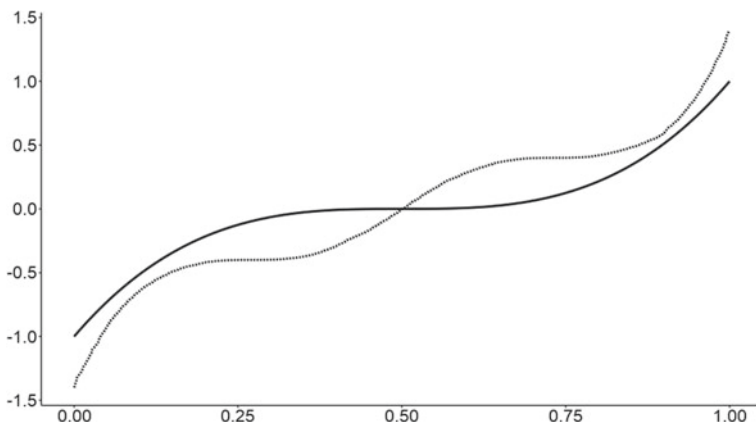


Fig. 1 For $m(x) = (2x - 1)^3$ (solid line), the figure shows the naive estimator (dotted line) based on ordering the observations Y_1, \dots, Y_n . Here the error variables ε_i have a discrete distribution with $P(\varepsilon_i = -0.4) = P(\varepsilon_i = 0.4) = 1/2$

conditions. Naively ordering the observations in an increasing manner does not lead to a consistent estimator, as illustrated in Fig. 1. Estimation in uncoupled isotonic regression is much harder than estimation in coupled isotonic regression where the permutation τ is known. This has been pointed out in Rigollet and Weed (2019), where an upper and lower bound has been established for uncoupled isotonic regression. They stated upper bounds of the order $\log \log n / \log n$ for the L_p error of estimates of m . Furthermore, they give a lower bound for Gaussian errors of the same order. We will come back to a comparison with coupled isotonic regression in Sect. 3.

In this note, we will show that much faster rates can be achieved in case of other error distributions. More precisely, we will show that in case of error distributions supported on a finite set of points, the function m can be estimated with rate $n^{-1/2}$ measured by the L_1 norm. We do not discuss practical applications of uncoupled isotonic regression. For such discussions, we refer to Rigollet and Weed (2019) where also further references to uncoupled isotonic regression can be found. See also Balabdaoui et al. (2020) where a strongly related estimation problem is discussed and Pananjady and Samworth (2020) where a multivariate generalization is studied.

As in Rigollet and Weed (2019), our estimator is based on minimum Wasserstein deconvolution. For our discussion below, we need the following definitions. First, we recall the definition of the Wasserstein-distance between probability distributions μ and ν on \mathbb{R} .

Definition 1 For $1 \leq p < \infty$, the Wasserstein-distance $W_p(\mu, \nu)$ between two probability distributions μ and ν with finite p -th moments on \mathbb{R} is defined as

$$W_p^p(\mu, \nu) = \inf_{\gamma} \int |x - y|^p d\gamma(x, y),$$

where the infimum is taken over all distributions γ on \mathbb{R}^2 with one-dimensional marginals μ and ν .

For probability measures on \mathbb{R} , their Wasserstein-distance can be easily calculated by the following formulas, see e.g. Bobkov and Ledoux (2019).

Proposition 1 *Let μ, ν be probability measures on \mathbb{R} with distribution functions F and G , respectively. If μ and ν have finite p -th moment, then*

$$W_p^p(\mu, \nu) = \int_0^1 |F^{-1}(x) - G^{-1}(x)|^p dx,$$

where F^{-1} and G^{-1} are the inverse distribution functions.

From this characterization, the following formula follows directly for $p = 1$.

Proposition 2 *Let μ, ν be probability measures on \mathbb{R} with distribution functions F and G , respectively. If μ and ν have finite first absolute moment, then*

$$W_1(\mu, \nu) = \int_{-\infty}^{\infty} |F(x) - G(x)| dx.$$

For the definition of minimum Wasserstein deconvolution, we need the following definition. Below, we will assume that the values x_1, \dots, x_n lie in a bounded subset of the real line. Without loss of generality, we assume that they lie in $[0, 1]$. We also assume that $x_1 \leq \dots \leq x_n$. For functions $f : [0, 1] \rightarrow \mathbb{R}$ and $p \geq 1$, we define the empirical norm $\|f\|_p$ by

$$\|f\|_p^p = \frac{1}{n} \sum_{i=1}^n |f(x_i)|^p.$$

Definition 2 For fixed values $x_1, \dots, x_n \in [0, 1]$ and any non-decreasing function $g : [0, 1] \rightarrow \mathbb{R}$, the discrete measure π_g is defined as

$$\pi_g = \frac{1}{n} \sum_{i=1}^n \delta_{g(x_i)},$$

where δ_z is the Dirac measure on the point $z \in \mathbb{R}$.

Additionally, for Y_1, \dots, Y_n from our model (1), we define $\hat{\pi}$ as

$$\hat{\pi} := \frac{1}{n} \sum_{i=1}^n \delta_{Y_i}.$$

With this in mind, we can now define the estimator based on minimum Wasserstein deconvolution (cf. Rigollet and Weed 2019).

Definition 3 For $V > 0$ and $q \geq 1$, the *minimum Wasserstein deconvolution estimator* is given by

$$\hat{m} \in \operatorname{argmin}_{g \in \mathcal{F}_V} W_q(\pi_g * \mathcal{D}, \hat{\pi}),$$

where \mathcal{F}_V is the class of non-decreasing functions from $[0, 1]$ to \mathbb{R} that are absolutely bounded by V , and $\pi_g * \mathcal{D}$ denotes the convolutions of the measures π_g and \mathcal{D} .

We suppose that in our model (1), the function m lies in \mathcal{F}_V . In Rigollet and Weed (2019), it was shown that the minimum Wasserstein deconvolution estimator with the choice $q = 2$ achieves the minimax optimal rate $\log \log n / \log n$ with respect to all L_p norms with $p \geq 1$. In this note, we will show for $p \geq 1$ that for discrete errors with a finite set of points as support, the optimal rate of convergence is of order $n^{-1/(2p)}$ with respect to the L_p norm. Furthermore, we will see that this rate of convergence is achieved by the minimum Wasserstein deconvolution estimator with the choice $q = 1$.

2 Estimation in Uncoupled Regression with Discrete Errors

From now on, we denote by \hat{m} the minimum Wasserstein deconvolution estimator with the choice $q = 1$. We get the following result for \hat{m} .

Theorem 1 *Suppose that \mathcal{D} is a discrete measure supported on a finite set of points. Then the minimum Wasserstein deconvolution estimator \hat{m} with $q = 1$ defined in Definition 3 fulfils the inequality*

$$\sup_{m \in \mathcal{F}_V} \mathbb{E} \|m - \hat{m}\|_1 \leq \frac{C(V, \mathcal{D})}{\sqrt{n}},$$

where $C(V, \mathcal{D})$ is a constant that depends only on V and \mathcal{D} .

Theorem 1 is our main result. Before we come to a further discussion of the theorem, we will now give the proof. For the proof of the theorem, we will make use of the following propositions.

Proposition 3 *Let μ, ν be probability measures supported on $[0, V]$ and suppose that \mathcal{D} is a discrete measure supported on a finite set of points. Then*

$$W_1(\mu, \nu) \leq C^*(V, \mathcal{D}) W_1(\mu * \mathcal{D}, \nu * \mathcal{D}),$$

where $C^*(V, \mathcal{D})$ is a constant only dependent on V and \mathcal{D} .

The following proposition is a modification of Proposition 3.1 in Rigollet and Weed (2019).

Proposition 4 *For the minimum Wasserstein-distance estimator defined in Definition 3 with $q = 1$, it holds that*

$$\mathbb{E}[W_1(\pi_{\hat{m}} * \mathcal{D}, \pi_m * \mathcal{D})] \leq \frac{4V + 2}{\sqrt{n}}.$$

We now come to the proof of Theorem 1.

Proof (of Theorem 1) According to Proposition 2.3 in Rigollet and Weed (2019), for non-decreasing functions $f, g \in \mathbf{L}_p([0, 1])$, it holds that $\|f - g\|_p = W_p(\pi_f, \pi_g)$. Application with $p = 1$ gives that

$$\|m - \hat{m}\|_1 = W_1(\pi_m, \pi_{\hat{m}}).$$

By Proposition 3, this can be bounded from above by

$$C^*(V, \mathcal{D}) W_1(\pi_m * \mathcal{D}, \pi_{\hat{m}} * \mathcal{D}).$$

By application of Proposition 4, we can bound the expectation of this term and we get that

$$\mathbb{E}\|m - \hat{m}\|_1 \leq \frac{C^*(V, \mathcal{D})}{\sqrt{n}}(4V + 2).$$

Thus, Theorem 1 holds with $C(V, \mathcal{D}) = C^*(V, \mathcal{D})(4V + 2)$. □

From Theorem 1, we get the following corollary.

Corollary 1 *Suppose that \mathcal{D} is a discrete measure supported on a finite set of points. Then the minimum Wasserstein deconvolution estimator \hat{m} with $q = 1$ defined in Definition 3 fulfils the inequality*

$$\sup_{m \in \mathcal{F}_V} \mathbb{E}\|m - \hat{m}\|_p \leq \sup_{m \in \mathcal{F}_V} (\mathbb{E}\|m - \hat{m}\|_p^p)^{1/p} \leq (2V)^{(p-1)/p} C(V, \mathcal{D})^{1/p} n^{-1/(2p)},$$

where $C(V, \mathcal{D})$ is the constant from Theorem 1 and where $p \geq 1$.

Proof Because m and \hat{m} are absolutely bounded by V , it holds that

$$\mathbb{E}\|m - \hat{m}\|_p^p \leq (2V)^{p-1} \mathbb{E}\|m - \hat{m}\|_1.$$

The corollary follows by application of Theorem 1. □

We now argue that the rates of convergence in Corollary 1 are minimax optimal. For this purpose, we suppose that the error variables are a Bernoulli sequence, i.e. the error variables fulfil $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$. We suppose that $x_i = i/n$ for $i = 1, \dots, n$. Furthermore, we consider the following binary testing problem $m = m_0$ versus $m = m_1$ where

$$m_0(x) = -1 + 2 \cdot 1_{x > \frac{1}{2}},$$

$$m_1(x) = -1 + 2 \cdot 1_{x > \frac{1}{2} - n^{-1/2}}.$$

Note that in this model Y_i takes only values in $\{-2, 0, 2\}$. For all i with $Y_i = -2$, it holds that $m(x_i) = -1$ and for all i with $Y_i = 2$ we have $m(x_i) = 1$. We define

$$N_1 = \#\{i : Y_i = -2\},$$

$$N_2 = \#\{i : Y_i = 2\}.$$

One can easily verify that (N_1, N_2) is a sufficient statistic in our testing problem. The variables N_1 and N_2 are independent and have Binomial distributions with parameters $(\frac{1}{2}, np_n)$ and $(\frac{1}{2}, n(1 - p_n))$, respectively, under the model m_0 and with parameter $(\frac{1}{2}, nq_n)$ and $(\frac{1}{2}, n(1 - q_n))$, respectively, under the model m_1 . Here p_n is chosen such that np_n is the largest natural number with $p_n \leq \frac{1}{2}$ and q_n is chosen such that nq_n is the largest natural number with $q_n \leq \frac{1}{2} - n^{-1/2}$. Under the model m_0 , the statistic $\sqrt{8}n^{-1/2}(N_1 - n/4, N_2 - n/4)$ has a limiting normal distribution with zero mean and identity covariance matrix. Under the model m_1 , the statistic has a limiting normal distribution with mean $(-\sqrt{2}, \sqrt{2})^T$ and identity covariance matrix. Thus we have a testing problem that does not degenerate in the limit. Furthermore note that we have

$$\|m_0 - m_1\|_p = ((p_n - q_n)2^p)^{1/p} = 2n^{-1/(2p)}(1 + o(1)). \quad (2)$$

By standard minimax theory arguments, we arrive at the following theorem.

Theorem 2 *Suppose that \mathcal{D} has Bernoulli distribution, that $0 \leq x_1 < \dots < x_n \leq 1$ and that $V \geq 1$. Then it holds that*

$$\inf_{\hat{m}_n} \sup_{m \in \mathcal{F}_V} \mathbb{E} \|m - \hat{m}_n\|_p \geq C_p n^{-1/(2p)},$$

where the infimum is taken over all estimators and where C_p is a constant depending on p .

At first sight, it may be surprising that the minimax rates depend so strongly on p . But it can be easily explained by the construction of our testing problem for the proof of Theorem 2. The functions m_0 and m_1 take only values in $\{-1, 1\}$ and they differ on an interval of decreasing length $n^{-1/2}$. For such two functions, the order of their L_p -distance strongly depends on p , see (2).

3 Comparison with Coupled Isotonic Regression

The rates of convergence in Theorem 1 and in Corollary 1 might seem startling at first, especially considering that it is faster than the rate of convergence for the general *coupled* isotonic regression problem, which is of order $n^{-1/3}$ (see Brunk 1970 for a first pointwise result, Van de Geer 1990, Nemirovskij et al. 1985 and Durot 2007 for results on global L_p -bounds and see Zhang 2002 for non-asymptotic risk bounds and for a summary on the history of error bounds in isotonic regression). The key difference here is that the uncoupled problem generally requires the error distribution \mathcal{D} to be known explicitly, while solutions of the coupled problem generally assume no explicit knowledge of \mathcal{D} . For a discussion of uncoupled isotonic regression with unknown error distribution, we refer to Balabdaoui et al. (2020). There is a well-established strand of literature on coupled isotonic regression and related estimation problems under shape constraints, starting in the fifties of the last century. For a detailed recent discussion of nonparametric estimation under shape constraints, see also Groeneboom and Jongbloed (2014).

To put the $1/\sqrt{n}$ rate into context, we shall investigate the behaviour of the *coupled* isotonic regression problem with Bernoulli noise. For this, we will examine a certain “reasonably well behaved” class of functions:

Definition 4 Let M be a class of monotone functions from $[0, 1]$ to $[0, V]$. We say that M is (δ, ε) -Lipschitz continuous, if $|f(x) - f(y)| < \varepsilon$ for all $f \in M$ and $x, y \in [0, 1]$ with $|x - y| < \delta$.

Proposition 5 Consider the coupled isotonic regression problem modelled by $Y_i = m(x_i) + \varepsilon_i$, where $\varepsilon_0, \dots, \varepsilon_n$ are i.i.d. with $\mathbb{P}(\varepsilon_i = -1) = \mathbb{P}(\varepsilon_i = 1) = \frac{1}{2}$ and x_i deterministic with $x_i = i/n$. If M is $(\delta, 1)$ -Lipschitz continuous for some $\delta > 0$, then there is an estimator \hat{m} such that for $n > \frac{1}{\delta}$ it holds that

$$\sup_{m \in M} \mathbb{E} \|m - \hat{m}\|_1 \leq \left(\frac{1}{2}\right)^n,$$

where $\|f - g\|_1 = \sum_{i=0}^n \frac{1}{n+1} |f(\frac{i}{n}) - g(\frac{i}{n})|$ denotes the empirical L_1 norm on an equidistant $n + 1$ point partition of $[0, 1]$.

Proof Suppose $\frac{1}{n} < \delta$. Imagine we know the value of $m(\frac{i}{n})$ for some $i \in \{0, \dots, n\}$. Since $\frac{i+1}{n} - \frac{i}{n} < \delta$ and m is monotone, we know that $m(\frac{i+1}{n}) \in [m(\frac{i}{n}), m(\frac{i}{n}) + 1]$. Thus, we can deduce the value of $\varepsilon_{i+1} = \text{sign}(Y_{i+1} - m(\frac{i}{n}))$ and therefore the value of $m(\frac{i+1}{n})$. Inductively we can use this method to reconstruct every value of m .

Consider now two values Y_i and Y_{i+1} for some i . If $Y_{i+1} - Y_i > 1$, we can deduce that $\varepsilon_i = -1$ and $\varepsilon_{i+1} = 1$ and thus we know the values of $m(\frac{i}{n})$ and $m(\frac{i+1}{n})$. If $Y_{i+1} < Y_i$, we can deduce that $\varepsilon_{i+1} = -1$ and $\varepsilon_i = 1$, and again we can recover the values of $m(\frac{i}{n})$ and $m(\frac{i+1}{n})$. Therefore, the only cases where none of the values of $m(\frac{i}{n})$ can be reliably recovered are the cases where $Y_i < Y_{i+1}$ and $Y_{i+1} - Y_i < 1$ hold for all $i = 0, \dots, n - 1$. There are precisely two such cases: The case where

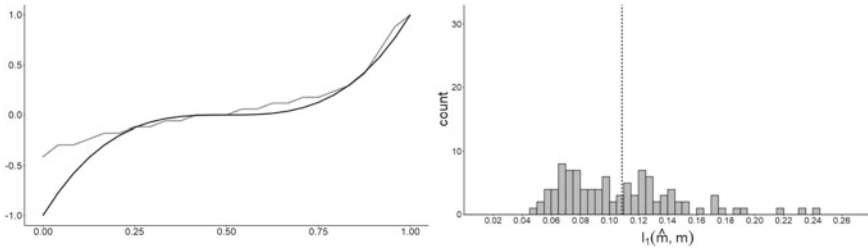


Fig. 2 Minimum Wasserstein deconvolution estimator for $n = 25$. The left plot shows the underlying curve m (solid line) and the minimum Wasserstein deconvolution estimator (dotted line). The plot shows one simulated estimator. The right plot shows a histogram of the L_1 errors of 100 simulated estimators. The dotted vertical line shows the mean value of the L_1 errors of the simulated minimum Wasserstein deconvolution estimators

$\varepsilon_0 = \dots = \varepsilon_n = -1$ and the case where $\varepsilon_0 = \dots = \varepsilon_n = 1$. Whenever we are in these cases, we put $\hat{m} \left(\frac{i}{n} \right) := Y_i$, and otherwise we put $\hat{m} := m$. We get

$$\mathbb{E} \|m - \hat{m}\|_1 \leq 1 \cdot (\mathbb{P}(\varepsilon_0, \dots, \varepsilon_n = -1) + \mathbb{P}(\varepsilon_0, \dots, \varepsilon_n = 1)) = \left(\frac{1}{2}\right)^n .$$

□

There is a relation between the Bernoulli noise case described above and the model of normally distributed noise covered in Rigollet and Weed (2019). In both cases, the uncoupled isotonic regression problems seem to be “logarithmically-worse” than their coupled counterparts. By *logarithmically-worse*, we mean that if r_n is an upper bound for the rate of convergence of the coupled problem, the upper bound for the uncoupled problem is worse than $\frac{1}{-\log(r_n)}$: In the Bernoulli case, we have something along the lines of $\exp(-n)$ as an upper bound for the coupled problem and $\frac{1}{\sqrt{n}}$ for the uncoupled problem, while in the normally distributed case, we have $(\frac{\log(n)}{n})^{1/3}$ as a bound for the coupled case and $\frac{\log(\log(n))}{\log(n)}$ as a bound for the uncoupled case.

We conclude this section by showing some simulations for the scenario we introduced in Fig. 1. In Figs. 2, 3 and 4, we show the performance of a discretized version of the minimum Wasserstein estimator for varying n . This computational estimator was originally introduced in Sect. 2.2 of Rigollet and Weed (2019) and an implementation of the algorithm to compute this estimator is available at <https://github.com/jan-imbi/UncoupledIsoReg> in the programming language R.

Again we use error variables with $\mathbb{P}(\varepsilon = 0.4) = \mathbb{P}(\varepsilon = -0.4) = \frac{1}{2}$. For numerical reasons, the space of functions over which we optimize the Wasserstein-distance was discretized, which is why the plots of our estimates show discontinuities.

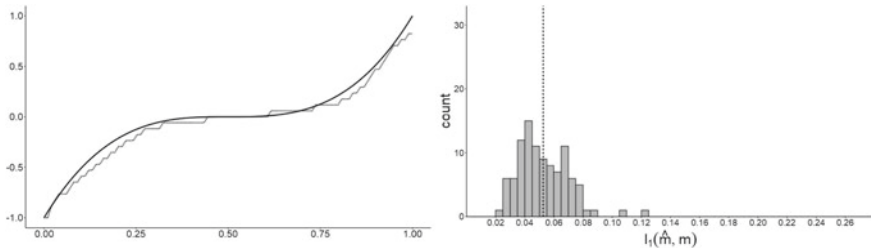


Fig. 3 Minimum Wasserstein deconvolution estimator as in Fig. 2 but for $n = 100$

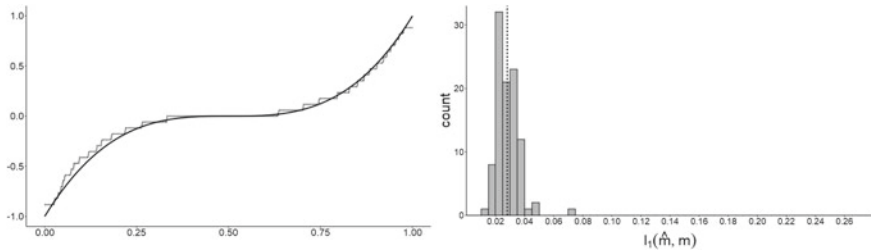


Fig. 4 Minimum Wasserstein deconvolution estimator as in Fig. 2 but for $n = 400$

4 Additional Proofs

Proof (of Proposition 3) Put $p_j = P(\varepsilon_i = u_j)$ for $j = 1, \dots, J$, where $\{u_1, \dots, u_J\}$ with $u_1 < \dots < u_J$ is the support of ε_i and J is the number of elements of the support. Denoting by F the cumulative distribution function of μ and by G the cumulative distribution function of ν , we get from Proposition 2 with $\Delta(x) = F(x) - G(x)$ and $\Gamma(x) = \sum_{j=1}^J p_j \Delta(x - u_j)$ that

$$W_1(\mu, \nu) = \int_{-\infty}^{\infty} |\Delta(x)| dx, \tag{3}$$

$$W_1(\mu * \mathcal{D}, \nu * \mathcal{D}) = \int_{-\infty}^{\infty} |\Gamma(x)| dx. \tag{4}$$

We now use that $\Delta(x - u_j) = 0$ for $x < u_j$ and for $x > u_j + V$ and that for $K \geq 1$ and $0 \leq x - u_j \leq V$

$$\begin{aligned}
 \Delta(x - u_J) &= \frac{1}{p_J} \Gamma(x) - \frac{1}{p_J} \sum_{j=1}^{J-1} p_j \Delta(x - u_j) & (5) \\
 &= \frac{1}{p_J} \Gamma(x) - \frac{1}{p_J^2} \sum_{j=1}^{J-1} p_j \Gamma(x + u_J - u_j) \\
 &\quad + \frac{1}{p_J^2} \sum_{j=1}^{J-1} \sum_{k=1}^{J-1} p_j p_k \Delta(x - u_j + u_J - u_k) \\
 &= \frac{1}{p_J} \Gamma(x) - \frac{1}{p_J^2} \sum_{j=1}^{J-1} p_j \Gamma(x + u_J - u_j) \pm \dots \\
 &\quad + (-1)^K \frac{1}{p_J^{K+1}} \sum_{j_1, \dots, j_K=1}^{J-1} p_{j_1} \cdot \dots \cdot p_{j_K} \Gamma(x + u_J - u_{j_1} + \dots + u_J - u_{j_K}) \\
 &\quad - (-1)^K \frac{1}{p_J^{K+1}} \sum_{j_1, \dots, j_{K+1}=1}^{J-1} p_{j_1} \cdot \dots \cdot p_{j_{K+1}} \Delta(x + K u_J - u_{j_1} - \dots - u_{j_{K+1}}).
 \end{aligned}$$

If we choose K with $K \geq \frac{V}{u_J - u_{J-1}} - 1$, we get for $1 \leq j_1, \dots, j_{K+1} \leq J - 1$ that for $0 \leq x - u_J \leq V$

$$\begin{aligned}
 x + K u_J - u_{j_1} - \dots - u_{j_{K+1}} &\geq (K + 1)(u_J - u_{J-1}) \\
 &\geq V.
 \end{aligned}$$

This implies that for such values of x, j_1, \dots, j_{K+1} and K ,

$$\Delta(x + K u_J - u_{j_1} - \dots - u_{j_{K+1}}) = 0.$$

Thus we get from (5) that

$$\begin{aligned}
 \Delta(x - u_J) &= \frac{1}{p_J} \Gamma(x) - \frac{1}{p_J^2} \sum_{j=1}^{J-1} p_j \Gamma(x + u_J - u_j) \pm \dots & (6) \\
 &\quad + (-1)^K \frac{1}{p_J^{K+1}} \sum_{j_1, \dots, j_K=1}^{J-1} p_{j_1} \cdot \dots \cdot p_{j_K} \Gamma(x + u_J - u_{j_1} + \dots + u_J - u_{j_K}).
 \end{aligned}$$

By application of (3) and (6), we get that

$$\begin{aligned}
 W_1(\mu, \nu) &= \int_{-\infty}^{\infty} |\Delta(x)| \, dx \\
 &= \int_{-\infty}^{\infty} |\Delta(x - u_J)| \, dx \\
 &\leq \frac{1}{p_J} \int_{-\infty}^{\infty} |\Gamma(x)| \, dx + \frac{1}{p_J^2} \sum_{j=1}^{J-1} p_j \int_{-\infty}^{\infty} |\Gamma(x + u_J - u_j)| \, dx + \dots \\
 &\quad + \frac{1}{p_J^{K+1}} \sum_{j_1, \dots, j_K=1}^{J-1} p_{j_1} \cdot \dots \cdot p_{j_K} \\
 &\quad \times \int_{-\infty}^{\infty} |\Gamma(x + u_J - u_{j_1} + \dots + u_J - u_{j_K})| \, dx \\
 &= \left(\frac{1}{p_J} + \dots + \frac{1}{p_J^{K+1}} \sum_{j_1, \dots, j_K=1}^{J-1} p_{j_1} \cdot \dots \cdot p_{j_K} \right) \int_{-\infty}^{\infty} |\Gamma(x)| \, dx \\
 &\leq \left(\frac{1}{p_J} + \dots + \frac{1}{p_J^{K+1}} \right) \int_{-\infty}^{\infty} |\Gamma(x)| \, dx \\
 &\leq C^*(V, \mathcal{D}) W_1(\mu * \mathcal{D}, \nu * \mathcal{D})
 \end{aligned}$$

with

$$C^*(V, \mathcal{D}) = \frac{\frac{1}{p_J^{K+2}} - \frac{1}{p_J}}{\frac{1}{p_J} - 1},$$

where in the last step (4) has been used. This concludes the proof of Proposition 3. □

Proof (of Proposition 4) In the proof of the proposition, we make use of the following classical result (see (Bobkov and Ledoux 2019, Theorem 3.2)): □

Lemma 1 *Let X_1, \dots, X_n be iid from μ and denote by $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ the empirical measure associated with this sample. Let $F : \mathbb{R} \rightarrow \mathbb{R}$ denote the cumulative distribution function of μ , and $F_n : \mathbb{R} \rightarrow \mathbb{R}$ the empirical cumulative distribution function of μ_n . It holds that*

$$\mathbb{E}[W_1(\mu, \hat{\mu}_n)] \leq \frac{1}{\sqrt{n}} \int \sqrt{F(x)(1 - F(x))} \, d\lambda(x).$$

Because $F(x)(1 - F(x))$ is bounded by $\frac{1}{4}$ and the domain of integration is bounded by the support of μ , we get from Lemma 1 that for μ supported on $[-V, V]$, it holds that

$$\mathbb{E}[W_1(\mu, \hat{\mu}_n)] \leq \frac{V}{\sqrt{n}}. \tag{7}$$

By application of the triangle inequality for Wasserstein distances and the choice of \hat{m} as the argmin from Definition 3, we get that

$$W_1(\pi_{\hat{m}} * \mathcal{D}, \pi_m * \mathcal{D}) \leq W_1(\pi_{\hat{m}} * \mathcal{D}, \hat{\pi}) + W_1(\hat{\pi}, \pi_m * \mathcal{D}) \leq 2 W_1(\pi_m * \mathcal{D}, \hat{\pi}).$$

We want to finish the argument by applying (7), but we need to employ the following trick to justify this: Recall that $\hat{\pi}$ is supported on $\{m(x_1) + \varepsilon_1, \dots, m(x_n) + \varepsilon_n\}$. Note that $m(x_1) + \varepsilon_1, m(x_2) + \varepsilon_2, \dots, m(x_n) + \varepsilon_n$ is not an iid sample from $\pi_m * \mathcal{D}$ because the x_i are not random. However, if we define w_1, \dots, w_n to be an iid sample from π_m , independent of $\varepsilon_1, \dots, \varepsilon_n$, then $w_1 + \varepsilon_1, \dots, w_n + \varepsilon_n$ is an iid sample from $\pi_m * \mathcal{D}$. Denote by $w_{(1)}, \dots, w_{(n)}$ the order statistic of w_1, \dots, w_n . By the triangle inequality, we get

$$W_1(\pi_m * \mathcal{D}, \hat{\pi}) \leq W_1\left(\pi_m * \mathcal{D}, \frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \varepsilon_i}\right) + W_1\left(\frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \varepsilon_i}, \hat{\pi}\right).$$

For the first summand, notice that $w_{(i)} + \varepsilon_i$ is in general not distributed according to $\pi_m * \mathcal{D}$. However, we can still apply (7) by the following argument: Since the ordering of the $w_{(i)}$ does not depend on the ε_i and because the w_i and ε_i are independent, we have that

$$\mathbb{E}[W_1(\pi_m * \mathcal{D}, \frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \varepsilon_i})] = \mathbb{E}[W_1(\pi_m * \mathcal{D}, \frac{1}{n} \sum_{i=1}^n \delta_{w_i + \varepsilon_i})].$$

Therefore, $W_1(\pi_m * \mathcal{D}, \frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \varepsilon_i})$ can indeed be bounded by $\frac{V+1}{\sqrt{n}}$.

For $W_1(\frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \varepsilon_i}, \hat{\pi})$, the coupling

$$\gamma := \frac{1}{n} \sum_{i=1}^n \delta_{(w_{(i)} + \varepsilon_i, Y_i)}$$

yields

$$W_1\left(\frac{1}{n} \sum_{i=1}^n \delta_{w_{(i)} + \varepsilon_i}, \hat{\pi}\right) \leq \frac{1}{n} \sum_{i=1}^n |w_{(i)} + \varepsilon_i - m(x_i) - \varepsilon_i| = W_1\left(\pi_m, \frac{1}{n} \sum_{i=1}^n \delta_{w_i}\right).$$

Here, the last equality follows from Proposition 1 applied to discrete measures μ and ν equal to the empirical measures π_m and $\frac{1}{n} \sum_{i=1}^n \delta_{w_i}$, respectively. Since the w_i are iid from π_m for $i = 1, \dots, n$, this term can also be bounded by $\frac{V}{\sqrt{n}}$ according to (7). \square

Acknowledgements This note generalizes results of the master thesis Meis (2019) written by the first author at Heidelberg University.

References

- Balabdaoui, F., Doss, C. R., & Durot, C. (2020). Unlinked monotone regression. [arXiv:2007.00830](https://arxiv.org/abs/2007.00830).
- Bobkov, S., & Ledoux, M. (2019). One-dimensional empirical measures, order statistics, and kantovich transport distances. *Memoirs of the American Mathematical Society*, 261(1259).
- Brunk, H. D. (1970). Estimation of isotonic regression. *Nonparametric techniques in statistical inference* (pp. 177–195). Cambridge: Cambridge University Press.
- Durot, C. (2007). On the L_p -error of monotonicity constrained estimators. *Annals of Statistics*, 35, 1080–1104.
- Groeneboom, P., & Jongbloed, G. (2014). *Nonparametric estimation under shape constraints*. Cambridge: Cambridge University Press.
- Meis, J. (2019). Uncoupled isotonic regression for Bernoulli-perturbed data, Master thesis, Faculty of Mathematics and Computer Science, Heidelberg University.
- Nemirovskij, A. S., Polyak, B., & Tsybakov, A. B. (1985). Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problems of Information Transmission*, 21(4), 258–272.
- Pananjady, A., Samworth, R. J. (2020). Isotonic regression with unknown permutations: Statistics, computation, and adaptation. [arXiv:2009.02609v1](https://arxiv.org/abs/2009.02609v1).
- Rigollet, P., & Weed, J. (2019). Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information and Inference: A Journal of the IMA*, 8(4), 691–717.
- Van de Geer, S. (1990). Estimating a regression function. *The Annals of Statistics*, 907–924.
- Zhang, C. H. (2002). Risk bounds in isotonic regression. *The Annals of Statistics*, 30(2), 528–555.

Quantiles and Expectiles

Partially Linear Expectile Regression Using Local Polynomial Fitting



Cécile Adam and Irène Gijbels

Abstract This chapter deals with partially linear expectile regression using local polynomial fitting as a basic smoothing technique in the various steps. The advantage of the estimation method is that an explicit expression for an optimal choice of the bandwidth (matrix) can be established, and based on this, a rule-of-thumb bandwidth selector is presented. A small simulation study demonstrates that the estimation method with this data-driven choice of the bandwidth performs very well. An illustration with a real data example is provided.

1 Introduction

In mean regression, the interest is in finding the impact that a vector of covariates $\mathbf{X} = (X_1, \dots, X_d)^T$ has, on average, on the variable of interest Y , i.e. the object of interest is $E(Y|\mathbf{X})$. The mean is however only one characteristic of the conditional distribution function $F_{Y|\mathbf{X}}$ of Y given \mathbf{X} . Conditional quantiles $q_\alpha(\mathbf{x}) = \inf_y \{y \in \mathbb{R} : F_{Y|\mathbf{X}}(y|\mathbf{x}) \geq \alpha\}$, with $\alpha \in (0, 1)$, provide a full description of the conditional distribution function of Y given \mathbf{x} . This property is shared by conditional expectiles, which are however quite different in nature. Expectiles are obtained by minimizing an asymmetrically weighted squared error criterion, whereas quantiles

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-73249-3_8) contains supplementary material, which is available to authorized users.

C. Adam

Department of Mathematics, KU Leuven, Celestijnenlaan 200B, B-3001 Leuven (Heverlee), Belgium

e-mail: cecile.adam@kuleuven.be

I. Gijbels (✉)

Department of Mathematics and Leuven Statistics Research Center (LStat), KU Leuven, Celestijnenlaan 200B, B-3001 Leuven (Heverlee), Belgium

e-mail: irene.gijbels@kuleuven.be

result from minimizing an asymmetrically weighted absolute error criterion. Consequently, conditional expectiles are sensitive to outlying observations, but exactly this lack of sensitivity to the magnitude of observations makes conditional quantiles less appropriate in, for example, risk management than expectiles. For a review on the advantages and inconveniences of quantiles and expectile, see e.g. Schulze Waltrup et al. (2015).

Newey and Powell (1987), following earlier work of Aigner et al. (1976), discussed expectiles in the context of linear regression models, in which the influence of the vector of covariates on the expectile function is of the form $\delta^T \mathbf{X}$, with $\delta \in \mathbb{R}^{d^*}$, and with d^* the dimension of \mathbf{X} . This linear influence is often too restrictive when it comes to applications, and in nonparametric settings, the influence of \mathbf{X} on the expectile curve is unspecified and involves, in general, a d^* -variate unknown function $g : \mathbb{R}^{d^*} \rightarrow \mathbb{R}$ that describes the influence $g(\mathbf{X})$ on the expectile curve. Yao and Tong (1996) studied nonparametric expectile regression in case of a univariate explanatory variable using local linear fitting. Recently, Adam and Gijbels (2021) extended this to local polynomial regression, studied in detail optimal bandwidth choice and investigated several data-driven bandwidth selectors. Other smoothing (nonparametric) techniques for estimating g are approximations by splines. See Schnabel and Eilers (2009) and Schulze Waltrup and Kauermann (2017), among others. Flexible expectile regression in a reproducing kernel Hilbert space context was studied in Yang et al. (2018).

The aim of this paper is to consider the setting in which part of a set of covariates \mathbf{X} have a linear effect on the expectile curve whereas for another set of covariates $\mathbf{Z} = (Z_1, \dots, Z_q)^T$, the form of the influence cannot be assumed to be linear, and hence is modelled in a nonparametric way through an unknown function $g : \mathbb{R}^q \rightarrow \mathbb{R}$. This leads to a combined influence of the vector $(\mathbf{X}^T, \mathbf{Z}^T)^T$ and the form $\delta^T \mathbf{X} + g(\mathbf{Z})$, and to a semiparametric model, more precisely a partially linear model. Sobotka et al. (2013) considered such a semiparametric model in a geoaddivitive setting when also spatial effects are involved.

When only a small set of predictors are significant in a linear expectile model, penalization techniques are used to select these in Zhao et al. (2018) and Liao et al. (2019), and extended to semiparametric expectile regression in Zhao et al. (2019). For a discussion on various approaches towards model selection for semiparametric expectile regression, see Spiegel and Sobotka (2017). When several plausible estimators are available Gu and Zou (2019), aggregate these using exponential weighting. The use of envelope models for estimation in expectile regression is studied in Chen et al. (2020).

In this chapter, we study multivariate partially linear expectile regression in which the nonparametric part is estimated using local polynomial techniques. A specific advantage of using this technique is that an explicit expression can be provided for an optimal choice of a bandwidth parameter. As in all smoothing techniques, good data-driven choices of smoothing parameters are crucial for practical use.

This chapter is organized as follows. In Sect. 2, we introduce the framework for partially linear expectile regression. The estimation methodology is exposed in Sect. 3, and Sect. 4 deals with the important issue of bandwidth selection. Numer-

ical illustrations are in Sects. 5 and 6. The Supplementary Material contains some theoretical derivations, further details on statistical methodology and finite-sample performance under heteroscedastic settings and additional simulation results.

2 Partially Linear Expectile Regression

Consider a random vector $(Y, \mathbf{X}^T, \mathbf{Z}^T)^T$ with $\mathbf{X} = (1, X_1, \dots, X_d)^T$ and $\mathbf{Z} = (Z_1, \dots, Z_q)^T$, where the notation \mathbf{A}^T denotes the transposed of the vector or matrix \mathbf{A} . The inclusion of a first component 1 into the covariate vector is needed for taking care of an intercept term. Denote $\mathbf{x} = (1, x_1, \dots, x_d)^T \in \mathbb{R}^{d+1}$ and $\mathbf{z} = (z_1, \dots, z_q)^T \in \mathbb{R}^q$. The interest here is in estimating the ω th conditional expectile of Y given $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$, with $\omega \in (0, 1)$, i.e.

$$\tau_\omega(\mathbf{x}, \mathbf{z}) = \arg \min_{a \in \mathbb{R}} E_{Y|\mathbf{X}, \mathbf{Z}} [Q_\omega(Y - a) | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}]$$

with Q_ω the expectile loss function

$$Q_\omega(y) = |\omega - \mathbb{1}\{y \leq 0\}| y^2, \quad (1)$$

which is to be distinguished from the quantile loss function (also called check function) in which the factor y^2 is replaced by $|y|$. For interpretations of expectiles, quantiles and their advantages and disadvantages, see, for example, Schulze Waltrup et al. (2015) and Adam and Gijbels (2021), among others.

A partially linear structured expectile curve takes the form

$$\tau_\omega(\mathbf{x}, \mathbf{z}) = \boldsymbol{\delta}_\omega^T \mathbf{x} + g_\omega(\mathbf{z}),$$

where possibly both parts, the parametric part $\boldsymbol{\delta}_\omega^T \mathbf{x}$ and the nonparametric part $g_\omega(\mathbf{z})$, may depend on ω . In this chapter, we focus on a regression model

$$Y = \boldsymbol{\delta}^T \mathbf{X} + g(\mathbf{Z}) + \sigma \epsilon, \quad (2)$$

with $\boldsymbol{\delta} = (\delta_0, \delta_1, \dots, \delta_d)^T$, $0 \leq \sigma < \infty$, and where the error term ϵ satisfies $E(\epsilon | \mathbf{X}, \mathbf{Z}) = 0$ and $\text{Var}(\epsilon | \mathbf{X}, \mathbf{Z}) = 1$. Note that, in order to ensure identifiability of all elements in model (2), we need to impose some constraint. Indeed, adding any constant to the function g and subtracting the same constant to the component $\delta_{0,\omega}$ lead to exactly the same expression for $\tau_\omega(\cdot, \cdot)$. A standard condition to guarantee identifiability is to impose that $E[g(\mathbf{Z})] = 0$.

An important property of unconditional expectiles is the following. Let $\tilde{Y} = a + bY$, with $a, b \in \mathbb{R}$, and denote the ω th expectile of Y by $\tau_{\omega,Y}$, then the ω th expectile of \tilde{Y} , denoted by $\tau_{\omega,\tilde{Y}}$, is given by

$$\tau_{\omega, \tilde{Y}} = \begin{cases} a + b\tau_{\omega, Y} & \text{if } b > 0 \\ a + b\tau_{1-\omega, Y} & \text{if } b \leq 0. \end{cases} \quad (3)$$

See, for example, Newey and Powell (1987) and Remillard and Abdous (1995).

Using a conditional version of (3), we obtain that under (2), the ω th conditional expectile of Y given $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$ equals

$$\tau_{\omega}(\mathbf{x}, \mathbf{z}) = \boldsymbol{\delta}^T \mathbf{x} + g(\mathbf{z}) + \sigma \tau_{\omega, \epsilon} \equiv \boldsymbol{\delta}_{\omega}^T \mathbf{x} + g(\mathbf{z}), \quad (4)$$

with $\tau_{\omega, \epsilon}$ the ω th unconditional expectile of the error term ϵ , and where we denoted $\boldsymbol{\delta}_{\omega} = (\delta_{0, \omega}, \delta_1, \dots, \delta_d)^T$ where $\delta_{0, \omega} = \delta_0 + \sigma \tau_{\omega, \epsilon}$. Hence only the first component of the linear part $\boldsymbol{\delta}_{\omega}^T \mathbf{x}$ in (4) is affected by the value of ω . All other d components of the regression vector $\boldsymbol{\delta}_{\omega}$ in the linear part as well as the nonparametric part $g(\mathbf{z})$ are not affected by the value of ω .

Consider an i.i.d. sample $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \dots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$ from $(Y, \mathbf{X}, \mathbf{Z})$ following model (2), i.e., for each $i = 1, \dots, n$, we have

$$Y_i = \boldsymbol{\delta}^T \mathbf{X}_i + g(\mathbf{Z}_i) + \sigma \epsilon_i,$$

where $\mathbf{X}_i = (1, X_{i1}, \dots, X_{id})^T$, and $\epsilon_1, \dots, \epsilon_n$ are i.i.d. from ϵ . The aim is then to estimate the expectile function $\tau_{\omega}(\mathbf{x}, \mathbf{z})$ in (4) which involves estimating the vector of regression coefficients $\boldsymbol{\delta}_{\omega}$ and the q -variate function $g(\cdot)$.

Homoscedasticity and Heteroscedasticity

Note that the error term in (2) is constant, i.e. not depending on the covariate vector $(\mathbf{X}^T, \mathbf{Z}^T)^T$. In case of heteroscedasticity, i.e. when the error variance σ^2 depends on the covariate vector, one can distinguish various cases: the error variance σ^2 is a function of only $\mathbf{X}_{(-1)} = (X_1, \dots, X_d)^T$, of only \mathbf{Z} or of the full covariate vector $(\mathbf{X}_{(-1)}^T, \mathbf{Z}^T)^T$. The vector $\mathbf{X}_{(-1)}$ differs from \mathbf{X} only by not containing the first component 1. That component need not be included when looking at into an error variance part. Consider first the case that σ is a function of \mathbf{Z} only, i.e. $\sigma(\mathbf{Z})$, and the underlying regression model is of the form

$$Y = \boldsymbol{\delta}^T \mathbf{X} + g(\mathbf{Z}) + \sigma(\mathbf{z}) \epsilon. \quad (5)$$

The ω th conditional expectile of Y given $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$ under model (5) equals

$$\tau_{\omega}(\mathbf{x}, \mathbf{z}) = \boldsymbol{\delta}^T \mathbf{x} + g(\mathbf{z}) + \sigma(\mathbf{z}) \tau_{\omega, \epsilon}.$$

Keeping in mind the constraint on the nonparametric function part (zero expectation), the expectile function is viewed as

$$\tau_{\omega}(\mathbf{x}, \mathbf{z}) = \boldsymbol{\delta}^T \mathbf{x} + \text{E}[\sigma(\mathbf{Z})] \tau_{\omega, \epsilon} + g(\mathbf{z}) + \{\sigma(\mathbf{z}) - \text{E}[\sigma(\mathbf{Z})]\} \tau_{\omega, \epsilon} \equiv \boldsymbol{\delta}_{\omega}^T \mathbf{x} + g_{\omega}(\mathbf{z}),$$

Table 1 Homoscedasticity and various heteroscedasticity structures

Error structure	Expectile function $\tau_\omega(\mathbf{x}, \mathbf{z})$	Elements depending on ω
$\sigma \epsilon$	$\delta_\omega^T \mathbf{x} + g(\mathbf{z})$	$\delta_{0,\omega} = \delta_0 + \sigma \tau_{\omega,\epsilon}$
$\sigma(\mathbf{Z}) \epsilon$	$\delta_\omega^T \mathbf{x} + g_\omega(\mathbf{z})$	$\delta_{0,\omega} = \delta_0 + E[\sigma(\mathbf{Z})] \tau_{\omega,\epsilon}$ $g_\omega(\mathbf{z}) = g(\mathbf{z}) + \{\sigma(\mathbf{z}) - E[\sigma(\mathbf{Z})]\} \tau_{\omega,\epsilon}$
$\sigma(\mathbf{X}_{(-1)}) \epsilon$	$\delta_\omega^T \mathbf{x} + g(\mathbf{z}) + g_\omega(\mathbf{x}_{(-1)})$	$\delta_{0,\omega} = \delta_0 + E[\sigma(\mathbf{X}_{(-1)})] \tau_{\omega,\epsilon}$ $g_\omega(\mathbf{x}_{(-1)}) = \{\sigma(\mathbf{x}_{(-1)}) - E[\sigma(\mathbf{X}_{(-1)})]\} \tau_{\omega,\epsilon}$
$\sigma(\mathbf{X}_{(-1)}, \mathbf{Z}) \epsilon$	$\delta_\omega^T \mathbf{x} + g_\omega(\mathbf{x}_{(-1)}, \mathbf{z})$	$\delta_{0,\omega} = \delta_0 + E[\sigma(\mathbf{X}_{(-1)}, \mathbf{Z})] \tau_{\omega,\epsilon}$ $g_\omega(\mathbf{x}_{(-1)}, \mathbf{z}) = g(\mathbf{z}) + \tilde{g}_\omega(\mathbf{x}_{(-1)}, \mathbf{z})$ $\tilde{g}_\omega(\mathbf{x}_{(-1)}, \mathbf{z}) = \{\sigma(\mathbf{x}_{(-1)}, \mathbf{z}) - E[\sigma(\mathbf{X}_{(-1)}, \mathbf{Z})]\} \tau_{\omega,\epsilon}$

with

$$\delta_{0,\omega} = \delta_0 + E[\sigma(\mathbf{Z})] \tau_{\omega,\epsilon} \quad \text{and} \quad g_\omega(\mathbf{z}) = g(\mathbf{z}) + \{\sigma(\mathbf{z}) - E[\sigma(\mathbf{Z})]\} \tau_{\omega,\epsilon},$$

and for which $E[g_\omega(Z)] = 0$, implied by $E[g(Z)] = 0$. Of importance is to note that under such a heteroscedastic structure, the dependence on ω shows up on two levels: the intercept term of the linear part, as well as in the nonparametric part. In real data applications, when there is no knowledge on homoscedasticity or a specific heteroscedasticity structure, some conclusions about these aspects might be drawn from the estimated linear part and nonparametric part. Similar (model) arguments can give insights into the other heteroscedasticity structures.

Table 1 summarizes the various error structures and the influence on the expectile surface $\tau_\omega(\mathbf{x}, \mathbf{z})$. Of importance is the impact of the heteroscedasticity structure on the linear and nonparametric part. Note for example that in the case of a heteroscedastic error structure $\sigma(\mathbf{X}_{(-1)}) \epsilon$, the nonparametric part of the expectile function depends on both covariates part, but that this is through an additive structure. If in real data applications such prior knowledge on the error structure is known, one should exploit this knowledge and use a bivariate-vector additive structure to estimate the nonparametric part $g(\mathbf{z}) + g_\omega(\mathbf{x}_{(-1)})$.

For simplicity of presentation, we focus in the sequel of the paper on the homoscedastic error setting in (2). However, the proposed statistical methodology is applicable to heteroscedastic error settings, as is demonstrated in Sections S2 and S3 of the Supplementary Material.

3 Statistical Estimation Methodology

Dealing with estimation in partially linear models is by now quite standard. Denoting $\tilde{\mathbf{X}}_i = \mathbf{X}_i - \mathbb{E}_{\mathbf{X}|\mathbf{Z}}[\mathbf{X}_i|\mathbf{Z} = \mathbf{Z}_i]$ and $\tilde{Y}_i = Y_i - \mathbb{E}_{Y|\mathbf{Z}}[Y_i|\mathbf{Z} = \mathbf{Z}_i]$, for $i = 1, \dots, n$, and using model (2), we first observe that

$$\begin{aligned} \mathbb{E}_{Y|\mathbf{Z}}[Y_i|\mathbf{Z} = \mathbf{Z}_i] &= \delta^T \mathbb{E}_{\mathbf{X}|\mathbf{Z}}[\mathbf{X}_i|\mathbf{Z} = \mathbf{Z}_i] + g(\mathbf{Z}_i) \\ \implies Y_i - \mathbb{E}_{Y|\mathbf{Z}}[Y_i|\mathbf{Z} = \mathbf{Z}_i] &= \delta^T (\mathbf{X}_i - \mathbb{E}_{\mathbf{X}|\mathbf{Z}}[\mathbf{X}_i|\mathbf{Z} = \mathbf{Z}_i]) + \sigma \epsilon_i \implies \tilde{Y}_i = \delta^T \tilde{\mathbf{X}}_i + \sigma \epsilon_i. \end{aligned}$$

Note that the first component of $\tilde{\mathbf{X}}_i$ equals zero, for all $i = 1, \dots, n$. In the sequel, we no longer write the subscripts in the conditional expectation, unless in case of possible confusion.

3.1 Estimation of the Vector of Regression Coefficients

Suppose first that we would know the conditional expectations $\mathbb{E}[\mathbf{X}|\mathbf{Z}]$ and $\mathbb{E}[Y|\mathbf{Z}]$. Then we could calculate $(\tilde{Y}_1, \tilde{\mathbf{X}}_1), \dots, (\tilde{Y}_n, \tilde{\mathbf{X}}_n)$ and based on the sample $(\tilde{Y}_1, \tilde{\mathbf{X}}_1, \mathbf{Z}_1), \dots, (\tilde{Y}_n, \tilde{\mathbf{X}}_n, \mathbf{Z}_n)$ we could, for given $\omega \in (0, 1)$, estimate δ_ω by minimizing

$$\sum_{i=1}^n Q_\omega(\tilde{Y}_i - \delta^T \tilde{\mathbf{X}}_i), \quad (6)$$

with respect to δ , where $Q_\omega(\cdot)$ is the expectile loss function in (1). The minimizer of (6) is the asymmetric least square estimator, studied by Newey and Powell (1987).

However, the conditional expectations $\mathbf{m}_\mathbf{X}(\mathbf{Z}) = \mathbb{E}[\mathbf{X}|\mathbf{Z}]$ and $m_Y(\mathbf{Z}) = \mathbb{E}[Y|\mathbf{Z}]$ are unknown, and hence need to be estimated. Note that $\mathbf{m}_\mathbf{X}(\mathbf{z})$ is a column vector of dimension $d + 1$ (with as first component 1), which will be estimated componentwise. We opt for using local linear techniques for estimating these two conditional expectations. For ease of presentation, we first discuss this when $q = 1$, i.e. $\mathbf{Z} = Z$ a univariate covariate, and i.i.d. observations Z_1, \dots, Z_n of Z are available. For a given point z , the local linear estimators for the j th component of $\mathbf{m}_\mathbf{X}(z)$ (denoted by $[\mathbf{m}_\mathbf{X}(z)]_j$) and $m_Y(z)$ are, respectively,

$$[\hat{\mathbf{m}}_\mathbf{X}(z)]_j = \sum_{i=1}^n K_{h_1}(Z_i - z) \frac{S_{n,2,h_1} - S_{n,1,h_1}(Z_i - z)}{S_{n,2,h_1} S_{n,0,h_1} - S_{n,1,h_1}^2} X_{ij}$$

and

$$\hat{m}_Y(z) = \sum_{i=1}^n K_{h_2}(Z_i - z) \frac{S_{n,2,h_2} - S_{n,1,h_2}(Z_i - z)}{S_{n,2,h_2} S_{n,0,h_2} - S_{n,1,h_2}^2} Y_i$$

where $K_{h_j}(\cdot) = K(\cdot/h_j)/h_j$, for $j = 1, 2$, is a kernel function, S_{n,ℓ,h_j} is defined as $S_{n,\ell,h_j} = \sum_{i=1}^n K_{h_j}(Z_i - z)(Z_i - z)^\ell$ and $h_j > 0$, for $j = 1, 2$, are bandwidth parameters. For the bandwidths h_j , $j = 1, 2$, one can use any of the practical bandwidth selectors that have been developed for local polynomial fitting, such as the rule-of-thumb bandwidth selector as discussed in Fan and Gijbels (1996).

When \mathbf{Z} is multivariate (i.e. $q > 1$) and as long as q is not too large, and the curse of dimensionality is not an issue, we use the multivariate local linear estimator as discussed in Fan and Gijbels (1996, Sect. 7.8). The observations for \mathbf{Z} are denoted $\mathbf{Z}_1, \dots, \mathbf{Z}_n$, where $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iq})^T$. Let K_q be a q -variate probability density function, satisfying thus $\int K_q(\mathbf{u})d\mathbf{u} = 1$ and $\int \mathbf{u}K_q(\mathbf{u})d\mathbf{u} = 0$. Further we assume that K_q has compact support and that

$$\int u_i u_j K_q(\mathbf{u})d\mathbf{u} = \delta_{ij} \mu_2(K_q),$$

with $\mu_2(K_q) \geq 0$. So the variance–covariance matrix of K_q is $\mu_2(K_q)\mathbf{I}_q$ with \mathbf{I}_q the $q \times q$ identity matrix. Define

$$K_{q,\mathbf{B}}(\mathbf{u}) = \frac{1}{|\mathbf{B}|} K_q(\mathbf{B}^{-1}\mathbf{u}), \tag{7}$$

where \mathbf{B} is a nonsingular $q \times q$ matrix, the bandwidth matrix, and $|\mathbf{B}|$ denotes its determinant. With $\mathbf{z} = (z_1, \dots, z_q)^T \in \mathbb{R}^q$ given, we denote the vector of Y -observations, the $n \times (q + 1)$ design matrix and the $n \times n$ diagonal weight matrix as respectively

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{Z}_D = \begin{pmatrix} 1 & Z_{11} - z_1 & \dots & Z_{1q} - z_q \\ 1 & Z_{21} - z_1 & \dots & Z_{2q} - z_q \\ \vdots & \vdots & & \vdots \\ 1 & Z_{n1} - z_1 & \dots & Z_{nq} - z_q \end{pmatrix} \text{ and } \mathbf{V} = \text{diag}(K_{q,\mathbf{B}}(\mathbf{Z}_i - \mathbf{z})). \tag{8}$$

Further we denote by $\mathbf{e}_1 = (1, 0, \dots, 0)^T$ the $(q + 1)$ -dimensional column vector with 1 at the first position and zero at all other positions. Further the column of all observations regarding X_j (for $j = 1, \dots, d + 1$) is denoted as $\mathbf{X}_{[j]} = (X_{1j}, \dots, X_{nj})^T$. Obviously for $j = 1$, this is the n -dimensional vector of one's. The local linear estimator for $[\mathbf{m}_X(\mathbf{z})]_j$ and $m_Y(\mathbf{z})$ is then given by

$$[\widehat{\mathbf{m}}_X(\mathbf{z})]_j = \mathbf{e}_1^T (\mathbf{Z}_D^T \mathbf{V} \mathbf{Z}_D)^{-1} \mathbf{Z}_D^T \mathbf{V} \mathbf{X}_{[j]} \quad \text{and} \quad \widehat{m}_Y(\mathbf{z}) = \mathbf{e}_1^T (\mathbf{Z}_D^T \mathbf{V} \mathbf{Z}_D)^{-1} \mathbf{Z}_D^T \mathbf{V} \mathbf{Y},$$

where $j = 1, \dots, d + 1$. For simplicity, we took the same bandwidth matrix \mathbf{B} for estimating $\mathbf{m}_X(\mathbf{z})$ and $m_Y(\mathbf{z})$. For background information on choices of a bandwidth matrix \mathbf{B} , see Fan and Gijbels (1996, Sect. 7.8), Wand and Jones (1995), Duong and Hazelton (2005) and references therein.

With these two estimates, we then obtain $\widehat{\mathbf{X}}_i = \mathbf{X}_i - \widehat{\mathbf{m}}_{\mathbf{X}}(\mathbf{Z}_i)$ and $\widehat{Y}_i = Y_i - \widehat{m}_Y(\mathbf{Z}_i)$. The estimator for the regression coefficient vector δ_ω is then

$$\widehat{\delta}_\omega := \arg \min_{\delta} \sum_{i=1}^n Q_\omega(\widehat{Y}_i - \delta^T \widehat{\mathbf{X}}_i).$$

3.2 Estimation of the Nonparametric Part

From (2) and expression (4), it is clear that

$$Y - \delta_\omega^T \mathbf{X} = g(\mathbf{Z}) + \sigma(\epsilon - \tau_{\omega, \epsilon}).$$

Denoting this residual $Y - \delta_\omega^T \mathbf{X}$ from the linear part by Y^* , we obtain that

$$\tau_{\omega, Y^*}(\mathbf{z}) = g(\mathbf{z}). \quad (9)$$

If the vector δ_ω would be known, one could consider the observations $(Y_1^*, \dots, Y_n^*) = (Y_1 - \delta_\omega^T \mathbf{X}_1, \dots, Y_n - \delta_\omega^T \mathbf{X}_n)$ and apply usual smoothing techniques to these data to obtain a mean, quantile or expectile curve.

Since the vector δ_ω is unknown, but can be estimated as described in Sect. 3.1, we thus instead consider

$$Y_i^* = Y_i - \widehat{\delta}_\omega^T \mathbf{X}_i \quad i = 1, \dots, n, \quad (10)$$

and perform a nonparametric step to estimate $g(\cdot)$ keeping in mind (9).

Let's look at $q = 1$ and $\mathbf{Z} = Z$ first. Using local polynomial fitting of order p (with $p \geq 0$ an integer) would result into minimizing

$$\sum_{i=1}^n Q_\omega \left(Y_i - \widehat{\delta}_\omega^T \mathbf{X}_i - \sum_{j=0}^p \gamma_j (Z_i - z)^j \right) K_h(Z_i - z), \quad (11)$$

with respect to $\boldsymbol{\gamma} = (\gamma_0, \dots, \gamma_p)^T$. Herein K is a univariate kernel function and $h > 0$ a bandwidth. Denoting by $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0, \dots, \widehat{\gamma}_p)^T$ the solution to optimization problem (11), the local polynomial estimator for $g(z)$ is $\widehat{\gamma}_0$. Adam and Gijbels (2021) studied this estimator, establishing its asymptotic properties and discussing in detail theoretical and practical choices for the bandwidth h . See also Sect. 4.

If $q > 1$ and we have a random vector \mathbf{Z} , we restrict this presentation to local linear fitting (i.e. $p = 1$ in the above paragraph). We now rely on approximating a q -variate function locally by applying a multivariate Taylor expansion of order 1 (i.e. up to first-order derivative terms). Due to the local modelling, we need to consider a multivariate function $K : \mathbb{R}^q \rightarrow \mathbb{R}$ (possibly different from K_q in Sect. 3.2) and

a bandwidth matrix \mathbf{H} . We assume that this is a symmetric and positive definite matrix. For the multivariate kernel function K , we make similar assumptions as for K_q , and we define its rescaled version $K_{\mathbf{H}}(\cdot) = (|\mathbf{H}|)^{-1} K(\mathbf{H}^{-1}\cdot)$ similarly as in (7). All together this leads to the optimization problem: minimize

$$\sum_{i=1}^n Q_{\omega} \left(Y_i^* - \beta_0 - \sum_{j=1}^q \beta_j (Z_{ij} - z_j) \right) K_{\mathbf{H}}(\mathbf{Z}_i - \mathbf{z}), \quad (12)$$

with respect to $\boldsymbol{\beta} = (\beta_0, \dots, \beta_q)^T$ where, for given ω , $\beta_0 = \tau_{\omega, Y^*}(\mathbf{z})$ and, $\beta_j = \frac{\partial \tau_{\omega, Y^*}(\mathbf{z})}{\partial z_j}$ for $j = 1, \dots, q$. Denote by $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \dots, \widehat{\beta}_q)^T$ the minimizer of (12).

It is convenient to introduce some further vector/matrix notations: denote

$$\mathbf{Y}^* = \begin{pmatrix} Y_1^* \\ Y_2^* \\ \vdots \\ Y_n^* \end{pmatrix} \text{ and } \mathbf{W} = \text{diag}(r_1(\omega)K_H(\mathbf{Z}_1 - \mathbf{z}), \dots, r_n(\omega)K_H(\mathbf{Z}_n - \mathbf{z})),$$

where \mathbf{W} is a diagonal matrix with as i th element

$$r_i(\omega) = \begin{cases} 1 - \omega & \text{if } Y_i^* \leq \beta_0 + \sum_{j=1}^q \beta_j (Z_{ij} - z_j) \\ \omega & \text{if } Y_i^* > \beta_0 + \sum_{j=1}^q \beta_j (Z_{ij} - z_j). \end{cases}$$

Note that these weight factors are induced by the form of the function $Q_{\omega}(y)$ in (1), which can be rewritten as $Q_{\omega}(y) = (1 - \omega)\mathbb{1}\{y \leq 0\}y^2 + \omega\mathbb{1}\{y > 0\}y^2$. Recalling also notations (8), the minimization problem in (12) can be rewritten in matrix form as

$$\underset{\boldsymbol{\beta}}{\text{minimize}} (\mathbf{Y}^* - \mathbf{Z}_D \boldsymbol{\beta})^T \mathbf{W} (\mathbf{Y}^* - \mathbf{Z}_D \boldsymbol{\beta}). \quad (13)$$

The complication in the setting of expectile estimation is that the diagonal weight matrix \mathbf{W} depends on the unknown vector $\boldsymbol{\beta}$ through the weight factors $r_i(\omega)$. As a consequence, the optimization problem (13) needs to be solved via an iterative procedure. See, for example, Yao and Tong (1996) and Adam and Gijbels (2021) for details on this iterative procedure. As a starting point for the iterative procedure, we use the vector of least squares regression estimators. After convergence of the iterative procedure, we obtain the local linear estimator for $\boldsymbol{\beta}$ denoted by $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \dots, \widehat{\beta}_q)^T$. The local linear estimator for $\tau_{\omega, Y^*}(\mathbf{z})$, denoted by $\widehat{\tau}_{\omega, Y^*}(\mathbf{z})$, is then $\widehat{\beta}_0$ and has the form

$$\widehat{\beta}_0 = \mathbf{e}_1^T \widehat{\boldsymbol{\beta}} = \mathbf{e}_1^T (\mathbf{Z}_D^T \mathbf{W} \mathbf{Z}_D)^{-1} \mathbf{Z}_D^T \mathbf{W} \mathbf{Y}^*. \quad (14)$$

This estimation procedure involves the choice of the bandwidth matrix \mathbf{H} . For mean regression, there are quite some papers that study how to optimally choose

a bandwidth matrix/value. This issue is far less studied in expectile regression. In Sect. 4, we derive an optimal choice for the bandwidth matrix \mathbf{H} and discuss some practical bandwidth selectors.

4 Asymptotic Properties and Bandwidth Selection

A detailed study of multivariate local linear fitting in mean regression can be found in Ruppert and Wand (1994). See also Cheng and Peng (2006). From the asymptotic bias and variance expressions in mean regression on the one hand, and in expectile regression using local polynomial fitting on the other hand (see Adam and Gijbels 2021), one can derive approximate expressions for the asymptotic bias and variance of the local linear estimator for $g(\cdot)$ in (14). These hold under standard assumptions which for brevity are not listed here (see e.g. Adam and Gijbels 2021). Before stating the expressions, we first introduce some notations:

$$\nu_0(K) = \int K^2(\mathbf{u})d\mathbf{u} \quad \text{and}$$

$$\mathbf{M}_2(\mathbf{z}) = \begin{pmatrix} \frac{\partial^2}{\partial z_1 \partial z_1} \tau_{\omega, Y^*}(\mathbf{z}) & \dots & \frac{\partial^2}{\partial z_1 \partial z_q} \tau_{\omega, Y^*}(\mathbf{z}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial z_q \partial z_1} \tau_{\omega, Y^*}(\mathbf{z}) & \dots & \frac{\partial^2}{\partial z_q \partial z_q} \tau_{\omega, Y^*}(\mathbf{z}) \end{pmatrix} = \begin{pmatrix} \frac{\partial^2}{\partial z_1 \partial z_1} g(\mathbf{z}) & \dots & \frac{\partial^2}{\partial z_1 \partial z_q} g(\mathbf{z}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial z_q \partial z_1} g(\mathbf{z}) & \dots & \frac{\partial^2}{\partial z_q \partial z_q} g(\mathbf{z}) \end{pmatrix}.$$

For a point \mathbf{z} in the interior of the support of $f_{\mathbf{Z}}$, the q -variate marginal density function of \mathbf{Z} , the main terms in the asymptotic conditional bias and the asymptotic conditional variance of $\widehat{\tau}_{\omega, Y^*}(\mathbf{z})$ are

$$\text{ABias}[\widehat{\tau}_{\omega, Y^*}(\mathbf{z})|\mathbf{Z}_1, \dots, \mathbf{Z}_n] = \frac{1}{2} \mu_2(K) \text{trace}(\mathbf{M}_2(\mathbf{z})\mathbf{H}\mathbf{H}^T)$$

$$\text{AVar}[\widehat{\tau}_{\omega, Y^*}(\mathbf{z})|\mathbf{Z}_1, \dots, \mathbf{Z}_n] = \frac{\nu_0(K)}{n|\mathbf{H}|} \frac{1}{\gamma^2(\omega, \mathbf{z})f_{\mathbf{Z}}(\mathbf{z})} \int (2L_{\omega}(y - \tau_{\omega, Y^*}(\mathbf{z})))^2 f_{Y^*|\mathbf{Z}}(y|\mathbf{z})dy$$

where $L_{\omega}(y) = |\omega - \mathbb{1}\{y \leq 0\}|y$, is, up to a factor 2, the first derivative of the expectile loss function $Q_{\omega}(\cdot)$, and

$$\gamma(\omega, \mathbf{z}) = 2(1 - \omega)\mathbb{P}\{Y^* \leq \tau_{\omega, Y^*}(\mathbf{Z})|\mathbf{Z} = \mathbf{z}\} + 2\omega\mathbb{P}\{Y^* > \tau_{\omega, Y^*}(\mathbf{Z})|\mathbf{Z} = \mathbf{z}\}. \quad (15)$$

4.1 Optimal Theoretical Bandwidth (Matrix)

With the aim to get to an optimal bandwidth (matrix), we first compute the Approximate Mean Squared Error (AMSE) which is defined as

$$\begin{aligned} \text{AMSE}(\widehat{\tau}_{\omega, Y^*}(\mathbf{z})) &= \text{ABias}^2 + \text{AVariance} \\ &= \frac{1}{4} \mu_2^2(K) \text{trace}^2(\mathbf{M}_2(\mathbf{z})\mathbf{H}\mathbf{H}^T) + \frac{\nu_0(K)}{n|\mathbf{H}|} \tilde{U}(\mathbf{z}), \end{aligned} \quad (16)$$

where we introduced the shorthand notation

$$\tilde{U}(\mathbf{z}) = \frac{1}{\gamma^2(\omega, \mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})} \left\{ \int (2L_{\omega}(y) - \tau_{\omega, Y^*}(\mathbf{z}))^2 f_{Y^*|\mathbf{Z}}(y|\mathbf{z}) dy \right\}, \quad (17)$$

a quantity containing unknown model elements.

From expression (16), one can proceed similarly as in Fan et al. (1997) to get to an expression for $\mathbf{H}\mathbf{H}^T$ that would constitute a minimum for the AMSE. Taking the derivative of (16) with respect to $\mathbf{H}\mathbf{H}^T$ (see Rao 1973, p. 72) and putting this derivative equal to 0, we obtain

$$\mu_2^2(K) \text{trace}(\mathbf{M}_2(\mathbf{z})\mathbf{H}\mathbf{H}^T)\mathbf{M}_2(\mathbf{z}) - \frac{\nu_0(K)}{n|\mathbf{H}|} (\mathbf{H}\mathbf{H}^T)^{-1} \tilde{U}(\mathbf{z}) = 0. \quad (18)$$

Suppose that the matrix of second-order partial derivative $\mathbf{M}_2(\mathbf{z})$ is positive or negative definite. Then we show in Section S1 of the Supplementary Material that the unique solution to (18) is

$$\mathbf{H}\mathbf{H}^T = \left(\frac{\nu_0(K)}{\mu_2^2(K)n} \tilde{U}(\mathbf{z}) |\mathbf{M}_2^*(\mathbf{z})|^{1/2} \right)^{2/(q+4)} (\mathbf{M}_2^*(\mathbf{z}))^{-1}, \quad (19)$$

with

$$\mathbf{M}_2^*(\mathbf{z}) = \begin{cases} \mathbf{M}_2(\mathbf{z}) & \text{for positive definite } \mathbf{M}_2(\mathbf{z}) \\ -\mathbf{M}_2(\mathbf{z}) & \text{for negative definite } \mathbf{M}_2(\mathbf{z}). \end{cases}$$

Any matrix \mathbf{H} that satisfies Eq. (19) is an optimal *local* bandwidth matrix.

One next can consider an Approximate weighted Mean Integrated Square Error (AMISE), obtained by integrating the AMSE using a weight factor, i.e.

$$\begin{aligned} \text{AMISE}(\widehat{\tau}_{\omega, Y^*}(\cdot)) \\ = \frac{1}{4} \mu_2^2(K) \int \text{trace}^2(\mathbf{M}_2(\mathbf{z})\mathbf{H}\mathbf{H}^T) k(\mathbf{z}) d\mathbf{z} + \frac{\nu_0(K)}{n|\mathbf{H}|} \int \tilde{U}(\mathbf{z}) k(\mathbf{z}) d\mathbf{z}, \end{aligned} \quad (20)$$

where $k(\cdot) \geq 0$ is a weight function. An optimal *global* bandwidth matrix \mathbf{H} would be a matrix which minimizes (20). Deriving something in general for this setting is quite involved, and not further elaborated here.

With the aim to come to some practical and simple rule for bandwidth selection, we restrict in the sequel to the setting where $\mathbf{H} = h \mathbf{I}_q$, with $h > 0$. In that case $|\mathbf{H}| = h^q |\mathbf{I}_q| = h^q$ and moreover $\mathbf{H}\mathbf{H}^\top = h^2 \mathbf{I}_q$, and the expression for AMISE reduces to

$$\begin{aligned} \text{AMISE}(\widehat{\tau}_{\omega, Y^*}(\cdot)) \\ = \frac{1}{4} \mu_2^2(K) h^4 \int \text{trace}^2(\mathbf{M}_2(\mathbf{z})) k(\mathbf{z}) d\mathbf{z} + \frac{v_0(K)}{n h^q} \int \widetilde{U}(\mathbf{z}) k(\mathbf{z}) d\mathbf{z}. \end{aligned} \quad (21)$$

An optimal value for the constant bandwidth parameter h is then easily found by minimizing (21) with respect to h , which leads to

$$h_{\text{opt}} = \left(\frac{v_0(K)q}{\mu_2^2(K)} \frac{\int \widetilde{U}(\mathbf{z}) k(\mathbf{z}) d\mathbf{z}}{\int \text{trace}(\mathbf{M}_2(\mathbf{z}))^2 k(\mathbf{z}) d\mathbf{z}} \right)^{1/(q+4)} n^{-1/(q+4)}. \quad (22)$$

4.2 Rule-of-Thumb (ROT) Bandwidth Selector

Looking at the expression of $\widetilde{U}(\mathbf{z})$ in (17), it is easily seen that a further simplification is obtained with taking the weight function $k(\mathbf{z}) = k_0(\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})$ with $k_0(\cdot) \geq 0$ a chosen weight function. From (22), we get

$$h_{\text{opt}} = \left(\frac{v_0(K)q}{\mu_2^2(K)} \frac{\int U(\mathbf{z}) k_0(\mathbf{z}) d\mathbf{z}}{\int \text{trace}(\mathbf{M}_2(\mathbf{z}))^2 f_{\mathbf{Z}}(\mathbf{z}) k_0(\mathbf{z}) d\mathbf{z}} \right)^{1/(q+4)} n^{-1/(q+4)}, \quad (23)$$

with now $U(\mathbf{z}) = \widetilde{U}(\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z}) = (\gamma^2(\omega, \mathbf{z}))^{-1} \{ \int (2L_\omega(y - \tau_{\omega, Y^*}(\mathbf{z})))^2 f_{Y^*|\mathbf{Z}}(y|\mathbf{z}) dy \}$.

Note first of all that

$$U(\mathbf{z}) = (\gamma^2(\omega, \mathbf{z}))^{-1} E_{Y^*|\mathbf{Z}} \left[(2L_\omega(Y^* - \tau_{\omega, Y^*}(\mathbf{z})))^2 \mid \mathbf{Z} = \mathbf{z} \right] \quad (24)$$

$$\int \text{trace}(\mathbf{M}_2(\mathbf{z}))^2 f_{\mathbf{Z}}(\mathbf{z}) k_0(\mathbf{z}) d\mathbf{z} = E_{\mathbf{Z}} \left[\text{trace}(\mathbf{M}_2(\mathbf{Z}))^2 k_0(\mathbf{Z}) \right]. \quad (25)$$

Inspecting (23), (24) and (25), we see that we are concerned essentially with the unknown quantities: $\tau_{\omega, Y^*}(\mathbf{z})$, the ω th expectile curve of Y^* given $\mathbf{Z} = \mathbf{z}$; the matrix of second-order partial derivatives of this function with respect to \mathbf{z} and the probability expression $\gamma(\omega, \mathbf{z})$ in (15) which also involves $\tau_{\omega, Y^*}(\mathbf{z})$. Since all these quantities require to know $\tau_{\omega, Y^*}(\mathbf{z})$, a first step in our proposed rule-of-thumb data-driven bandwidth procedure consists of fitting *globally* a polynomial function to the pseudo observations $(Y_1^*, \mathbf{Z}_1), \dots, (Y_n^*, \mathbf{Z}_n)$, where Y_i^* , $i = 1, \dots, n$, is as in (10). We then

employ this global parametric fit to get rough approximations of the unknown quantities in (15), (24) and (25), appearing in the expression for the optimal bandwidth in (23).

More precisely, the rule-of-thumb (ROT) bandwidth selector is obtained via the following procedure.

- Fit globally a parametric polynomial of order 5 to the pseudo observations $(Y_1^*, \mathbf{Z}_1), \dots, (Y_n^*, \mathbf{Z}_n)$, using the expectile loss function $Q_\omega(\cdot)$, and obtain the fit

$$\tau_\omega(\mathbf{z}) = \alpha_0 + \alpha_1^\top \mathbf{z} + \dots + \alpha_5^\top \mathbf{z}^5$$

where $\alpha_\ell = (\alpha_{\ell 1}, \dots, \alpha_{\ell q})^\top$, with $\mathbf{z}^\ell = (z_1^\ell, \dots, z_q^\ell)^\top$, for $\ell = 1, \dots, 5$.

- This leads to the following “rough” approximations of the unknown quantities. We approximate the quantity $E_{Y^*|\mathbf{Z}}[(2L_\omega(Y^* - \tau_\omega(\mathbf{z})))^2 | \mathbf{Z} = \mathbf{z}]$ in (24) roughly by

$$\frac{4}{n} \sum_{i=1}^n L_\omega^2(Y_i^* - \tau_\omega(\mathbf{Z}_i)),$$

the probability in (15) by

$$2(1 - \omega) \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{Y_i^* \leq \tau_\omega(\mathbf{Z}_i)\} + 2\omega \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{Y_i^* > \tau_\omega(\mathbf{Z}_i)\}$$

and the expectation in (25) by

$$\frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^q \left(\frac{\partial^2 \tau_\omega(\mathbf{z})}{\partial z_j^2} \Big|_{\mathbf{z}=\mathbf{Z}_i} \right) \right\}^2 k_0(\mathbf{Z}_i).$$

With these rough approximations, we calculate the rule-of-thumb (ROT) bandwidth

$$\begin{aligned} \check{h}_{\text{opt}} &= \left(\frac{v_0(K)q}{n\mu_2^2(K)} \right)^{1/(q+4)} \\ &\times \left(\frac{\frac{4}{n} \sum_{i=1}^n L_\omega^2(Y_i^* - \tau_\omega(\mathbf{Z}_i))}{4((1-\omega)\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{Y_i^* \leq \tau_\omega(\mathbf{Z}_i)\} + \omega\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{Y_i^* > \tau_\omega(\mathbf{Z}_i)\})^2} \int k_0(\mathbf{z}) d\mathbf{z}}{\frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^q \left(\frac{\partial^2 \tau_\omega(\mathbf{z})}{\partial z_j^2} \Big|_{\mathbf{z}=\mathbf{Z}_i} \right) \right)^2 k_0(\mathbf{Z}_i)} \right)^{1/(q+4)}. \end{aligned} \quad (26)$$

There are other approaches possible to obtain data-driven bandwidth selection procedures. In a fully nonparametric setting, and in case $q = 1$, several of these approaches have been discussed in detail by Adam and Gijbels (2021), and their

performances investigated via a simulation study. In the setting of that paper, a similar rule-of-thumb bandwidth selector showed a very good overall performance, which is the main reason why we focus on a similar approach in this paper. Another valid approach would also be to exploit the location-scale model in combination with the relationships that exist between expectiles and quantiles. We refer to the above paper for a discussion on other data-driven bandwidth procedures.

5 Simulation Study

In this section, we provide a small simulation study to investigate the local polynomial estimation method of Sect. 3 to estimate the ω th expectile function in (4). We investigate the finite-sample quality of the estimators for the parameter vector δ_ω , for the nonparametric part $g(\cdot)$ as well as for the target quantity $\tau_{\omega,Y}(\cdot, \cdot)$.

We consider two simulation models, with $\mathbf{X} = (X_1, X_2)^T$ (i.e. $d = 2$, no intercept term); one model with a univariate Z (i.e. $q = 1$) and a second model with a bivariate vector \mathbf{Z} (i.e. $q = 2$). These models were inspired by those considered in Zhu et al. (2013). The different model elements are given in Table 2.

Note that in these models, \mathbf{X} and \mathbf{Z} are dependent. The error term in Model 1 has a normal distribution, whereas the error term in Model 2 has a Student-t distribution with three degrees of freedom, and hence a heavier tail.

Models 1 and 2 have a homoscedastic error structure. In Section S3 of the Supplementary Material, one can find simulation results for models with various heteroscedastic error structures.

For each model, we simulated 100 samples of size n (with $n = 100$ for Model 1, and $n = 200$ for Model 2 involving the heavier-tailed Student-t error distribution). For each sample, we calculate the estimates for (i) the regression parameter vector $\delta_\omega = (\tau_{\omega,\epsilon}, \delta^T)^T = (\tau_{\omega,\epsilon}, \delta_1, \delta_2)^T$; (ii) the functional part $g(\cdot)$ and finally for (iii) the ω th expectile curve $\tau_\omega(\mathbf{x}, \mathbf{z})$ in (4).

Table 2 The two simulation models

Model	Location-scale model	Details
1	$Y = \delta^T \mathbf{X} + 10 \sin(0.9Z) + \epsilon$	$Z \sim \mathcal{U}[-3, 3]$, $\delta = (0.8, -0.8)^T$ $X_j = 0.9Z + 1.5\eta_j$ for $j = 1, 2$, $\eta_j \sim \mathcal{N}(0, j)$, $\epsilon \sim \mathcal{N}(0, 5^2)$
2	$Y = \delta^T \mathbf{X} + 0.2 \exp(1.5\mathbf{y}^T \mathbf{Z}) + \epsilon$	$\delta = (0, -0.8)^T$ $Z_1, Z_2 \sim \mathcal{U}[0, 1]$, Z_1 and Z_2 independent $X_j = \mathbf{y}^T \mathbf{Z} + 1.5\eta_j$ for $j = 1, 2$, $\mathbf{y} = (3, -0.4)^T$ $\eta_j \sim \mathcal{N}(0, j)$, $\epsilon \sim \text{Student-t}(3)$

To compute the local linear estimates for the conditional expectations $E[\mathbf{X}|\mathbf{Z} = \mathbf{z}]$ and $E[Y|\mathbf{Z} = \mathbf{z}]$, we use the R package `locpol` or, since this can only deal with univariate covariates, the `lsfit` command in R. Herein we use a Gaussian kernel and the rule-of-thumb bandwidth procedure developed by Fan and Gijbels (1996), applying the command `pluginBw` in the package `locpol` (see Cabrera 2018). For estimation of the regression parameter vector δ_ω , we use the R package `expectreg` (see Otto-Sobotka et al. 2019), and in particular the command `expectreg.ls`, to compute the linear expectile part.

When estimating the function $g(\cdot)$, we solve optimization problem (13) for an equispaced grid of values of \mathbf{z} . In the multivariate case, we use a product kernel, based on univariate Gaussian kernels (for both parts of the estimation procedure, in Sects. 3.1 and 3.2). For Model 1, we take 200 equispaced grid-values, denoted by $\{z_1, \dots, z_{200}\}$ on the domain $[-3, 3]$ of the variable Z . For Model 2, we take a grid of 200 equispaced points in each dimension, on the domain of each component of \mathbf{Z} . For estimating the nonparametric part, we use the ROT bandwidth selector \check{h}_{opt} in (26), as described in Sect. 4.2. For the weight function $k_0(\mathbf{z})$, we take $k_0(z) = \mathbb{1}\{-2.9 \leq z \leq 2.9\}$ in Model 1, and in Model 2 $k_0(z_1, z_2) = \mathbb{1}\{0.1 \leq z_1 \leq 0.9\} \cdot \mathbb{1}\{0.1 \leq z_2 \leq 0.9\}$, for which $\int k_0(\mathbf{z})d\mathbf{z} = 0.8^2$.

For summarizing the results with respect to the estimation of the nonparametric part $g(\cdot)$, we proceed as follows. For each sample, we calculate the estimator $\hat{g}(\cdot)$ over the fixed grid of points and compute the Approximate Integrated Square Error (AISE)

$$AISE = \frac{1}{N_{grid}} \sum_{j=1}^{N_{grid}} (\hat{g}(\mathbf{z}_j) - g(\mathbf{z}_j))^2, \tag{27}$$

where N_{grid} denotes the number of grid points. After ordering these 100 AISE values, we obtain the 0.05th, 0.50th and 0.95th percentile value and depict the corresponding estimates $\hat{g}(\cdot)$ as representative estimates among the 100 estimated curves. The scatterplot that is shown in the concerned plots is that of the sample with the median performance (0.50th percentile among AISE values).

5.1 Simulation Results for Model 1

In Fig. 1, the true expectile function $\tau_\omega(\mathbf{x}, \mathbf{z})$ is depicted, where on one of coordinate axes we present $\delta^T \mathbf{x}$ and on the other coordinate \mathbf{z} . The surface $\tau_\omega(\mathbf{x}, \mathbf{z})$ is shown for five values of ω : $\omega = 0.1, 0.3, 0.5, 0.7, 0.9$, where the surface for the smallest (highest) value of ω is the lowest (highest) situated surface. Since we have a location-scale model, the five expectile surfaces are parallel. One can clearly see the sinusoidal type of influence of the variable Z .

Boxplots of the 100 estimated values for each of the components of $\delta_\omega = (\tau_{\omega,\epsilon}, \delta_1, \delta_2)^T$ are presented in Fig. 2, and this for each of the five ω values. The horizontal (red) dotted line indicated the true values of the parameter components,

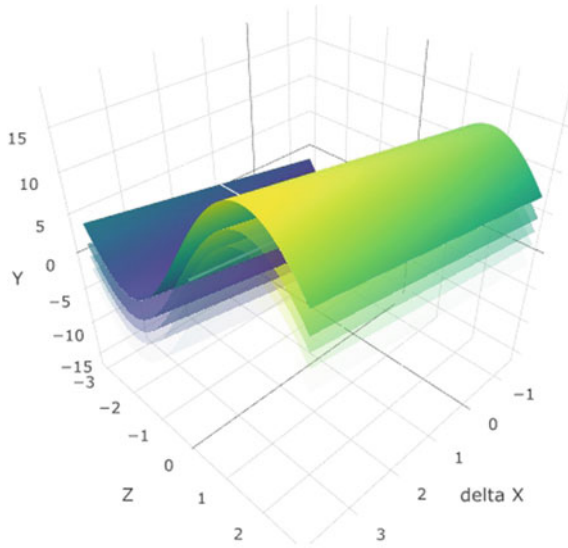


Fig. 1 Model 1. 3D surface plot showing the true expectile curves. The expectile surfaces for the subsequent values $\omega = 0.1, 0.3, 0.5, 0.7, 0.9$ are the lowest to highest situated surfaces

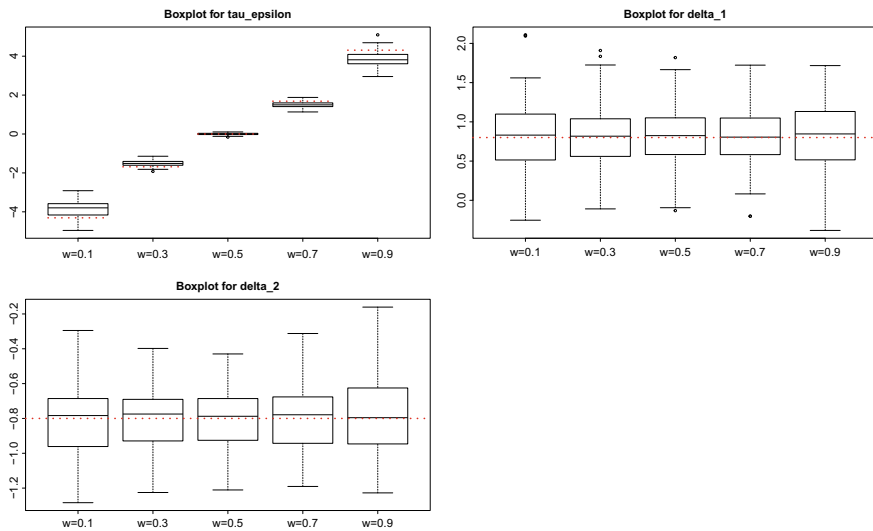


Fig. 2 Model 1. Boxplots of the estimates $\widehat{\tau}_{\omega,\epsilon}$, $\widehat{\delta}_1$ and $\widehat{\delta}_2$ for $\omega = 0.1, 0.3, 0.5, 0.7, 0.9$. The dotted horizontal lines indicate the true values

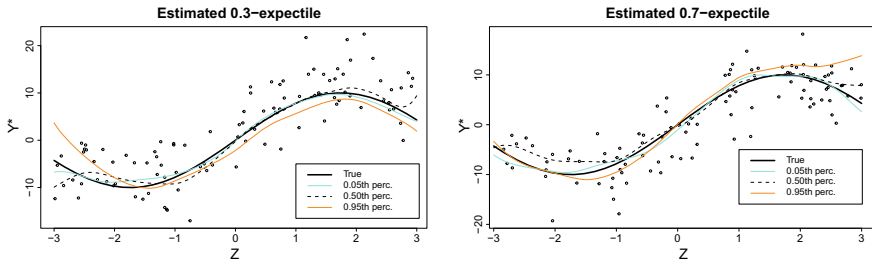


Fig. 3 Model 1. True expectile curve $\tau_{0.3,Y^*}(\cdot)$ (left) and $\tau_{0.7,Y^*}(\cdot)$ (right) in black and three representative local linear estimates: 0.05th AISE-percentile (light-grey; colour blue), 0.5th AISE-percentile (dashed line) and 0.95th AISE-percentile (grey; colour ochre yellow)

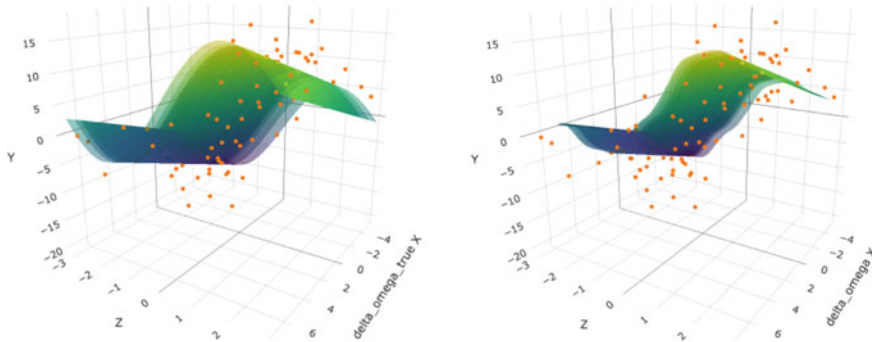


Fig. 4 Model 1. 3D surface plots showing the true 0.7th expectile curve (left) and the estimated 0.7th expectile curve (right)

i.e. $\tau_{\omega,\epsilon}, \delta_1 = 0.8$ and $\delta_2 = -0.8$. All estimators are of good quality, although estimation in case of small or large values of ω typically is a bit more difficult.

Figure 3 presents the representative estimates of g when $\omega = 0.3$ (left) and $\omega = 0.7$ (right). Obviously $g(\cdot)$ does not depend on ω , but the estimation of g is influenced by the quality of the estimation of the parametric part (that depends on ω). Figure 3 gives a graphical idea about the quality of the estimator by presenting the true expectile curve $\tau_{\omega,Y^*}(\cdot) = g(\cdot)$ together with the three representative estimates for $\omega = 0.3$ and 0.7 .

Finally, Fig. 4 shows the true 0.7th expectile curve and the estimate with median performance (according to a corresponding AISE criterion). The estimated surfaces are relatively smooth, confirming the quality of the ROT bandwidth selector, and clearly reveal the sinusoidal behaviour of $g(\cdot)$.

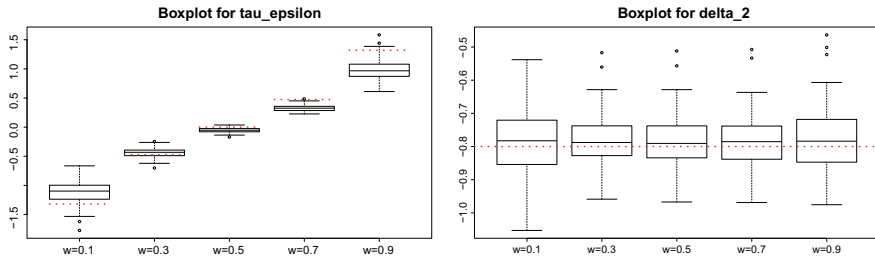


Fig. 5 Model 2. Boxplot of the estimates of $\widehat{\tau}_{\omega,\epsilon}$ and $\widehat{\delta}_2$ for $\omega = 0.1, 0.3, 0.5, 0.7, 0.9$. The dotted horizontal lines indicate the true values

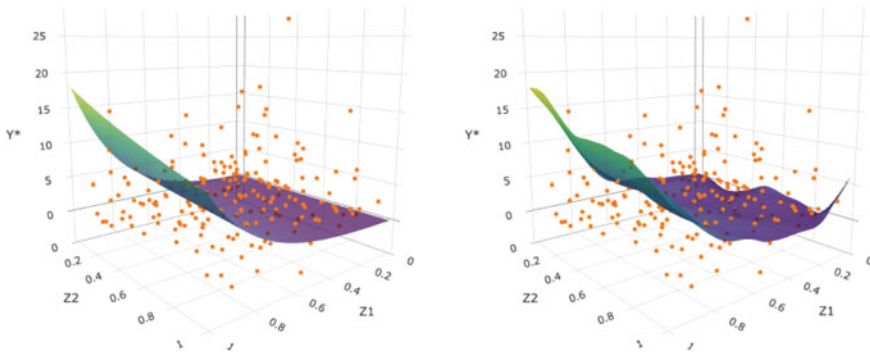


Fig. 6 Model 2. 3D surface plot showing the true $\tau_{0.3,Y^*}(\cdot) = g(\cdot)$ curve (left) and the estimated curve (right)

5.2 Simulation Results for Model 2

Figure 5 presents the boxplots of the estimates for $\tau_{\omega,\epsilon}$ and δ_2 (with true value -0.8). Note the presence of outlying values in the boxplots, related to the Student-t error, as opposed to the normal error structure in Model 1.

Figure 6 presents the true (left) and the estimated (right) $\tau_{0.3,Y^*}(\cdot) = g(\cdot)$ surfaces, where the scatterplot and the estimated surface are those corresponding to a median performance across simulations.

6 Real Data Application

The data we consider are measurements on air quality in New York in the period May 1, 2013 till September 30, 2013, downloaded from <https://globalweather.tamu.edu> and <https://www.epa.gov/outdoor-air-quality-data/download-daily-data>. There are 153 observations and the variables we consider here are *Ozone* concentration (in parts per million), solar radiation (*Solar.R*) (in MegaJoule per square metre),

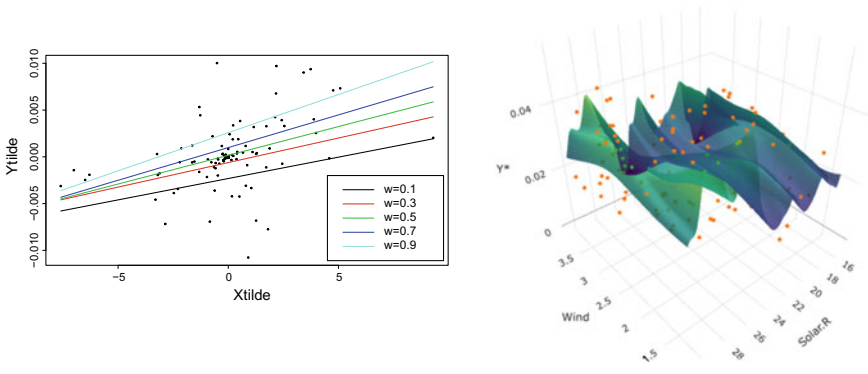


Fig. 7 Air quality data. Analysis $q = 2$. Left: $\hat{\delta}_\omega^T \tilde{\mathbf{X}}$ of the parametric part for ω values 0.1, 0.3, 0.5, 0.7, 0.9. Right: Estimated expectile regression surface $\hat{g}(\cdot, \cdot)$ for $\omega = 0.7$

Wind (in metres/second) and *Temp* (in degrees Celsius). The response variable Y is *Ozone* concentration. Due to too few observations in certain regions of the covariates domain, we restricted the dataset to observations for which *Solar.R* is larger or equal to 15 and *Wind* does not exceed 4. This led to a reduced dataset of 119 observations.

A scatterplot matrix of all bivariate scatterplots indicated that *Temp* appears to have a linear influence, whereas for *Wind* and *Solar.R*, the effect appears as possibly nonlinear. In a first instance, we consider *Temp* as the sole element in the set of covariates \mathbf{X} , with \mathbf{Z} (with $q = 2$) containing *Wind* and *Solar.R*.

Figure 7 (left) shows the estimate $\hat{\delta}_\omega^T \tilde{\mathbf{x}}$ for ω values 0.1, 0.3, 0.5, 0.7 and 0.9, i.e. the estimated parametric part. Note that the lines are not parallel which indicates that there is likely heteroscedasticity in the data. Table 3 (first block of rows) gives the estimated values for $\hat{\tau}_{\omega, \epsilon}$ and $\hat{\delta}_1$. Note the positive values for $\hat{\delta}_1$, with an average around 0.0006218, indicating a small positive effect of *Temp* on the expectile of *Ozone*. Figure 7 (right) depicts the estimate \hat{g} , for $\omega = 0.7$. The impact of *Solar.R* on the estimated conditional expectile of *Ozone* appears as nonlinear. The influence of *Wind* on the other hand seems not too far from linear. Therefore, in the next step of our analysis, we also include *Wind* in the linear part.

For this case of $q = 1$, Fig. 8 (left) shows the estimated parametric part $\hat{\delta}_\omega^T \tilde{\mathbf{x}}$ for $\omega = 0.3$ and 0.7. Note the non-parallel estimated surfaces. The estimated values $\hat{\delta}_\omega$ are given in the second block of rows in Table 3. Note the overall negative values for $\hat{\delta}_2$, indicating that more wind reduces the ozone concentration expectile. Figure 8 (right) shows the estimated expectile surface $\hat{\tau}_{0.7}(\cdot, \cdot)$, with coordinate axis $\hat{\delta}_\omega \mathbf{x}$ and z . In $\delta^T \mathbf{x}$, the pattern is linear and in Z , we can observe the increasing nonlinear trend.

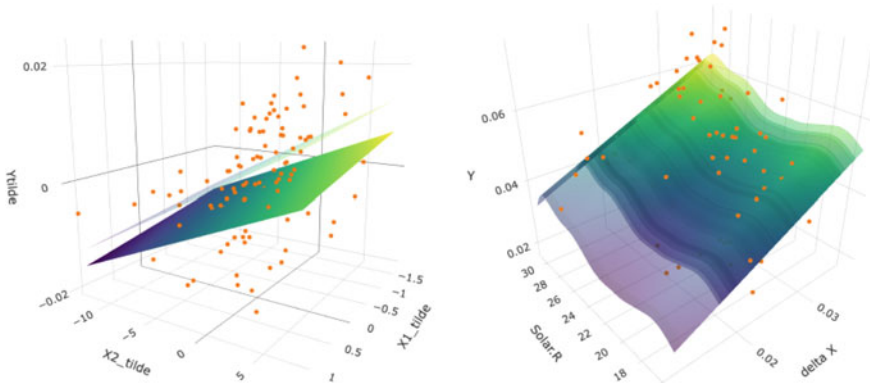


Fig. 8 Air quality data. Analysis $q = 1$. Left: 3D surface plot showing the estimates $\widehat{\delta}_\omega^T \widetilde{\mathbf{X}}$ of the parametric part for values of ω equal to 0.3 and 0.7. Right: 3D surface plot showing the estimate $\widehat{\tau}_{0.7}(\cdot, \cdot)$ surface, with on the coordinate axis $\widehat{\delta}_\omega^T \widetilde{\mathbf{X}}$ and *Solar.R*

Table 3 Values of the two (for $q = 2$) or three (for $q = 1$) estimators $\widehat{\tau}_{\omega, \epsilon}$, $\widehat{\delta}_1$ and $\widehat{\delta}_2$ for $\omega = 0.1, 0.3, 0.5, 0.7, 0.9$. These estimated parameters constitute the estimated linear part, revealing the estimated linear influence of temperature ($q = 2$) and temperature and wind ($q = 1$) in the expectile function

Dimension	ω	0.1	0.3	0.5	0.7	0.9
$q = 2$	$\widehat{\tau}_{\omega, \epsilon}$	-0.0023181	-0.0006114	0.0001647	0.0009888	0.0025815
	$\widehat{\delta}_1$	0.0004568	0.0005246	0.0006137	0.0006986	0.0008154
$q = 1$	$\widehat{\tau}_{\omega, \epsilon}$	-0.0070013	-0.0027328	0.0000191	0.0027633	0.0067605
	$\widehat{\delta}_1$	0.0005939	0.0008373	0.0009688	0.0010814	0.0012171
	$\widehat{\delta}_2$	-0.0033660	-0.0022400	-0.0017538	-0.0014734	-0.0013471

7 Further Reading

In applications, one might have the natural restriction that the curve of interest has some qualitative properties, such as being monotone (increasing or decreasing), being convex or concave. There is a vast literature on estimation under shape constraints. In mean and quantile regression, for example, the assumption of monotonicity of the mean regression or the quantile regression curves is often quite justifiable. See Mammen and Thomas-Agnan (1999), Gijbels (2006), Poiraud-Casenova and Thomas-Agnan (2000) and Groeneboom and Jongbloed (2014), among others.

Expectiles are of particular interest in risk measures, in particular since it was argued by Ziegel (2016) that expectiles lead to the only coherent and elicitable law-invariant risk measure. Among the recent contributions in this area is the paper by Daouia et al. (2020).

Expectile regression when the response Y is multivariate is a real challenge. Some recent work on this is Herrmann et al. (2018) and Daouia and Paindaveine (2019), among others.

Acknowledgements The authors gratefully acknowledge the support of Research Grant FWO G0D6619N from the Flemish Science Foundation and of projects GOA/12/014 and C16/20/002 from the Research Fund KU Leuven.

References

- Adam, C., & Gijbels, I. (2021). Local polynomial expectile regression. *The Annals of the Institute of Statistical Mathematics*. <https://doi.org/10.1007/s10463-021-00799-y>.
- Aigner, D. J., Amemiya, T., & Poirier, D. J. (1976). On the estimation of production frontiers: Maximum likelihood estimation of the parameters of a discontinuous density function. *International Economic Review*, 17(2), 377–396.
- Cabrera, J. L. O. (2018). *Kernel local polynomial regression*. R package version 0.7-0.
- Chen, T., Su, Z., Yang, Y., & Ding, S. (2020). Efficient estimation in expectile regression using envelope models. *Electronic Journal of Statistics*, 14, 143–173.
- Cheng, M.-Y., & Peng, L. (2006). Simple and efficient improvements of multivariate local linear regression. *Journal of Multivariate Analysis*, 97(7), 1501–1524.
- Daouia, A., Girard, S., & Stupfler, G. (2020). Tail expectile process and risk assessment. *Bernoulli*, 26(1), 531–556.
- Daouia, A., & Paindaveine, D. (2019). From halspace M-depth to multiple-output expectile regression. [arXiv:1905.12718v1](https://arxiv.org/abs/1905.12718).
- Duong, T., & Hazelton, M. L. (2005). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis*, 93, 417–433.
- Fan, J., Gasser, T., Gijbels, I., Brockmann, M., & Engel, J. (1997). Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics*, 49(1), 79–99.
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. Number 66 in Monographs on statistics and applied probability series. London: Chapman & Hall.
- Gijbels, I. (2006). Monotone regression. In S. Kotz, N. L. Johnson, C. B. Read, N. Balakrishnan, & B. Vidakovic (Eds.), *Encyclopedia of statistical sciences* (pp. 4951–4968). New York: Wiley.
- Groeneboom, P., & Jongbloed, G. (2014). *Nonparametric estimation under shape constraints*. Cambridge: Cambridge University Press.
- Gu, Y., & Zou, H. (2019). Aggregate expectile regression by exponential weighting. *Statistica Sinica*, 29(2), 671–692.
- Herrmann, K., Hofert, M., & Mailhot, M. (2018). Multivariate geometric expectiles. *Scandinavian Actuarial Journal*, 7, 629–659.
- Liao, L., Park, C., & Choi, H. (2019). Penalized expectile regression: an alternative to penalized quantile regression. *Annals of the Institute of Statistical Mathematics*, 71, 409–438.
- Mammen, E., & Thomas-Agnan, C. (1999). Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics*, 26(2), 239–252.
- Newey, W., & Powell, J. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55(4), 819–847.
- Otto-Sobotka, F., Spiegel, E., Schnabel, S., Waltrup, L. S., Eilers, P. (contrib.), Kneib, T. (contrib.), & Kauermann, G. (contrib.). (2019). Expectile and Quantile Regression. *R package version*, 50.
- Poiraud-Casenova, S., & Thomas-Agnan, C. (2000). About monotone regression quantiles. *Statistics & Probability Letters*, 48, 101–104.

- Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.
- Remillard, B., & Abdous, B. (1995). Relating quantiles and expectiles under weighted-symmetry. *Annals of the Institute of Statistical Mathematics*, 47, 371–384.
- Ruppert, D., & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22(3), 1346–1370.
- Schnabel, S., & Eilers, P. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, 53, 4168–4177.
- Schulze Waltrup, L., & Kauermann, G. (2017). Smooth expectiles for panel data using penalized splines. *Statistics and Computing*, 27, 271–282.
- Schulze Waltrup, L., Sobotka, F., Kneib, T., & Kauermann, G. (2015). Expectile and quantile regression—David and Goliath? *Statistical Modelling*, 15(5), 433–456.
- Sobotka, F., Kauermann, G., Schulze Waltrup, L., & Kneib, T. (2013). On confidence intervals for semiparametric expectile regression. *Statistics and Computing*, 23(2), 135–148.
- Spiegel, E., Sobotka, T., & Kneib, F. (2017). Model selection in semiparametric expectile regression. *Electronic Journal of Statistics*, 11(2), 3008–3038.
- Wand, M., & Jones, M. (1995). *Kernel smoothing*. London: Chapman and Hall.
- Yang, Y., Zhang, T., & Zou, H. (2018). Flexible expectile regression in reproducing kernel Hilbert spaces. *Technometrics*, 60(1), 26–35.
- Yao, Q., & Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics*, 6, 273–292.
- Zhao, J., Chen, Y., & Zhang, Y. (2018). Expectile regression for analyzing heteroscedasticity in high dimension. *Statistics and Probability Letters*, 137, 304–311.
- Zhao, J., Yan, G., & Zhang, Y. (2019). Semiparametric expectile regression for high-dimensional heavy-tailed and heterogeneous data. [arXiv:1908.06431v1](https://arxiv.org/abs/1908.06431v1).
- Zhu, L., Li, R., & Cui, H. (2013). Robust estimation for partially linear models with large-dimensional covariates. *Science China Mathematics*, 56(10), 2069–2088.
- Ziegel, J. (2016). Coherence and elicibility. *Mathematical Finance*, 26(4), 901–918.

Piecewise Linear Continuous Estimators of the Quantile Function



Delphine Blanke and Denis Bosq

Abstract In Blanke and Bosq (2018), families of piecewise linear estimators of the distribution function F were introduced. It was shown that they reduce the mean integrated squared error (MISE) of the empirical distribution function F_n and that the minimal MISE was reached by connecting the midpoints $(\frac{X_k^* + X_{k+1}^*}{2}, \frac{k}{n})$, with X_1^*, \dots, X_n^* the order statistics. In this contribution, we consider the reciprocal estimators, built respectively for known and unknown support of distribution, for estimating the quantile function F^{-1} . We prove that these piecewise linear continuous estimators again strictly improve the MISE of the classical sample quantile function F_n^{-1} .

1 Introduction

If X_1, X_2, \dots, X_n are independent and identically distributed (i.i.d.) real random variables with absolutely continuous distribution function F , the quantile function is defined as $F^{-1}(t) = \inf\{x : F(x) \geq t\}$. The sample (or empirical) quantile function is then $F_n^{-1}(t) = \inf\{x : F_n(x) \geq t\}$, with $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{]-\infty, x]}(X_i)$, $x \in \mathbb{R}$ and \mathbb{I}_A denotes the indicator function of the set A . This is equivalent to $F_n^{-1}(t) = X_k^*$ for $t \in]\frac{k-1}{n}, \frac{k}{n}]$, $k = 1, \dots, n$ and where $X_1^* < \dots < X_n^*$ (almost surely) denotes the ordered sample. We study the properties of two piecewise linear alternatives of F_n^{-1} that respectively address the cases of known and unknown support of the density f . Actually, these estimators are the reciprocals of two particular estimators considered in Blanke and Bosq (2018) to estimate F . More precisely, in this last cited

D. Blanke (✉)

Laboratoire de Mathématiques d'Avignon, LMA, Avignon Université,
F-84029 Avignon, France
e-mail: delphine.blanke@univ-avignon.fr

D. Bosq

Laboratoire de Probabilités, Statistique et Modélisation, LPSM, CNRS,
Sorbonne Universités, F-75005 Paris, France
e-mail: denis.bosq@upmc.fr

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_9

reference, the MISE of a general family of polygonal estimators of the distribution function is studied. These estimators consist in linearly interpolating F_n at different points, namely $(X_k^* + p(X_{k+1}^* - X_k^*), \frac{k}{n})$, $k = 1, \dots, n - 1$, where p is a chosen parameter in $[0, 1]$. For example, $p = 0$ corresponds to a piecewise linear interpolation at $(X_k^*, \frac{k}{n})$, $p = 1$ at $(X_{k+1}^*, \frac{k}{n})$, and the choice $p = \frac{1}{2}$ connects the midpoints $(\frac{X_k^* + X_{k+1}^*}{2}, \frac{k}{n})$. It is then shown in Blanke and Bosq (2018) that for all p chosen in $]0, 1[$, the MISE of F_n is strictly improved and that it is minimal at $p = \frac{1}{2}$ (while the choices $p = 0$ or 1 cannot be recommended). The reciprocals of estimators connecting the midpoints of F_n are studied in this contribution to estimate the quantile function, they join the midpoints of F_n^{-1} , and their formulation depends on whether or not the support is known.

A large literature exists on quantile estimation or L -statistics (linear functions of order statistics): we may refer to the review proposed by Poiraud-Casanova and Thomas-Agnan (1998) or to the detailed introductions in Sheather and Marron (1990) and Zelterman (1990) for smooth quantile estimation, and to Cheng and Parzen (1997) for an unified approach. The origin of the quantile estimators studied in this paper can go back to Hazen (1914) in hydrology (see Harter 1984, for a discussion about plotting positions). Even if their good behavior had been outlined by Parzen (1979) and Dielman et al. (1994), as far as we can judge, there has been no theoretical study of their statistical behavior until now.

The paper is organized as follows. In Sect. 2, we introduce our two piecewise quantile estimators and give their first properties deduced from their proximity to the sample quantile function F_n^{-1} . The main result of this paper is the derivation of their MISE established in Theorem 1. It appears that the piecewise quantile estimators strictly improve the sample quantile function and have an equivalent MISE up to the second order. A conclusion and discussion about possible extensions of our results appear in Sect. 3. Finally, the proof of the most technical results is postponed to the appendix.

2 The Piecewise Quantile Estimators

2.1 Definition

For independent and identically distributed (i.i.d.) random variables X_1, \dots, X_n with compact support $[a, b]$ and absolutely continuous distribution function F , we introduce two continuous piecewise linear estimators of the quantile function F^{-1} . These estimators are the reciprocals of the two estimators of F , studied in Blanke and Bosq (2018), which linearly interpolate the empirical cumulative distribution function F_n at its midpoints, and which minimize the MISE among the set of polygonal estimators considered in the latter reference.

Our first quantile estimator, G_{n1}^{-1} , addresses the case of a known support $[a, b]$ by using this support in its construction. The second estimator, G_{n2}^{-1} , modifies G_{n1}^{-1} at its

both ends and handles by this way the case of an unknown support. Note that even if the results of this article are established for distributions with compact support, the definition of G_{n2}^{-1} is adapted for the case where no information on the support of the distribution is available (and so can even be infinite).

Definition 1 (1) For known support $[a, b]$, we define

$$G_{n1}^{-1}(t) = \begin{cases} 2nt(X_1^* - a) + a & \text{if } t \in [0, \frac{1}{2n}], \\ (nt - k + \frac{1}{2})(X_{k+1}^* - X_k^*) + X_k^* & \text{if } t \in]\frac{2k-1}{2n}, \frac{2k+1}{2n}], k = 1, \dots, n - 1, \\ b - 2n(1 - t)(b - X_n^*) & \text{if } t \in]1 - \frac{1}{2n}, 1]. \end{cases} \tag{1}$$

(2) In the general case, we set $G_{n2}^{-1}(t) = G_{n1}^{-1}(t)$ for $t \in]\frac{1}{2n}, 1 - \frac{1}{2n}]$, and for $n \geq 2$,

$$G_{n2}^{-1}(t) = \begin{cases} (nt - \frac{1}{2})(X_2^* - X_1^*) + X_1^* & \text{if } t \in [0, \frac{1}{2n}], \\ (nt - n + \frac{1}{2})(X_n^* - X_{n-1}^*) + X_n^* & \text{if } t \in]1 - \frac{1}{2n}, 1]. \end{cases} \tag{2}$$

Let us recall that the classical sample quantile function is the generalized inverse function of F_n defined by

$$F_n^{-1}(t) = \inf\{x : F_n(x) \geq t\}$$

and is equivalent to $F_n^{-1}(t) = X_k^*$ for $t \in]\frac{k-1}{n}, \frac{k}{n}]$, $k = 1, \dots, n$. Our estimators simply regularize F_n^{-1} by connecting its midpoints on $[\frac{1}{2n}, 1 - \frac{1}{2n}]$ and are extended in a natural way at both ends (toward the support for G_{n1}^{-1} and by lengthening the last segments for G_{n2}^{-1}).

Let us notice that connecting the midpoints of F_n^{-1} on $[\frac{1}{2n}, 1 - \frac{1}{2n}]$ is an old proposition, suggested in Hazen (1914), which remains popular and used in hydrology. Such an estimator also appears in Harter (1984), Parzen (1979), Parrish (1990), and Dielman et al. (1994) and is implemented in statistical packages (Hyndman and Fan 1996). But according to these authors, even with good performance in simulations and good properties of construction, it presents several problems:

- not being justified on the basis of an estimation argument (Hyndman and Fan 1996),
- being restricted on the support $[\frac{1}{2n}, 1 - \frac{1}{2n}]$ (Dielman et al. 1994),
- and only suited to symmetric distributions (Parzen 1979; Dielman et al. 1994).

The results presented in this contribution address the above-mentioned drawbacks. We establish the asymptotic behavior of the estimators G_{n1}^{-1} and G_{n2}^{-1} defined on $[0,1]$, and we show that they are always better than the sample quantile function in terms of MISE.

2.2 First Properties

First, note that estimators G_{nj}^{-1} , $j = 1, 2$, are examples of linear functions of order statistics. Such L -estimators have been extensively studied and share natural properties expected for the quantile function. They are defined as weighted averages of consecutive order statistics. Those of them involving one- or two-order statistics can be written as: $(1 - \gamma)X_k^* + \gamma X_{k+1}^*$, where $(k - \ell)/n \leq t < (k - \ell + 1)/n$ and $\gamma = nt + \ell - k$ with $\ell \in \mathbb{R}$ a constant determined by the considered estimator (see Hyndman and Fan 1996, for values taken by ℓ according to the chosen sample quantile). In our case, the choice $\ell = \frac{1}{2}$ gives $G_{nj}^{-1}(t)$, $j = 1, 2$, for $t \in [\frac{1}{2n}, 1 - \frac{1}{2n}]$. For the intervals $[0, \frac{1}{2n}]$ and $[1 - \frac{1}{2n}, 1]$, the same expression holds for G_{n2}^{-1} by setting $k = 1$ and $k = n - 1$, respectively, on the valid definition over $[\frac{1}{2n}, 1 - \frac{1}{2n}]$. The following proposition reviews some of other natural properties of our estimators.

Proposition 1 For $j = 1, 2$, we get that the estimators G_{nj}^{-1} are

- (a) continuous on $[0, 1]$,
- (b) symmetric,
- (c) invariant by translation (only on $[\frac{1}{2n}, 1 - \frac{1}{2n}]$ for G_{n1}^{-1}),
- (d) equal to the usual sample median for $t = \frac{1}{2}$.

Proof (a) Clear by construction.

(b) We have to check that $G_{-x,nj}^{-1}(t) = -G_{nj}^{-1}(1 - t)$ for $G_{-x,nj}^{-1}$ built with $(-X_1, \dots, -X_n)$. Symmetry is obtained by substituting X_k^* by $-X_{n-k+1}^*$ for $k = 1, \dots, n$ and $[a, b]$ by $[-b, -a]$ in (1)–(2).

(c) For $Y_k = X_k + c$, $k = 1, \dots, n$ with some constant c , we have to establish that $G_{Y,nj}^{-1}(t) = G_{nj}^{-1}(t) + c$ if $G_{Y,nj}^{-1}$ is the sample quantile estimator built with Y_1, \dots, Y_n . The result is clear with $Y_k^* = X_k^* + c$ for all $k = 1, \dots, n$ in (1)–(2). The property is no longer true for $G_{n1}^{-1}(t)$ with $t \in [0, \frac{1}{2n}]$ or $t \in [1 - \frac{1}{2n}, 1]$.

(d) For $j = 1, 2$ and $n = 1$, $G_{nj}^{-1}(\frac{1}{2}) = X_1^*$. For $n \geq 2$ and $n = 2p$, $G_{nj}^{-1}(\frac{1}{2}) = \frac{X_p^* + X_{p+1}^*}{2}$ while for $n = 2p + 1$, $G_{nj}^{-1}(\frac{1}{2}) = X_{p+1}^*$. □

The next immediate lemma specifies the proximity between F_n^{-1} and G_{nj}^{-1} and will be useful for establishing the convergence of our estimators. Note that from now on, we set $a = 0$ and $b = 1$ to simplify the presentation of the results.

Lemma 1 (1) For $j = 1, 2$ and $k = 1, \dots, n - 1$,

$$G_{nj}^{-1}(t) - F_n^{-1}(t) = \begin{cases} (nt - k + \frac{1}{2})(X_{k+1}^* - X_k^*) & \text{if } t \in]\frac{k}{n} - \frac{1}{2n}, \frac{k}{n}] \\ (nt - k - \frac{1}{2})(X_{k+1}^* - X_k^*) & \text{if } t \in]\frac{k}{n}, \frac{k}{n} + \frac{1}{2n}]. \end{cases}$$

(2) For $t \in [0, \frac{1}{2n}]$, $G_{n1}^{-1}(t) - F_n^{-1}(t) = (2nt - 1)X_1^*$ while $G_{n2}^{-1}(t) - F_n^{-1}(t) = (nt - \frac{1}{2})(X_2^* - X_1^*)$.

(3) For $t \in [1 - \frac{1}{2n}, 1]$, $G_{n1}^{-1}(t) - F_n^{-1}(t) = (2nt - 2n + 1)(1 - X_n^*)$ while $G_{n2}^{-1}(t) - F_n^{-1}(t) = (nt - n + \frac{1}{2})(X_n^* - X_{n-1}^*)$.

Corollary 1 *If F is absolutely continuous with density f such that f is continuous on $[0,1]$ and $\inf_{x \in [0,1]} f(x) \geq c_0$ for some positive constant c_0 , one obtains that $\sup_{t \in [0,1]} |G_{nj}^{-1}(t) - F_n^{-1}(t)| = O_p(n^{-1})$, $j = 1, 2$.*

Proof First recall that the joint density of (X_k^*, X_{k+1}^*) (see e.g. David and Nagaraja 2003, p. 12) is given by

$$f_{(X_k^*, X_{k+1}^*)}(x, y) = \frac{n!}{(k-1)!(n-k-1)!} F^{k-1}(x) f(x) f(y) (1-F(y))^{n-k-1} \mathbb{I}_{[0,y]}(x) \mathbb{I}_{[0,1]}(y).$$

Next, integrations by parts imply that $\mathbb{E}(X_{k+1}^* - X_k^*) = C_n^k \int_0^1 F^k(x) (1-F(x))^{n-k} dx$ so that

$$\mathbb{E}(X_{k+1}^* - X_k^*) = C_n^k \int_0^1 u^k (1-u)^{n-k} \frac{1}{f(F^{-1}(u))} du \leq \frac{C_n^k}{c_0} \int_0^1 u^k (1-u)^{n-k} du.$$

From the standard result $\int_0^1 u^k (1-u)^{n-k} du = \frac{k!(n-k)!}{(n+1)!}$, we may deduce that $\mathbb{E}(X_{k+1}^* - X_k^*) = O(\frac{1}{n+1})$ uniformly in k . One easily concludes with Lemma 1 and Lemma 3.2(a)–(b) in Blanke and Bosq (2018) (recalled in the Appendix, see Lemma 3) together with Markov inequality. \square

We may deduce that the two estimators are asymptotically equivalent to F_n^{-1} ; for example, they get the same limit in distribution since $\sqrt{n}(G_{nj}^{-1}(t) - F_n^{-1}(t)) \xrightarrow[n \rightarrow \infty]{P} 0$ for $j = 1, 2$.

2.3 Mean Integrated Squared Error

We now give the main result of this contribution showing that the estimators strictly improve the sample quantile function in terms of MISE and are equivalent up to second order.

Theorem 1 *If F is absolutely continuous with density f such that f is C^1 on $[0,1]$ and $\inf_{x \in [0,1]} f(x) > 0$, we get that, for $j = 1, 2$,*

$$\begin{aligned} & \int_0^1 \mathbb{E}(G_{nj}^{-1}(t) - F^{-1}(t))^2 dt \\ &= \int_0^1 \mathbb{E}(F_n^{-1}(t) - F^{-1}(t))^2 dt - \frac{1}{4n^2} \int_0^1 \frac{1}{f(x)} dx + O(n^{-\frac{5}{2}}). \end{aligned}$$

The proof of Theorem 1 is based on the decomposition of $(G_{nj}^{-1}(t) - F^{-1}(t))^2$ into

$$(G_{nj}^{-1}(t) - F_n^{-1}(t))^2 + (F_n^{-1}(t) - F^{-1}(t))^2 + 2(G_{nj}^{-1}(t) - F_n^{-1}(t))(F_n^{-1}(t) - F^{-1}(t))$$

with the following proposition proved in the appendix.

Proposition 2 *Under the assumptions of Theorem 1, we obtain*

(1) for $j = 1, 2$,

$$\mathbb{E} \int_0^1 (G_{nj}^{-1}(t) - F_n^{-1}(t))^2 dt = \frac{1}{6n(n+1)} \int_0^1 \frac{1}{f(x)} dx + \mathcal{O}(n^{-3}).$$

(2) for $j = 1$,

$$\begin{aligned} \mathbb{E} \int_0^1 (G_{n1}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t) dt \\ = \frac{\mathbb{E}(1 - X_n^*)}{4n} - \frac{1}{4n(n+1)} \int_0^1 \frac{1}{f(x)} dx + \mathcal{O}(n^{-3}) \end{aligned}$$

while, for $j = 2$,

$$\begin{aligned} \mathbb{E} \int_0^1 (G_{n2}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t) dt \\ = \frac{\mathbb{E}(X_n^* - X_{n-1}^*)}{8n} - \frac{1}{4n(n+1)} \int_0^1 \frac{1}{f(x)} dx + \mathcal{O}(n^{-3}). \end{aligned}$$

(3) for $j = 1$,

$$\mathbb{E} \int_0^1 (G_{n1}^{-1}(t) - F_n^{-1}(t))F^{-1}(t) dt = \frac{\mathbb{E}(1 - X_n^*)}{4n} - \frac{1}{24n^2} \int_0^1 \frac{1}{f(x)} dx + \mathcal{O}(n^{-\frac{5}{2}})$$

while, for $j = 2$,

$$\begin{aligned} \mathbb{E} \int_0^1 (G_{n2}^{-1}(t) - F_n^{-1}(t))F^{-1}(t) dt \\ = \frac{\mathbb{E}(X_n^* - X_{n-1}^*)}{8n} - \frac{1}{24n^2} \int_0^1 \frac{1}{f(x)} dx + \mathcal{O}(n^{-\frac{5}{2}}). \end{aligned}$$

Moreover, in the proof of Proposition 2, the following result is established, see Eq. (7) in the appendix. We highlight this result because it might be useful for other applications.

Lemma 2 *If F is absolutely continuous with density f such that f is C^1 on $[0,1]$ and $\inf_{x \in [0,1]} f(x) > 0$, we get that*

$$\int_0^1 \int_x^1 (1 - F(y) + F(x))^n dy dx = \frac{1}{n+1} \int_0^1 \frac{1}{f(y)} dy - \frac{1}{2(n+1)(n+2)f^2(1)} - \frac{1}{2(n+1)(n+2)f^2(0)} + O(n^{-3}).$$

Note that Theorem 1 illustrates the classical phenomenon of deficiency with the dominant term given by the MISE of the sample quantile function. This phenomenon appears also for kernel quantile estimators, see Falk (1984), Sheather and Marron (1990), as well as for estimators inverting kernel estimators of the distribution function, see Azzalini (1981). In these works, an optimal choice of the bandwidth allows a gain compared to the MISE of F_n that is a $O(n^{-\frac{4}{3}})$ for the term of the second order. In this way, these estimators are more efficient than our piecewise linear ones, that have a gain of only $O(n^{-2})$, but $G_{n,j}^{-1}$, $j = 1, 2$, present the immediate advantages to not depend on any smoothing parameter and can be plotted directly in an easy way. As indicated in Proposition 1, they also meet the qualities expected for empirical quantiles (see also Hyndman and Fan 1996).

To conclude this part, we complete Theorem 1 with the MISE of the sample quantile function (in accordance with the Bahadur representation) as we did not find the explicit result in the literature.

Proposition 3 *Under the assumptions of Theorem 1, we have*

$$\int_0^1 \mathbb{E} (F_n^{-1}(t) - F^{-1}(t))^2 dt = \frac{1}{n} \int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))} dt + O(n^{-\frac{3}{2}}).$$

If we suppose moreover that f is C^2 on $[0,1]$, we get that

$$\int_0^1 \mathbb{E} (F_n^{-1}(t) - F^{-1}(t))^2 dt = \frac{1}{n} \int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))} dt + O(n^{-2}).$$

Proof We start from

$$\begin{aligned} \int_0^1 (F_n^{-1}(t) - F^{-1}(t))^2 dt &= \sum_{k=1}^n \int_{\frac{k-1}{n}}^{\frac{k}{n}} (X_k^* - F^{-1}(t))^2 dt \\ &= \frac{1}{n} \sum_{k=1}^n X_k^{*2} + \mathbb{E} (X_1^2) - 2 \sum_{k=1}^n X_k^* \int_{\frac{k-1}{n}}^{\frac{k}{n}} F^{-1}(t) dt \end{aligned}$$

so that $\mathbb{E} \int_0^1 (F_n^{-1}(t) - F^{-1}(t))^2 dt = 2(\mathbb{E} (X_1^2) - \sum_{k=1}^n \mathbb{E} (X_k^*) \int_{\frac{k-1}{n}}^{\frac{k}{n}} F^{-1}(t) dt)$. The main task is the evaluation of the last term. Taylor formula and continuity of f' give that, uniformly over k ,

$$\int_{\frac{k-1}{n}}^{\frac{k}{n}} F^{-1}(t) dt = \frac{1}{n} F^{-1}\left(\frac{k-1}{n}\right) + \frac{1}{2n^2} \frac{1}{f(F^{-1}(\frac{k-1}{n}))} + O(n^{-3})$$

which in turn gives a $O(n^{-2})$ for the term of rest after multiplying it by $\sum_{k=1}^n \mathbb{E}(X_k^*) = n\mathbb{E}(X_1)$. Next, again using Taylor formula to control the remaining terms, we may write

$$\begin{aligned} & \frac{1}{n} \sum_{k=1}^n \mathbb{E}(X_k^*) F^{-1}\left(\frac{k-1}{n}\right) \\ &= \frac{1}{n} \sum_{k=1}^n \mathbb{E}(X_k^*) \left[F^{-1}\left(\frac{k}{n+1}\right) - \frac{n-(k-1)}{n(n+1)} \frac{1}{f(F^{-1}(\frac{k-1}{n}))} \right] + O(n^{-2}) \\ &= \frac{1}{n} \sum_{k=1}^n \mathbb{E}(X_k^*) \left[F^{-1}\left(\frac{k}{n+1}\right) - \frac{1}{nf(F^{-1}(\frac{k}{n+1}))} + \frac{k}{n(n+1)f(F^{-1}(\frac{k}{n+1}))} \right] + O(n^{-2}). \end{aligned} \tag{3}$$

Next, we apply results concerning the expectation of linear combinations of order statistics, $\frac{1}{n} \sum_{k=1}^n J(\frac{k}{n+1}) X_k^*$, given in Stigler (1974) and Helmers (1980). First, since F^{-1} is twice differentiable on $[0, 1]$, we may adapt Eq. (5.11) in the proof of Theorem 2.2 in Helmers (1980) to obtain that

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^n \mathbb{E}(X_k^*) F^{-1}\left(\frac{k}{n+1}\right) &= \int_0^1 (F^{-1}(t))^2 dt \\ &\quad - \frac{1}{2n} \int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))} dt + \frac{1}{n} \int_0^1 \left(\frac{1}{2} - t\right) \frac{F^{-1}(t)}{f(F^{-1}(t))} dt + O(n^{-\frac{3}{2}}). \end{aligned} \tag{4}$$

For the two last terms in (3) involving the density f , we may apply Theorem 4 of Stigler (1974) that does not require the existence of f'' . This allows to obtain that

$$\frac{1}{n} \sum_{k=1}^n \mathbb{E}(X_k^*) \frac{1}{f(F^{-1}(\frac{k}{n+1}))} = \int_0^1 \frac{F^{-1}(t)}{f(F^{-1}(t))} dt + o(n^{-\frac{1}{2}}) \tag{5}$$

and

$$\frac{1}{n} \sum_{k=1}^n \mathbb{E}(X_k^*) \frac{k}{n+1} \frac{1}{f(F^{-1}(\frac{k}{n+1}))} = \int_0^1 \frac{tF^{-1}(t)}{f(F^{-1}(t))} dt + o(n^{-\frac{1}{2}}). \tag{6}$$

To conclude the proof, one may note that if f'' exists and is continuous, one may apply the Helmers (1980)'s result to get a $O(n^{-2})$ instead of $O(n^{-\frac{3}{2}})$ in (4) and a $O(n^{-1})$ instead of a $o(n^{-\frac{1}{2}})$ in (5)–(6). \square

3 Discussion

We have studied two smoothed quantile estimators, G_{n1}^{-1} and G_{n2}^{-1} , and have derived their properties as well as exact expansions for the MISE at the second order for compactly supported distributions. These estimators present several advantages: they are simple generalizations of the quantile process $F_n^{-1}(t)$, they do not depend on any smoothing parameter, and can be plotted directly without any computation. They also meet the standard properties expected for quantiles. Our main result points ahead that they strictly improve the MISE of F_n^{-1} . Moreover, the two estimators have equivalent MISE up to the order n^{-2} . The first one uses the support of the density in its construction while the second one does not require the knowledge of this support. Even if their good numerical properties had already been pointed out in the literature, see for example Parrish (1990) and Dielman et al. (1994), their theoretical study had not been carried out until now. We hope that this paper can give additional motivation for their study and their use in practical problems.

Our results are in agreement with those established for equivalent estimators of the distribution function in Blanke and Bosq (2018). In this reference, general families of estimators of F have been considered, by joining linearly the empirical distribution function F_n at some defined points. It is established that estimators joining the midpoints (as their reciprocal forms studied in the present article for quantile estimation) reach a minimal MISE at the order n^{-2} . In addition, it is shown that piecewise linear estimators joining the order statistics (either the points $(X_k^*, \frac{k}{n})$ or $(X_{k+1}^*, \frac{k}{n})$) do not improve the MISE of F_n at this order. It would be interesting to see if this bad behavior can also be established also for quantile estimation since the simple kernel estimator (joining two consecutive order statistics) is still popular among practitioners. Indeed, more general sample quantiles of the type $Q_n(t) = (1 - \gamma)X_k^* + \gamma X_{k+1}^*$ when $(k - \ell)/n \leq t < (k - \ell + 1)/n$ for some ℓ could be studied with the techniques of our article in order to compare their asymptotic behaviors (work in progress).

Various other extensions of our results may be envisaged. The first one is to relax the assumption of bounded support and then, to consider a weighted mean integrated squared error to ensure the existence of the integrals. As noted previously, the estimator G_{n2}^{-1} seems naturally suitable for such a framework. We may also remark, see Babu et al. (2002), that a monotone transformation like $Y = X/(1 + X)$ may handle the case of random variables with support $[0; \infty[$, and $Y = (1/2) + (\tan^{-1} X/\pi)$ can be taken for real random variables. It should be interesting to look at the transformation of our estimators in these cases. Also, some numerical studies on F_n^{-1} , G_{n1}^{-1} , and G_{n2}^{-1} , not exposed here, have been conducted for Gaussian mixtures (in the same way as in Blanke and Bosq 2018) and give good results even in these unbounded cases.

Finally, conditional quantile estimation is now a large field of research; we can refer to the nice survey of Poiraud-Casanova and Thomas-Agnan (1998) or to the more recent handbooks of Koenker (2005) and Koenker et al. (2018). It allows to take into account the influence of covariates on the studied distribution and has

multiple applications in medicine, economics, and finance. Also, they represent a robust alternative to the conditional mean and are involved in curve estimation, see e.g. Aragon et al. (2005) for frontier estimation and Leconte et al. (2002) for random censorship. It would be interesting to see how our estimators could be written in these frameworks, and see if their easy implementation could offer an interesting alternative to existing usual estimators.

Acknowledgements We thank the reviewers for their careful reading and helpful comments. We are also grateful to the editors for giving us the opportunity to contribute to this volume in honor of Christine Thomas-Agnan, a great mathematician and very nice person.

Appendix

The two following lemmas are useful for calculations.

Lemma 3 (Blanke and Bosq 2018, Lemma 3.2) *If f is continuous on $[0,1]$ and $\inf_{x \in [0,1]} f(x) \geq c_0$ for some positive constant c_0 then, for all integers $r \geq 0$ and $m \geq 1$, not depending on n , we get*

(a)

$$\mathbb{E} \left(\inf_{i=1, \dots, n+r} X_i \right)^m = \frac{a_m}{n^m} + \mathcal{O} \left(\frac{1}{n^{m+1}} \right), \quad a_m > 0,$$

(b)

$$\mathbb{E} \left(1 - \sup_{i=1, \dots, n+r} X_i \right)^m = \frac{b_m}{n^m} + \mathcal{O} \left(\frac{1}{n^{m+1}} \right), \quad b_m > 0,$$

(c)

$$\mathbb{E} (X_2^* - X_1^*) = \frac{d_1}{n} + \mathcal{O} \left(\frac{1}{n^2} \right), \quad d_1 > 0, \quad \text{and} \quad \mathbb{E} (X_2^* - X_1^*)^m = \mathcal{O} \left(\frac{1}{n^m} \right),$$

(d)

$$\mathbb{E} (X_n^* - X_{n-1}^*) = \frac{e_1}{n} + \mathcal{O} \left(\frac{1}{n^2} \right), \quad e_1 > 0, \quad \text{and} \quad \mathbb{E} (X_n^* - X_{n-1}^*)^m = \mathcal{O} \left(\frac{1}{n^m} \right).$$

Lemma 4 (Blanke and Bosq 2018, Proposition A1) *If h is measurable and integrable on $[0, 1]^2$, then*

$$\sum_{k=1}^{n-1} \mathbb{E} (h(X_k^*, X_{k+1}^*)) = n(n-1) \int_0^1 \int_0^y h(x, y) f(x) f(y) (1 - F(y) + F(x))^{n-2} dx dy.$$

Proof of Proposition 2

(1) We start from Lemma 1 and simple integrations give

$$\mathbb{E} \int_0^1 (G_{n1}^{-1}(t) - F_n^{-1}(t))^2 dt = \frac{\mathbb{E}(X_1^*)^2}{6n} + \frac{\mathbb{E}(1 - X_n^*)^2}{6n} + \frac{\sum_{k=1}^{n-1} \mathbb{E}(X_{k+1}^* - X_k^*)^2}{12n}$$

for $j = 1$, while for $j = 2$, one gets

$$\begin{aligned} \mathbb{E} \int_0^1 (G_{n2}^{-1}(t) - F_n^{-1}(t))^2 dt \\ = \frac{\mathbb{E}(X_2^* - X_1^*)^2}{24n} + \frac{\mathbb{E}(X_n^* - X_{n-1}^*)^2}{24n} + \frac{\sum_{k=1}^{n-1} \mathbb{E}(X_{k+1}^* - X_k^*)^2}{12n}. \end{aligned}$$

Lemma 3 implies that the two first terms in these expressions are negligible in $O(n^{-3})$. Lemma 4 implies that

$$\sum_{k=1}^{n-1} \mathbb{E}(X_{k+1}^* - X_k^*)^2 = n(n-1) \int_0^1 \int_0^y (y-x)^2 f(x)f(y)(1-F(y)+F(x))^{n-2} dx dy.$$

Next, integrations by parts give that

$$\begin{aligned} \sum_{k=1}^{n-1} \mathbb{E}(X_{k+1}^* - X_k^*)^2 = -2 \int_0^1 y \mathbb{P}(X_1^* > y) dy - 2 \int_0^1 (1-x) \mathbb{P}(X_n^* \leq x) dx \\ + 2 \int_0^1 \int_x^1 (1-F(y)+F(x))^n dy dx. \end{aligned}$$

Setting $t = y^2$ and $t = (1-x)^2$ in the two first integrals give a $O(n^{-2})$ for these terms with Lemma 3. For the term $2 \int_0^1 \int_x^1 (1-F(y)+F(x))^n dy dx$, we perform the change of variables $y = F^{-1}(t)$, $x = F^{-1}(s)$ to get

$$\int_0^1 \int_x^1 (1-F(y)+F(x))^n dy dx = \int_0^1 \int_s^1 (1-t+s)^n \frac{1}{f(F^{-1}(t))} \frac{1}{f(F^{-1}(s))} ds dt.$$

Again multiple integrations by parts lead to

$$\begin{aligned} \int_0^1 \int_x^1 (1-F(y)+F(x))^n dy dx = \frac{1}{n+1} \int_0^1 \frac{1}{f(y)} dy \\ - \frac{1}{2(n+1)(n+2)f^2(1)} - \frac{1}{2(n+1)(n+2)f^2(0)} + O(n^{-3}). \quad (7) \end{aligned}$$

Now, one may conclude that

$$\sum_{k=1}^{n-1} \mathbb{E} (X_{k+1}^* - X_k^*)^2 = \frac{2}{n+1} \int_0^1 \frac{1}{f(x)} dx + \mathcal{O}(n^{-2}) \quad (8)$$

and the result follows.

(2) From Lemma 1 and $F_n^{-1}(t) = X_k^*$ for $t \in]\frac{k-1}{n}, \frac{k}{n}]$, $k = 1, \dots, n$, we may calculate each integral to obtain, for $j = 1$,

$$\int_0^1 (G_{n1}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t) dt = -\frac{(X_1^*)^2}{4n} + \frac{X_n^*(1 - X_n^*)}{4n} - \sum_{k=1}^{n-1} \frac{(X_{k+1}^* - X_k^*)^2}{8n}$$

and since $X_n^*(1 - X_n^*) = (1 - X_n^*) - (1 - X_n^*)^2$, Lemma 3 implies that

$$\mathbb{E} \int_0^1 (G_{n1}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t) dt = \frac{\mathbb{E}(1 - X_n^*)}{4n} - \sum_{k=1}^{n-1} \frac{\mathbb{E}(X_{k+1}^* - X_k^*)^2}{8n} + \mathcal{O}\left(\frac{1}{n^3}\right)$$

and one may conclude with the relation (8). For $j = 2$, we obtain

$$\begin{aligned} \int_0^1 (G_{n2}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t) dt \\ = -\frac{(X_2^* - X_1^*)X_1^*}{8n} + \frac{X_n^*(X_n^* - X_{n-1}^*)}{8n} - \sum_{k=1}^{n-1} \frac{(X_{k+1}^* - X_k^*)^2}{8n} \end{aligned}$$

and, since $(X_n^* - X_{n-1}^*)X_n^* = -(X_n^* - X_{n-1}^*)(1 - X_n^*) + (X_n^* - X_{n-1}^*)$, Cauchy-Schwarz inequality and Lemma 3 imply that

$$\mathbb{E} \int_0^1 (G_{n2}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t) dt = \frac{\mathbb{E}(X_n^* - X_{n-1}^*)}{8n} - \sum_{k=1}^{n-1} \frac{\mathbb{E}(X_{k+1}^* - X_k^*)^2}{8n} + \mathcal{O}(n^{-3})$$

and again the relation (8) gives the result.

(3) This is the most technical term to handle. For $j = 1$, we decompose it into

$$\begin{aligned} \mathbb{E}(X_1^*) \int_0^{\frac{1}{2}} (2nt - 1)F^{-1}(t) dt + \mathbb{E}(1 - X_n^*) \int_{1-\frac{2}{n}}^1 (2nt - 2n + 1)F^{-1}(t) dt \\ + \sum_{k=1}^{n-1} \mathbb{E}(X_{k+1}^* - X_k^*) \int_{\frac{k}{n}-\frac{1}{2n}}^{\frac{k}{n}} (nt - k + \frac{1}{2})F^{-1}(t) dt \\ + \sum_{k=1}^{n-1} \mathbb{E}(X_{k+1}^* - X_k^*) \int_{\frac{k}{n}}^{\frac{k}{n}+\frac{1}{2n}} (nt - k - \frac{1}{2})F^{-1}(t) dt. \quad (9) \end{aligned}$$

We introduce $K_0(t)$ and $K_1(t)$ the primitives of $F^{-1}(t)$ and $tF^{-1}(t)$ and we use Taylor expansions with integral remainder with $K_1''(t) = F^{-1}(t) + \frac{t}{f(F^{-1}(t))}$, $K_1^{(3)}(t) = \frac{2}{f(F^{-1}(t))} - \frac{tf'(F^{-1}(t))}{f^3(F^{-1}(t))}$; and $K_0''(t) = \frac{1}{f(F^{-1}(t))}$, $K_0^{(3)}(t) = -\frac{f'(F^{-1}(t))}{f^3(F^{-1}(t))}$.

By integrating by parts, we arrive at

$$\begin{aligned} \int_0^{\frac{1}{2n}} (2nt - 1)F^{-1}(t) dt &= 2n \int_0^{\frac{1}{2n}} \frac{(\frac{1}{2n} - t)^2}{f(F^{-1}(t))} dt - \frac{1}{8n^2 f(0)} + O(n^{-3}) \\ &= \frac{1}{12n^2 f(0)} - \frac{1}{8n^2 f(0)} + O(n^{-3}) = -\frac{1}{24n^2 f(0)} + O(n^{-3}) \end{aligned}$$

so that Lemma 3-(a) gives that the first term of (9) is a $O(n^{-3})$. We use the same methodology for the second term to obtain that

$$\begin{aligned} \int_{1-\frac{1}{2n}}^1 (2nt - 2n + 1)F^{-1}(t) dt &= \frac{F^{-1}(1 - \frac{1}{2n})}{4n} + \frac{1}{12n^2 f(F^{-1}(1 - \frac{1}{2n}))} + O(n^{-3}) \\ &= \frac{1}{4n} - \frac{1}{24n^2 f(1)} + O(n^{-3}). \end{aligned}$$

We may deduce that the second term of (9) is equal to $\frac{\mathbb{E}(1-X_n^*)}{4n} + O(n^{-3})$ with the help of Lemma 3-(d). We use again Taylor expansions with integral remainder together with integration by parts for the two terms depending on k . This allows to get, uniformly in k , that

$$\int_{\frac{k}{n} - \frac{1}{2n}}^{\frac{k}{n}} (nt - k + \frac{1}{2})F^{-1}(t) dt = \frac{F^{-1}(\frac{k}{n} - \frac{1}{2n})}{8n} + \frac{1}{24n^2 f(F^{-1}(\frac{k}{n} - \frac{1}{2n}))} + O(n^{-3})$$

and

$$\int_{\frac{k}{n}}^{\frac{k}{n} + \frac{1}{2n}} (nt - k - \frac{1}{2})F^{-1}(t) dt = -\frac{F^{-1}(\frac{k}{n})}{8n} - \frac{1}{48n^2 f(F^{-1}(\frac{k}{n}))} + O(n^{-3}).$$

By this way,

$$\begin{aligned} \sum_{k=1}^{n-1} \mathbb{E}(X_{k+1}^* - X_k^*) \left(\int_{\frac{k}{n} - \frac{1}{2n}}^{\frac{k}{n}} (nt - k + \frac{1}{2})F^{-1}(t) dt + \int_{\frac{k}{n}}^{\frac{k}{n} + \frac{1}{2n}} (nt - k - \frac{1}{2})F^{-1}(t) dt \right) \\ = -\frac{1}{24n^2} \mathbb{E} \sum_{k=1}^{n-1} \frac{(X_{k+1}^* - X_k^*)}{f(F^{-1}(\frac{k}{n}))} + O(n^{-3}). \end{aligned}$$

The last task is now to study $\mathbb{E} \left(\sum_{k=1}^{n-1} \frac{X_{k+1}^* - X_k^*}{f(F^{-1}(\frac{k}{n}))} \right)$. As $F_n(X_k^*) = \frac{k}{n}$, we have

$$\begin{aligned} \mathbb{E} \sum_{k=1}^{n-1} \frac{(X_{k+1}^* - X_k^*)}{f(F^{-1}(\frac{k}{n}))} &= \mathbb{E} \left(\sum_{k=1}^{n-1} \frac{X_{k+1}^* - X_k^*}{f(X_k^*)} \right) \\ &+ \mathbb{E} \left(\sum_{k=1}^{n-1} \frac{(X_{k+1}^* - X_k^*)(f(X_k^*) - f(F^{-1}(\frac{k}{n})))}{f(F^{-1}(\frac{k}{n}))f(X_k^*)} \right). \end{aligned}$$

The first term is evaluated with Lemma 4 yielding to

$$\begin{aligned} \mathbb{E} \left(\sum_{k=1}^{n-1} \frac{X_{k+1}^* - X_k^*}{f(X_k^*)} \right) &= n(n-1) \int_0^1 \int_x^1 (y-x)f(y)(1-F(y)+F(x))^{n-2} dy dx \\ &= -n \int_0^1 (1-x)F^{n-1}(x) dx + n \int_0^1 \int_x^1 (1-F(y)+F(x))^{n-1} dy dx. \end{aligned}$$

Using (7) and the relation $\int_0^1 (1-x)F^{n-1}(x) dx = \frac{(1-\mathbb{E}(\sup_{i=1,\dots,n-1} X_i))^2}{2}$ together with Lemma 3-(b), we arrive at

$$\begin{aligned} \mathbb{E} \left(\sum_{k=1}^{n-1} \frac{X_{k+1}^* - X_k^*}{f(X_k^*)} \right) &= \int_0^1 \frac{1}{f(y)} dy - \frac{1}{2(n+1)f^2(0)} - \frac{1}{2(n+1)f^2(1)} \\ &- \frac{n(1-\mathbb{E}(\sup_{i=1,\dots,n-1} X_i))^2}{2} + \mathcal{O}(n^{-2}). \end{aligned}$$

For the second term, using Cauchy–Schwarz inequality implies that it may be bounded by $C(\sum_{k=1}^{n-1} \mathbb{E}(X_{k+1}^* - X_k^*)^2)^{\frac{1}{2}} (\sum_{k=1}^{n-1} \mathbb{E}(F_n^{-1}(\frac{k}{n}) - F^{-1}(\frac{k}{n}))^2)^{\frac{1}{2}}$ with C some positive constant. From relation (8), Riemann approximation, and Proposition 2, this term is of order $\mathcal{O}(n^{-\frac{1}{2}})$. Collecting all the results, the assertion holds for $j = 1$ and is unchanged for $j = 2$, details are omitted. \square

References

Aragon, Y., Daouia, A., & Thomas-Agnan, C. (2005). Nonparametric frontier estimation: A conditional quantile-based approach. *Econometric Theory*, 21(2), 358–389.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68(1), 326–328.

Babu, G. J., Canty, A. J., & Chaubey, Y. P. (2002). Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, 105(2), 377–392.

Blanke, D., & Bosq, D. (2018). Polygonal smoothing of the empirical distribution function. *Statistical Inference for Stochastic Processes*, 21(2), 263–287.

Cheng, C., & Parzen, E. (1997). Unified estimators of smooth quantile and quantile density functions. *Journal of Statistical Planning and Inference*, 59(2), 291–307.

David, H. A., & Nagaraja, H. N. (2003). *Order statistics*. Wiley Series in probability and statistics (3rd ed.). Hoboken: Wiley-Interscience [John Wiley & Sons].

- Dielman, T., Lowry, C., & Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in Statistics - Simulation and Computation*, 23(2), 355–371.
- Falk, M. (1984). Relative deficiency of kernel type estimators of quantiles. *Annals of Statistics*, 12(1), 261–268.
- Harter, H. L. (1984). Another look at plotting positions. *Communications in Statistics - Theory and Methods*, 13(13), 1613–1633.
- Hazen, A. (1914). Storage to be providing in impounding reservoirs for municipal water supply (with discussion). *Transaction of the American Society of Civil Engineers*, 77, 1539–1669.
- Helmers, R. (1980). Edgeworth expansions for linear combinations of order statistics with smooth weight functions. *Annals of Statistics*, 8(6), 1361–1374.
- Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4), 361–365.
- Koenker, R. (2005). *Quantile regression* (Vol. 38). Econometric society monographs. New York: Cambridge University Press.
- Koenker, R., Chernozhukov, V., He, X., & Peng, L. (Eds.). (2018). *Handbook of quantile regression*. Boca Raton: Chapman & Hall/CRC. Handbooks of modern statistical methods. CRC Press.
- Lecote, E., Poiraud-Casanova, S., & Thomas-Agnan, C. (2002). Smooth conditional distribution function and quantiles under random censorship. *Lifetime Data Analysis*, 8(3), 229–246.
- Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, 46(1), 247–257.
- Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association*, 74(365), 105–131.
- Poiraud-Casanova, S., & Thomas-Agnan, C. (1998). Quantiles conditionnels. *Journal de la société française de statistique*, 139(4), 31–44.
- Sheather, S. J., & Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410), 410–416.
- Stigler, S. M. (1974). Linear functions of order statistics with smooth weight functions. *Annals of Statistics*, 2, 676–693.
- Zelnerman, D. (1990). Smooth nonparametric estimation of the quantile function. *Journal of Statistical Planning and Inference*, 26(3), 339–352.

Single-Index Quantile Regression Models for Censored Data



Axel Bücher, Anouar El Ghouch, and Ingrid Van Keilegom

Abstract When the dimension of the covariate space is high, semiparametric regression models become indispensable to gain flexibility while avoiding the curse of dimensionality. These considerations become even more important for incomplete data. In this work, we consider the estimation of a semiparametric single-index model for conditional quantiles with right-censored data. Iteratively applying the local-linear smoothing approach, we simultaneously estimate the linear coefficients and the link function. We show that our estimating procedure is consistent and we study its asymptotic distribution. Numerical results are used to show the validity of our procedure and to illustrate the finite-sample performance of the proposed estimators.

1 Introduction

Quantile regression is a very attractive alternative to the classical mean-regression model based on the quadratic loss. While the latter provides only information about the central behavior of the data, by varying the quantile level, the former provides a more complete picture, both in the center and in the tails. At the same time, one

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-73249-3_10) contains supplementary material, which is available to authorized users.

A. Bücher
Mathematisches Institut, Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, 40225
Düsseldorf, Germany
e-mail: axel.buecher@hhu.de

A. El Ghouch
ISBA, UCLouvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium
e-mail: anouar.elghouch@uclouvain.be

I. Van Keilegom (✉)
ORSTAT, KU Leuven, Naamsestraat 69, box 3500, 3000 Leuven, Belgium
e-mail: ingrid.vankeilegom@kuleuven.be

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_10

does not need to impose restrictive assumptions about the unknown data generating process. There are many cases where studying the conditional mean is uninformative compared to the conditional upper or lower quantiles representing more extreme situations. A nice illustration can be found in Elsner et al. (2008), where the interest lies in the lifetime-maximum wind speeds of tropical cyclones. The authors found that trends are near zero for the mean and lower quantiles (median and below), but are upward for higher quantiles.

With the objective of providing a robust yet easily computable alternative to linear mean models, Koenker and Bassett (1978) propose a method to estimate a linear quantile model using the so-called check loss function. This seminal work inspired many researchers from different fields and the method has been generalized and adapted to a wide range of statistical applications including fully nonparametric methods like local-polynomial or spline smoothing; see, e.g., Yu and Jones (1998) and Koenker et al. (1994). Although a completely nonparametric approach is flexible, its application requires a large amount of data in order to overcome the curse of dimensionality. While retaining much flexibility, semiparametric models avoid the curse of dimensionality by imposing some structure on the model. One such structure is the single-index model in which one assumes that the objective function depends linearly on the covariates through an unknown link function. Many widely used parametric models can be seen as particular cases of the single-index model. Examples are the linear regression model and the generalized linear model. In a single-index model, no matter the number of covariates, the curse of dimensionality is avoided because the nonparametric part (link function) is of dimension one. This model was investigated and successfully applied to many objective functions, including the conditional mean and conditional quantiles. For some related papers, see, for example, Ichimura (1993), Klein and Spady (1993), Härdle et al. (1993), Carroll et al. (1997), Delecroix et al. (2003), Wu et al. (2010), and Kong and Xia (2012) to cite just some of the relevant papers.

The majority of the available literature is devoted to the case where the variable of interest, say Y , is completely observed. This is not the case in many interesting applications including survival analysis where censoring prevents the direct application of “classical” semiparametric methods because instead of observing Y , one only observes the minimum of Y and a censoring variable. For general results on (linear) quantile regression within such a setting, see, e.g., Portnoy (2003), Wang and Wang (2009), and references therein. Compared to the uncensored case, the literature on single-index models dealing with censoring is very sparse. To the best of our knowledge, the only paper so far is the one of Christou and Akritas (2019) who studied a non-iterative approach based on a combination of four local smoothing estimators: the local Kaplan–Meier estimator for estimating the conditional distribution function of the censoring variable, the nonparametric estimator of Kong et al. (2013), a Nadaraya–Watson-type estimator for estimating the link function, and a local-linear estimator for estimating the desired conditional quantile. For the case of the conditional mean, we refer to Lopez et al. (2013) and the references therein.

In this paper, we study the single-index model for the conditional quantile function when the data are right-censored. We estimate the parameters of interest by

constructing a weighted check function in a way similar to the method of El Ghouch and Van Keilegom (2009). The main difficulties here are the non-differentiability of the check loss function and the fact that the weight function depends on the censoring distribution, which is unknown and needs to be estimated and then plugged-in in the estimating equation. Our proposed local-linear estimation method is based on an iterative procedure involving a \sqrt{n} -consistent estimator of the single-index parameters. In every iteration, we need to maximize a large number of local equations. We derive the asymptotic properties of the resulting quantile regression function under some suitable sufficient conditions. The practical performance of the proposed method is examined via Monte Carlo experiments. The estimator is shown to perform very well for data of moderate size, even when the percentage of censoring is relatively high.

The remainder of the paper is organized as follows. Section 2 describes the estimation procedure. The asymptotic properties such as the consistency and the asymptotic normality of our semiparametric estimator are obtained in Sect. 3. The problem of selecting the bandwidth parameter is tackled in Sect. 4. Simulation studies are presented in Sects. 5, and 6 highlights a brief application to real data. Proofs and technical lemmas are deferred to an online supplement.

2 Model and Estimation

Suppose that Y is a non-negative response depending on a d -dimensional covariate X . The object of interest in this paper is the τ th conditional quantile of Y given $X = x$, $\tau \in (0, 1)$, which we denote by $Q_\tau(x)$. We impose a single-index structure on Q_τ , i.e., we suppose that

$$Q_\tau(x) = m_\tau(x^T \beta_{0,\tau}), \quad (1)$$

where $m_\tau : \mathbb{R} \rightarrow \mathbb{R}$ is an unknown smooth link function and where $\beta_{0,\tau}$ is a vector of unknown coefficients in the unit sphere $S^{d-1} = \{\beta \in \mathbb{R}^d : \|\beta\| = 1\}$, where $\|\cdot\|$ denotes the Euclidean norm on \mathbb{R}^d . For identifiability reasons, we suppose that the first coordinate of $\beta_{0,\tau}$ is positive. As long as it will not cause any ambiguity, we suppress the index τ and write $m = m_\tau$ and $\beta_0 = \beta_{0,\tau}$. In model (1), estimating Q_τ boils down to estimating m and β_0 .

For $u \in \mathbb{R}$, let $\rho_\tau(u) = u\{\tau - \mathbb{1}(u < 0)\}$ denote the check function. Then, it is well known that β_0 is given by

$$\begin{aligned} \beta_0 &= \operatorname{argmin}_{\beta \in \mathbb{R}^d} \mathbb{E}[\rho_\tau\{Y - m(X^T \beta)\}] \\ &= \operatorname{argmin}_{\beta \in \mathbb{R}^d} \mathbb{E}[\mathbb{E}[\rho_\tau\{Y - m(X^T \beta)\} | X^T \beta]]. \end{aligned} \quad (2)$$

The expressions $\mathbb{E}[\rho_\tau\{Y - m(X^T \beta)\}]$ and $\mathbb{E}[\rho_\tau\{Y - m(X^T \beta)\} | X^T \beta]$ can be interpreted as the expected and the conditional expected loss, respectively.

For the moment, let us suppose that there is no censoring and that we observe an i.i.d. sample $(X_i, Y_i)_{i=1}^n$ from (X, Y) . The following procedure for estimating β_0 and $m(v)$, where $v \in \mathbb{R}$ is arbitrary, stems from Wu et al. (2010). The main idea is

to define an empirical analog of the expected loss in (2), which can be minimized subsequently. Let $\beta \in S^{d-1}$ be given. Then, assuming that m is sufficiently smooth and that $X_i^T \beta$ is close to v , a Taylor expansion yields

$$m(X_i^T \beta) \approx m(v) + m'(v)(X_i^T \beta - v) = a + b(X_i^T \beta - v),$$

where $a = m(v)$ and $b = m'(v)$. Thus,

$$\sum_{i=1}^n \rho_\tau \{Y_i - a - b(X_i^T \beta - v)\} K\{(X_i^T \beta - v)/h\} \quad (3)$$

with some kernel function K and a bandwidth h represents an empirical analog of the conditional expected loss in (2). Note that, for given $\beta = \beta_0$, minimizing (3) with respect to a and b yields oracle estimators for $m(v)$ and $m'(v)$, respectively. To get an empirical analog of $\mathbb{E}[\rho_\tau\{Y - m(X^T \beta)\}]$, we need to average (3) over v . Hence, setting $v = v_j = X_j^T \beta$, we obtain

$$\sum_{j=1}^n \sum_{i=1}^n \rho_\tau \{Y_i - a_j - b_j(X_{ij}^T \beta)\} w_{ij}(\beta), \quad (4)$$

where $X_{ij} = X_i - X_j$ and where

$$w_{ij}(\beta) = \left\{ \sum_{i=1}^n K \left(\frac{X_{ij}^T \beta}{h} \right) \right\}^{-1} K \left(\frac{X_{ij}^T \beta}{h} \right).$$

By minimizing the expression in (4) with respect to $(a_j, b_j)_{j=1}^n$ and β , we obtain estimators of $(m(v_j), m'(v_j))_{j=1}^n$ and β_0 . To simplify this minimization problem, Wu et al. (2010) proposed an iterative procedure based on successive estimation of β_0 and $(m(v), m'(v))$, for any given $v \in \mathbb{R}$. In the present paper, we adapt their approach to the case where the observations of the response variable may be censored.

In the presence of censoring, we do not fully observe the response variables Y_i . Instead, we observe a sequence of i.i.d. triplets $(X_i, Z_i, \Delta_i)_{i=1}^n$ from (X, Z, Δ) , where $Z = \min(Y, C)$, $\Delta = \mathbb{1}(Y \leq C)$, and $C \geq 0$ denotes a censoring variable.

Assume for the moment that C is independent of Y given $X^T \beta$ and let $F_{C|X^T \beta}(z|x^T \beta) = \Pr(C \leq z | X^T \beta = x^T \beta)$ denote the conditional distribution of C given $X^T \beta = x^T \beta$. Then, some simple calculations based on the tower property of conditional expectations show that, for any measurable function $h : \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$\mathbb{E}[h(Y, X^T \beta) | X^T \beta] = \mathbb{E} \left[\frac{h(Z, X^T \beta) \Delta}{1 - F_{C|X^T \beta}(Z - |X^T \beta)} \middle| X^T \beta \right]. \quad (5)$$

Therefore, we can write $\mathbb{E}[\rho_\tau\{Y - a - b(X^T \beta - v)\} | X^T \beta]$ as

$$\begin{aligned} & \mathbb{E} \left[Q(\beta) \rho_\tau \{ Z - a - b(X^T \beta - v) \} \mid X^T \beta \right] \\ &= \tau \mathbb{E} \left[Y - Z \mid X^T \beta \right] + \mathbb{E} \left[\{ Z - a - b(X^T \beta - v) \} \left[\tau \right. \right. \\ & \quad \left. \left. - Q(\beta) \mathbb{1} \{ Z < a + b(X^T \beta - v) \} \right] \mid X^T \beta \right], \end{aligned}$$

where $Q(\beta) = \Delta / \{1 - F_{C|X^T \beta}(Z - |X^T \beta)\}$. This suggests to replace (3) by either

$$\sum_{i=1}^n \hat{Q}_i(\beta) \rho_\tau \{ Z_i - a - b(X_i^T \beta - v) \} K \left(\frac{X_i^T \beta - v}{h} \right), \quad (6)$$

or

$$\sum_{i=1}^n \{ Z_i - a - b(X_i^T \beta - v) \} \left[\tau - \hat{Q}_i(\beta) \mathbb{1} \{ Z_i < a + b(X_i^T \beta - v) \} \right] K \left(\frac{X_i^T \beta - v}{h} \right), \quad (7)$$

with $\hat{Q}_i(\beta) = \Delta_i / \{1 - \hat{F}_{C|X^T \beta}(Z_i - |X_i^T \beta)\}$, where $\hat{F}_{C|X^T \beta}$ is a suitable estimator of $F_{C|X^T \beta}$. For instance, one may use the local Kaplan–Meier estimator given by

$$\hat{F}_{C|X^T \beta}(z | x^T \beta) = 1 - \prod_{Z_i \leq z} \left(1 - \frac{B_i(\beta, x)}{\sum_{Z_j \geq Z_i} B_i(\beta, x)} \right)^{1 - \Delta_i},$$

with

$$B_i(\beta, x) = \frac{K \left(\frac{\beta^T X_i - \beta^T x}{a_n} \right)}{\sum_{j=1}^n K \left(\frac{\beta^T X_j - \beta^T x}{a_n} \right)},$$

and where a_n is a bandwidth sequence converging to zero as n tends to infinity. When $B_i = n^{-1}$ for all i , $\hat{F}_{C|X^T \beta}$ reduces to the classical (unconditional) Kaplan–Meier estimator, subsequently simply denoted by \hat{F}_C . Note that, for any given β , both (6) and (7) are convex functions. Although the numerical minimization of (6) may be easier than that of (7), in this work we opt for the latter because, as is well known, the Kaplan–Meier estimator is very unstable at the right tail and this problem can be adequately and automatically dealt with through (7). In fact, in (6), the Kaplan–Meier estimator needs to be calculated for every Z_i whereas in (7), using the fact that $\hat{Q}_i(\beta) \mathbb{1} \{ Z_i < a + b(X_i^T \beta - v) \} = 0$ if $Z_i \geq a + b(X_i^T \beta - v)$, the observations beyond $m(x^T \beta)$ would have no or a very small impact (depending on the bandwidth) on the resulting estimator. A very similar approach was used in El Ghouch and Van Keilegom (2009) for the case of one covariate. An approach based on minimizing a quantity closely related to (7) can be found in He et al. (2013) for analyzing high-dimensional survival data.

For simplicity, and to avoid some technical difficulties, in the present paper, we assume that

(C1) C is independent of Y given X and C are independent of X

(a different assumption, also used for instance by Bouaziz and Lopez (2010) recently, under which the asymptotic results in this paper remain valid is given in Remark 1 below). In such a case, Y and C are independent given $X^T\beta$, and $F_{C|X^T\beta}(z|x^T\beta) = \Pr(C \leq z) = F_C(z)$ so that the unconditional Kaplan–Meier estimator can be used. To sum up, we estimate $m(v)$ and $m'(v)$ by $\hat{m}(v, \beta) = \hat{a}(v, \beta)$ and $\hat{m}'(v, \beta) = \hat{b}(v, \beta)$, respectively, where

$$\begin{aligned} (\hat{a}(v, \beta), \hat{b}(v, \beta)) = \operatorname{argmin}_{a, b \in \mathbb{R}} \sum_{i=1}^n \{Z_i - a - b(X_i^T \beta - v)\} [\tau \\ - \hat{Q}_i \mathbb{1}\{Z_i < a + b(X_i^T \beta - v)\}] K\left(\frac{X_i^T \beta - v}{h}\right), \end{aligned} \quad (8)$$

and where $\hat{Q}_i = \Delta_i / \{1 - \hat{F}_C(Z_i -)\}$ with the unconditional Kaplan–Meier estimator \hat{F}_C . Still, it remains to construct an estimator for β_0 . To do so, we proceed as in the uncensored case and define the following empirical analog of (4):

$$\sum_{j=1}^n \sum_{i=1}^n \{Z_i - a_j - b_j(X_{ij}^T \beta)\} [\tau - \hat{Q}_i \mathbb{1}\{Z_i < a_j + b_j(X_{ij}^T \beta)\}] w_{ij}(\beta).$$

The joint minimization of the resulting expression with respect to $(a_j, b_j)_{j=1}^n$ and β is complicated and likely to lead to unstable estimates, hence we propose the following iterative procedure adapted from Wu et al. (2010).

Step 1. Start with an initial estimator $\hat{\beta}^{(0)}$ of β_0 and set $\beta_{iter} = \hat{\beta}^{(0)}$ (see below for a suitable example on how to obtain $\hat{\beta}^{(0)}$).

Step 2. For $j = 1, \dots, n$, let

$$\begin{aligned} (\hat{a}_j, \hat{b}_j) = \operatorname{argmin}_{a, b \in \mathbb{R}} \sum_{i=1}^n \{Z_i - a - b(X_{ij}^T \beta_{iter})\} [\tau - \\ \hat{Q}_i \mathbb{1}\{Z_i < a + b(X_{ij}^T \beta_{iter})\}] w_{ij}(\beta_{iter}). \end{aligned}$$

Step 3. Using the estimates $(\hat{a}_j, \hat{b}_j)_{j=1}^n$, set

$$\begin{aligned} \beta^* = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \sum_{j=1}^n \sum_{i=1}^n \{Z_i - \hat{a}_j - \hat{b}_j(X_{ij}^T \beta)\} [\tau - \\ \hat{Q}_i \mathbb{1}\{Z_i < \hat{a}_j + \hat{b}_j(X_{ij}^T \beta)\}] w_{ij}(\beta_{iter}) \end{aligned}$$

and update β_{iter} by setting $\beta_{iter} = \operatorname{sgn}(\beta_1^*) \beta^* / \|\beta^*\|$.

- Step 4. Repeat Steps 2 and 3 until the difference between two consecutive estimations of β is smaller than a given threshold and define the final estimate $\hat{\beta}$ by setting $\hat{\beta} = \beta_{iter}$.
- Step 5. For any desired index value $v \in \mathbb{R}$, estimate $m(v)$ and $m'(v)$ by $\hat{m}(v, \hat{\beta}) = \hat{a}(\hat{\beta})$ and $\hat{m}'(v, \hat{\beta}) = \hat{b}(\hat{\beta})$, the latter estimators being defined in (8). For any desired index value $x \in \mathbb{R}^d$, estimate $Q_\tau(x)$ by $\hat{m}(x^T \hat{\beta}, \hat{\beta})$.

Step 1 requires an initial estimator for β_0 . We propose to use an estimator adapted from the OPG (outer product of gradients) method in the mean-regression context in Xia et al. (2002). The method requires that X has a density, and the underlying idea is as follows: For any $x \in \mathbb{R}^d$, we have $\partial m(x^T \beta_0) / \partial x = m'(x^T \beta_0) \beta_0$. Hence, the partial derivatives of $m(x^T \beta_0)$ with respect to x are parallel to β_0 . For $j = 1, \dots, n$, let $b_j = m'(X_j^T \beta_0) \beta_0$. One can easily see that the (standardized) eigenvector corresponding to the largest eigenvalue of $V_n = n^{-1} \sum_{i=1}^n b_j b_j^T$ is given by β_0 , which suggests to estimate β_0 by replacing b_j in the definition of V_n by suitable estimators \hat{b}_j , that is, we define $\hat{\beta}_0$ as the (standardized) eigenvector corresponding to the largest eigenvalue of $\hat{V}_n = n^{-1} \sum_{j=1}^n \hat{b}_j \hat{b}_j^T$. For the estimation of b_j , we propose to use the local-polynomial estimators

$$(\hat{a}_j, \hat{b}_j^T) = \operatorname{argmin}_{(a, b^T) \in \mathbb{R}^{d+1}} \sum_{i=1}^n \{Z_i - a - b^T X_{ij}\} \left[\tau - \hat{Q}_i \mathbb{1}\{Z_i < a + b^T X_{ij}\} \right] K(X_{ij}/h),$$

where K denotes a d -dimensional kernel.

3 Asymptotic Results

In this section, we present asymptotic results for the final estimator $\hat{m} = \hat{m}(\hat{\beta})$ arising from Step 5 of the procedure described in the preceding section. In particular, we show that the estimator for m does not depend on the specific form (or asymptotic distribution) of the parametric estimator $\hat{\beta}$, as long as it is \sqrt{n} -consistent for β_0 . In a non-censored case, the latter assumption has for instance been shown for a similar recursively defined estimator in Kong and Xia (2012). In a censored case, it is satisfied for the maximum likelihood estimator proposed by Strzalkowska-Kominiak and Cao (2013) and for the regression-like semiparametric estimator of Bouaziz and Lopez (2010).

We begin by describing technical conditions. For fixed $v \in \mathbb{R}$, suppose that there exist neighborhoods U_{β_0} , $U_{m(v)}$, and U_v of β_0 , $m(v)$ and v , respectively, such that:

- (A1) The kernel K is a density function on \mathbb{R} which is symmetric around 0, has a compact support denoted by $\text{supp}(K)$, and is differentiable with a bounded derivative.
- (A2) The function m is twice continuously differentiable on U_v with bounded derivatives.
- (A3) (i) The support of X , denoted by $\text{supp}(X)$, is contained in a compact subset D_X of \mathbb{R}^d .
 (ii) For any $\beta \in U_{\beta_0}$, the random variable $X^T \beta$ has a density $f_{X^T \beta}$. The function $U_{\beta_0} \times U_v \rightarrow \mathbb{R}$, $(\beta, u) \mapsto f_{X^T \beta}(u)$ is bounded and Lipschitz-continuous at (β_0, v) . In addition, $f_{X^T \beta_0}(v) > 0$.
- (A4) (i) The conditional distribution $F_{Y|X}$ of Y given X has a conditional density $f_{Y|X}(\cdot|\cdot)$ that is bounded on $U_{m(v)} \times \text{supp}(X)$.
 (ii) For any $\beta \in U_{\beta_0}$, the conditional distribution of Y given $X^T \beta$ has a conditional density $f_{Y|X^T \beta}(\cdot|\cdot)$. The function $U_{\beta_0} \times U_{m(v)} \times U_v \rightarrow \mathbb{R}$, $(\beta, y, u) \mapsto f_{Y|X^T \beta}(y|u)$ is bounded and Lipschitz-continuous at $(\beta_0, m(v), v)$. In addition, $f_{Y|X^T \beta_0}(m(v)|v) > 0$.
 (iii) $U_{\beta_0} \times U_{m(v)} \times U_v \rightarrow \mathbb{R}$, $(\beta, y, u) \mapsto f_{Y|X^T \beta}(y|u)$ is partially differentiable with respect to y and the derivative, denoted by $f'_{Y|X^T \beta}(y|u)$, is bounded.
- (A5) The point $v \in \mathbb{R}$ satisfies $F_Z\{m(v)\} < 1$, where F_Z denotes the c.d.f. of Z .

Before we formulate the main results, let us introduce some additional notations. For $\beta \in \mathbb{R}^d$ and $u \in \mathbb{R}$, let $\mathcal{X}_i(\beta, u) = (1, (X_i^T \beta - u)/h)^T$, $\mathcal{Z}_i(\beta, u) = Z_i - m(u) - m'(u)(X_i^T \beta - u)$, and $\mathcal{K}_i(\beta, u) = K\{(X_i^T \beta - u)/h\}$. Moreover, set $\bar{K}_j = \int_{\mathbb{R}} u^j K(u) du$ and $\bar{K}'_j = \int_{\mathbb{R}} u^j K^2(u) du$ for $j \in \{0, 1, 2, 3\}$ and let

$$\bar{K} = \begin{pmatrix} \bar{K}_0 & \bar{K}_1 \\ \bar{K}_1 & \bar{K}_2 \end{pmatrix}, \quad \bar{K}' = \begin{pmatrix} \bar{K}'_0 & \bar{K}'_1 \\ \bar{K}'_1 & \bar{K}'_2 \end{pmatrix}.$$

For some constant $M > 0$, let U_M denote the closed d -dimensional ball of radius M with center 0, i.e., $U_M = \{\gamma \in \mathbb{R}^d : \|\gamma\| \leq M\}$. Finally, for $\beta \in \mathbb{R}^d$ and $u \in \mathbb{R}$ (usually considered to be close to β_0 and v), let

$$\mathbb{M}_n(u, \beta) = \sqrt{nh} \left\{ \begin{pmatrix} \hat{m}(u, \beta) - m(v) \\ h\{\hat{m}'(u, \beta) - m'(v)\} \end{pmatrix} - \frac{h^2}{2} \bar{K}^{-1} \begin{pmatrix} \bar{K}_2 \\ \bar{K}_3 \end{pmatrix} m''(v) \right\}$$

with $\hat{m}(u, \beta)$ and $\hat{m}'(u, \beta)$ as defined in (8).

Theorem 1 *Suppose that (C1) is met and that $h = h(n) \rightarrow 0$ satisfies $\lim_{n \rightarrow \infty} nh^3 = \infty$ and $nh^5 = O(1)$ as $n \rightarrow \infty$. Then, for any $v \in \mathbb{R}$ that satisfies conditions (A1)–(A5) and for any $M > 0$,*

$$\sup_{(\gamma, \kappa) \in U_M \times [-M, M]} \left\| \mathbb{M}_n(v_n^\kappa, \beta_n^\gamma) - V^{-1} \frac{1}{\sqrt{nh}} \sum_{i=1}^n \left[\tau - Q_i \mathbf{1}\{Z_i < m(X_i^T \beta_0)\} \right] \times \mathcal{X}_i(\beta_0, v) \mathcal{K}_i(\beta_0, v) \right\| = o_P(1),$$

where $v_n^\kappa = v + \kappa/\sqrt{n}$ and $\beta_n^\gamma = \beta_0 + \gamma/\sqrt{n}$, where $Q_i = \Delta_i/\{1 - F_C(Z_i-)\}$ and where $V = [f_{Y|X^T\beta_0}\{m(v) | v\} f_{X^T\beta_0}(v)]\bar{K}$.

Note that the sum between the norm signs in Theorem 1 consists of centered summands as a consequence of (5). The uniformity in γ and κ in Theorem 1 is essential for the next corollary which can be regarded as the main result of this paper: it states that the final estimator for $Q_\tau(x)$ in Step 5 is asymptotically normally distributed.

Corollary 1 *Let $\hat{\beta}_n \in S^{d-1}$ be an estimator for β_0 such that $\hat{\gamma}_n = \sqrt{n}(\hat{\beta}_n - \beta_0) = O_P(1)$. Suppose that (C1) and the conditions on the bandwidth of Theorem 1 are met. Then, for any $v \in \mathbb{R}$ that satisfies conditions (A1)–(A5) and for any $x \in \mathbb{R}^d$ such that $v = x^T \beta_0$ satisfies conditions (A1)–(A5),*

$$\begin{aligned} \mathbb{M}_n(v, \hat{\beta}_n) &\rightsquigarrow \mathcal{N}_2(0, \sigma^2(v)\bar{K}^{-1}\bar{K}'\bar{K}^{-1}), \text{ and} \\ \mathbb{M}_n(x^T \hat{\beta}_n, \hat{\beta}_n) &\rightsquigarrow \mathcal{N}_2(0, \sigma^2(x^T \beta_0)\bar{K}^{-1}\bar{K}'\bar{K}^{-1}), \end{aligned}$$

where, for any $v \in \mathbb{R}$,

$$\sigma^2(v) = \frac{\Phi_{\beta_0}\{m(v) | v\} - \tau^2}{f_{Y|X^T\beta_0}^2\{m(v) | v\} f_{X^T\beta_0}(v)}$$

and where, for any $u, v \in \mathbb{R}$,

$$\Phi_{\beta_0}(u | v) = \mathbb{E} \left[\frac{\mathbb{1}(Y < u)}{1 - F_C(Y-)} \mid X^T \beta_0 = v \right].$$

Remark 1 The results of Theorem 1 and Corollary 1 remain valid provided we replace Condition (C1) by the following Condition (C2) originating from Stute (1993). Note that it is also imposed in Bouaziz and Lopez (2010).

(C2) Δ is independent of X given Y and C are independent of Y .

We also refer to Lopez et al. (2013), where assumption (C1) is replaced by a weaker assumption involving independence between C and Y conditional on $g(X)$ for some function g . For the sake of brevity, we omit further details.

4 Bandwidth Selection

The practical performance of any nonparametric regression technique depends crucially on the choice of smoothing parameters. A (theoretical) local optimal bandwidth can be derived from the result in Corollary 1 by minimizing the asymptotic mean squared error of $\hat{m}(v, \hat{\beta})$ with respect to h , yielding

$$h_n^{opt} = h_n^{opt}(v) = \left\{ \frac{\sigma^2(v)\bar{K}_0}{\{m''(v)\}^2\bar{K}_2^2} \right\}^{1/5} n^{-1/5}.$$

Unfortunately, this expression is not directly applicable in practice, since it depends on several unknown quantities. Even in the simpler non-censored case, the derivation of reliable estimators for the respective quantities is delicate. For that reason, alternative procedures for the bandwidth selection have been proposed, see, e.g., Yu and Jones (1998) or Kong and Xia (2012) for procedures relying on the mean-regression case. However, these procedures are not directly applicable in the presence of censoring. For that reason, we propose to use the following leave-one out cross-validation (CV) procedure (see also Zheng and Yang 1998; Leung 2005; El Ghouch and Van Keilegom 2009):

(CV1) For a given h , estimate $\hat{\beta} = \hat{\beta}(h)$ as in Steps 1–4.

(CV2) For any $j = 1, \dots, n$, set $\hat{m}_{-j,h}(X_j^T \hat{\beta}) = \hat{a}_{-j}(X_j^T \hat{\beta}, \hat{\beta})$, where, for any $v \in \mathbb{R}$ and $\beta \in S^{d-1}$,

$$\begin{aligned} (\hat{a}_{-j}(v, \beta), \hat{b}_{-j}(v, \beta)) &= \operatorname{argmin}_{a, b \in \mathbb{R}} \sum_{\substack{i=1, \dots, n \\ i \neq j}} \{Z_i - a - b(X_i^T \beta - v)\} \\ &\quad \times \hat{Q}_{i,-j} \left[\tau - \mathbb{1}\{Z_i < a + b(X_i^T \beta - v)\} \right] K \left(\frac{X_i^T \beta - v}{h} \right) \end{aligned}$$

denotes the estimator based on all observations except the j th.

(CV3) For $j \in \{1, \dots, n\}$ such that $\Delta_j = 1$, set $\hat{c}_{-j,h} = |\hat{m}_{-j,h}(X_j^T \hat{\beta}) - Z_j|$. Let $CV(h)$ denote either the median or the mean or the $m\%$ -trimmed mean of that sample (referred to as MAE, MSE, or trimmed MSE in the following).

(CV4) Repeat the first three steps for several bandwidths and set $h_n^{CV} = \operatorname{argmin}_h CV(h)$.

We consider 10%-trimmed MSE, which, together with the MSE and the MAE, yields three different criteria.

5 Numerical Results

In this section, we assess the finite-sample performance of the 5-step estimator for $m(v)$. For reasons of numerical stability, we constrain all minimizations to a compact set $[-M, M]^p$, with $M = 10$. Additionally, we stop the algorithm in Step 4 after at most 25 iterations, if convergence has not occurred until then. We perform 500 repetitions for two different models, two sample sizes ($n = 200, 400$), two levels of censoring (on average 25% and 50%), three values of $\tau \in \{0.3, 0.5, 0.7\}$, two dimensions $d \in \{3, 6\}$ and 61 values for $v \in \{0.05, 0.075, 0.1, \dots, 1.525, 1.55\}$. We consider 15 different bandwidths $h \in \{0.1, 0.15, \dots, 0.75, 0.8\}$. Additionally, we

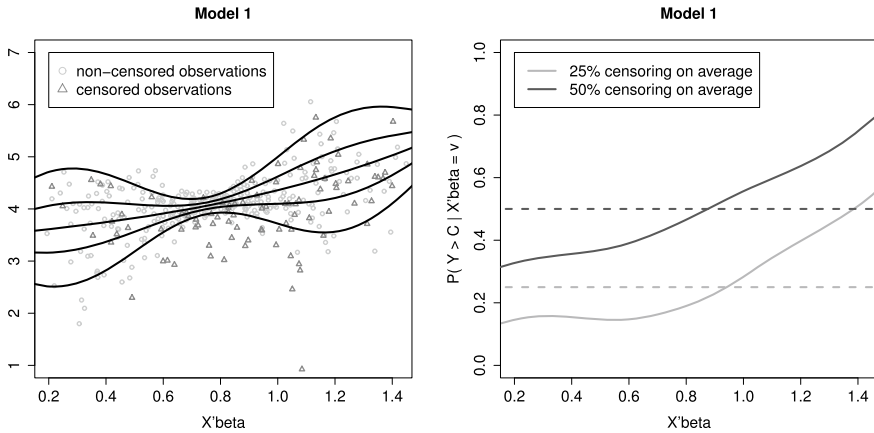


Fig. 1 Left: True quantile curves for $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$ (black curves, in increasing order) and a simulated sample of size $n = 400$ (for $d = 3$, with 25% censoring on average). Right: Probability of censoring $v \mapsto \Pr(Y > C \mid X^T \beta_0 = v)$ for Model 1. The average probability of censoring $\Pr(Y > C)$ is 25% for the black curve and 50% for the gray curve

investigate the performance of the cross-validation method described in Sect. 4. The considered models are as follows.

Model 1 (location-scale model, censoring independent of the covariate)

For $i = 1, \dots, n$, we consider

$$Y_i = 3 + \frac{1}{2} \exp(X_i^T \beta_0) + \{1 + \frac{3}{4} \sin(2\pi X_i^T \beta_0)\} \varepsilon_i, \quad X_i = (X_{i,1}, \dots, X_{i,d}),$$

where $X_{i,j}$ is i.i.d. uniform on $(0, 1)$ for $i = 1, \dots, n$ and $j = 1, \dots, d$, and where ε_i is i.i.d. normal with mean 0 and variance 0.25. During the simulation study, we consider the vector $\beta_0 = \|(d, d - 1, \dots, 1)\|_2^{-1} \times (d, d - 1, \dots, 1)$. Note that the support of $X^T \beta_0$ is the interval $[0, \|\beta_0\|_1]$, with $\|\beta_0\|_1 = 1.60$ for $d = 3$ and $\|\beta_0\|_1 = 2.20$ for $d = 6$. The τ th conditional quantile of Y_i given $X_i = x$ is given by

$$Q_\tau(x) = q_\tau\left(\frac{1}{2} \exp(x^T \beta_0), \frac{1}{2} \{1 + \frac{3}{4} \sin(2\pi x^T \beta_0)\}\right), \tag{9}$$

where $q_\tau(\mu, \sigma)$ denotes the τ th-quantile of the normal distribution with mean μ and standard deviation σ . The curves are depicted in the left panel of Fig. 1, for $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

The censoring variables are i.i.d. normal with mean μ_C and variance $\sigma_C^2 = 1$, independent of X_i and ε_i . We consider two choices for the mean μ_C , which result in either a proportion of censoring of about 50% or of about 25% (for instance, for $d = 3$ the choices are $\mu_C = 4.2$ to obtain a proportion of censoring of about 50%, and $\mu_C = 5$ for proportion of censoring of about 25%). A sample of size $n = 400$ with $d = 3$ and 25% censoring is depicted in the left panel of Fig. 1.

Note that the probability of censoring given $X = x$ is given by

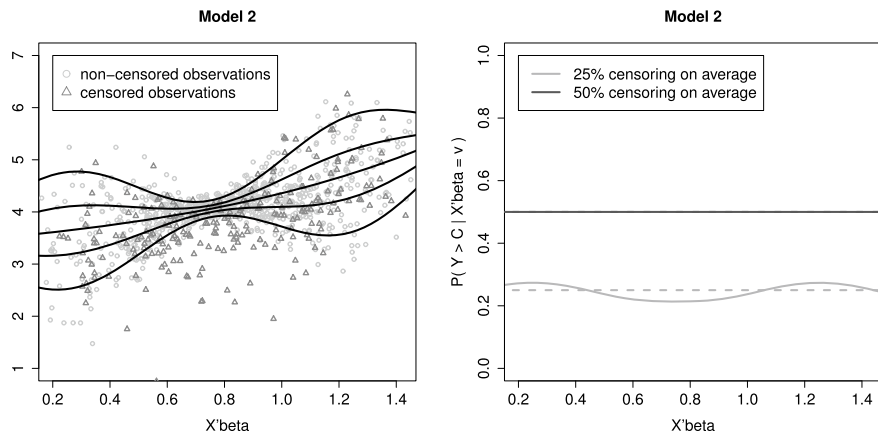


Fig. 2 Left: True quantile curves for $\tau = 0.1, 0.3, 0.5, 0.7, 0.9$ (black curves, in increasing order) and a simulated sample of size $n = 400$ (for $d = 3$, with 25% censoring on average). Right: Probability of censoring $v \mapsto \Pr(Y > C \mid X^T \beta_0 = v)$ for Model 2. The average probability of censoring $\Pr(Y > C)$ is 25% for the black curve and 50% for the gray curve

$$\Pr(Y > C \mid X = x) = \Phi\left(\frac{3 + \frac{1}{2} \exp(x^T \beta_0) - \mu_C}{\sqrt{1 + \frac{1}{4} \{1 + \frac{3}{4} \sin(2\pi x^T \beta_0)\}^2}}\right),$$

where Φ is the standard normal cumulative distribution function. The corresponding curves $v \mapsto \Pr(Y > C \mid X^T \beta_0 = v)$ are depicted in the right panel of Fig. 1 for $\mu_C \in \{4.2, 5\}$ (which, for $d = 3$, yields a proportion of censoring of about 50% and 25%, respectively). From these graphs, we expect the estimator $\hat{m}(v, \hat{\beta})$ to have worse performance for large values of v .

Model 2 (location-scale model, censoring depending on the covariate)

We consider the same data generating mechanism for Y_i as for Model 1. In particular, the conditional quantile curves are given by (9).

The censoring variables are i.i.d. normal with mean $\mu_C + \frac{1}{2} \exp(X^T \beta_0)$ and variance $\sigma_C^2 = 1$, independent of ε_i . We consider two choices for the mean μ_C , which result in either a proportion of censoring of about 50% or of about 25% (for instance, for $d = 3$ the choices are $\mu_C = 3$ to obtain a proportion of censoring of about 50%, and $\mu_C = 3.8$ for proportion of censoring of about 25%). A sample of size $n = 400$ with $d = 3$ and 25% censoring is depicted in the left panel of Fig. 2.

The probability of censoring given $X = x$ is given by

$$\Pr(Y > C \mid X = x) = \Phi\left(\frac{3 - \mu_C}{\sqrt{1 + \frac{1}{4} \{1 + \frac{3}{4} \sin(2\pi x^T \beta_0)\}^2}}\right).$$

The corresponding curves $v \mapsto \Pr(Y > C \mid X^T \beta_0 = v)$ are depicted in the right panel of Fig. 2 for $\mu_C \in \{3, 3.8\}$ (which, for $d = 3$, yields a proportion of censoring of about 50% and 25%, respectively). The curves are much flatter than in Model 1, whence we may expect the estimator to perform similarly throughout the support of $X^T \beta_0$.

The results of our simulation study for the fixed bandwidth case are reported in Table 1 and Figs. 3 and 4. The results in Table 1 concern both the performance of the estimator of β and the estimator of $m(v)$ for various values of v . We state the minimal MSE (for $\hat{\beta}$: the minimal summed MSE over the coordinates of β), over all 15 bandwidth choices $h \in \{0.1, 0.15, \dots, 0.8\}$, alongside with the value realizing that minimum. The results in Figs. 3 and 4 illustrate the performance of the estimator $\hat{m}(v)$ in dependence of the bandwidth parameter h , for a fixed value of $v = 0.85$. The reported boxplots concern the empirical squared estimation error over $N = 500$ simulation runs, and are only reported for $d = 3$ (the results for $d = 6$ look very similar and are not presented here for the sake of brevity).

Overall, the results are as to be expected: for both models, they (greatly) improve with larger sample sizes and a smaller proportion of censoring. Concerning the quantile level, the results are in most cases best for $\tau = 0.5$, closely followed by $\tau = 0.3$ and then $\tau = 0.7$. Despite the fact that the estimator for Model 2 (lower half of Table 1 and Fig. 4) is more complicated (being based on the local Kaplan–Meier estimator for the censoring distribution), the performance of the estimator is often better than for the Model 1, in particular for the parametric estimator $\hat{\beta}$.

Finally, Table 2 shows simulation results on the cross-validation method based on the 10%-trimmed MSE for choosing the optimal bandwidth as described in Sect. 4. For the sake of brevity, we only consider Model 1 with $d = 6$. We measure the quality of the cross-validation method in terms of the relative efficiency:

$$RE = \frac{MSE(\hat{t}, h^{gl.opt})}{MSE(\hat{t}, h_n^{CV})},$$

where $h^{gl.opt} = \min_{h \in \{0.1, \dots, 0.8\}} \{MSE(\hat{\beta}, h) + MSE(\hat{m}(0.7), h) + MSE(\hat{m}(1), h) + MSE(\hat{m}(1.3), h)\}$ and where $\hat{t} \in \{\hat{\beta}, \hat{m}(0.7), \hat{m}(1), \hat{m}(1.3)\}$.

The results in Table 2 show that, overall, the cross-validation method works reasonably well but we also noticed that in some cases, the method may lead to unsatisfactory results. Therefore more work is needed to develop a better solution for this challenging problem of bandwidth selection.

6 Case Study

In this section, we fit the single-index quantile regression model to a subset of the data from the University of Massachusetts AIDS Research Unit IMPACT Study (called U_{IS}-dataset), available online at the John Wiley & Sons website, ftp://ftp.wiley.com/public/sci_tech_med/survival. This dataset has been extensively studied

Table 1 Minimal summed MSE of β and minimal MSE of \hat{m} for four values of v in Model 1 (upper half) and Model 2 (lower half), multiplied by 10^3 , over all bandwidths $h \in \{0.1, 0.15, \dots, 0.75, 0.8\}$, alongside with the bandwidth realizing that minimum. The first and third quarter are for $d = 3$, while the the second and fourth quarter are for dimension $d = 6$

n	Cens.	τ	$\hat{\beta}$	h_{opt}	$\hat{m}(0.4)$	h_{opt}	$\hat{m}(0.7)$	h_{opt}	$\hat{m}(1)$	h_{opt}	$\hat{m}(1.3)$	h_{opt}
200	0.25	0.3	20.3	0.50	24.4	0.45	3.0	0.25	8.5	0.50	34.5	0.80
200	0.50	0.3	38.1	0.55	32.7	0.55	5.4	0.35	12.3	0.70	60.4	0.80
200	0.25	0.5	17.3	0.75	10.0	0.80	2.5	0.55	6.2	0.80	39.2	0.80
200	0.50	0.5	37.5	0.75	16.0	0.80	4.3	0.70	9.5	0.80	77.8	0.80
200	0.25	0.7	23.3	0.55	22.9	0.50	7.2	0.30	12.1	0.80	75.3	0.75
200	0.50	0.7	68.0	0.55	36.3	0.60	20.0	0.45	24.9	0.80	149.1	0.80
400	0.25	0.3	8.3	0.45	13.0	0.45	1.8	0.30	3.6	0.45	15.7	0.80
400	0.50	0.3	13.0	0.50	17.8	0.50	2.1	0.30	5.1	0.45	28.5	0.80
400	0.25	0.5	7.8	0.75	5.5	0.80	0.9	0.50	2.7	0.80	21.3	0.80
400	0.50	0.5	13.9	0.75	8.8	0.80	1.6	0.50	4.7	0.80	44.1	0.80
400	0.25	0.7	9.6	0.55	14.0	0.45	2.5	0.20	6.0	0.55	35.2	0.70
400	0.50	0.7	23.6	0.55	23.3	0.50	7.1	0.30	12.9	0.75	83.5	0.80
200	0.25	0.3	109.6	0.80	137.1	0.50	5.9	0.55	26.2	0.30	126.0	0.80
200	0.50	0.3	189.5	0.80	174.3	0.80	11.4	0.70	59.1	0.80	170.9	0.10
200	0.25	0.5	60.8	0.80	18.7	0.80	7.9	0.80	22.1	0.75	40.4	0.80
200	0.50	0.5	132.4	0.80	30.0	0.80	21.7	0.80	51.7	0.80	108.4	0.80
200	0.25	0.7	67.4	0.80	73.1	0.55	21.2	0.30	27.9	0.80	33.0	0.80
200	0.50	0.7	163.5	0.75	66.2	0.55	63.0	0.40	87.2	0.80	98.3	0.45
400	0.25	0.3	45.4	0.30	75.6	0.50	3.0	0.60	4.7	0.25	47.4	0.25
400	0.50	0.3	94.3	0.80	111.7	0.45	3.9	0.60	17.0	0.40	106.0	0.80
400	0.25	0.5	28.2	0.80	11.6	0.80	3.7	0.70	9.0	0.45	18.3	0.80
400	0.50	0.5	61.4	0.80	14.4	0.80	7.0	0.80	18.7	0.35	39.1	0.80
400	0.25	0.7	32.0	0.80	47.2	0.45	7.7	0.20	10.7	0.80	16.8	0.80
400	0.50	0.7	82.9	0.80	48.9	0.50	18.7	0.25	31.0	0.55	41.8	0.80
200	0.25	0.3	17.1	0.45	20.9	0.50	3.0	0.30	7.6	0.45	28.6	0.80
200	0.50	0.3	26.2	0.50	26.7	0.60	4.0	0.35	10.5	0.50	43.9	0.80
200	0.25	0.5	13.7	0.80	12.4	0.80	2.1	0.60	5.6	0.80	24.9	0.80
200	0.50	0.5	23.7	0.75	21.0	0.80	3.5	0.65	8.4	0.80	46.4	0.80
200	0.25	0.7	16.1	0.55	32.1	0.45	5.7	0.30	8.2	0.80	42.5	0.80
200	0.50	0.7	32.2	0.60	57.0	0.50	10.2	0.45	14.7	0.75	87.1	0.80
400	0.25	0.3	7.4	0.40	11.2	0.50	1.5	0.20	3.1	0.40	11.6	0.80
400	0.50	0.3	11.0	0.40	15.1	0.55	2.3	0.25	4.1	0.40	17.3	0.80
400	0.25	0.5	5.9	0.80	8.0	0.70	0.7	0.60	2.3	0.80	13.5	0.80
400	0.50	0.5	10.0	0.70	14.4	0.80	1.3	0.55	3.8	0.80	21.1	0.80
400	0.25	0.7	6.7	0.55	16.8	0.40	2.0	0.20	4.2	0.65	22.3	0.70
400	0.50	0.7	12.8	0.55	36.6	0.40	3.7	0.25	7.7	0.70	49.5	0.75
200	0.25	0.3	74.5	0.30	118.8	0.55	4.8	0.60	11.5	0.30	90.3	0.30
200	0.50	0.3	109.0	0.40	157.9	0.80	7.1	0.45	19.3	0.40	141.8	0.30
200	0.25	0.5	48.4	0.80	20.6	0.80	5.2	0.65	11.5	0.40	27.5	0.80

(continued)

Table 1 (continued)

n	Cens.	τ	$\hat{\beta}$	h_{opt}	$\hat{m}(0.4)$	h_{opt}	$\hat{m}(0.7)$	h_{opt}	$\hat{m}(1)$	h_{opt}	$\hat{m}(1.3)$	h_{opt}
200	0.50	0.5	81.3	0.80	31.6	0.80	10.0	0.70	21.4	0.55	44.5	0.80
200	0.25	0.7	47.6	0.80	103.5	0.55	11.9	0.25	13.5	0.80	22.0	0.80
200	0.50	0.7	89.0	0.80	120.5	0.65	26.3	0.80	28.9	0.80	41.7	0.80
400	0.25	0.3	27.5	0.25	61.7	0.50	2.9	0.55	2.9	0.30	26.6	0.20
400	0.50	0.3	39.8	0.30	85.3	0.50	3.3	0.60	4.3	0.35	37.2	0.25
400	0.25	0.5	23.0	0.80	13.8	0.80	2.9	0.70	5.3	0.35	12.8	0.45
400	0.50	0.5	38.3	0.80	17.5	0.80	4.2	0.70	7.1	0.40	18.6	0.50
400	0.25	0.7	23.4	0.80	68.4	0.50	4.5	0.20	6.1	0.80	13.6	0.35
400	0.50	0.7	41.9	0.80	91.0	0.55	10.0	0.80	11.3	0.80	25.3	0.35

Table 2 Relative Efficiency of $\hat{\beta}$ and of \hat{m} in Model 1 ($d = 6$) based on the 10% trimmed MSE cross-validation criterion

n	Cens.	τ	$\hat{\beta}$	$\hat{m}(0.7)$	$\hat{m}(1)$	$\hat{m}(1.3)$
200	0.25	0.3	0.88	0.91	0.88	0.65
200	0.50	0.3	0.61	0.45	0.55	0.63
200	0.25	0.7	0.79	0.98	0.72	0.61
200	0.50	0.7	0.72	0.83	0.88	0.64
400	0.25	0.3	0.91	1.04	0.46	0.78
400	0.50	0.3	0.82	0.79	0.99	0.72
400	0.25	0.7	0.81	0.99	0.90	0.76
400	0.50	0.7	0.74	0.87	0.76	0.68

in the textbook Hosmer et al. (2008), see in particular Section 1.3 and the references therein.

The censored, dependent variable of interest Y is the number of days from admission of a drug abusing patient until his/her self-reported return to drug use. While the entire UIS-dataset from the above website consists of (incomplete) data on 628 subjects, we only consider a subsample of size $n = 202$, consisting of patients receiving one particular treatment (long term) and stemming from one particular treatment site (site A). The proportion of censoring, i.e., the proportion of patients that did not return to drug use, is about 21%. We are interested in the effects of 4 (approximately continuous) covariates on the dependent variable: length of treatment in days (X_1), age at enrollment (X_2), Beck Depression Score at admission (X_3), and number of prior drug treatments (X_4).

To preprocess the data, we take logarithms of the number of days to return to drug use. The four covariates are standardized to have mean 0 and variance 1. Denote the estimated values of the single-index parameter by $\hat{\beta}(\tau) = (\hat{\beta}_1(\tau), \dots, \hat{\beta}_4(\tau))' \in S^3$, where $\tau \in \{0.1, 0.3, 0.5, 0.7\}$. Note that due to the proportion of censoring of about 21%, higher quantiles cannot be expected to give any insight into the relationship

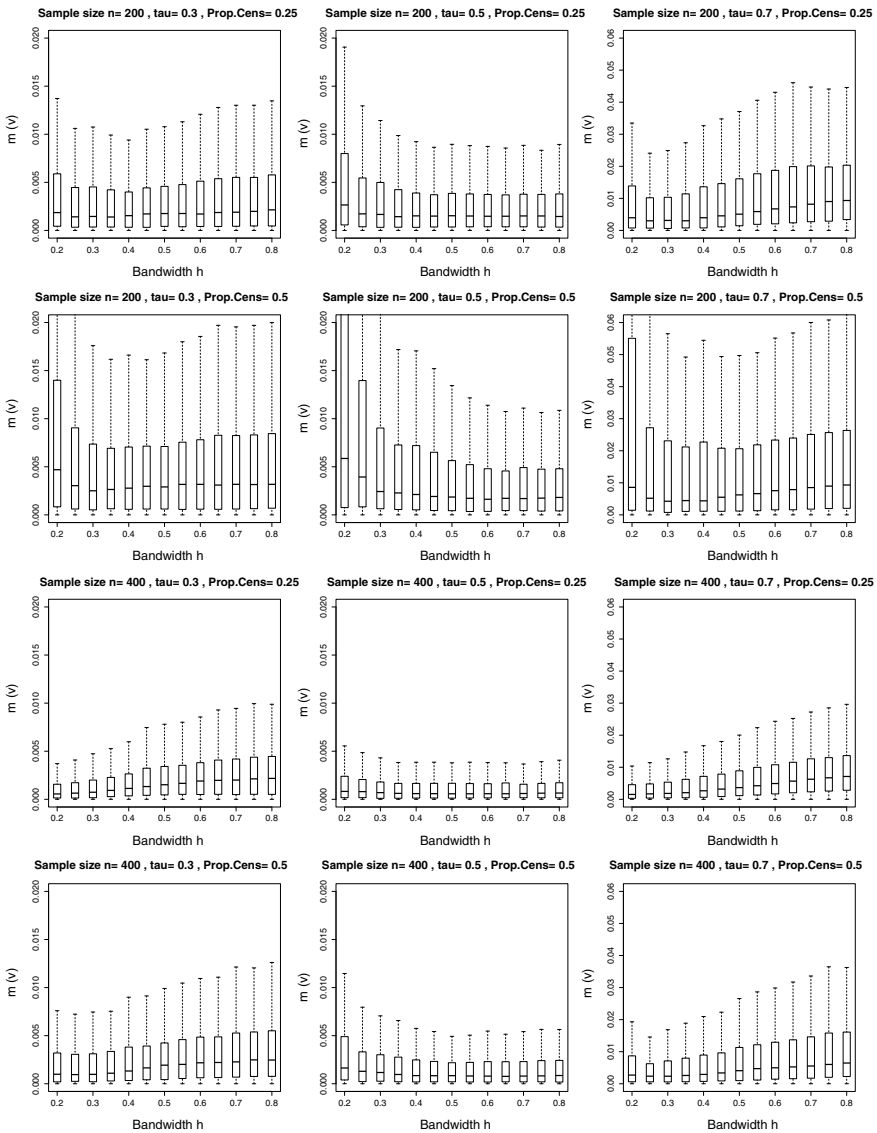


Fig. 3 Squared estimation error of $\hat{m}(v)$ for $v = 0.85$ against the bandwidth h in Model 1 for $d = 3$. Upper six pictures: $n = 200$, lower six pictures: $n = 400$. Note the different scale in the last column (corresponding to $\tau = 0.7$)

between the dependent variable and the covariates (see also the plot of the observations in Fig. 5). The bandwidth parameters are chosen based on the 10%-trimmed MSE-criterion.

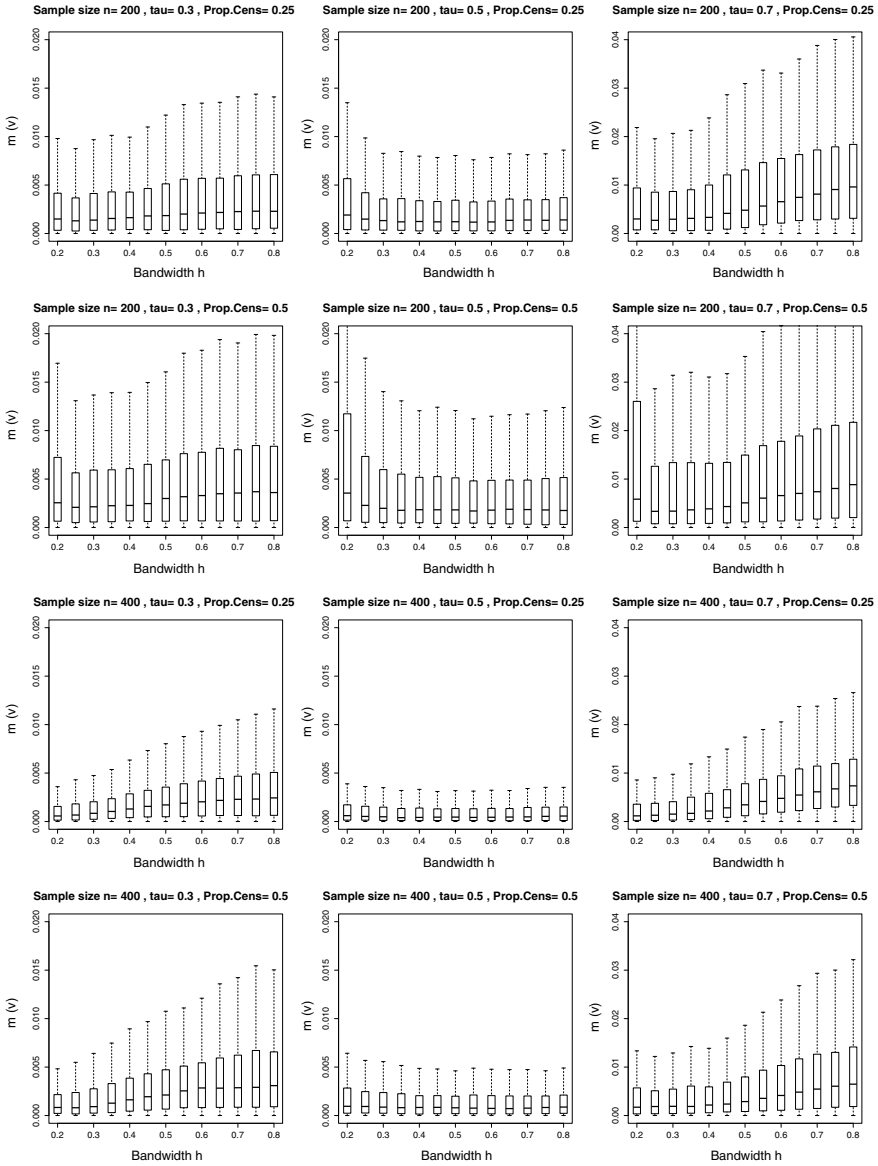


Fig. 4 Squared estimation error of $\hat{m}(v)$ for $v = 0.85$ against the bandwidth h in Model 2 for $d = 3$. Upper six pictures: $n = 200$, lower six pictures: $n = 400$. Note the different scale in the last column (corresponding to $\tau = 0.7$)

The estimated link functions, based on the 10%-trimmed-mean criterion, are shown in Fig. 5, whereas the estimated single-index parameters are given in Table 3.

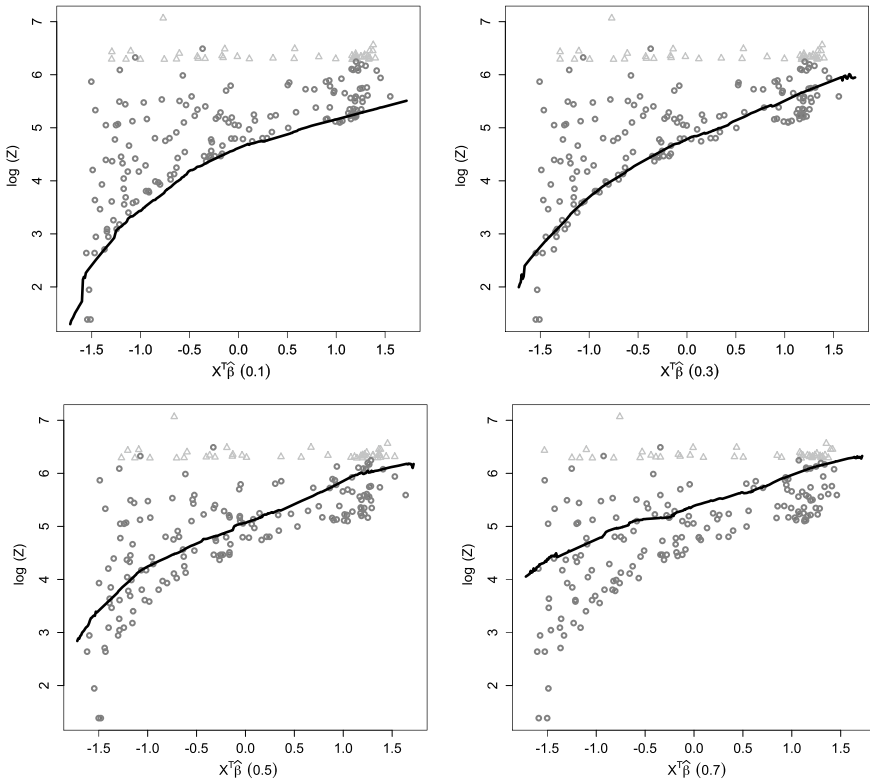


Fig. 5 Estimated link function $x^T \hat{\beta} \mapsto \hat{m}(x^T \hat{\beta})$, for $\tau \in \{0.1, 0.3, 0.5, 0.7\}$ (from upper left to lower right)

Table 3 Estimated single-index parameter for the UIS-dataset

τ	$\hat{\beta}_1(\tau)$	$\hat{\beta}_2(\tau)$	$\hat{\beta}_3(\tau)$	$\hat{\beta}_4(\tau)$
0.1	0.999	0.005	-0.040	-0.001
0.3	0.999	0.007	-0.041	0.004
0.5	0.996	0.045	-0.034	-0.073
0.7	0.994	-0.052	-0.095	0.021

The triangles and circles in Fig. 5 are the censored and uncensored observations, respectively.

The results reveal some interesting features about the effects of the covariates on the response. First of all, we observe that for all quantile levels under consideration, the covariate “length of treatment in days” seems to have a more important impact than the three other covariates, since the coefficients of the standardized variables are very different in size, as can be seen from Table 3. As a general conclusion, a longer treatment period results in a longer time until drug abusers return to drug use. The

estimated link function is strictly increasing for all quantile levels and non-linear and strictly concave for $\tau \in \{0.1, 0.3, 0.5\}$. Furthermore, it is interesting to note that the strength of concavity increases with decreasing quantile. Hence, the marginal utility of an increase of X_1 in its left tail is largest for those patients which generally tend to return to drug abuse rather quickly (i.e., small quantiles of the response—these may be considered as the most interesting group of patients).

Acknowledgements The third author acknowledges financial support from the European Research Council (2016–2021, Horizon 2020, and grant agreement 694409). Computational resources have been provided by the supercomputing facilities of the UCLouvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fonds de la Recherche Scientifique de Belgique under convention 2.5020.11.

References

- Bouaziz, O., & Lopez, O. (2010). Conditional density estimation in a censored single-index regression model. *Bernoulli*, *16*(2), 514–542.
- Carroll, R. J., Fan, J., Gijbels, I., & Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*(438), 477–489.
- Christou, E., & Akritas, M. G. (2019). Single index quantile regression for censored data. *Statistical Methods & Applications*, *28*, 655–678.
- Delecroix, M., Härdle, W., & Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, *86*(2), 213–226.
- El Ghouch, A., & Van Keilegom, I. (2009). Local linear quantile regression with dependent censored data. *Statistica Sinica*, *19*(4), 1621–1640.
- Elsner, J. B., Kossin, J. P., & Jagger, T. H. (2008). The increasing intensity of the strongest tropical cyclones. *Nature*, *455*(7209), 92–95.
- Härdle, W., Hall, P., & Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, *21*(1), 157–178.
- He, X., Wang, L., & Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, *41*(1), 342–369.
- Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis* (Second ed.). Wiley series in probability and statistics. Hoboken: Wiley-Interscience [Wiley]. Regression modeling of time-to-event data.
- Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, *58*(1–2), 71–120.
- Klein, R. W., & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, *61*(2), 387–421.
- Koenker, R., & Bassett, G. Jr. (1978). Regression quantiles. *Econometrica*, *46*(1), 33–50.
- Koenker, R., Ng, P., & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, *81*(4), 673–680.
- Kong, E., Linton, O., & Xia, Y. (2013). Global bahadur representation for nonparametric censored regression quantiles and its applications. *Econometric Theory*, 941–968.
- Kong, E., & Xia, Y. (2012). A single-index quantile regression model and its estimation. *Econometric Theory*, *28*(4), 730–768.
- Leung, D. H.-Y. (2005). Cross-validation in nonparametric regression with outliers. *Annals of Statistics*, *33*(5), 2291–2310.
- Lopez, O., Patilea, V., & Van Keilegom, I. (2013). Single index regression models in the presence of censoring depending on the covariates. *Bernoulli*, *19*(3), 721–747.

- Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association*, 98(464), 1001–1012.
- Strzalkowska-Kominiak, E., & Cao, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. *Journal of Multivariate Analysis*, 114, 74–98.
- Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *Journal of Multivariate Analysis*, 45(1), 89–103.
- Wang, H. J., & Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association* 104(487), 1117–1128.
- Wu, T. Z., Yu, K., & Yu, Y. (2010). Single-index quantile regression. *Journal of Multivariate Analysis*, 101(7), 1607–1621.
- Xia, Y., Tong, H., Li, W. K., & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B* 64(3), 363–410.
- Yu, K., & Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93(441), 228–237.
- Zheng, Z. G., & Yang, Y. (1998). Cross-validation and median criterion. *Statistica Sinica*, 8(3), 907–921.

Extreme L^p -quantile Kernel Regression



Stéphane Girard, Gilles Stupfler, and Antoine Usseglio-Carleve

Abstract Quantiles are recognized tools for risk management and can be seen as minimizers of an L^1 -loss function, but do not define coherent risk measures in general. Expectiles, meanwhile, are minimizers of an L^2 -loss function and define coherent risk measures; they have started to be considered as good alternatives to quantiles in insurance and finance. Quantiles and expectiles belong to the wider family of L^p -quantiles. We propose here to construct kernel estimators of extreme conditional L^p -quantiles. We study their asymptotic properties in the context of conditional heavy-tailed distributions, and we show through a simulation study that taking $p \in (1, 2)$ may allow to recover extreme conditional quantiles and expectiles accurately. Our estimators are also showcased on a real insurance data set.

1 Introduction

The quantile, also called Value-at-Risk in actuarial and financial areas, is a widespread tool for risk measurement, due to its simplicity and interpretability: if Y is a random variable with a cumulative distribution function F , the quantile at level $\alpha \in (0, 1)$ is defined as $q(\alpha) = \inf \{y \in \mathbb{R} | F(y) \geq \alpha\}$. As pointed out in Koenker and Bassett (1978), quantiles may also be seen as a solution of the following minimization problem:

$$q(\alpha) = \arg \min_{t \in \mathbb{R}} \mathbb{E} [\rho_{\alpha}^{(1)}(Y - t) - \rho_{\alpha}^{(1)}(Y)], \quad (1)$$

S. Girard (✉) · A. Usseglio-Carleve
University of Grenoble-Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
e-mail: stephane.girard@inria.fr

A. Usseglio-Carleve
e-mail: antoine.usseglio-carleve@inria.fr

G. Stupfler
University of Rennes, Ensai, CNRS, CREST - UMR 9194, 35000 Rennes, France
e-mail: gilles.stupfler@ensai.fr

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_11

where $\rho_\alpha^{(1)}(y) = |\alpha - \mathbb{1}_{\{y \leq 0\}}||y|$ is the quantile check function. However, the quantile is not subadditive in general and so is not a coherent risk measure in the sense of Artzner et al. (1999). An alternative risk measure gaining popularity is the expectile, introduced in Newey and Powell (1987). This is the solution of (1), with the new loss function $\rho_\alpha^{(2)}(y) = |\alpha - \mathbb{1}_{\{y \leq 0\}}|y^2$ in place of $\rho_\alpha^{(1)}$. Expectiles larger than the mean are coherent risk measures, and have started to be used in actuarial and financial practice (see for instance Cai and Weng 2016). A pioneering paper for the estimation of extreme expectiles in heavy-tailed settings is Daouia et al. (2018).

Quantiles and expectiles may be generalized by considering the family of L^p -quantiles. Introduced in Chen (1996), this class of risk measures is defined, for all $p \geq 1$, by

$$q^{(p)}(\alpha) = \arg \min_{t \in \mathbb{R}} \mathbb{E} [\rho_\alpha^{(p)}(Y - t) - \rho_\alpha^{(p)}(Y)], \quad (2)$$

where $\rho_\alpha^{(p)}(y) = |\alpha - \mathbb{1}_{\{y \leq 0\}}||y|^p$ is the L^p -quantile loss function; the case $p = 1$ leads to the quantile and $p = 2$ gives the expectile. Note that, for $p > 1$, using the formulation (2) and through the subtraction of the (at first sight unimportant) term $\rho_\alpha^{(p)}(Y)$, it is a straightforward consequence of the mean value theorem applied to the function $\rho_\alpha^{(p)}$ that the L^p -quantile $q^{(p)}(\alpha)$ is well defined as soon as $\mathbb{E}(|Y|^{p-1}) < \infty$. While the expectile is the only coherent L^p -quantile (see Bellini et al. 2014), Daouia et al. (2019) showed that for extreme levels of quantiles or expectiles ($\alpha \rightarrow 1$), it may be better to estimate L^p -quantiles first (where typically p is between 1 and 2) and exploit an asymptotic proportionality relationship to estimate quantiles or expectiles. An overview of the potential applications of this kind of statistical assessment of extreme risk may for instance be found in Embrechts et al. (1997).

The contribution of this work is to propose a methodology to estimate extreme L^p -quantiles of $Y|\mathbf{X} = \mathbf{x}$, where the random covariate vector $\mathbf{X} \in \mathbb{R}^d$ is recorded alongside Y . In this context, the case $p = 1$ (quantile) has been considered in Daouia et al. (2011) and Daouia et al. (2013), and the case $p = 2$ (expectile) has recently been studied in Girard et al. (2021). For the general case $p \geq 1$, only Usseglio-Carleve (2018) proposes an estimation procedure under the strong assumption that the vector (\mathbf{X}, Y) is elliptically distributed. The present paper avoids this modeling assumption by constructing a kernel estimator.

The paper is organized as follows. Section 2 introduces an estimator of conditional L^p -quantiles. Section 3 gives the asymptotic properties of the estimator previously introduced, at extreme levels. Finally, Sect. 4 proposes a simulation study in order to assess the accuracy of our estimator which is then showcased on a real insurance data set in Sect. 5. Proofs are postponed to the Appendix.

2 L^p -quantile Kernel Regression

Let (\mathbf{X}_i, Y_i) , $i = 1, \dots, n$ be independent realizations of a random vector $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$. For the sake of simplicity we assume that $Y \geq 0$ with probability 1. We denote by g the density function of \mathbf{X} and let, in the sequel, \mathbf{x} be a fixed point in \mathbb{R}^d such that $g(\mathbf{x}) > 0$. We denote by $\bar{F}^{(1)}(y|\mathbf{x}) = \mathbb{P}(Y > y|\mathbf{X} = \mathbf{x})$ the conditional survival function of Y given $\mathbf{X} = \mathbf{x}$ and assume that this survival function is continuous and regularly varying with index $-1/\gamma(\mathbf{x})$:

$$\forall t > 0, \lim_{y \rightarrow \infty} \frac{\bar{F}^{(1)}(ty|\mathbf{x})}{\bar{F}^{(1)}(y|\mathbf{x})} = t^{-1/\gamma(\mathbf{x})}. \tag{3}$$

Such a distribution belongs to the Fréchet maximum domain of attraction (de Haan and Ferreira 2006). Note that for any $k < 1/\gamma(\mathbf{x})$, $\mathbb{E}[Y^k|\mathbf{X} = \mathbf{x}] < \infty$. Since the definition of L^p -quantiles in (2) requires $\mathbb{E}[|Y|^{p-1}|\mathbf{X} = \mathbf{x}] < \infty$, our minimal assumption will be that $p - 1 < 1/\gamma(\mathbf{x})$. From Eq. (2), L^p -quantiles of level $\alpha \in (0, 1)$ of Y given $\mathbf{X} = \mathbf{x}$ may also be seen as the solution of the following equation:

$$\frac{\mathbb{E}[|Y - y|^{p-1} \mathbb{1}_{\{Y > y\}}|\mathbf{X} = \mathbf{x}]}{\mathbb{E}[|Y - y|^{p-1}|\mathbf{X} = \mathbf{x}]} = 1 - \alpha.$$

In other terms, as noticed in Jones (1994), (conditional) L^p -quantiles can be equivalently defined as quantiles

$$q^{(p)}(\alpha|\mathbf{x}) = \inf \{y \in \mathbb{R} \mid \bar{F}^{(p)}(y|\mathbf{x}) \leq 1 - \alpha\}$$

of the distribution associated with the survival function

$$\bar{F}^{(p)}(y|\mathbf{x}) = \frac{\varphi^{(p-1)}(y|\mathbf{x})}{m^{(p-1)}(y|\mathbf{x})},$$

where, for all $k \geq 0$,

$$m^{(k)}(y|\mathbf{x}) = \mathbb{E}[|Y - y|^k|\mathbf{X} = \mathbf{x}] g(\mathbf{x})$$

$$\text{and } \varphi^{(k)}(y|\mathbf{x}) = \mathbb{E}[|Y - y|^k \mathbb{1}_{\{Y > y\}}|\mathbf{X} = \mathbf{x}] g(\mathbf{x}).$$

Obviously, if $p = 1$, we get the survival function introduced above. The case $p = 2$ leads to the function introduced in Jones (1994) and used in Girard et al. (2021). To estimate $\bar{F}^{(p)}(y|\mathbf{x})$, we let K be a probability density function on \mathbb{R}^d and we introduce the kernel estimators

$$\hat{m}_n^{(k)}(y|\mathbf{x}) = \frac{\sum_{i=1}^n |Y_i - y|^k K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right)}{nh_n^d}, \hat{\varphi}_n^{(k)}(y|\mathbf{x}) = \frac{\sum_{i=1}^n |Y_i - y|^k K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right) \mathbb{1}_{\{Y_i > y\}}}{nh_n^d}.$$

Note that $\hat{m}_n^{(0)}(0|\mathbf{x})$ is the kernel density estimator of $g(\mathbf{x})$, and $\hat{m}_n^{(1)}(0|\mathbf{x})/\hat{m}_n^{(0)}(0|\mathbf{x})$ is the standard kernel regression estimator (since the Y_i are nonnegative). The kernel estimators of $\bar{F}^{(p)}(y|\mathbf{x})$ and $q^{(p)}(\alpha|\mathbf{x})$ are then easily deduced:

$$\hat{F}_n^{(p)}(y|\mathbf{x}) = \frac{\hat{\varphi}_n^{(p-1)}(y|\mathbf{x})}{\hat{m}_n^{(p-1)}(y|\mathbf{x})}, \hat{q}_n^{(p)}(\alpha|\mathbf{x}) = \inf \left\{ y \in \mathbb{R} \mid \hat{F}_n^{(p)}(y|\mathbf{x}) \leq 1 - \alpha \right\}. \quad (4)$$

The case $p = 1$ gives the kernel quantile estimator introduced in Daouia et al. (2013), while $p = 2$ leads to the conditional expectile estimator of Girard et al. (2021). We study here the asymptotic properties of $\hat{q}_n^{(p)}(\alpha|\mathbf{x})$ for an arbitrary $p \geq 1$, when $\alpha = \alpha_n \rightarrow 1$.

3 Main Results

We first make a standard assumption on the kernel. We fix a norm $\|\cdot\|$ on \mathbb{R}^d .

(K) The density function K is bounded and its support S is contained in the unit ball.

To be able to analyze extreme conditional L^p -quantiles in a reasonably simple way, we make a standard second-order regular variation assumption (for a survey of those conditions, see Sect. 2 in de Haan and Ferreira (2006)).

$\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot|\mathbf{x}))$ There exist $\gamma(\mathbf{x}) > 0$, $\rho(\mathbf{x}) \leq 0$ and a positive or negative function $A(\cdot|\mathbf{x})$ converging to 0 such that

$$\forall t > 0, \lim_{y \rightarrow \infty} \frac{1}{A(y|\mathbf{x})} \left(\frac{q^{(1)}(1 - 1/(ty)|\mathbf{x})}{q^{(1)}(1 - 1/y|\mathbf{x})} - t^{\gamma(\mathbf{x})} \right) = \begin{cases} t^{\gamma(\mathbf{x})} \frac{t^{\rho(\mathbf{x})} - 1}{\rho(\mathbf{x})} & \text{if } \rho(\mathbf{x}) < 0, \\ t^{\gamma(\mathbf{x})} \log(t) & \text{if } \rho(\mathbf{x}) = 0. \end{cases}$$

Our last assumption is a local Lipschitz condition which may be found for instance in Daouia et al. (2013); El Methni et al. (2014). We denote by $B(\mathbf{x}, r)$ the ball with center \mathbf{x} and radius r .

(L) We have $g(\mathbf{x}) > 0$ and there exist $c, r > 0$ such that

$$\forall \mathbf{x}' \in B(\mathbf{x}, r), |g(\mathbf{x}) - g(\mathbf{x}')| \leq c\|\mathbf{x} - \mathbf{x}'\|.$$

To be able to control the local oscillations of $(\mathbf{x}, y) \mapsto \bar{F}^{(1)}(y|\mathbf{x})$, we let, for any nonnegative $y_n \rightarrow \infty$,

$$\begin{aligned} \omega_{h_n}^{(1)}(y_n|\mathbf{x}) &= \sup_{\mathbf{x}' \in B(\mathbf{x}, h_n)} \sup_{z \geq y_n} \frac{1}{\log(z)} \left| \log \left(\frac{\bar{F}^{(1)}(z|\mathbf{x}')}{\bar{F}^{(1)}(z|\mathbf{x})} \right) \right|, \\ \omega_{h_n}^{(2)}(y_n|\mathbf{x}) &= \sup_{\mathbf{x}' \in B(\mathbf{x}, h_n)} \sup_{0 < y \leq y_n} |\bar{F}^{(1)}(y|\mathbf{x}') - \bar{F}^{(1)}(y|\mathbf{x})|, \end{aligned}$$

$$\text{and } \omega_{h_n}^{(3)}(y_n|\mathbf{x}) = \sup_{\mathbf{x}' \in B(\mathbf{x}, h_n)} \sup_{\lambda \geq 1} \sup_{b_n, b'_n \rightarrow 0} \left| \frac{\bar{F}^{(1)}(\lambda y_n(1 + b_n)|\mathbf{x}')}{\bar{F}^{(1)}(\lambda y_n(1 + b'_n)|\mathbf{x}')} - 1 \right|.$$

The quantity $\omega_{h_n}^{(1)}(y_n|\mathbf{x})$, discussed for instance in Girard et al. (2021), controls the oscillation of the conditional survival function with respect to \mathbf{x} in its right tail, while $\omega_{h_n}^{(2)}(y_n|\mathbf{x})$ and $\omega_{h_n}^{(3)}(y_n|\mathbf{x})$ are introduced to be able to deal with the case $p \notin \{1, 2\}$ specifically. Let us highlight that $\omega_{h_n}^{(3)}(y_n|\mathbf{x})$ is again geared toward controlling an oscillation of the right tail of the conditional distribution; however, $\omega_{h_n}^{(2)}(y_n|\mathbf{x})$ focuses on the oscillation of the center of the conditional distribution with respect to \mathbf{x} . For $p > 1$, the introduction of a quantity such as $\omega_{h_n}^{(2)}(y_n|\mathbf{x})$ is in some sense natural, since we will have to deal with the local oscillation of the conditional moment $m^{(p-1)}(y|\mathbf{x})$, appearing in the denominator of $\bar{F}^{(p)}(y|\mathbf{x})$, and this conditional moment indeed depends on the whole of the conditional distribution rather than merely on its right tail. Typically $\omega_{h_n}^{(1)}(y_n|\mathbf{x}) = O(h_n)$, $\omega_{h_n}^{(2)}(y_n|\mathbf{x}) = O(h_n)$ and $\omega_{h_n}^{(3)}(y_n|\mathbf{x}) = o(1)$ under reasonable assumptions; we give examples below.

Remark 1 Assume that $Y|\mathbf{X} = \mathbf{x}$ has a Pareto distribution with tail index $\gamma(\mathbf{x}) > 0$:

$$\forall y \geq 1, \bar{F}^{(1)}(y|\mathbf{x}) = y^{-1/\gamma(\mathbf{x})}.$$

If γ is locally Lipschitz continuous, we clearly have $\omega_{h_n}^{(1)}(y_n|\mathbf{x}) = O(h_n)$. Furthermore, for any $y \geq 1$, the mean value theorem yields

$$|\bar{F}^{(1)}(y|\mathbf{x}') - \bar{F}^{(1)}(y|\mathbf{x})| \leq \left| \frac{1}{\gamma(\mathbf{x}')} - \frac{1}{\gamma(\mathbf{x})} \right| \times y^{-1/[\gamma(\mathbf{x}) \vee \gamma(\mathbf{x}')] } \log y.$$

(Here and below \vee denotes the maximum operator.) Under this same local Lipschitz assumption, one then finds $\omega_{h_n}^{(2)}(y_n|\mathbf{x}) = O(h_n)$ as well. Finally, for any $y, y' > 1$,

$$\left| \frac{\bar{F}^{(1)}(y'|\mathbf{x}')}{\bar{F}^{(1)}(y|\mathbf{x}')} - 1 \right| = \left| \left(\frac{y}{y'} \right)^{1/\gamma(\mathbf{x}')} - 1 \right| \leq \frac{|y - y'|}{y'} \times \frac{1 + (y/y')^{1/\gamma(\mathbf{x}')-1}}{\gamma(\mathbf{x}')}$$

by the mean value theorem again. This inequality yields $\omega_{h_n}^{(3)}(y_n|\mathbf{x}) = o(1)$.

The same arguments, and asymptotic bounds on $\omega_{h_n}^{(1)}(y_n|\mathbf{x})$, $\omega_{h_n}^{(2)}(y_n|\mathbf{x})$ and $\omega_{h_n}^{(3)}(y_n|\mathbf{x})$, apply to the conditional Fréchet model

$$\forall y > 0, \bar{F}^{(1)}(y|\mathbf{x}) = 1 - \exp(-y^{-1/\gamma(\mathbf{x})}).$$

Analogous results are easily obtained for the conditional Burr model

$$\forall y > 0, \bar{F}^{(1)}(y|\mathbf{x}) = (1 + y^{-\rho(\mathbf{x})/\gamma(\mathbf{x})})^{1/\rho(\mathbf{x})}$$

when $\rho < 0$ is assumed to be locally Lipschitz continuous, and the conditional mixture Pareto model

$$\forall y \geq 1, \bar{F}^{(1)}(y|\mathbf{x}) = y^{-1/\gamma(\mathbf{x})} [c(\mathbf{x}) + (1 - c(\mathbf{x}))y^{\rho(\mathbf{x})/\gamma(\mathbf{x})}]$$

when $\rho < 0$ and $c \in (0, 1)$ are assumed to be locally Lipschitz continuous. □

3.1 Intermediate L^p -quantile Regression

In this paragraph, we assume that $\sigma_n^{-2} = nh_n^d(1 - \alpha_n) \rightarrow \infty$. Such an assumption means that the L^p -quantile level α_n tends to 1 slowly (by extreme value standards), hence the denominations *intermediate sequence* and *intermediate L^p -quantiles*. This assumption is widespread in the literature of risk measure regression: see, among others, Daouia et al. (2013, 2011); El Methni et al. (2014); Girard et al. (2021). Throughout, we let $\|K\|_2^2 = \int_S K(\mathbf{u})^2 d\mathbf{u}$ be the squared L^2 -norm of K , $\Psi(\cdot)$ denote the digamma function and $IB(t, x, y) = \int_0^t u^{x-1}(1-u)^{y-1} du$ be the incomplete Beta function. Note that $IB(1, x, y) = B(x, y)$ is the standard Beta function.

We now give our first result on the joint asymptotic normality of a finite number J of empirical conditional quantiles with an empirical conditional L^p -quantile ($p > 1$).

Theorem 1 *Assume that (\mathcal{K}) , (\mathcal{L}) and $\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot|\mathbf{x}))$ hold. Let $\alpha_n \rightarrow 1$, $h_n \rightarrow 0$ and $a_n = 1 - \tau(1 - \alpha_n)(1 + o(1))$, where $\tau > 0$. Assume further that $\sigma_n^{-2} = nh_n^d(1 - \alpha_n) \rightarrow \infty$, $nh_n^{d+2}(1 - \alpha_n) \rightarrow 0$, $\sigma_n^{-1}A((1 - \alpha_n)^{-1}|\mathbf{x}) = O(1)$, $\omega_{h_n}^{(3)}(q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x}) \rightarrow 0$ and there exists $\delta \in (0, 1)$ such that*

$$\sigma_n^{-1} \omega_{h_n}^{(1)}((1 - \delta)(\theta \wedge 1)q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x}) \log(1 - \alpha_n) \rightarrow 0, \tag{5}$$

where $\theta = (\tau\gamma(\mathbf{x})/B(p, \gamma(\mathbf{x})^{-1} - p + 1))^{-\gamma(\mathbf{x})}$. Let further $\alpha_{n,j} = 1 - \tau_j(1 - \alpha_n)$, for $0 < \tau_1 < \tau_2 < \dots < \tau_J \leq 1$ such that

$$\sigma_n^{-1} \omega_{h_n}^{(2)}((1 + \delta)(\theta \vee \tau_1^{-\gamma(\mathbf{x})})q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x}) \rightarrow 0. \tag{6}$$

Then, for all $p \in (1, \gamma(\mathbf{x})^{-1}/2 + 1)$, one has

$$\sigma_n^{-1} \left\{ \left(\frac{\hat{q}_n^{(1)}(\alpha_{n,j}|\mathbf{x})}{q^{(1)}(\alpha_{n,j}|\mathbf{x})} - 1 \right)_{1 \leq j \leq J}, \left(\frac{\hat{q}_n^{(p)}(a_n|\mathbf{x})}{q^{(p)}(a_n|\mathbf{x})} - 1 \right) \right\} \xrightarrow{d} \mathcal{N} \left(\mathbf{0}_{J+1}, \frac{\|K\|_2^2}{g(\mathbf{x})} \gamma(\mathbf{x})^2 \Sigma(\mathbf{x}) \right), \tag{7}$$

where $\Sigma(\mathbf{x})$ is the symmetric matrix having entries

$$\left\{ \begin{array}{l} \Sigma_{j,\ell}(\mathbf{x}) = \\ \Sigma_{j,J+1}(\mathbf{x}) = \tau_j^{-1} \left[\gamma(\mathbf{x}) \frac{{}^{(p-1)I}B\left(\left(1 \vee \frac{\tau_j^{-\gamma(\mathbf{x})}}{\theta}\right)^{-1}, \gamma(\mathbf{x})^{-1-p+1}, p-1\right)}{B(p, \gamma(\mathbf{x})^{-1-p+1})} + \left(\left(1 \vee \frac{\tau_j^{-\gamma(\mathbf{x})}}{\theta}\right) - 1\right)^{p-1} \right] \\ \Sigma_{J+1,J+1}(\mathbf{x}) = \frac{B(2p-1, \gamma(\mathbf{x})^{-1-2p+2})}{\tau B(p, \gamma(\mathbf{x})^{-1-p+1})} \end{array} \right. \quad (8)$$

Theorem 1, which will be useful to introduce estimators of the tail index $\gamma(\mathbf{x})$ as part of our extrapolation methodology, generalizes and adapts to the conditional setup several results already found in the literature: see Theorem 1 in Daouia et al. (2013), Theorem 1 in Daouia et al. (2019) and Theorem 3 in Daouia et al. (2020b). Note however that, although they are in some sense related, Theorem 1 does not imply Theorem 1 of Girard et al. (2021), because the latter is stated under weaker regularity conditions warranted by the specific context $p = 2$ of extreme conditional expectile estimation. On the technical side, assumptions (5) and (6) ensure that the bias introduced by smoothing in the \mathbf{x} direction is negligible compared to the standard deviation σ_n of the estimator. The aim of the next paragraph is now to extrapolate our intermediate estimators to properly extreme levels.

3.2 Extreme L^p -quantile Regression

We consider here a level $\beta_n \rightarrow 1$ such that $nh_n^d(1 - \beta_n) \rightarrow c < \infty$. The estimators previously introduced no longer work at such an extreme level. In order to overcome this problem, we first recall a result of Daouia et al. (2019) (see also Lemma 5 below)

$$\forall p \geq 1, \lim_{\alpha \rightarrow 1} \frac{q^{(p)}(\alpha|\mathbf{x})}{q^{(1)}(\alpha|\mathbf{x})} = \left(\frac{\gamma(\mathbf{x})}{B(p, \gamma(\mathbf{x})^{-1} - p + 1)} \right)^{-\gamma(\mathbf{x})}. \quad (9)$$

In the sequel, we shall use the notation $g_p(\gamma) = \gamma/B(p, \gamma^{-1} - p + 1)$. A first consequence of this result is that the L^p -quantile function is regularly varying, i.e.,

$$\forall t > 0, \lim_{y \rightarrow \infty} \frac{q^{(p)}(1 - 1/(ty)|\mathbf{x})}{q^{(p)}(1 - 1/y|\mathbf{x})} = t^{\gamma(\mathbf{x})}. \quad (10)$$

This suggests then that, by considering an intermediate sequence (α_n) , our conditional extreme L^p -quantile may be approximated (and estimated) as follows:

$$q^{(p)}(\beta_n|\mathbf{x}) \approx \left(\frac{1-\alpha_n}{1-\beta_n}\right)^{\gamma(\mathbf{x})} q^{(p)}(\alpha_n|\mathbf{x}),$$

estimated by $\tilde{q}_{n,\alpha_n}^{(p)}(\beta_n|\mathbf{x}) = \left(\frac{1-\alpha_n}{1-\beta_n}\right)^{\hat{\gamma}_{\alpha_n}(\mathbf{x})} \hat{q}_n^{(p)}(\alpha_n|\mathbf{x})$.

Here, $\hat{q}_n^{(p)}(\alpha_n|\mathbf{x})$ is the kernel estimator introduced in Eq. (4), and $\hat{\gamma}_{\alpha_n}(\mathbf{x})$ is a consistent estimator of the conditional tail index $\gamma(\mathbf{x})$. This is a class of Weissman-type estimators (see Weissman 1978) of which we give the asymptotic properties.

Theorem 2 Assume that (\mathcal{K}) , (\mathcal{L}) and $C_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot|\mathbf{x}))$ hold with $\rho(\mathbf{x}) < 0$. Let $\alpha_n, \beta_n \rightarrow 1$, $h_n \rightarrow 0$ be such that $\sigma_n^{-2} = nh_n^d(1-\alpha_n) \rightarrow \infty$ and $nh_n^d(1-\beta_n) \rightarrow c < \infty$. Assume further that $nh_n^{d+2}(1-\alpha_n) \rightarrow 0$, $\omega_{h_n}^{(3)}(q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x}) \rightarrow 0$ and

- (i) $\sigma_n^{-1}A((1-\alpha_n)^{-1}|\mathbf{x}) = O(1)$, $\sigma_n^{-1}(1-\alpha_n) = O(1)$ and $\sigma_n^{-1}\mathbb{E}[Y\mathbb{1}_{\{0 < Y < q^{(1)}(\alpha_n|\mathbf{x})\}}|\mathbf{x}]q^{(1)}(\alpha_n|\mathbf{x})^{-1} = O(1)$,
- (ii) For some $\delta \in (0, 1)$, $\sigma_n^{-1}\omega_{h_n}^{(1)}((1-\delta)[g_p(\gamma(\mathbf{x}))]^{-\gamma(\mathbf{x})}q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x})\log(1-\alpha_n) \rightarrow 0$ and $\sigma_n^{-1}\omega_{h_n}^{(2)}((1+\delta)q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x}) \rightarrow 0$,
- (iii) $\sigma_n^{-1}/\log((1-\alpha_n)/(1-\beta_n)) \rightarrow \infty$.

Take $p \in (1, \gamma(\mathbf{x})^{-1}/2 + 1)$. If in addition $\sigma_n^{-1}(\hat{\gamma}_{\alpha_n}(\mathbf{x}) - \gamma(\mathbf{x})) \xrightarrow{d} \Gamma$, then

$$\frac{\sigma_n^{-1}}{\log((1-\alpha_n)/(1-\beta_n))} \left(\frac{\tilde{q}_{n,\alpha_n}^{(p)}(\beta_n|\mathbf{x})}{q^{(p)}(\beta_n|\mathbf{x})} - 1 \right) \xrightarrow{d} \Gamma.$$

We notice, as is classical in the analysis of heavy tails, that the asymptotic distribution of the extrapolated estimator $\tilde{q}_{n,\alpha_n}^{(p)}(\beta_n|\mathbf{x})$ is exactly that of the purely empirical estimator $\hat{\gamma}_{\alpha_n}(\mathbf{x})$ with a slightly slower rate of convergence. Technically speaking, assumption (i) controls the bias due to the asymptotic approximation (9), while assumption (ii) is used to deal with the bias due to smoothing.

Our aim is now to propose some estimators of $\gamma(\mathbf{x})$ solely based on intermediate L^p -quantiles, in order to carry out the extrapolation step.

3.3 L^p -quantile-Based Estimation of the Conditional Tail Index

The aim of this paragraph is to discuss the estimation of the conditional tail index $\gamma(\mathbf{x})$. A local Pickands estimator is studied in Daouia et al. (2013, 2011). This estimator however has a large variance, which is why Daouia et al. (2011) propose a simplified, conditional, and local version of the Hill estimator:

$$\hat{\gamma}_{\alpha_n}^{(H)}(\mathbf{x}) = \frac{1}{\log(J!)} \sum_{j=1}^J \log \left(\hat{q}_n \left(\frac{j-1 + \alpha_n}{j} | \mathbf{x} \right) / \hat{q}_n(\alpha_n | \mathbf{x}) \right). \quad (11)$$

They also mentioned that taking $J = 9$ is an optimal choice, and leads to an asymptotic variance close to $1.25 \|K\|_2^2 \gamma(\mathbf{x})^2 / g(\mathbf{x})$. Recently, Daouia et al. (2020a); Girard et al. (2021) have shown that replacing the quantile by the expectile in tail index estimators can lead to a significant variance reduction. Our idea here is to propose an estimator based on L^p -quantiles rather than quantiles. In this context, we propose to follow the approach of Girard et al. (2019) and exploit the asymptotic relationship (9) by introducing the following estimator, valid for all $1 < p < \gamma(\mathbf{x})^{-1} + 1$:

$$\hat{\gamma}_{\alpha_n}^{(p)}(\mathbf{x}) = \inf \left\{ \gamma > 0 : g_p(\gamma) \leq \frac{\hat{F}_n^{(1)}(\hat{q}_n^{(p)}(\alpha_n | \mathbf{x}) | \mathbf{x})}{1 - \alpha_n} \right\}. \quad (12)$$

This class of estimators is introduced in Girard et al. (2019) in an unconditional setting, and the (explicit) estimator $\hat{\gamma}_{\alpha_n}^{(2)}(\mathbf{x})$ is introduced in Girard et al. (2021). Using the results previously obtained, we can give the asymptotic distribution of $\hat{\gamma}_{\alpha_n}^{(p)}(\mathbf{x})$ for all $1 < p < \gamma(\mathbf{x})^{-1}/2 + 1$.

Theorem 3 Assume that (\mathcal{K}) , (\mathcal{L}) and $\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot | \mathbf{x}))$ hold with $\gamma(\mathbf{x}) < 1$. Let $\alpha_n \rightarrow 1$ and $h_n \rightarrow 0$. Assume further that $\sigma_n^{-2} = nh_n^d(1 - \alpha_n) \rightarrow \infty$, $nh_n^{d+2}(1 - \alpha_n) \rightarrow 0$, $\omega_{h_n}^{(3)}(q^{(1)}(\alpha_n | \mathbf{x}) | \mathbf{x}) \rightarrow 0$ and

- (i) $\sigma_n^{-1} A((1 - \alpha_n)^{-1} | \mathbf{x}) \rightarrow 0$,
- (ii) $\sigma_n^{-1} q^{(1)}(\alpha_n | \mathbf{x})^{-1} \rightarrow \lambda \in \mathbb{R}$,
- (iii) For some $\delta \in (0, 1)$, $\sigma_n^{-1} \omega_{h_n}^{(1)}((1 - \delta)(g_p(\gamma(\mathbf{x}))^{-\gamma(\mathbf{x})} q^{(1)}(\alpha_n | \mathbf{x}) | \mathbf{x}) \log(1 - \alpha_n) \rightarrow 0$ and $\sigma_n^{-1} \omega_{h_n}^{(2)}((1 + \delta)(q^{(1)}(\alpha_n | \mathbf{x}) | \mathbf{x}) \rightarrow 0$.

Then, for all $p \in (1, \gamma(\mathbf{x})^{-1}/2 + 1)$, one has

$$\sigma_n^{-1} \left(\hat{\gamma}_{\alpha_n}^{(p)}(\mathbf{x}) - \gamma(\mathbf{x}), \frac{\hat{q}_n^{(p)}(\alpha_n | \mathbf{x})}{q^{(p)}(\alpha_n | \mathbf{x})} - 1 \right) \xrightarrow{d} \Theta, \quad (13)$$

where Θ is a bivariate Gaussian distribution with mean vector $(b_p(\mathbf{x}), 0)$ and covariance matrix $\|K\|_2^2 \gamma(\mathbf{x})^2 g(\mathbf{x})^{-1} \Omega(\mathbf{x})$ such that

$$\left\{ \begin{aligned} b_p(\mathbf{x}) &= \frac{(1-p)\gamma(\mathbf{x})g_p(\gamma(\mathbf{x}))^{\gamma(\mathbf{x})}\mathbb{E}[Y|\mathbf{X}=\mathbf{x}]}{1 - \frac{1}{\gamma(\mathbf{x})}(\Psi(\gamma(\mathbf{x})^{-1}+1) - \Psi(\gamma(\mathbf{x})^{-1}-p+1))} \lambda \\ \Omega_{11}(\mathbf{x}) &= \frac{B(p, \gamma(\mathbf{x})^{-1}-p+1)}{\left(1 - \frac{1}{\gamma(\mathbf{x})}(\Psi(\gamma(\mathbf{x})^{-1}+1) - \Psi(\gamma(\mathbf{x})^{-1}-p+1))\right)^2} \left(\frac{B(2p-1, \gamma(\mathbf{x})^{-1}-2p+2)}{B(p, \gamma(\mathbf{x})^{-1}-p+1)^2} - \frac{1}{\gamma(\mathbf{x})} \right) \\ \Omega_{12}(\mathbf{x}) &= \frac{B(p, \gamma(\mathbf{x})^{-1}-p+1)}{1 - \frac{1}{\gamma(\mathbf{x})}(\Psi(\gamma(\mathbf{x})^{-1}+1) - \Psi(\gamma(\mathbf{x})^{-1}-p+1))} \left(\frac{1}{\gamma(\mathbf{x})} - \frac{B(2p-1, \gamma(\mathbf{x})^{-1}-2p+2)}{B(p, \gamma(\mathbf{x})^{-1}-p+1)^2} \right) \\ \Omega_{22}(\mathbf{x}) &= \frac{B(2p-1, \gamma(\mathbf{x})^{-1}-2p+2)}{B(p, \gamma(\mathbf{x})^{-1}-p+1)} \end{aligned} \right. \quad (14)$$

Let us remark here that although Theorem 3 can be seen as a version of Theorem 4 of Girard et al. (2021), the latter is stated under weaker regularity assumptions and applies to further examples of estimators developed specifically in the conditional expectile setup.

Note that condition $\gamma(\mathbf{x}) < 1$ entails $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] < \infty$ and leads to a simple expression of the bias term $b_p(\mathbf{x})$. A result dropping this assumption is available in the unconditional setting in Girard et al. (2019); here, our motivation for this condition is that we shall use extreme regression L^p -quantiles as a way to estimate extreme regression expectiles, for the existence of which a natural condition is that $\mathbb{E}[|Y||\mathbf{X} = \mathbf{x}] < \infty$. The bias term $b_p(\mathbf{x})$ is related to $\gamma(\mathbf{x})$, $q^{(1)}(\alpha_n|\mathbf{x})$ and $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. All these quantities may be easily estimated (the latter two by kernel regression estimators) to construct a bias-reduced conditional tail index estimator as follows:

$$\tilde{\gamma}_{\alpha_n}^{(p)}(\mathbf{x}) = \hat{\gamma}_{\alpha_n}^{(p)}(\mathbf{x}) \left(1 + \frac{(p-1) \left(\frac{\sum_{i=1}^n Y_i K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{h_n}\right)} \right) \hat{q}_n^{(p)}(\alpha_n|\mathbf{x})^{-1}}{1 + \frac{1}{\hat{\gamma}_{\alpha_n}^{(p)}(\mathbf{x})} \left(\Psi\left(1/\hat{\gamma}_{\alpha_n}^{(p)}(\mathbf{x}) - p + 1\right) - \Psi\left(1/\hat{\gamma}_{\alpha_n}^{(p)}(\mathbf{x}) + 1\right) \right)} \right).$$

Under the conditions of Theorem 3, it is clear that $\sigma_n^{-1}(\tilde{\gamma}_{\alpha_n}^{(p)}(\mathbf{x}) - \gamma(\mathbf{x})) \xrightarrow{d} \mathcal{N}(0, \Omega_{11}(\mathbf{x}))$ where $\Omega_{11}(\mathbf{x})$ is given in Eq. (14). This bias reduction improves significantly the numerical results, and is used in the finite-sample study below.

Even though L^p -quantiles with $1 < p < 2$ are more widely estimable than expectiles and take into account the whole tail information, they are neither easy to interpret nor coherent as risk measures. Recent work in Daouia et al. (2019) has shown that extreme L^p -quantiles can be used as vehicles for extreme quantile and expectile estimation; see also Gardes et al. (2020) for an analogous study of the estimation of (a compromise between) Median Shortfall and Conditional Tail Expectation at extreme levels, using tail L^p -medians. Our focus in the following finite-sample study is to analyze the potential of extreme regression L^p -quantiles for the estimation of extreme regression quantiles and expectiles.

4 Simulation Study

We consider here a one-dimensional covariate ($d = 1$), uniformly distributed on $[0, 1]$, and a Burr-type distribution for Y given $X = x$:

$$\bar{F}^{(1)}(y|x) = \left(1 + y^{-\rho(x)/\gamma(x)}\right)^{1/\rho(x)}, \gamma(x) = \frac{4 + \sin(2\pi x)}{10} \text{ and } \rho(x) \equiv -1.$$

Such a distribution fulfills Assumption $\mathcal{C}_2(\gamma(x), \rho(x), A(\cdot|x))$ with auxiliary function $A(y|x) = \gamma(x)y^{\rho(x)}$. We simulate $N = 500$ samples of size $n = 1,000$ independent replications of (X, Y) , and propose to estimate the conditional quantiles and expectiles of level $\beta_n = 1 - 1/n = 0.999$ using our extreme regression L^p -quantile estimators. Note that the quantiles may be calculated explicitly:

$$q(\alpha|x) = [(1 - \alpha)^{\rho(x)} - 1]^{-\gamma(x)/\rho(x)}.$$

Expectiles have to be approximated numerically, since they do not have a simple closed form. In order to estimate these two quantities, we propose to compare different approaches (called either direct or indirect):

- (i) Use the conditional Weissman-type estimators, respectively, based on empirical quantiles and the estimator $\hat{\gamma}_{\alpha_n}^{(H)}(x)$ (direct quantile estimator) and on empirical expectiles and $\tilde{\gamma}_{\alpha_n}^{(2)}(x)$ (direct expectile estimator), i.e.

$$\left(\frac{1 - \alpha_n}{1 - \beta_n}\right)^{\hat{\gamma}_{\alpha_n}^{(H)}(x)} \hat{q}_n^{(1)}(\alpha_n|x), \left(\frac{1 - \alpha_n}{1 - \beta_n}\right)^{\tilde{\gamma}_{\alpha_n}^{(2)}(x)} \hat{q}_n^{(2)}(\alpha_n|x).$$

- (ii) Indirect quantile estimator: estimate first the conditional L^p -quantile using estimator (4), and exploit asymptotic relationship (9) to recover the extreme conditional quantile,

$$\left(\frac{1 - \alpha_n}{1 - \beta_n}\right)^{\tilde{\gamma}_{\alpha_n}^{(p)}(x)} \hat{q}_n^{(p)}(\alpha_n|x) \left(\frac{\tilde{\gamma}_{\alpha_n}^{(p)}(x)}{B(p, \tilde{\gamma}_{\alpha_n}^{(p)}(x)^{-1} - p + 1)}\right)^{\tilde{\gamma}_{\alpha_n}^{(p)}(x)}.$$

- (iii) Indirect expectile estimator: use Eq. (9) to get a connection between L^p -quantile and quantile, and quantile and expectile, resulting in the extreme conditional expectile estimator

$$\left(\frac{1 - \alpha_n}{1 - \beta_n}\right)^{\tilde{\gamma}_{\alpha_n}^{(p)}(x)} \hat{q}_n^{(p)}(\alpha_n|x) \left(\frac{B(2, \tilde{\gamma}_{\alpha_n}^{(p)}(x)^{-1} - 1)}{B(p, \tilde{\gamma}_{\alpha_n}^{(p)}(x)^{-1} - p + 1)}\right)^{\tilde{\gamma}_{\alpha_n}^{(p)}(x)}.$$

The choice of p is discussed in Girard et al. (2019) using the MSE of (the unconditional version of) $\tilde{\gamma}_{\alpha_n}^{(p)}(x)$ as a criterion. Cross-validation choices of the bandwidth h_n and intermediate quantile level α_n , meanwhile, are discussed in Daouia et al. (2013); Girard et al. (2021). For the sake of simplicity, we choose here common parameters $p = 1.7$ following the guidelines of Girard et al. (2019)), $h_n = 0.15$ and $\alpha_n = 1 - 1/\sqrt{n} \approx 0.968$ across all replications and K is the Epanechnikov kernel defined by $K(t) = 0.75(1 - t^2)\mathbb{1}_{\{|t| < 1\}}$. Results are shown in Fig. 1.

We can notice that an indirect estimation of extreme quantiles or expectiles with a L^p -quantile (with p between 1 and 2) leads to a trade-off between bias and variance:

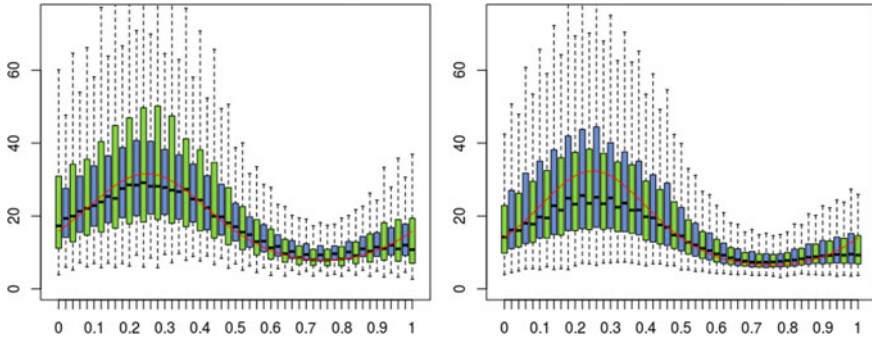


Fig. 1 Left: Boxplots of 500 estimates of $q^{(1)}(\beta_n|x)$ with the direct (green) and indirect (blue) quantile estimators. Right: Boxplots of 500 estimates of $q^{(2)}(\beta_n|x)$ with the direct (green) and indirect (blue) expectile estimators. True values are in red

the indirect L^p –estimator of an extreme regression quantile is less variable than the direct estimator but slightly more biased, and the indirect L^p –estimator of an extreme regression expectile is more variable than the direct estimator but less biased. For conditional quantiles, an explanation is that using the asymptotic approximation (9) in the construction of the indirect estimator adds a source of bias, while the reduced variance stems from the use of $p = 1.7$ in the estimator $\tilde{\gamma}_{\alpha_n}^{(p)}(x)$, providing an estimator with lower variance compared to the simple Hill estimator in our case (see Girard et al. 2019). The case of conditional expectiles is less clear, although the increased variability observed for $x \in [0, 0.5]$ seems to originate in the use of the estimated constant $B(2, \tilde{\gamma}_{\alpha_n}^{(p)}(x)^{-1} - 1)/B(p, \tilde{\gamma}_{\alpha_n}^{(p)}(x)^{-1} - p + 1)$: when $\tilde{\gamma}_{\alpha_n}^{(p)}(x)$ gets close to 1, which is sometimes the case in this zone where $\gamma(x) \in [0.4, 0.5]$, this estimated constant tends to explode, while the direct estimator is less affected. A similar observation, in the context of extreme Wang distortion risk measure estimation, is made by El Methni and Stupfler (2017).

5 Real Data Example

We study here a data set on motorcycle insurance, collected from the former Swedish insurance provider Wasa. Our data is on motorcycle insurance policies and claims over the period 1994–1998 and is available from www.math.su.se/GLMbook or the R packages `insuranceData` and `CASdatasets`, and analyzed in Ohlsson and Johansson (2010). We concentrate here on the relationship between the claim severity Y (defined as the ratio of claim cost by number of claims for each given policyholder) in Swedish kroner (SEK), and the number of years X of exposure of a policyholder. Data for $X > 3$ are very sparse, so we restrict our attention to the case $Y > 0$ and $X \in [0, 3]$, resulting in $n = 593$ pairs (X_i, Y_i) .

Our goal in this section is to estimate extreme conditional quantiles and expectiles of Y given X , at a level $\beta_n = 1 - 3/n \approx 0.9949$. This level is slightly less extreme than the more standard $\beta_n = 1 - 1/n \approx 0.9985$, but is an appropriately extreme level in this conditional context where less data are available locally for the estimation. A preliminary diagnostic using a local version of the Hill estimator (which we do not show here) suggests that the data is indeed heavy-tailed with $\gamma(x) \in [0.25, 0.6]$. Following again the guidelines in Girard et al. (2019), we choose $p = 1.7$ for our indirect extreme conditional quantile and expectile estimators. These are, respectively, compared to

- the estimator $\hat{q}_n^W(\beta_n|x)$ of Girard et al. (2021), calculated as in Sect. 5 therein, and our direct quantile estimator presented in Sect. 4 (i),
- the estimator $\hat{e}_n^{W,BR}(\beta_n|x)$ of Girard et al. (2021), calculated as in Sect. 5 therein, and our direct expectile estimator presented in Sect. 4 (i).

For the direct and indirect estimators presented in Sect. 4 (ii)–(iii), the parameters α_n and h_n are chosen by a cross-validation procedure analogous to that of Girard et al. (2021). The Epanechnikov kernel is adopted. Results are given in Fig. 2. In each case, all three estimators reassuringly point to roughly the same results, with slight differences; in particular, for quantile estimation and when data is scarce, the direct estimator in Sect. 4 (i) appears to be more sensitive to the local shape of the tail than the indirect, L^p -quantile-based estimator in Sect. 4 (ii), resulting in less stable estimates.

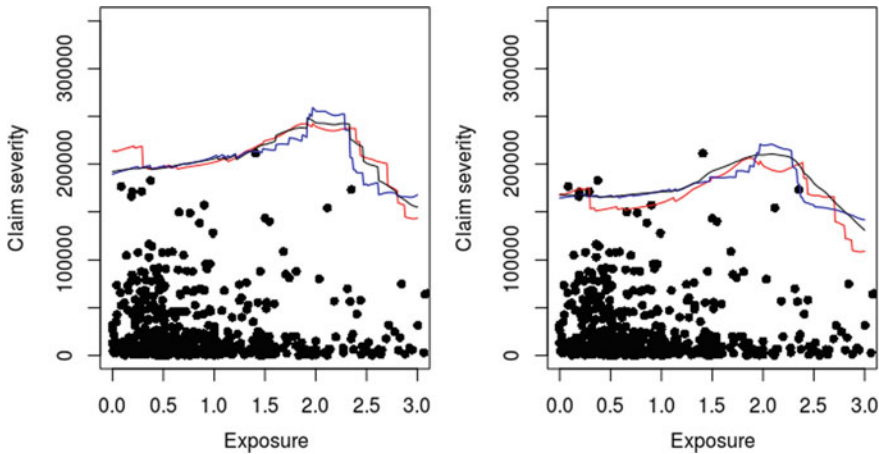


Fig. 2 Swedish motorcycle insurance data. Left panel: extreme conditional quantile estimation, black curve: estimator $\hat{q}_n^W(\beta_n|x)$ of Girard et al. (2021), blue curve: direct quantile estimator (i) of Sect. 4, red curve: indirect quantile estimator (ii) of Sect. 4. Right panel: extreme conditional expectile estimation, black curve: estimator $\hat{e}_n^{W,BR}(\beta_n|x)$ of Girard et al. (2021), blue curve: direct expectile estimator (i) of Sect. 4, red curve: indirect expectile estimator (iii) of Sect. 4. In each panel, x-axis: number of years of exposure of policyholder, y-axis: claim severity

6 Appendix

6.1 Preliminary Results

Lemma 1 Assume that (\mathcal{L}) and $\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot|\mathbf{x}))$ hold, and let $y_n \rightarrow \infty$ and $h_n \rightarrow 0$ be such that $\omega_{h_n}^{(1)}(y_n|\mathbf{x}) \log(y_n) \rightarrow 0$ and $\omega_{h_n}^{(2)}(y_n|\mathbf{x}) \rightarrow 0$. Then for all $0 \leq k < \gamma(\mathbf{x})^{-1}$ we have, uniformly in $\mathbf{x}' \in B(\mathbf{x}, h_n)$,

$$m^{(k)}(y_n|\mathbf{x}') = m^{(k)}(y_n|\mathbf{x}) \left(1 + O(h_n) + o\left(\omega_{h_n}^{(1)}(y_n|\mathbf{x})\right) + O\left(\omega_{h_n}^{(2)}(y_n|\mathbf{x})\right) \right).$$

In particular $m^{(k)}(y_n|\mathbf{x}') = y_n^k g(\mathbf{x}) (1 + o(1))$ uniformly in $\mathbf{x}' \in B(\mathbf{x}, h_n)$.

Proof Let us first write

$$m^{(k)}(y_n|\mathbf{x}) = \mathbb{E} \left[(Y - y_n)^k \mathbb{1}_{\{Y > y_n\}} | \mathbf{X} = \mathbf{x} \right] g(\mathbf{x}) + \mathbb{E} \left[(y_n - Y)^k \mathbb{1}_{\{Y \leq y_n\}} | \mathbf{X} = \mathbf{x} \right] g(\mathbf{x}).$$

By the arguments of the proof of Lemma 3 in Girard et al. (2021),

$$\frac{\mathbb{E} \left[(Y - y_n)^k \mathbb{1}_{\{Y > y_n\}} | \mathbf{X} = \mathbf{x}' \right] g(\mathbf{x}')}{\mathbb{E} \left[(Y - y_n)^k \mathbb{1}_{\{Y > y_n\}} | \mathbf{X} = \mathbf{x} \right] g(\mathbf{x})} = 1 + O(h_n) + O\left(\omega_{h_n}^{(1)}(y_n|\mathbf{x}) \log(y_n)\right).$$

Besides, an integration by parts yields

$$\mathbb{E} \left[(y_n - Y)^k \mathbb{1}_{\{Y \leq y_n\}} | \mathbf{X} = \mathbf{x} \right] = \int_0^{y_n} kt^{k-1} F^{(1)}(y_n - t | \mathbf{x}) dt.$$

It clearly follows that

$$\left| \mathbb{E} \left[(y_n - Y)^k \mathbb{1}_{\{Y \leq y_n\}} | \mathbf{X} = \mathbf{x}' \right] - \mathbb{E} \left[(y_n - Y)^k \mathbb{1}_{\{Y \leq y_n\}} | \mathbf{X} = \mathbf{x} \right] \right| \leq y_n^k \omega_{h_n}^{(2)}(y_n|\mathbf{x}).$$

Now

$$\mathbb{E} \left[(y_n - Y)^k \mathbb{1}_{\{Y \leq y_n\}} | \mathbf{X} = \mathbf{x} \right] = y_n^k \mathbb{E} \left[\left(1 - \frac{Y}{y_n} \right)^k \mathbb{1}_{\{Y \leq y_n\}} | \mathbf{X} = \mathbf{x} \right] = y_n^k (1 + o(1))$$

by the dominated convergence theorem, and

$$\mathbb{E} \left[(Y - y_n)^k \mathbb{1}_{\{Y > y_n\}} | \mathbf{X} = \mathbf{x} \right] = \frac{g(\mathbf{x}) B(k+1, \gamma(\mathbf{x})^{-1} - k)}{\gamma(\mathbf{x})} y_n^k \bar{F}^{(1)}(y_n|\mathbf{x}) (1 + o(1)), \quad (15)$$

see for instance Lemma 1(i) in Daouia et al. (2019). The result follows from direct calculations.

Lemma 2 Assume that (\mathcal{K}) , (\mathcal{L}) and $\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot|\mathbf{x}))$ hold, and let $y_n \rightarrow \infty$ and $h_n \rightarrow 0$ be such that $nh_n^d \rightarrow \infty$, $\omega_{h_n}^{(1)}(y_n|\mathbf{x}) \log(y_n) \rightarrow 0$ and $\omega_{h_n}^{(2)}(y_n|\mathbf{x}) \rightarrow 0$. Then for all $0 \leq k < \gamma(\mathbf{x})^{-1}/2$,

$$\mathbb{E} \left[\hat{m}_n^{(k)}(y_n|\mathbf{x}) \right] = m^{(k)}(y_n|\mathbf{x}) \left(1 + O(h_n) + o\left(\omega_{h_n}^{(1)}(y_n|\mathbf{x})\right) + O\left(\omega_{h_n}^{(2)}(y_n|\mathbf{x})\right) \right)$$

and $\text{Var} \left[\hat{m}_n^{(k)}(y_n|\mathbf{x}) \right] = \frac{\|K\|_2^2}{nh_n^d} g(\mathbf{x}) y_n^{2k} (1 + o(1)).$

Proof Note that $\mathbb{E} \left[\hat{m}_n^{(k)}(y_n|\mathbf{x}) \right] = \int_S m^{(k)}(y_n|\mathbf{x} - \mathbf{u}h_n) K(\mathbf{u}) d\mathbf{u}$ by Assumption (\mathcal{K}) and a change of variables, and use Lemma 1 to get the first result. The second result is obtained through similar calculations. \square

Lemma 3 Assume that (\mathcal{K}) , (\mathcal{L}) and $\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot|\mathbf{x}))$ hold. Let $y_n \rightarrow \infty$, $h_n \rightarrow 0$ be such that $nh_n^d \rightarrow \infty$ and $\omega_{h_n}^{(1)}(y_n|\mathbf{x}) \log(y_n) \rightarrow 0$. Then for all $0 \leq k < \gamma(\mathbf{x})^{-1}/2$,

$$\begin{cases} \mathbb{E} \left[\hat{\varphi}_n^{(k)}(y_n|\mathbf{x}) \right] = \varphi^{(k)}(y_n|\mathbf{x}) \left(1 + O(h_n) + O\left(\omega_{h_n}^{(1)}(y_n|\mathbf{x}) \log(y_n)\right) \right), \\ \text{Var} \left[\hat{\varphi}_n^{(k)}(y_n|\mathbf{x}) \right] = \|K\|_2^2 g(\mathbf{x}) \frac{B(2k+1, \gamma(\mathbf{x})^{-1}-2k)}{\gamma(\mathbf{x})} \frac{y_n^{2k} \bar{F}^{(1)}(y_n|\mathbf{x})}{nh_n^d} (1 + o(1)). \end{cases}$$

Proof See Lemma 5 of Girard et al. (2021).

Lemma 4 Assume that $\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot|\mathbf{x}))$ holds. Let $\lambda \geq 1$, $y_n \rightarrow \infty$, $y'_n = \lambda y_n (1 + o(1))$ and $0 < k < \gamma(\mathbf{x})^{-1}$.

(i) Then the following asymptotic relationship holds:

$$\begin{aligned} & \mathbb{E} \left[|Y - y_n|^k \mathbb{1}_{\{Y > y'_n\}} | \mathbf{X} = \mathbf{x} \right] \\ &= y_n^k \bar{F}^{(1)}(y_n|\mathbf{x}) \left[kIB(\lambda^{-1}, \gamma(\mathbf{x})^{-1} - k, k) + (\lambda - 1)^k \lambda^{-1/\gamma(\mathbf{x})} \right] (1 + o(1)). \end{aligned}$$

\square

(ii) Assume further that $\omega_{h_n}^{(1)}(y_n \wedge y'_n|\mathbf{x}) \log(y_n) \rightarrow 0$ and $\omega_{h_n}^{(3)}(y_n|\mathbf{x}) \rightarrow 0$. Then, uniformly in $\mathbf{x}' \in B(\mathbf{x}, h_n)$,

$$\mathbb{E} \left[|Y - y_n|^k \mathbb{1}_{\{Y > y'_n\}} | \mathbf{X} = \mathbf{x}' \right] = \mathbb{E} \left[|Y - y_n|^k \mathbb{1}_{\{Y > y'_n\}} | \mathbf{X} = \mathbf{x} \right] (1 + o(1)).$$

Proof (i) Straightforward calculations entail

$$\begin{aligned} & \mathbb{E} [|Y - y_n|^k \mathbb{1}_{\{Y > y'_n\}} | \mathbf{X} = \mathbf{x}] \\ &= y_n^k \mathbb{E} \left[\left\{ \left(\frac{Y}{y_n} - 1 \right)^k - (\lambda - 1)^k \right\} \mathbb{1}_{\{Y > \lambda y_n\}} | \mathbf{X} = \mathbf{x} \right] (1 + o(1)) \\ &+ y_n^k (\lambda - 1)^k \bar{F}^{(1)}(\lambda y_n | \mathbf{x}) (1 + o(1)), \end{aligned}$$

with $y'_n = \lambda y_n (1 + o(1))$. The result then comes directly from the regular variation property of $\bar{F}^{(1)}(\cdot | \mathbf{x})$ and Lemma 1 in Daouia et al. (2019) with $H(t) = (t - 1)^k$ and $b = \lambda$.

(ii) Note first that for n large enough

$$\begin{aligned} & \left| \mathbb{E} [|Y - y_n|^k \mathbb{1}_{\{Y > y'_n\}} | \mathbf{X} = \mathbf{x}'] - \mathbb{E} [|Y - y_n|^k \mathbb{1}_{\{Y > \lambda y_n\}} | \mathbf{X} = \mathbf{x}'] \right| \\ & \leq [|y'_n - y_n|^k + (\lambda - 1)^k y_n^k] [\bar{F}^{(1)}(y'_n \wedge \lambda y_n | \mathbf{x}') - \bar{F}^{(1)}(y'_n \vee \lambda y_n | \mathbf{x}')] \\ & \leq 3(\lambda - 1)^k y_n^k \times \bar{F}^{(1)}(y'_n | \mathbf{x}') \times \omega_{h_n}^{(3)}(y_n | \mathbf{x}). \end{aligned}$$

Write $(Y - y_n)^k = ((Y - y_n)^k - (\lambda - 1)^k y_n^k) + (\lambda - 1)^k y_n^k$. It then follows from the assumption $\omega_{h_n}^{(3)}(y_n | \mathbf{x}) \rightarrow 0$ that, uniformly in $\mathbf{x}' \in B(\mathbf{x}, h_n)$,

$$\begin{aligned} \mathbb{E} [|Y - y_n|^k \mathbb{1}_{\{Y > y'_n\}} | \mathbf{X} = \mathbf{x}'] &= (\lambda - 1)^k y_n^k \bar{F}^{(1)}(y'_n | \mathbf{x}') (1 + o(1)) \\ &+ k \int_{\lambda y_n}^{\infty} (z - y_n)^{k-1} \bar{F}^{(1)}(z | \mathbf{x}') dz (1 + o(1)). \end{aligned}$$

Remark now $\bar{F}^{(1)}(y'_n | \mathbf{x}') (y'_n)^{-\omega_{h_n}^{(1)}(y'_n | \mathbf{x}')} \leq \bar{F}^{(1)}(y'_n | \mathbf{x}') \leq \bar{F}^{(1)}(y'_n | \mathbf{x}') (y'_n)^{\omega_{h_n}^{(1)}(y'_n | \mathbf{x}')}.$ Then condition $\omega_{h_n}^{(1)}(y'_n | \mathbf{x}') \log(y_n) \rightarrow 0$ entails, uniformly in $\mathbf{x}' \in B(\mathbf{x}, h_n)$, $\bar{F}^{(1)}(y'_n | \mathbf{x}') = \bar{F}^{(1)}(y'_n | \mathbf{x}') (1 + o(1)) = \bar{F}^{(1)}(\lambda y_n | \mathbf{x}') (1 + o(1)).$ Besides, for any $z \geq \lambda y_n \geq y_n$, $\bar{F}^{(1)}(z | \mathbf{x}') z^{-\omega_{h_n}^{(1)}(y_n | \mathbf{x}')} \leq \bar{F}^{(1)}(z | \mathbf{x}') \leq \bar{F}^{(1)}(z | \mathbf{x}') z^{\omega_{h_n}^{(1)}(y_n | \mathbf{x}')}.$ Following the proof of Lemma 3 in Girard et al. (2021), we get, uniformly in $\mathbf{x}' \in B(\mathbf{x}, h_n)$,

$$\left| \frac{\int_{\lambda y_n}^{\infty} (z - y_n)^{k-1} \bar{F}^{(1)}(z | \mathbf{x}') dz}{\int_{\lambda y_n}^{\infty} (z - y_n)^{k-1} \bar{F}^{(1)}(z | \mathbf{x}) dz} - 1 \right| = O(\omega_{h_n}^{(1)}(y_n | \mathbf{x}) \log(y_n)) \rightarrow 0.$$

Since $\int_{\lambda y_n}^{\infty} (z - y_n)^{k-1} \bar{F}^{(1)}(z | \mathbf{x}) dz$ is of order $y_n^k \bar{F}^{(1)}(y_n | \mathbf{x})$ (by regular variation of $\bar{F}^{(1)}(\cdot | \mathbf{x})$), the conclusion follows.

Lemma 5 Assume that $\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot | \mathbf{x}))$ holds. For all $1 \leq p < \gamma(\mathbf{x})^{-1} + 1$,

$$\frac{\bar{F}^{(p)}(y | \mathbf{x})}{\bar{F}^{(1)}(y | \mathbf{x})} = \frac{B(p, \gamma(\mathbf{x})^{-1} - p + 1)}{\gamma(\mathbf{x})} [1 + r(y | \mathbf{x})]$$

where there are constants $C_1(\mathbf{x}), C_2(\mathbf{x}), C_3(\mathbf{x})$ such that

$$r(y|\mathbf{x}) = C_1(\mathbf{x}) \frac{\mathbb{E}(Y \mathbb{1}_{\{0 < Y < y\}} | \mathbf{X} = \mathbf{x})}{y} (1 + o(1)) + C_2(\mathbf{x}) \bar{F}^{(1)}(y|\mathbf{x})(1 + o(1)) + C_3(\mathbf{x}) A(1/\bar{F}^{(1)}(y|\mathbf{x})|\mathbf{x})(1 + o(1)) \text{ as } y \rightarrow \infty.$$

Similarly

$$\frac{q^{(p)}(\alpha|\mathbf{x})}{q^{(1)}(\alpha|\mathbf{x})} = \left(\frac{\gamma(\mathbf{x})}{B(p, \gamma(\mathbf{x})^{-1} - p + 1)} \right)^{-\gamma(\mathbf{x})} [1 + R(\alpha|\mathbf{x})]$$

where there are constants $D_1(\mathbf{x})$, $D_2(\mathbf{x})$, $D_3(\mathbf{x})$ such that

$$R(\alpha|\mathbf{x}) = D_1(\mathbf{x}) \frac{\mathbb{E}(Y \mathbb{1}_{\{0 < Y < q^{(1)}(\alpha|\mathbf{x})\}} | \mathbf{X} = \mathbf{x})}{q^{(1)}(\alpha|\mathbf{x})} (1 + o(1)) + D_2(\mathbf{x})(1 - \alpha)(1 + o(1)) + D_3(\mathbf{x}) A((1 - \alpha)^{-1}|\mathbf{x})(1 + o(1)) \text{ as } \alpha \rightarrow 1.$$

□

Proof We start by focusing on the ratio $\bar{F}^{(p)}(y|\mathbf{x})/\bar{F}^{(1)}(y|\mathbf{x})$. By Lemma 1 in Girard et al. (2019), the function $\bar{F}^{(p)}(\cdot|\mathbf{x})$ is continuous and strictly decreasing on the support of Y given $\mathbf{X} = \mathbf{x}$. It is therefore enough to show the announced formula for $y = q^{(p)}(\alpha|\mathbf{x})$ with $\alpha \rightarrow 1$; this, in turn, is a simple corollary of Proposition 2 in Daouia et al. (2019). To show the analogous formula on $q^{(p)}(\alpha|\mathbf{x})/q^{(1)}(\alpha|\mathbf{x})$, we define $U^{(1)}(t|\mathbf{x}) = q^{(1)}(1 - t^{-1}|\mathbf{x})$; $U^{(1)}(\cdot|\mathbf{x})$ also satisfies a (local uniform) second-order regular variation condition, see Theorem 2.3.9 p.48 in de Haan and Ferreira (2006). Consequently, we note that the asymptotic expansion on $\bar{F}^{(p)}(y|\mathbf{x})/\bar{F}^{(1)}(y|\mathbf{x})$ entails a similar expansion on

$$\frac{U^{(1)}(1/\bar{F}^{(1)}(y|\mathbf{x})|\mathbf{x})}{U^{(1)}(1/\bar{F}^{(p)}(y|\mathbf{x})|\mathbf{x})} = \frac{y}{q^{(1)}(F^{(p)}(y|\mathbf{x}))} (1 + o(A(1/\bar{F}^{(1)}(y|\mathbf{x})|\mathbf{x})))$$

as $y \rightarrow \infty$, with different constants (here Lemma 1 in Daouia et al. (2020b) was used). Setting $y = q^{(p)}(\alpha|\mathbf{x})$, with $\alpha \rightarrow 1$, gives the announced result.

Lemma 6 Assume that (\mathcal{K}) , (\mathcal{L}) and $\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot|\mathbf{x}))$ hold. Let $y_n \rightarrow \infty$, $h_n \rightarrow 0$ and $z_n = \theta y_n(1 + o(1))$, where $\theta > 0$. Assume further that $\epsilon_n^{-2} = nh_n^d \bar{F}^{(1)}(y_n|\mathbf{x}) \rightarrow \infty$, $nh_n^{d+2} \bar{F}^{(1)}(y_n|\mathbf{x}) \rightarrow 0$, there exists $\delta \in (0, 1)$ such that $\epsilon_n^{-1} \omega_{h_n}^{(1)}((1 - \delta)(\theta \wedge 1)y_n|\mathbf{x}) \log(y_n) \rightarrow 0$, and $\omega_{h_n}^{(3)}(z_n|\mathbf{x}) \rightarrow 0$. Letting, for all $j \in \{1, \dots, J\}$, $y_{n,j} = \tau_j^{-\gamma(\mathbf{x})} y_n(1 + o(1))$ with $0 < \tau_1 < \tau_2 < \dots < \tau_J \leq 1$, and $p \in (1, \gamma(\mathbf{x})^{-1}/2 + 1)$, one has

$$\epsilon_n^{-1} \left\{ \left(\frac{\hat{\varphi}_n^{(0)}(y_{n,j}|\mathbf{x})}{\varphi^{(0)}(y_{n,j}|\mathbf{x})} - 1 \right)_{1 \leq j \leq J}, \left(\frac{\hat{\varphi}_n^{(p-1)}(z_n|\mathbf{x})}{\varphi^{(p-1)}(z_n|\mathbf{x})} - 1 \right) \right\} \xrightarrow{d} \mathcal{N} \left(\mathbf{0}_{J+1}, \frac{\|K\|_2^2}{g(\mathbf{x})} \mathbf{\Lambda}(\mathbf{x}) \right),$$

where $\mathbf{\Lambda}(\mathbf{x})$ is a symmetric matrix having entries:

$$\left\{ \begin{array}{l} \Lambda_{j,\ell}(\mathbf{x}) = \\ \Lambda_{j,J+1}(\mathbf{x}) = \\ \Lambda_{J+1,J+1}(\mathbf{x}) = \end{array} \right. \frac{(\tau_j \vee \tau_\ell)^{-1}}{\gamma(\mathbf{x})} \frac{(p-1)B \left(\left(1 \vee \frac{\tau_j}{\theta} \right)^{-1}, \gamma(\mathbf{x})^{-1-p+1, p-1} \right) + \left(1 \vee \frac{\tau_j}{\theta} - 1 \right)^{p-1} \left(1 \vee \frac{\tau_j}{\theta} \right)^{-1/\gamma(\mathbf{x})}}{\tau_j B(p, \gamma(\mathbf{x})^{-1-p+1})} \cdot \quad (16)$$

$$\frac{\tau_j B(p, \gamma(\mathbf{x})^{-1-p+1})}{\gamma(\mathbf{x}) \frac{B(2p-1, \gamma(\mathbf{x})^{-1-2p+2})}{B(p, \gamma(\mathbf{x})^{-1-p+1})^2} \theta^{1/\gamma(\mathbf{x})}}$$

Proof Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J, \beta_{J+1}) \in \mathbb{R}^{J+1}$. Set

$$\mathcal{Z}_n = \epsilon_n^{-1} \sum_{j=1}^J \beta_j \left(\frac{\hat{\varphi}_n^{(0)}(y_{n,j}|\mathbf{x})}{\varphi^{(0)}(y_{n,j}|\mathbf{x})} - 1 \right) + \epsilon_n^{-1} \beta_{J+1} \left(\frac{\hat{\varphi}_n^{(p-1)}(z_n|\mathbf{x})}{\varphi^{(p-1)}(z_n|\mathbf{x})} - 1 \right).$$

Clearly $\omega_{h_n}^{(1)}(y_{n,j}|\mathbf{x}) \leq \omega_{h_n}^{(1)}((1-\delta)y_n|\mathbf{x})$ and $\omega_{h_n}^{(1)}(z_n|\mathbf{x}) \leq \omega_{h_n}^{(1)}((1-\delta)\theta y_n|\mathbf{x})$ for n large enough. Lemma 3 then provides $\mathbb{E}(\mathcal{Z}_n) = o(1)$. It thus remains to focus on the asymptotic distribution of the centered variable $Z_n = \mathcal{Z}_n - \mathbb{E}(\mathcal{Z}_n)$. Note that $\mathbb{V}\text{ar}[Z_n] = \epsilon_n^{-2} \boldsymbol{\beta}^\top \mathbf{B}^{(n)} \boldsymbol{\beta}$, where $\mathbf{B}^{(n)}$ is the symmetric matrix having entries:

$$\left\{ \begin{array}{l} B_{j,\ell}^{(n)}(\mathbf{x}) = \frac{\text{cov}(\hat{\varphi}_n^{(0)}(y_{n,j}|\mathbf{x}), \hat{\varphi}_n^{(0)}(y_{n,\ell}|\mathbf{x}))}{\varphi^{(0)}(y_{n,j}|\mathbf{x})\varphi^{(0)}(y_{n,\ell}|\mathbf{x})}, \quad j, \ell \in \{1, \dots, J\}, \quad j \leq \ell, \\ B_{j,J+1}^{(n)}(\mathbf{x}) = \frac{\text{cov}(\hat{\varphi}_n^{(0)}(y_{n,j}|\mathbf{x}), \hat{\varphi}_n^{(p-1)}(z_n|\mathbf{x}))}{\varphi^{(0)}(y_{n,j}|\mathbf{x})\varphi^{(p-1)}(z_n|\mathbf{x})}, \quad j \in \{1, \dots, J\}, \\ B_{J+1,J+1}^{(n)}(\mathbf{x}) = \frac{\mathbb{V}\text{ar}[\hat{\varphi}_n^{(p-1)}(z_n|\mathbf{x})]}{\varphi^{(p-1)}(z_n|\mathbf{x})^2}. \end{array} \right.$$

We recall $z_n = \theta y_n(1 + o(1))$, hence $\bar{F}^{(1)}(z_n|\mathbf{x}) = \theta^{-1/\gamma(\mathbf{x})} \bar{F}^{(1)}(y_n|\mathbf{x})(1 + o(1))$ and Lemma 3 combined with Eq. (15) immediately gives

$$B_{J+1,J+1}^{(n)}(\mathbf{x}) = \frac{\|K\|_2^2}{g(\mathbf{x})} \gamma(\mathbf{x}) \frac{B(2p-1, \gamma(\mathbf{x})^{-1} - 2p+2)}{B(p, \gamma(\mathbf{x})^{-1} - p+1)^2} \theta^{1/\gamma(\mathbf{x})} \epsilon_n^2 (1 + o(1)).$$

The calculation of $B_{j,\ell}^{(n)}(\mathbf{x})$ gives, through straightforward calculations and the use of Lemma 3 and Eq. (15),

$$B_{j,\ell}^{(n)}(\mathbf{x}) = \frac{\|K\|_2^2}{nh_n^d} \frac{\bar{F}^{(1)}(y_{n,j} \vee y_{n,\ell}|\mathbf{x})}{g(\mathbf{x}) \bar{F}^{(1)}(y_{n,j}|\mathbf{x}) \bar{F}^{(1)}(y_{n,\ell}|\mathbf{x})} (1 + o(1)).$$

The regular variation property of $\bar{F}^{(1)}$ gives $B_{j,\ell}^{(n)}(\mathbf{x}) = \frac{\|K\|_2^2}{g(\mathbf{x})} (\tau_j \vee \tau_\ell)^{-1} \epsilon_n^2 (1 + o(1))$. It remains to calculate $B_{j,J+1}^{(n)}(\mathbf{x})$. Using Eq. (15), with $Q(\cdot) = K(\cdot)^2 / \|K\|_2^2$ a kernel satisfying (K) , this term equals

$$\frac{\frac{1}{nh_n^{2d}} \|K\|_2^2 \mathbb{E} \left[|Y - z_n|^{p-1} Q \left(\frac{\mathbf{x} - \mathbf{X}}{h_n} \right) \mathbb{1}_{\{Y > z_n \vee y_{n,j}\}} \right]}{g(\mathbf{x})^2 B(p, \gamma(\mathbf{x})^{-1} - p + 1) z_n^{p-1} \bar{F}^{(1)}(y_{n,j} | \mathbf{x}) \bar{F}^{(1)}(z_n | \mathbf{x}) / \gamma(\mathbf{x}) (1 + o(1))} - \frac{\frac{1}{n} \mathbb{E} \left[\frac{1}{h_n^d} K \left(\frac{\mathbf{x} - \mathbf{X}}{h_n} \right) \mathbb{1}_{\{Y > y_{n,j}\}} \right] \mathbb{E} \left[|Y - z_n|^{p-1} \frac{1}{h_n^d} K \left(\frac{\mathbf{x} - \mathbf{X}}{h_n} \right) \mathbb{1}_{\{Y > z_n\}} \right]}{g(\mathbf{x})^2 B(p, \gamma(\mathbf{x})^{-1} - p + 1) z_n^{p-1} \bar{F}^{(1)}(y_{n,j} | \mathbf{x}) \bar{F}^{(1)}(z_n | \mathbf{x}) / \gamma(\mathbf{x}) (1 + o(1))}.$$

Clearly, as a direct consequence of Lemma 3, the first term dominates. Remark that $z_n \vee y_{n,j} = (1 \vee \tau_j^{-\gamma(\mathbf{x})} / \theta) z_n (1 + o(1))$ and combine Assumption (\mathcal{K}) , the results of Lemma 4 (with $\lambda = (1 \vee \tau_j^{-\gamma(\mathbf{x})} / \theta)$), and the regular variation property of $\varphi^{(k)}(\cdot)$ (see Eq. (15)) to find that the numerator of this first term is asymptotically equivalent to

$$\frac{\|K\|_2^2}{nh_n^d} g(\mathbf{x}) z_n^{p-1} \bar{F}^{(1)}(z_n | \mathbf{x}) \left[(p-1) IB \left((1 \vee \tau_j^{-\gamma(\mathbf{x})} / \theta)^{-1}, \gamma(\mathbf{x})^{-1} - p + 1, p-1 \right) + ((1 \vee \tau_j^{-\gamma(\mathbf{x})} / \theta) - 1)^{p-1} \left(1 \vee \tau_j^{-\gamma(\mathbf{x})} / \theta \right)^{-1/\gamma(\mathbf{x})} \right].$$

And finally $B_{j,J+1}^{(n)}(\mathbf{x})$ is asymptotically equivalent to

$$\frac{\tau_j^{-1} \gamma(\mathbf{x}) \frac{\|K\|_2^2}{g(\mathbf{x})} \epsilon_n^2}{B(p, \gamma(\mathbf{x})^{-1} - p + 1)} \left[(p-1) IB \left((1 \vee \tau_j^{-\gamma(\mathbf{x})} / \theta)^{-1}, \gamma(\mathbf{x})^{-1} - p + 1, p-1 \right) + ((1 \vee \tau_j^{-\gamma(\mathbf{x})} / \theta) - 1)^{p-1} \left(1 \vee \tau_j^{-\gamma(\mathbf{x})} / \theta \right)^{-1/\gamma(\mathbf{x})} \right].$$

Therefore, $\text{Var}[Z_n] = \|K\|_2^2 \boldsymbol{\beta}^\top \boldsymbol{\Lambda}(\mathbf{x}) \boldsymbol{\beta} / g(\mathbf{x}) (1 + o(1))$, where $\boldsymbol{\Lambda}(\mathbf{x})$ is given in Eq. (16). It remains to prove the asymptotic normality of Z_n . For that purpose, we denote $Z_n = \sum_{i=1}^n Z_{i,n}$, where

$$Z_{i,n} = \frac{\epsilon_n^{-1}}{nh_n^d} \sum_{j=1}^J \beta_j \frac{K \left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n} \right) \mathbb{1}_{\{Y_i > y_{n,j}\}} - \mathbb{E} \left[K \left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n} \right) \mathbb{1}_{\{Y_i > y_{n,j}\}} \right]}{\varphi^{(0)}(y_{n,j} | \mathbf{x})} + \frac{\epsilon_n^{-1}}{nh_n^d} \beta_{J+1} \frac{|Y_i - z_n|^{p-1} K \left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n} \right) \mathbb{1}_{\{Y_i > z_n\}} - \mathbb{E} \left[|Y_i - z_n|^{p-1} K \left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n} \right) \mathbb{1}_{\{Y_i > z_n\}} \right]}{\varphi^{(p-1)}(z_n | \mathbf{x})}.$$

Taking $\delta > 0$ sufficiently small and arguing as in the closing stages of the proof of Lemma 6 in Girard et al. (2021), we find that $n \mathbb{E} [|Z_{1,n}|^{2+\delta}] = O(\epsilon_n^\delta) = o(1)$. Applying the classical Lyapunov central limit theorem concludes the proof. \square

Proposition 1 *Assume that (\mathcal{K}) , (\mathcal{L}) and $\mathcal{C}_2(\gamma(\mathbf{x}), \rho(\mathbf{x}), A(\cdot | \mathbf{x}))$ hold. Let $y_n \rightarrow \infty$, $h_n \rightarrow 0$ and $z_n = \theta y_n (1 + o(1))$, where $\theta > 0$. Assume further that $\epsilon_n^{-2} = nh_n^d \bar{F}^{(1)}(y_n | \mathbf{x}) \rightarrow \infty$, $nh_n^{d+2} \bar{F}^{(1)}(y_n | \mathbf{x}) \rightarrow 0$, $\omega_{h_n}^{(3)}(y_n | \mathbf{x}) \rightarrow 0$ and there exists $\delta \in$*

$(0, 1)$ such that $\epsilon_n^{-1} \omega_{h_n}^{(1)}((1 - \delta)(\theta \wedge 1)y_n|\mathbf{x}) \log(y_n) \rightarrow 0$. If, for all $j \in \{1, \dots, J\}$, the $y_{n,j} = \tau_j^{-\gamma(\mathbf{x})} y_n(1 + o(1))$ with $0 < \tau_1 < \tau_2 < \dots < \tau_J \leq 1$ are such that $\epsilon_n^{-1} \omega_{h_n}^{(2)}((1 + \delta)(\theta \vee \tau_1^{-\gamma(\mathbf{x})})y_n|\mathbf{x}) \rightarrow 0$, then, for all $p \in (1, \gamma(\mathbf{x})^{-1}/2 + 1)$, one has

$$\epsilon_n^{-1} \left\{ \left(\frac{\hat{F}_n^{(1)}(y_{n,j}|\mathbf{x})}{\bar{F}^{(1)}(y_{n,j}|\mathbf{x})} - 1 \right)_{1 \leq j \leq J}, \left(\frac{\hat{F}_n^{(p)}(z_n|\mathbf{x})}{\bar{F}^{(p)}(z_n|\mathbf{x})} - 1 \right) \right\} \xrightarrow{d} \mathcal{N} \left(\mathbf{0}_{J+1}, \frac{\|K\|_2^2}{g(\mathbf{x})} \mathbf{\Lambda}(\mathbf{x}) \right),$$

where $\mathbf{\Lambda}(\mathbf{x})$ is given in Eq. (16).

Proof Notice that

$$\frac{\hat{F}_n^{(p)}(u_n|\mathbf{x})}{\bar{F}^{(p)}(u_n|\mathbf{x})} - 1 = \left(\frac{\hat{\varphi}_n^{(p-1)}(u_n|\mathbf{x})}{\varphi^{(p-1)}(u_n|\mathbf{x})} - 1 \right) \frac{m^{(p-1)}(u_n|\mathbf{x})}{\hat{m}_n^{(p-1)}(u_n|\mathbf{x})} + \left(\frac{m^{(p-1)}(u_n|\mathbf{x})}{\hat{m}_n^{(p-1)}(u_n|\mathbf{x})} - 1 \right).$$

Lemma 2 and the Chebyshev inequality ensure that for all $p \in (1, \gamma(\mathbf{x})^{-1}/2 + 1)$ and $u_n \in \{y_{n,1}, \dots, y_{n,J}, z_n\}$, $\hat{m}_n^{(p-1)}(u_n|\mathbf{x})/m^{(p-1)}(u_n|\mathbf{x}) - 1 = O_{\mathbb{P}}(1/\sqrt{nh_n^d})$, so that

$$\epsilon_n^{-1} \left(\frac{\hat{F}_n^{(p)}(u_n|\mathbf{x})}{\bar{F}^{(p)}(u_n|\mathbf{x})} - 1 \right) = \epsilon_n^{-1} \left(\frac{\hat{\varphi}_n^{(p-1)}(u_n|\mathbf{x})}{\varphi^{(p-1)}(u_n|\mathbf{x})} - 1 \right) + o_{\mathbb{P}}(1).$$

Applying Lemma 6 concludes the proof. \square

6.2 Proofs of Main Results

Proof of Theorem 1 Let us denote $\mathbf{t} = (t_1, \dots, t_J, t_{J+1})$ and focus on the probability

$$\Phi_n(\mathbf{t}) = \mathbb{P} \left(\bigcap_{j=1}^J \left\{ \sigma_n^{-1} \left(\frac{\hat{q}_n^{(1)}(\alpha_{n,j}|\mathbf{x})}{q^{(1)}(\alpha_{n,j}|\mathbf{x})} - 1 \right) \leq t_j \right\} \cap \left\{ \sigma_n^{-1} \left(\frac{\hat{q}_n^{(p)}(a_n|\mathbf{x})}{q^{(p)}(a_n|\mathbf{x})} - 1 \right) \leq t_{J+1} \right\} \right).$$

Set $y_n = q^{(1)}(\alpha_n|\mathbf{x})$, $y_{n,j} = q^{(1)}(\alpha_{n,j}|\mathbf{x})(1 + \sigma_n t_j)$ and $z_n = q^{(p)}(a_n|\mathbf{x})(1 + \sigma_n t_{J+1})$. The technique of proof of Proposition 1 in Girard et al. (2019) yields

$$\Phi_n(\mathbf{t}) = \mathbb{P} \left(\bigcap_{j=1}^J \left\{ \sigma_n^{-1} \left(\frac{\hat{F}_n^{(1)}(y_{n,j}|\mathbf{x})}{\bar{F}^{(1)}(y_{n,j}|\mathbf{x})} - 1 \right) \leq \sigma_n^{-1} \left(\frac{\bar{F}^{(1)}(q^{(1)}(\alpha_{n,j}|\mathbf{x})|\mathbf{x})}{\bar{F}^{(1)}(y_{n,j}|\mathbf{x})} - 1 \right) \right\} \right. \\ \left. \cap \left\{ \sigma_n^{-1} \left(\frac{\hat{F}_n^{(p)}(z_n|\mathbf{x})}{\bar{F}^{(p)}(z_n|\mathbf{x})} - 1 \right) \leq \sigma_n^{-1} \left(\frac{\bar{F}^{(p)}(q^{(p)}(a_n|\mathbf{x})|\mathbf{x})}{\bar{F}^{(p)}(z_n|\mathbf{x})} - 1 \right) \right\} \right).$$

Second-order regular variation arguments similar to those of the proof of Proposition 1 in Girard et al. (2019) give, for all $j \in \{1, \dots, J\}$,

$$\sigma_n^{-1} \left(\frac{\bar{F}^{(1)}(q^{(1)}(\alpha_{n,j}|\mathbf{x})|\mathbf{x})}{\bar{F}^{(1)}(y_{n,j}|\mathbf{x})} - 1 \right) = \frac{t_j}{\gamma(\mathbf{x})} (1 + o(1))$$

and similarly

$$\sigma_n^{-1} \left(\frac{\bar{F}^{(p)}(q^{(p)}(a_n|\mathbf{x})|\mathbf{x})}{\bar{F}^{(p)}(z_n|\mathbf{x})} - 1 \right) = \frac{t_{J+1}}{\gamma(\mathbf{x})} (1 + o(1)).$$

Finally, notice that $y_{n,j} = \tau_j^{-\gamma(\mathbf{x})} y_n (1 + o(1))$ and $z_n = \theta y_n (1 + o(1))$ (see (9)). Moreover, for n large enough, $\omega_{h_n}^{(1)}(y_{n,j}|\mathbf{x}) \leq \omega_{h_n}^{(1)}((1 - \delta)q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x})$ and $\omega_{h_n}^{(1)}(z_n|\mathbf{x}) \leq \omega_{h_n}^{(1)}((1 - \delta)\theta q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x})$. Similarly, $\omega_{h_n}^{(2)}(y_{n,j}|\mathbf{x}) \leq \omega_{h_n}^{(2)}((1 + \delta)\tau_1^{-\gamma(\mathbf{x})} q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x})$ and $\omega_{h_n}^{(2)}(z_n|\mathbf{x}) \leq \omega_{h_n}^{(2)}((1 + \delta)\theta q^{(1)}(\alpha_n|\mathbf{x})|\mathbf{x})$. Conclude using Proposition 1. \square

Proof of Theorem 2 We recall $\sigma_n^{-2} = nh_n^d(1 - \alpha_n)$. Write

$$\begin{aligned} \frac{\sigma_n^{-1}}{\log\left(\frac{1-\alpha_n}{1-\beta_n}\right)} \log\left(\frac{\tilde{q}_{n,\alpha_n}^{(p)}(\beta_n|\mathbf{x})}{q^{(p)}(\beta_n|\mathbf{x})}\right) &= \sigma_n^{-1}(\hat{\gamma}_{\alpha_n}(\mathbf{x}) - \gamma(\mathbf{x})) + \frac{\sigma_n^{-1}}{\log\left(\frac{1-\alpha_n}{1-\beta_n}\right)} \log\left(\frac{\hat{q}_n^{(p)}(\alpha_n|\mathbf{x})}{q^{(p)}(\alpha_n|\mathbf{x})}\right) \\ &\quad + \frac{\sigma_n^{-1}}{\log\left(\frac{1-\alpha_n}{1-\beta_n}\right)} \log\left(\left(\frac{1-\alpha_n}{1-\beta_n}\right)^{\gamma(\mathbf{x})} \frac{q^{(p)}(\alpha_n|\mathbf{x})}{q^{(p)}(\beta_n|\mathbf{x})}\right). \end{aligned}$$

The first term converges in distribution to Γ . The second one converges to 0 in probability, by Theorem 1. To control the third one, write

$$\left(\frac{1-\alpha_n}{1-\beta_n}\right)^{\gamma(\mathbf{x})} \frac{q^{(p)}(\alpha_n|\mathbf{x})}{q^{(p)}(\beta_n|\mathbf{x})} = \left(\frac{1-\alpha_n}{1-\beta_n}\right)^{\gamma(\mathbf{x})} \frac{q^{(1)}(\alpha_n|\mathbf{x})}{q^{(1)}(\beta_n|\mathbf{x})} \frac{q^{(p)}(\alpha_n|\mathbf{x})}{q^{(1)}(\alpha_n|\mathbf{x})} \frac{q^{(1)}(\beta_n|\mathbf{x})}{q^{(p)}(\beta_n|\mathbf{x})}.$$

In view of Theorem 4.3.8 in de Haan and Ferreira (2006) and its proof, $((1 - \alpha_n)/(1 - \beta_n))^{\gamma(\mathbf{x})} q^{(1)}(\alpha_n|\mathbf{x}) = q^{(1)}(\beta_n|\mathbf{x}) (1 + O(A((1 - \alpha_n)^{-1}|\mathbf{x}))) = q^{(1)}(\beta_n|\mathbf{x}) (1 + O(\sigma_n))$. By Lemma 5 then,

$$\left(\frac{1-\alpha_n}{1-\beta_n}\right)^{\gamma(\mathbf{x})} \frac{q^{(p)}(\alpha_n|\mathbf{x})}{q^{(p)}(\beta_n|\mathbf{x})} = 1 + O(\sigma_n).$$

The third term therefore converges to 0. Conclude using Slutsky's lemma and the delta-method. \square

Proof of Theorem 3 This proof is similar to those of Theorem 4 in Girard et al. (2021) (where $p = 2$) and Theorem 1 in Girard et al. (2019) (an unconditional version) and is thus left to the reader.

Acknowledgements This research was supported by the French National Research Agency under the grant ANR-19-CE40-0013/ExtremReg project. S. Girard gratefully acknowledges the support of the Chair Stress Test, Risk Management and Financial Steering, led by the French Ecole Polytechnique and its Foundation and sponsored by BNP Paribas, and the support of the French National Research Agency in the framework of the Investissements d’Avenir program (ANR-15-IDEX-02).

References

- Artzner, P., Delbaen, F., Eber, J., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), 203–228.
- Bellini, F., Klar, B., Muller, A., & Gianin, E. R. (2014). Generalized quantiles as risk measures. *Insurance: Mathematics and Economics* 54, 41–48 (2014).
- Cai, J., & Weng, C. (2016). Optimal reinsurance with expectile. *Scandinavian Actuarial Journal*, 2016(7), 624–645.
- Chen, Z. (1996). Conditional L_p -quantiles and their application to the testing of symmetry in non-parametric regression. *Statistics & Probability Letters* 29(2), 107–115.
- Daouia, A., Gardes, L., & Girard, S. (2013). On kernel smoothing for extremal quantile regression. *Bernoulli*, 19(5B), 2557–2589.
- Daouia, A., Gardes, L., Girard, S., & Lekina, A. (2011). Kernel estimators of extreme level curves. *TEST*, 20(2), 311–333.
- Daouia, A., Girard, S., & Stupfler, G. (2018). Estimation of tail risk based on extreme expectiles. *Journal of the Royal Statistical Society: Series B*, 80(2), 263–292.
- Daouia, A., Girard, S., & Stupfler, G. (2019). Extreme M-quantiles as risk measures: from L^1 to L^p optimization. *Bernoulli*, 25(1), 264–309.
- Daouia, A., Girard, S., & Stupfler, G. (2020). ExpectHill estimation, extreme risk and heavy tails. *Journal of Econometrics*, 221(1), 97–117.
- Daouia, A., Girard, S., & Stupfler, G. (2020). Tail expectile process and risk assessment. *Bernoulli*, 26(1), 531–556.
- de Haan, L., & Ferreira, A. (2006). *Extreme value theory: An introduction*. New York: Springer.
- El Methni, J., Gardes, L., & Girard, S. (2014). Non-parametric estimation of extreme risk measures from conditional heavy-tailed distributions. *Scandinavian Journal of Statistics*, 41(4), 988–1012.
- El Methni, J., & Stupfler, G. (2017). Extreme versions of Wang risk measures and their estimation for heavy-tailed distributions. *Statistica Sinica*, 27(2), 907–930.
- Embrechts, P., Kluppelberg, C., & Mikosch, T. (1997). *Modelling extremal events*. Berlin: Springer.
- Gardes, L., Girard, S., & Stupfler, G. (2020). Beyond tail median and conditional tail expectation: extreme risk estimation using tail L^p -optimisation. *Scandinavian Journal of Statistics*, 47(3), 922–949.
- Girard, S., Stupfler, G., & Usseglio-Carleve, A. (2019). An L^p -quantile methodology for estimating extreme expectiles, preprint. <https://hal.inria.fr/hal-02311609v3/document>.
- Girard, S., Stupfler, G., & Usseglio-Carleve, A. (2021). Nonparametric extreme conditional expectile estimation, To appear in Scandinavian Journal of Statistics. <https://hal.archives-ouvertes.fr/hal-02114255>.

- Jones, M. (1994). Expectiles and M-quantiles are quantiles. *Statistics & Probability Letters* 20, 149–153.
- Koenker, R., & Bassett, G. J. (1978). Regression quantiles. *Econometrica*, 46(1), 33–50.
- Newey, W., & Powell, J. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55(4), 819–847.
- Ohlsson, E., & Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. Berlin: Springer.
- Usseglio-Carleve, A. (2018). Estimation of conditional extreme risk measures from heavy-tailed elliptical random vectors. *Electronic Journal of Statistics*, 12(2), 4057–4093.
- Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364), 812–815.

Robust Efficiency Analysis of Public Hospitals in Queensland, Australia



Bao Hoang Nguyen and Valentin Zelenyuk

We dedicate our modest contribution to Professor Christine Thomas-Agnan—a great Scholar who together with various colleagues have originated, developed and inspired many interesting directions in research, among which is the concept of partial α -frontier modelling that we use in this work.

Abstract In this study, we utilize various approaches for efficiency analysis to explore the state of efficiency of public hospitals in Queensland, Australia, in the year 2016/17. Besides the traditional nonparametric approaches like DEA and FDH, we also use a more recent and very promising robust approach—order- α quantile frontier estimators (Aragon et al. 2005). Upon obtaining the individual estimates from various approaches, we also analyze performance on a more aggregate level—the level of Local Hospital Networks by using an aggregate efficiency measure constructed from the estimated individual efficiency scores. Our analysis suggests that the relatively low efficiency of some Local Hospital Networks in Queensland can be partially explained by the fact that the majority of their hospitals are small and located in remote areas.

1 Introduction

In Australia, the provision of free public hospital services is the responsibility of the state and territory governments. The management of public hospitals in states and territories is usually geographically based. Since the National Health Reform

B. H. Nguyen

School of Economics, University of Queensland, Brisbane, QLD 4072, Australia

e-mail: bao.nguyen3@uq.net.au

V. Zelenyuk (✉)

School of Economics and Centre for Efficiency and Productivity Analysis,

University of Queensland, Brisbane, QLD 4072, Australia

e-mail: v.zelenyuk@uq.edu.au

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_12

221

Agreement in 2012 (Council of Australian Governments 2011), the governance of public hospitals in Australia has become more decentralized with the establishment of Local Hospital Networks. The Local Hospital Network is an independent statutory body established by each Australian state/territory government. Local Hospital Networks directly operate a group of public hospitals and are directly responsible for their performance.

In Queensland, Local Hospital Networks are known as Hospital and Health Services (HHSs). There are 16 HHSs in the state, of these 15 HHSs are geographically based, the remaining one is a specialist statewide HHS dedicated to caring for children and young people. Each HHS is independently and locally controlled by a Hospital and Health Board and operated by a Health Service Chief Executive. HHSs relate to the Queensland Department of Health (Queensland Health) through a service agreement. Queensland Health acts as a system manager who has responsibility for purchasing healthcare services to cover the healthcare needs of citizens as well as monitoring the performance of HHSs. Meanwhile, each HHS acts as a provider whose function is to deliver healthcare services to its local community.

Although state and territory governments are responsible for delivering public hospital services, funding for public hospitals is provided by both federal and state/territory governments based on taxes collected from all states/territories across Australia. In the year 2016/17, 50% of expenditure on public hospital services in Queensland came from the state government, while 40% of the expenditure was provided by the Australian government (Australian Institute of Health and Welfare 2018). Public hospitals are funded either via Activity-Based Funding or a Block Funding model. In Queensland, 36 hospitals (predominantly large and urban hospitals) are funded by Activity Based Funding.¹ Meanwhile, 87 hospitals (mainly small and rural hospitals) are funded by Block Funding.

Public hospitals in Queensland are widely dispersed geographically with a relatively high proportion in regional and remote areas, which in part reflects the share of the state's population living outside the major cities and the obligation of the state government to provide equitable access to public hospital services for all residents. Public hospitals in the state are also diverse in terms of size: 91 out of 123 hospitals have 50 beds or fewer, yet 19 out of 123 hospitals have more than 200 beds and account for 75% of Queensland's total hospital beds (Australian Institute of Health and Welfare 2019).

As public hospitals are the key institution in the acute healthcare sector where the majority of healthcare expenditure occurs, improving hospital efficiency has been viewed as a fundamentally important means to contain healthcare costs in Australia.² In the study published in 2010, the Productivity Commission (2010) pointed out that

¹Under Activity Based Funding, hospitals are reimbursed based on the number and the complexity of patient care episodes they provide. Hospitals receive a fixed rate for each episode, and the value of the fixed rate is determined by the DRG to which the episode belongs.

²In the fiscal year 2016/17, Australia spent \$181 billion on healthcare (more than \$7,400 per person and 10% of its GDP), about a 57% increase since 2006/07 (after adjusting for inflation). This turns out to be an average annual growth rate of 4.67% over the decade: around 2% higher than average growth of GDP (Australian Institute of Health and Welfare 2018).

the average inefficiency level of Australian hospitals is around 10% and they would decrease operating expenditures by about 7% if the inefficiency was eliminated. Given that the efficiency of public hospitals is an important issue of public concern and has now become the main responsibility of HHSs, it is important to analyze hospital performance, not only at the individual level but also at the HHS level. These analyses will provide useful information about the relative performance of HHSs and possibly identify sources of efficiency differentials, which are imperatively needed for any plan to promote hospital efficiency.

This study will provide such an analysis by exploring the state of efficiency of public hospitals at the level of HHSs in Queensland, Australia in the year 2016/17. To analyze performance on the aggregate level, we utilize an aggregate efficiency measure constructed from individual efficiency scores estimated using various approaches. Besides the traditional nonparametric approaches like DEA and FDH, we also use a more recent and very promising robust approach—order- α quantile frontier estimators (Aragon et al. 2005). The order- α quantile frontier estimators appear to be more appealing than the conventional nonparametric approaches because they are more robust with respect to extreme values and/or outliers in a finite sample and do not suffer from the well-known curse of dimensionality (Simar and Wilson 2013).³

Based on the robust estimates of aggregate efficiency, we use k-mean clustering technique (an unsupervised machine learning algorithm) to classify HHSs in Queensland into three groups, namely relatively low, medium and high efficiency. Moreover, our analysis also suggests that the relatively low efficiency of some HHSs in Queensland can be partially explained by the fact that the majority of their hospitals are small and located in remote areas.

Our paper is organized as follows. Section 2 presents theoretical frameworks for efficiency measures and their nonparametric estimators. Section 3 provides a description of the data sources and variables used in this study. Section 4 discusses the results, and Sect. 5 provides concluding remarks.

2 Methodology

2.1 Theoretical Concepts

Let us consider a production process in which a production unit uses a set of p inputs, denoted as $x = (x_1, \dots, x_p)' \in \mathfrak{R}_+^p$, to produce a univariate output, denoted as $y \in \mathfrak{R}_+$.⁴ According to the production theory (Shephard 1953, 1970), the production technology can be mathematically characterized by a technology set defined as

³Although the order- α quantile frontier estimators can provide new insights from the data compared to the traditional nonparametric estimators, the traditional approach, especially the CRS-DEA, still has its merits and value in itself (see more discussion in Sect. 4).

⁴For the cases of multiple-output, one can either follow the multivariate conditional quantile approach proposed by Daouia and Simar (2007) or utilize aggregation techniques to aggregate outputs. In this study, we adopt Daraio and Simar's (2007) approach (the approach based on Princi-

$$\Psi = \{(x, y) \in \mathfrak{R}_+^p \times \mathfrak{R}_+ : x \text{ can produce } y\}. \tag{1}$$

Some regularity conditions are usually assumed for the technology set, among those the three most common assumptions are as follows⁵:

- A1. Ψ is closed.
- A2. The output sets (defined in (2) below) are bounded, $\forall x \in \mathfrak{R}_+^p$.
- A3. All inputs and outputs are strongly disposable, i.e. $(x_0, y_0) \in \Psi \Rightarrow (x, y) \in \Psi, \forall x \geq x_0, y \leq y_0$.

The production technology can also be described mathematically in terms of its sections: input requirement set and output attainable set. In this paper, we measure efficiency in output direction, thus our discussion here focuses on the output attainable set.⁶ It is defined as

$$P(x) = \{y \in \mathfrak{R}_+ : (x, y) \in \Psi\}, x \in \mathfrak{R}_+^p. \tag{2}$$

When efficiency is a concern, the boundary of the technology set is of interest. For the case of univariate output, the upper bound of the output attainable set (the production frontier) is also referred to as production function and defined as

$$\partial P(x) = \max_y \{y \mid y \in P(x)\}. \tag{3}$$

The Farrell-type output-oriented technical efficiency for the production unit is then defined as a radial distance from a point in output space representing the production unit toward the boundary and is defined mathematically as

$$\lambda(x, y) = \sup_{\lambda} \{\lambda > 0 \mid \lambda y \in P(x)\} = \sup_{\lambda} \{\lambda > 0 \mid (x, \lambda y) \in \Psi\}. \tag{4}$$

One might find it more convenient to look at the reciprocal of the output-oriented efficiency (also known as the Shephard distance function) since it gives an efficiency measure between 0 and 1, where 1 stands for 100% efficiency.

Now let us look at a more aggregate level, consider an industry of n production units, $\mathcal{X}_n = \{(X_i, Y_i) \mid i = 1, \dots, n\}$, which can be partitioned into L groups (according some external economic criteria) with the input-output allocation of each

pal Component Analysis) to aggregate hospital outputs into a single output measure. An alternative approach would be to use a price-based aggregation approach (Zelenyuk 2020).

⁵Other standard regularity conditions are “No Free Lunch” and “Producing Nothing is Possible” (see more details in Sickles and Zelenyuk 2019).

⁶Being similar to recent studies in the literature (e.g. Clement et al. 2008; Hu et al. 2012; Besstremyannaya 2013; Chowdhury and Zelenyuk 2016), we measure efficiency in output direction because the level of inputs used in public hospitals is usually fixed and influenced by external factors (the budget of hospitals are usually planned in advance with relatively fixed (typically 12+ months) labour contracts and huge investment in fixed inputs). Moreover, an output-oriented model is consistent with the aim of Queensland Health, which is to maximize healthcare services delivered to local community from given resources (see Queensland Health 2016).

group, say group ℓ , denoted as $X^\ell = \{(X_i^\ell, Y_i^\ell) \mid i = 1, \dots, n_\ell\}$, $\ell \in \{1, \dots, L\}$. One can measure the efficiency of each group ℓ in the industry by using the aggregate efficiency measure proposed by Färe and Zelenyuk (2003), extended by Simar and Zelenyuk (2007) and further elaborated on in Simar and Zelenyuk (2018). The main advantage of this measure is that it uses meaningful weights derived from the economic optimization principle to aggregate individual efficiency scores in order to construct a group measure (see detail in Färe and Zelenyuk 2003). In the case of univariate output, the aggregate efficiency for group ℓ is the weighted average of individual efficiency scores, where weights are output shares of individual production units in the group and is defined as

$$\overline{TE}^\ell = \sum_{i=1}^{n_\ell} \lambda (X_i^\ell, Y_i^\ell) \times S_i^\ell, \quad S_i^\ell = \frac{Y_i^\ell}{\sum_{i=1}^{n_\ell} Y_i^\ell}. \quad (5)$$

2.2 Nonparametric Estimators

2.2.1 DEA and FDH

In practice, Ψ is unknown and thus needs to be estimated from a sample of production units, say X_n . There have been two widely used approaches to estimate the production frontiers in the literature, usually referred to as the ‘deterministic frontier models’ and the ‘stochastic frontier models’. The deterministic frontier models assume all observed production units belong to the technology set with probability one, whereas the stochastic frontier models allow some observations to be outside of the technology set by including two-sided random noise. The traditional Stochastic Frontier Approach (SFA) requires parametric restrictions on the shape of the production frontier and on the data generating process to estimate the frontier and to identify the inefficiency term from the random noise component.⁷ Recently, semiparametric and nonparametric estimators have been developed for stochastic frontier models to mitigate the parameterization of the approach (see more details in Parmeter and Zelenyuk 2019).

The deterministic frontier models appear to be more appealing because they are usually handled via nonparametric estimators and rely on less restrictive assumptions. The most flexible deterministic frontier model is the Free Disposal Hull (FDH) estimator introduced by Deprins et al. (1984), which requires only the strong disposability assumption on the technology set. If, in addition, one imposes the convexity assumption on the technology set, one can use the Data Envelopment Analysis (DEA) estimator, which was initiated by Farrell (1957) and popularized by Charnes et al. (1978). For DEA models, one can further impose Constant Returns to Scale (CRS) or Variable Returns to Scale (VRS) on the technology set to obtain CRS-DEA or

⁷The traditional stochastic frontier approach was proposed independently by Aigner et al. (1977) and Meeusen and van Den Broeck (1977).

VRS-DEA estimators (Färe et al. 1983; Banker et al. 1984). The three estimators can be formulated respectively as follows

$$\widehat{\Psi}_{FDH} \equiv \left\{ (x, y) : y \leq \sum_{i=1}^n \zeta_i Y_i, x \geq \sum_{i=1}^n \zeta_i X_i, \sum_{i=1}^n \zeta_i = 1, \zeta_i \in \{0, 1\}, i = 1, \dots, n \right\}, \tag{6}$$

$$\widehat{\Psi}_{CRS-DEA} \equiv \left\{ (x, y) : y \leq \sum_{i=1}^n \zeta_i Y_i, x \geq \sum_{i=1}^n \zeta_i X_i, \zeta_i \geq 0, i = 1, \dots, n \right\}, \tag{7}$$

$$\widehat{\Psi}_{VRS-DEA} \equiv \left\{ (x, y) : y \leq \sum_{i=1}^n \zeta_i Y_i, x \geq \sum_{i=1}^n \zeta_i X_i, \sum_{i=1}^n \zeta_i = 1, \zeta_i \geq 0, i = 1, \dots, n \right\}. \tag{8}$$

The FDH/DEA estimators of technical efficiency are obtained by plugging $\widehat{\Psi}_{FDH}$ or $\widehat{\Psi}_{CRS-DEA}$ or $\widehat{\Psi}_{VRS-DEA}$ in (4). The asymptotic properties of FDH/DEA estimators have been well-established in the literature (e.g. see Kneip et al. 1998, 2008; Park et al. 2000, 2010). In summary, under appropriate assumptions, the estimators are consistent (converging to the true values when sample sizes go to infinity) and have limiting distributions. Convergence rates depend on the type of estimators and the dimension of input-output space (the number of inputs, p , plus the number of outputs, q). To be more specific, the convergence rates for FDH, CRS-DEA, VRS-DEA estimators are n^κ , where $\kappa = 1/(p + q)$, $2/(p + q)$, or $2/(p + q + 1)$, respectively (e.g. see more discussion in Simar and Wilson 2015; Sickles and Zelenyuk 2019).

The envelopment estimators of aggregate efficiency of group ℓ then can be obtained by plugging the envelopment estimators of individual efficiency into equation (5)

$$\widehat{TE}^\ell = \sum_{i=1}^{n_\ell} \hat{\lambda} (X_i^\ell, Y_i^\ell | X_n) \times S_i^\ell, \quad S_i^\ell = \frac{Y_i^\ell}{\sum_{i=1}^{n_\ell} Y_i^\ell}. \tag{9}$$

2.2.2 Partial Frontiers

The deterministic frontier models, however, are particularly sensitive to extreme values and/or outliers because by construction, they fully envelop all observed data. Various techniques have been proposed to deal with the disadvantage. One approach is to identify and possibly delete any outliers in the data, but the approach, to some extent, depends on how the researcher defines an ‘outlier’ (Simar and Wilson 2015). As an alternative, one can also use the stochastic versions of DEA and FDH, where data is prewhitened from the noise and outliers using nonparametric SFA in the first stage and DEA/FDH is applied to estimate efficiency in the second stage (e.g. see Simar 2007; Simar Zelenyuk 2011).

Another approach is to use robust partial frontier estimators. There are mainly two types of partial frontiers, which are (i) order- m frontiers introduced by Cazals et al. (2002) and (ii) order- α quantile frontiers introduced by Aragon et al. (2005) and extended by Daouia and Simar (2007). The idea of partial frontier estimators is to estimate something “close” to but not the boundary of the technology set (Simar and Wilson 2013). For example, in output orientation, order- m frontiers are defined as the expected maximum obtainable outputs among m production units drawn randomly from the population of those using at most a given level of inputs. Meanwhile, order- α quantile frontiers represent the expected maximum output levels that are exceeded by $100(1 - \alpha)\%$ of production units using less than or equal to a given level of inputs.

The nonparametric estimators of these frontiers turn out to be more appealing than the conventional deterministic frontier models because they do not suffer from the well-known curse of dimensionality and achieve the standard parametric root- n (\sqrt{n}) rate of convergence (for a fixed value of order α) (Cazals et al. 2002; Aragon et al. 2005; Daouia and Simar 2007). Moreover, both the estimators are also consistent estimators of the full frontier and share asymptotic properties with FDH estimators but are more robust with respect to extreme values and/or outliers in finite sample than the conventional FDH or DEA estimators (Simar and Wilson 2013).

Among the two above-mentioned partial frontier approaches, the order- α quantile frontier estimators are argued to have better robustness properties than the order- m frontier estimators. For example, Aragon et al. (2005) compared the two estimators using various simulated data sets, and reached the same conclusion with all the data sets that the order- m frontier estimators are less resistant to outliers than the order- α quantile frontier estimators. Daouia and Simar (2007) examined the robustness properties of the two estimators from the theoretical points of view using the concept of influence function, and came up with the same conclusion. Thus, we will use the order- α quantile frontier estimators in our analysis and focus our discussion on these estimators.

Let us define the technology set Ψ as the support of the joint distribution of a random variable (X, Y) , which generates the random sample \mathcal{X}_n . Here, we focus on the interior of the set, $\Psi^* = \{(x, y) \in \Psi \mid F_X(x) > 0\}$, where $F_X(\cdot)$ represents the marginal distribution of X . As in Cazals et al. (2002), the production function defined in (3) can be rewritten in a probabilistic representation as

$$\partial P(x) = \sup_y \{y \mid F_{Y|X}(y|x) < 1\}, \tag{10}$$

where $F_{Y|X}(y|x)$ is the conditional distribution of Y given $X \leq x$, i.e.

$$F_{Y|X}(y|x) = \frac{F_{XY}(x, y)}{F_X(x)}, \tag{11}$$

where $F_{XY}(x, y) = \text{Prob}(X \leq x, Y \leq y)$ is the joint distribution of (X, Y) .

Equivalently, $\partial P(x)$ can be formulated as the order one quantile of the distribution of Y given $X \leq x$ as

$$q_1(x) = \inf_y \{y \geq 0 \mid F_{Y|X}(y|x) = 1\}. \tag{12}$$

One can interpret $q_1(x)$ as the minimum output level not exceeded by any production unit using at most x inputs. Based on the formulation, Aragon et al. (2005) introduced a concept of order- α quantile frontiers as the quantile functions of order α , $\alpha \in [0, 1]$, of the distribution of Y given that X is less than or equal to a given level of inputs and defined as

$$q_\alpha(x) = \inf_y \{y \geq 0 \mid F_{Y|X}(y|x) \geq \alpha\}. \tag{13}$$

The order- α quantile frontier, $q_\alpha(x)$, represents the output threshold exceeded by 100(1 - α) % of production units using at most x inputs. The efficiency measure with respect to the frontier is referred to as the order- α quantile efficiency and defined as

$$\lambda_\alpha(x, y) = \inf_\lambda \{\lambda \mid F_{Y|X}(\lambda y|x) \geq \alpha\}. \tag{14}$$

The order- α quantile efficiency represents the radial distance from a point in output space representing the production unit toward the order- α quantile frontier. The measure $\lambda_\alpha(x, y)$ can have values between 0 and $+\infty$, where $\lambda_\alpha(x, y) < 1$ indicates that the production unit with input-output allocation (x, y) is above the order- α quantile frontier (i.e. super-efficient production unit).

To estimate order- α quantile frontiers and order- α quantile efficiency, we can apply the plug-in principle by replacing $F_{Y|X}(\cdot|x)$ in (13) and (14), respectively, by its empirical analogue

$$\widehat{q}_{\alpha,n}(x) = \inf_y \{y \geq 0 \mid \widehat{F}_{Y|X}(y|x) \geq \alpha\} \tag{15}$$

and

$$\widehat{\lambda}_{\alpha,n}(x, y) = \inf_\lambda \{\lambda \mid \widehat{F}_{Y|X}(\lambda y|x) \geq \alpha\}. \tag{16}$$

As an extension of Theorem 4.1 in Aragon et al. (2005), Daouia and Simar (2007) show that under appropriate assumptions, order- α quantile efficiency estimators have asymptotic normality with the standard parametric root- n (\sqrt{n}) rate of convergence for a fixed value of order α . Moreover, the order- α quantile efficiency estimators converge to the FDH estimator as $\alpha \rightarrow 1$

$$\lim_{\alpha \rightarrow 1} \widehat{\lambda}_{\alpha,n}(x, y) = \widehat{\lambda}_{FDH}(x, y). \tag{17}$$

More details on this interesting method can be found in Aragon et al. (2005) and Daouia and Simar (2007), while in the next section we will apply it to analyze the efficiency of public hospitals in Queensland, Australia.

Before going to discuss the empirical results, it is worth noting here that individual efficiency of each hospital in this study is estimated using the entire industry sample (a sample of size n). It is also important to emphasize that while we recognize that each hospital may use different technologies (and potentially much more complex than any model can handle), the goal of this study is to measure the relative efficiency with respect to the frontier of the unconditional technology set, where the so-called ‘separability assumption’ (Simar and Wilson 2007, 2011) is satisfied by definition. In a sense, it is similar to a grand competition, like Olympics or country-wide student evaluation, where everyone is measured with respect to the same criteria, regardless of their backgrounds.

It is very well possible that further stratification of the sample is needed to account for various conditions that each of the hospitals may face and potentially may prevent them from reaching the frontier of the unconditional technology set. In such cases, an alternative (as well as complementary) strategy would be to take the so-called conditional frontier approach, where the frontier may vary across different groups or even depend on the values of various continuous variables (and thus allow for uncountably infinite possibilities of different frontiers).⁸

To be more precise, the conditional technology can be defined (following Simar and Wilson 2007, 2011) as

$$\Psi^z = \{ (X, Y) \mid X \text{ can produce } Y \text{ when } Z = z \},$$

where Z is a vector of conditioning variables (potentially very large one, and in reality possibly endogenous to the production unit), and z is a particular value it may take out of the all the possibilities, denoted by the set \mathcal{Z} . Meanwhile, the unconditional technology (which we focus on here) is then defined as

$$\Psi = \bigcup_{z \in \mathcal{Z}} \Psi^z,$$

which does not depend on Z .

Recently, rigorous statistical tests have been developed to verify which of the pre-selected models fit a given data ‘tighter’ according to a selected statistical criterion, for a pre-selected set of conditional variables, Z , and a pre-selected order of the frontier and other assumptions.⁹ In our case, there are several dozens of potential variables that one may hypothesize as potentially determining the conditional frontiers alone or in various combinations with each other and in different combinations with other assumptions on the model and different orders (α) of the frontier. The number of all such unique possible combinations is very large and there is no theory

⁸E.g. one could use Badin et al. (2012) approach or, alternatively, a nonparametric stochastic approach (e.g. see Simar et al. 2017; Parmeter and Zelenyuk 2019, and references therein).

⁹E.g. see Daraio et al. (2018) and Simar and Wilson (2020) for details. Similar tests can be also explored for the nonparametric and semiparametric stochastic frontiers mentioned above, e.g. see Simar et al. (2017).

that dictates which combinations shall or shall not be considered. And so, in principle, once this strategy is taken, it is fair to consider all the relevant combinations in the testing to arrive to a robust conclusion, which is well beyond the scope of this paper. In principle, such testing and selection of a tightest-fit model out of the myriad of possibilities can be made automatic by adapting various machine learning techniques (LASSO, best subset selection, forward step-wise selection, etc.), which is also a subject in itself and so is left for future endeavours.

3 Variables and Data

In this study, we compare the technical efficiency of public hospitals across 15 geographically based HHSs in Queensland in the financial year 2016/17.¹⁰ Our sample includes 111 public acute hospitals.¹¹ The hospital data are sourced from two data collections of Queensland Health, namely Financial and Residential Activity Collection (FRAC) and Monthly Activity Collection (MAC). We obtained the information about hospital staffing and drug, surgical and medical supply expenditures from the FRAC, while the MAC provided us with the data on the number of beds, non-admitted patient activities, and admitted patient episodes of care by Diagnosis-Related Groups (DRGs).

Following the common practice in the literature,¹² to model the production process of hospitals, we use three inputs, which are labour input, capital input and consumable input. Regarding the labour input, the data on hospital staffing is provided in the form of Full-Time Equivalent (FTE) staff in six major categories including salaried medical officers, nurses, diagnostic and health professionals, other personal care staff, administrative and clerical staff, and domestic and other staff. To increase the discriminant power of the nonparametric envelopment estimator, but still cover the information contained in all the labour categories, we reduce the dimensions using Principal Component Analysis (PCA). In particular, we adopt the variant of PCA proposed by Daraio and Simar (2007) to aggregate the six labour categories into a single measure of labour input, called labour factor (denoted as FLABOUR). For the other two inputs, we use the number of beds (BEDS) as a proxy for capital input, and

¹⁰There are 16 HHSs in Queensland, but only 15 HHSs directly manage and operate public hospitals in defined local geographical areas, the remaining HHS is a specialist statewide HHS dedicated to caring for children and young people from across Queensland.

¹¹Public hospitals in Queensland include acute hospitals, mixed sub- and non-acute hospitals, early parenting centres, women's and children's hospitals, and psychiatric hospitals. We only consider public acute hospitals, which account for more than 90% of inpatient cases treated. Our sample does not include hospitals that were just opened in 2017 and hospitals that are not operated by a HHS.

¹²See the reviews in O'Neill et al. (2008); Kohl et al. (2019).

drug, surgical and medical supply expenditure (DMSEXP) as a proxy for consumable input.¹³

Similarly, using Daraio and Simar's (2007) approach, we aggregate two widely used measures of hospital outputs, namely (i) non-admitted occasions of services and (ii) casemix weighted inpatient episodes, into a single output measure, called output factor (we denote as FOUTPUT). The information about non-admitted occasions of services, which include both outpatient visits and emergency department services, is readily available in our datasets. Meanwhile, the casemix weighted inpatient episode is constructed as the weighted sum of the number of inpatient episodes by DRG, where the weight is the inlier DRG cost weight obtained from the Independent Hospital Pricing Authority.¹⁴

In addition, we obtain information about hospital peer groups and geographic location from Australian Institute of Health and Welfare (2015). Based on hospital peer groups, in our study, hospitals are classified as large hospitals if they are principal referral hospitals, public acute group A hospitals, or public acute group B hospitals, and classified as small hospitals if they are public acute group C hospitals, or public acute group D hospitals.¹⁵ Moreover, hospitals in our sample are also categorized into two groups based on their geographic location, namely remote hospitals (located in remote and very remote areas), and non-remote hospitals (located in major cities, inner regional, or outer regional areas).¹⁶

Table 1 provides information about the number of hospitals, proportion of remote and small hospitals as well as total inputs utilized and total outputs provided by all hospitals belonging to each HHS in our sample. We can see that HHS 402, HHS 403 and HHS 436 are the only HHSs where all hospitals are small hospitals.¹⁷ Moreover, almost all of their hospitals are located in remote areas. Meanwhile, the majority of hospitals managed by HHS 408, HHS 431, HHS 487, and HHS 494 are large hospitals and located in non-remote areas. Since the distributions of inputs and output are highly right-skewed with the large and non-remote hospitals being on the right tails of the distributions (see Fig. 1), the utilization of inputs and the provision of services varies significantly across the HHSs in our sample. For instance, HHS 436 has the total number of beds of only 69. Meanwhile, the total number of beds

¹³See more discussion about the selection and construction of hospital inputs and outputs in Chowdhury et al. (2014); Chowdhury and Zelenyuk (2016).

¹⁴Ideally, outputs of hospitals should be measured by the improvement in medical condition of patients. However, it is technically difficult to obtain this measure in practice, thus most of the hospital efficiency studies use quantities of services as an alternative measure of hospital outputs (Hollingsworth 2008).

¹⁵Public acute hospitals in Australia are divided into five groups listed in descending order of activity volume and service diversification, as follows: principal referral hospitals, public acute group A hospitals, public acute group B hospitals, public acute group C hospitals, public acute group D hospitals. According to Australian Institute of Health and Welfare (2015), hospitals in the first three groups are generally larger than hospitals in the last two groups.

¹⁶The classification is based on the remoteness area information provided in the Australian hospital peer groups in which the remoteness of a hospital is measured by the physical road distance to its nearest urban centre.

¹⁷Note that the IDs here are not the real ID but randomly generated for each HHS.

Table 1 Descriptive Statistics of variables by HHSs

Random ID	No. of Hospitals	Proportion of Remote hospitals	Proportion of Small hospitals	No. of Beds (Total)	DMSEXP (Total) (\$ millions)	FLABOUR (Total) FTE	FOUTPUT ^a (Total)
402	4	1.00	1.00	104	3.33	159.42	0.39
403	11	0.91	1.00	166	5.09	190.88	0.53
408	2	0.00	0.00	1167	170.77	3849.69	8.34
418	5	0.00	0.80	509	44.25	1270.69	3.19
423	8	0.25	0.88	335	50.44	1033.18	2.74
431	5	0.00	0.20	2354	387.34	5516.82	16.14
435	6	1.00	0.83	119	8.76	331.07	1.04
436	5	1.00	1.00	69	2.95	110.71	0.27
442	12	0.17	0.83	526	64.35	1211.94	3.17
451	7	0.00	0.86	843	112.95	2420.74	5.04
468	8	0.38	0.88	823	97.46	2207.21	5.25
478	10	0.00	0.70	584	64.53	1406.52	4.01
481	19	0.11	0.95	701	70.22	1624.25	3.98
487	4	0.00	0.25	300	90.02	1772.83	4.23
494	5	0.00	0.20	1937	316.03	5647.32	13.65

^aSince the unit of measurement of non-admitted occasions of services and casemix weighted inpatient episodes are different, we normalize them by their standard deviations before the aggregation.

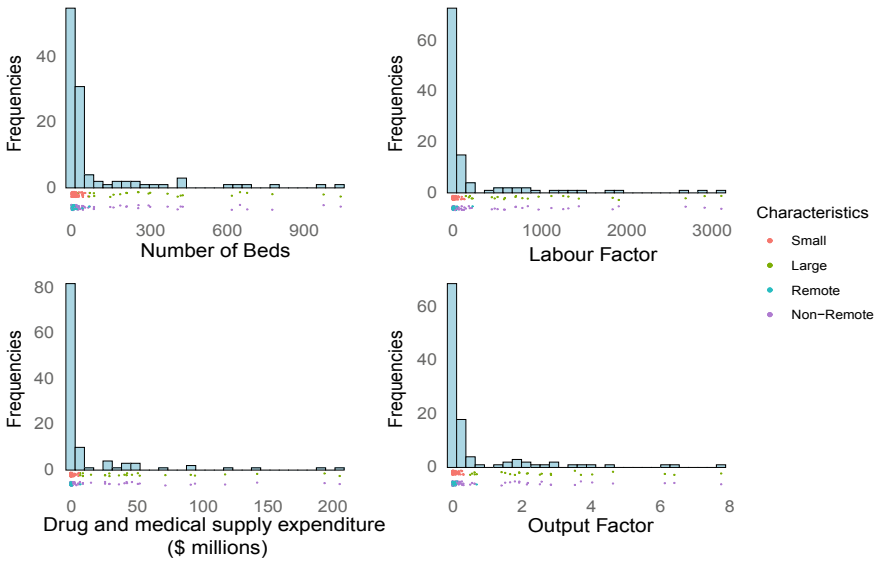


Fig. 1 Histograms (with jittered points representing individual hospitals) of input and output variables

operated by HHS 431 is 2354, being nearly 35 times higher than that of HHS 436. The similar pattern is also observed on the output side, where the highest figure of output factor is around 60 times higher than its lowest figure.

One issue with frontier estimators is the possibility of ‘gaps’ between small and large production units on the output or input axes, which might cause difficulty in estimating the frontiers.¹⁸ The histograms (with jittered points representing individual hospitals) in Fig. 1 help us to visualize the distributions of the inputs and output in our sample to identify the possible ‘gaps’. Looking at Fig. 1, it seems to be that there do not exist such “gaps” between small and large hospitals in our dataset.

4 Results and Discussions

4.1 Univariate Input-Output Illustration

In this subsection, we aim at providing a graphical illustration of different types of frontier estimators. To do so, we utilize the same technique as discussed in Sect. 3 to aggregate inputs further into a single variable, we denote as FINPUT, representing all resources utilized by hospitals. In the case of univariate input-output production technology, we can present the estimated production frontiers (i.e. production functions) on a 2-D graph together with data points as shown in Fig. 2. As we can see, DEA and FDH estimators envelope all the data points, whereas some data points are above the estimated order- α quantile frontiers (even for a relatively high value of α , say, $\alpha = 0.99$). Moreover, when α increases to 1, the estimated order- α quantile frontiers get closer to the estimated FDH frontier. Actually, as pointed out in Sect. 2, the FDH frontier is a special case of order- α quantile frontiers when $\alpha = 1$.

4.2 Main Analysis: Multiple Inputs Case

Before discussing the results, it is worth mentioning here that the results in this study are reported based on the reciprocal of the output-oriented efficiency score, which gives an efficiency measure between 0 and 1 for FDH and DEA estimators, and an efficiency measure between 0 and $+\infty$ for order- α quantile frontier estimators. As a result, if a hospital has an efficiency score from the order- α quantile frontier estimators in $(0, 1)$, $\{1\}$, or $(1, +\infty)$, then it is interpreted, respectively, as “below”, “on”, or “above” the corresponding order- α quantile frontier.

Figure 3 shows how $p(\alpha)$, the percentage of hospitals being above the estimated order- α quantile frontier, changes when the order α increases. It is remarkable that when the order α increases from 0 to 0.8, $p(\alpha)$ decreases slowly, indicating that the

¹⁸We thank the anonymous referee for this insight.

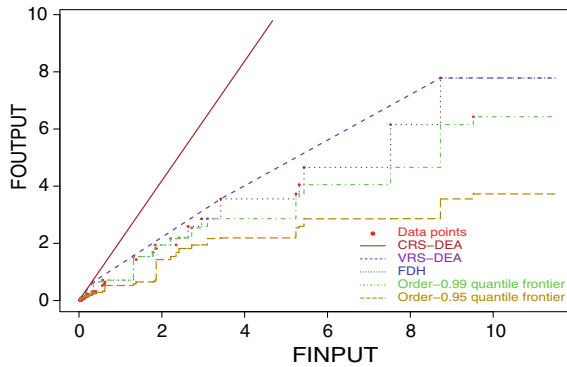


Fig. 2 Estimated frontiers for the case of univariate input and output

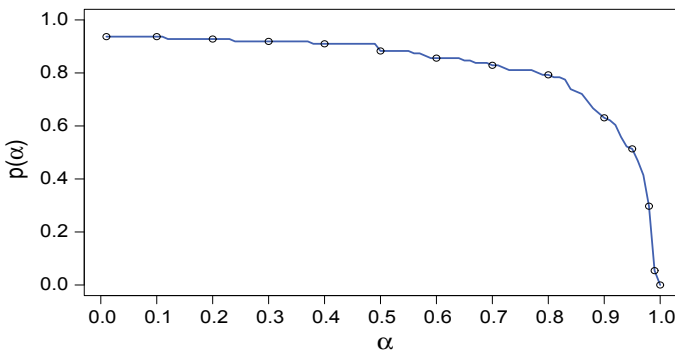


Fig. 3 Evolution of the proportion of hospitals being above the estimated order- α quantile frontier

quantile frontiers of orders α in this range are very tight. From the order of around 0.8, $p(\alpha)$ decreases with a faster rate, showing that the quantile frontiers become more spaced. The values of $p(\alpha)$ are, however, still relatively high for the values of α close to one. For example, $p(\alpha)$ is 51% for $\alpha = 0.95$, 30% for $\alpha = 0.98$ and still 5% for $\alpha = 0.99$. This fact suggests that only quantile frontiers of orders extremely close to one are possibly influenced by extreme values.

In the following analysis, we measure hospital efficiency with respect to the quantile frontiers of order $\alpha = 0.99$. This quantile frontier is less likely to be affected by extreme values/outliers and still represents the output threshold exceeded by only 1% of hospitals in population using at most a given level of inputs. In our sample, all of those super-efficient hospitals are large and non-remote hospitals.

We are interested in comparing hospital efficiency across HHSs, thus after obtaining the individual estimates from various estimators (including order-0.99 quantile frontier, FDH, VRS-DEA and CRS-DEA), we utilize the aggregate efficiency measure discussed in Sect. 2 to analyze the performance of HHSs. Table 2 reports the

Table 2 Estimated Aggregate Efficiencies

Random ID	Efficiency Estimators				Clusters
	Order-0.99 quantile frontiers	FDH	VRS-DEA	CRS-DEA	
436	0.82	0.82	0.57	0.53	3
403	0.87	0.87	0.69	0.62	3
402	0.89	0.89	0.68	0.54	3
442	0.93	0.93	0.76	0.59	2
481	0.97	0.97	0.80	0.56	2
478	0.98	0.98	0.88	0.64	2
423	0.98	0.98	0.93	0.63	2
435	0.98	0.98	0.91	0.74	2
418	0.99	0.99	0.93	0.57	2
487	0.99	0.99	0.96	0.84	2
451	1.04	1.00	0.76	0.48	2
494	1.06	0.92	0.91	0.57	2
408	1.24	1.00	0.99	0.50	1
468	1.27	0.99	0.95	0.54	1
431	1.36	1.00	0.96	0.64	1

Notes

*Results are reported based on the reciprocals of the output-oriented efficiency scores

*Computations are done in R (R Core Team 2019) using 'frontiles' package (Daouia and Laurent 2013) and 'Benchmarking' package (Bogetoft and Otto 2019)

estimated aggregate efficiencies, and Fig. 4 presents the estimated aggregate efficiencies by different types of estimators on a parallel coordinate plot.

For both VRS-DEA and CRS-DEA estimators, some HHSs turn out to be very inefficient, especially for CRS-DEA estimators, where 9 out of 15 HHSs are at least 40% inefficient (see Fig. 4). On the other hand, these models give a high variation in efficiency (or have high discriminative power) that might be explained through additional analysis. The prevalence of inefficient HHSs might be attributed to the fact that the frontiers estimated by DEA estimators are particularly sensitive to extreme values. A very few super-efficient production units can possibly shift the whole estimated frontiers outward and substantially change the distribution of the estimated efficiency scores. Identifying and removing these outliers from the sample (and studying them separately) may be useful for further analysis with the CRS-DEA model since it has value in itself. Indeed, provided there are no outliers, CRS-DEA can be considered as the most appropriate benchmark from a social point of view to evaluate the performance of production units in the public sector because it identifies the level of highest utilization of inputs into outputs (or highest average productivity) and the best practice socially optimal scale.¹⁹ In our sample, five hospitals are on the

¹⁹See more discussion in Grosskopf et al. (2020) and Nguyen and Zelenyuk (2021).

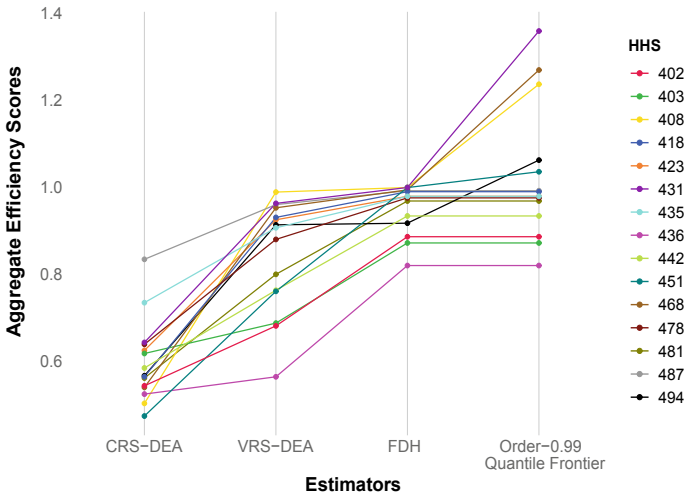


Fig. 4 Aggregate efficiencies by different types of estimators

CRS-DEA frontier, which are hospital 1119, hospital 1031, hospital 1035, hospital 1095 and hospital 1001.²⁰ Among these five hospitals, four hospitals are small and located in remote areas. It might be an indicator that large hospitals in our sample are not operating near the socially optimal scale, and this can be explored in future research.

Compared to DEA estimators, FDH estimators are less sensitive to extreme values. The estimated aggregate efficiencies obtained from FDH estimators are relatively reasonable ranging from 0.8 to 1. However, the FDH model has low discriminative power with many observations attaining high or 100% efficiency scores, of these some appear to be very inefficient when benchmarking using DEA. Moreover, for some HHSs, the evaluation of relative performance seems still to be influenced by the presence of super-efficient production units. For example, looking at Fig. 4, we can see that HHS 494 is in the top highest performance HHSs based on order-0.99 quantile aggregate efficiency, but it is in the bottom lowest performance HHSs based on FDH aggregate efficiencies. Due to the limited space, in the following discussion, we focus exclusively on the results obtained from order-0.99 quantile frontier estimators.

Based on order-0.99 quantile aggregate efficiencies, we use k-mean clustering technique to classify HHSs in Queensland into three groups, namely relatively low, medium and high efficiency (denoted as clusters 3, 2 and 1, respectively, in Table 2).²¹ The relatively low-efficiency group includes HHS 402, HHS 403 and HHS 436. The relatively high-efficiency group includes HHS 408, HHS 431 and HHS 468. The relatively medium efficiency group is composed of the remaining HHSs.

²⁰Note that the IDs here are not the real ID but randomly generated for each hospital.

²¹K-mean clustering is an unsupervised machine learning algorithm helping cluster data into a predetermined number of clusters so as to minimize the within-cluster sum of squares.

To further investigate the differences in efficiency of HHSs, we look at characteristics of their hospitals. As discussed in Sect. 3, HHS 402, HHS 403 and HHS 436 are the only HHSs with all their hospitals being small hospitals. Moreover, almost all of their hospitals are located in remote areas. The boxplots in Fig. 5 provide some insights about the relative performance of hospitals according to these characteristics. In our sample, large hospitals and hospitals in non-remote areas are relatively more efficient than small hospitals and hospitals in remote areas, respectively.

The above explanatory analysis suggests that the relatively low efficiency of HHS 402, HHS 403 and HHS 436 with respect to the order-0.99 quantile frontier can be partially explained by the fact that the majority of their hospitals are small and located in remote areas. Rural hospitals are argued to face many disadvantageous conditions (e.g. shortages of medical staff, high chronic illness rate in the rural population and stagnation in the rural economy); thus they might not provide health services as efficiently as urban hospitals do (Weisgrau 1995). Similarly, compared to large hospitals, small hospitals might be less efficient because they usually have a lower level of standardization and specialization, resulting in weaker communication and coordination between hospital facilities (Munson and Zuckerman 1983).

The evidence about the relative inefficiency in utilizing healthcare resources of small and remote hospitals might have useful policy implications for managers of relevant HHSs as well as Queensland Health. The presence of public hospitals in remote and very remote areas is an important vehicle to ensure equitable access to health services for all residents in Queensland given its geographically dispersed population. However, given the inefficiency of small and remote hospitals, other models of health service delivery, such as Telehealth, perhaps should be given a

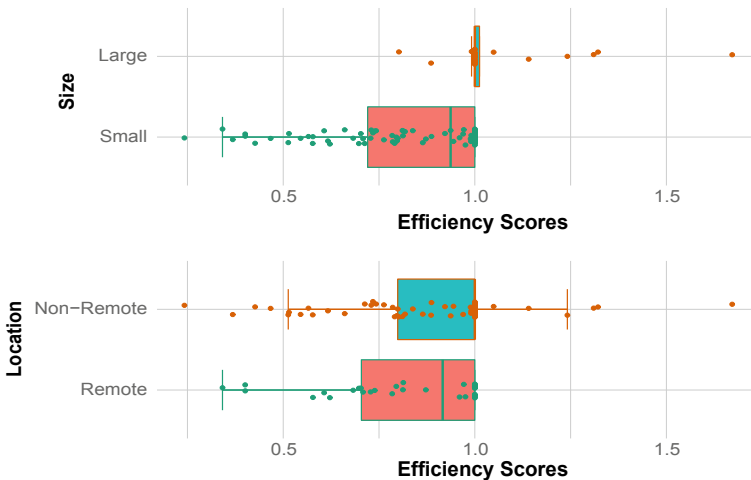


Fig. 5 Boxplots (with jittered points representing individual hospitals) of estimates of order-0.99 quantile efficiencies by size and location

higher priority to develop as an alternative measure to better meet the healthcare needs of communities in rural areas.

It is worth recalling here that with the robust order- α quantile frontier estimator, hospitals are benchmarked relative to the frontier nonparametrically estimated from its closest peers, without imposing any assumptions of returns to scale, monotonicity or convexity. This flexibility may be viewed both as an advantage in some respects as well as a limitation in other respects. For example, the estimated efficiency scores could be very high, e.g. 100% or higher, for some very large hospitals perhaps because there are not many (or even any) peers to compare them with and reveal their inefficiency. In particular, all such large hospitals could be very large and very inefficient relative to the socially optimal scale frontier (see more discussion in Nguyen and Zelenyuk 2021).

In previous studies in the Australian context, large and urban hospitals were also found to be more efficient than small and rural hospitals (Paul 2002; Productivity Commission 2010). However, as in the current paper, the constant returns to scale assumption is not imposed in all these studies, and thus hospitals are not benchmarked with respect to the socially optimal scale frontier. In future research, it might be useful to explore hospital efficiency with respect to the socially optimal scale frontier using a CRS-DEA model, since the scale efficiency might possibly be substantially different between small and large hospitals and might influence their relative efficiency.

5 Concluding Remarks

In this study, we explored the state of the efficiency of public hospitals at the level of Hospital and Health Services—independent statutory bodies who directly operate a group of public hospitals in a defined geographical area, in Queensland, Australia. To analyze their performance on the aggregate level, we utilize an aggregate efficiency measure constructed from individual efficiency scores which were estimated using various approaches. Besides the traditional nonparametric approaches like DEA and FDH, we also use a more recent and very promising robust approach—order- α quantile frontier estimators (Aragon et al. 2005). Our analysis suggests that efficiency scores of some Local Hospital Networks in Queensland are relatively low, which can be partially explained by the fact that the majority of their hospitals are small and located in remote areas.

Care is, however, needed when interpreting the results. High-efficiency scores of large hospitals with respect to the order- α quantile frontier do not necessarily mean that they are efficient from a social point of view. These hospitals might utilize too many resources to deliver what can be otherwise done by smaller sized hospitals that operate at a socially optimal scale (see more discussion with intuitive examples in Nguyen and Zelenyuk 2021). Indeed, operating at a socially optimal scale is of vital importance for the healthcare systems, particularly in urgent circumstances,

like pandemics. It allows hospitals to flexibly expand their operations to efficiently deliver the necessary healthcare services to society.

Moreover, the relatively low aggregate efficiency scores of some HHSs do not necessarily mean that they are not as efficient as other HHSs in operating public hospitals. There might possibly be other factors beyond the control of managers that are negatively affecting the performance of their hospitals. Remoteness and size are just two among many factors that are necessary to take into account.²² Moreover, although the above explanatory analysis is an important step to identify sources of efficiency differentials, more analysis that will account for other confounding factors is needed. For example, this can be done by using the truncated regression with the bootstrap approach of Simar and Wilson (2007) or the conditional efficiency framework of Badin et al. (2012).

Similar to many other studies in the literature, due to data availability, this study does not take into account the quality dimension when estimating hospital efficiency and comparing the performance of HHSs. This might be unfair for those who have to utilize more resources to maintain the high quality of services. Therefore, a natural recommendation is to gather more data to incorporate the output quality indicator(s) in the analysis. It is also worth remarking here that the aggregation of inputs and outputs in this study helps to increase discrimination power and to mitigate the curse of dimensionality issue for nonparametric estimators, but it may come at a cost of losing some information and incurring aggregation bias. As a result, considering different aggregation strategies and sensitivity of results across them could be a direction for future research. Another fruitful direction of research would be to develop and apply statistical tests based on Central Limit Theorems for aggregate efficiency recently developed by Simar and Zelenyuk (2018) to statistically compare the performance of hospitals at HHS level.

Acknowledgements We thank the Editor and two anonymous referees for many fruitful comments that helped improving this paper substantially. We acknowledge the support from our institution. We also acknowledge the financial support from the Australian Research Council (from the ARC Future Fellowship grant FT170100401). We thank Dan O'Halloran for his fruitful comments. We also thank David Du, Hong Ngoc Nguyen, Zhichao Wang and Evelyn Smart for their feedback from proofreading. We acknowledge and thank Queensland Health for providing part of the data that we used in this study. These individuals and organizations are not responsible for the views expressed in this paper.

²²A deeper analysis on hospital efficiency based on geographical location (e.g. with some spatial maps) could be a fruitful research direction. Some hospitals in major cities may benefit from the presence of other hospitals to adjust their capacities or to select their patients, while this may be not possible for hospitals in remote areas. Moreover, some hospitals in urban areas may be in intensive competition, while others in rural areas may be local monopolies. We thank the anonymous referee for this insight.

References

- Aigner, D., Lovell, C. A. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 64(6), 1263–1297. [https://doi.org/10.1016/0304-4076\(77\)90052-5](https://doi.org/10.1016/0304-4076(77)90052-5).
- Aragon, Y., Daouia, A., & Thomas-Agnan, C. (2005). Nonparametric frontier estimation: A conditional quantile-based approach. *Econometric Theory*, 21(2), 358–389. <https://doi.org/10.1017/S0266466605050206>.
- Australian Institute of Health and Welfare. (2015). Australian hospital peer groups (tech. rep. No. 66). Australian Institute of Health and Welfare. Canberra, ACT. <https://www.aihw.gov.au/getmedia/79e7d756-7cfe-49bf-b8c0-0bbb0daa2430/14825.pdf.aspx?inline=true>.
- Australian Institute of Health and Welfare. (2018). Health expenditure Australia 2016–17 (tech. rep. No. 64). Australian Institute of Health and Welfare. Canberra, ACT. <https://www.aihw.gov.au/getmedia/e8d37b7d-2b52-4662-a85f-01eb176f6844/aihw-hwe-74.pdf.aspx?inline=true>.
- Australian Institute of Health and Welfare. (2019). Hospital resources 2017–18: Australian hospital statistics (tech. rep. No. 233). Australian Institute of Health and Welfare. Canberra, ACT. <https://www.aihw.gov.au/reports/hospitals/hospital-resources-2017-18-ahs/contents/summary>.
- Badin, L., Daraio, C., & Simar, L. (2012). How to measure the impact of environmental factors in a nonparametric production model. *European Journal of Operational Research*, 223(3), 818–833. <https://doi.org/10.1016/j.ejor.2012.06.028>.
- Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30(9), 1078–1092. <https://doi.org/10.1287/mnsc.30.9.1078>.
- Beststremyannaya, G. (2013). The impact of Japanese hospital financing reform on hospital efficiency: A difference-in-difference approach. *The Japanese Economic Review*, 64(3), 337–362. <https://doi.org/10.1111/j.1468-5876.2012.00585.x>.
- Bogetoft, P., & Otto, L. (2019). Benchmarking: Benchmark and frontier analysis using DEA and SFA. *R package version*, 28. <https://cran.r-project.org/web/packages/Benchmarking>.
- Cazals, C., Florens, J.-P., & Simar, L. (2002). Nonparametric frontier estimation: A robust approach. *Journal of Econometrics*, 106(1), 1–25. [https://doi.org/10.1016/S0304-4076\(01\)00080-X](https://doi.org/10.1016/S0304-4076(01)00080-X).
- Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6), 429–444. [https://doi.org/10.1016/0377-2217\(78\)90138-8](https://doi.org/10.1016/0377-2217(78)90138-8).
- Chowdhury, H., & Zelenyuk, V. (2016). Performance of hospital services in Ontario: DEA with truncated regression approach. *Omega*, 63, 111–122. <https://doi.org/10.1016/j.omega.2015.10.007>.
- Chowdhury, H., Zelenyuk, V., Laporte, A., & Wodchis, W. P. (2014). Analysis of productivity, efficiency and technological changes in hospital services in Ontario: How does case-mix matter? *International Journal of Production Economics*, 150, 74–82. <https://doi.org/10.1016/j.pe.2013.12.003>.
- Clement, J. P., Valdmanis, V. G., Bazzoli, G. J., Zhao, M., & Chukmaitov, A. (2008). Is more better? an analysis of hospital outcomes and efficiency with a DEA model of output congestion. *Health Care Management Science*, 11(1), 67–77. <https://doi.org/10.1007/s10729-007-9025-8>.
- Council of Australian Governments. (2011). National health reform agreement. http://www.federalfinancialrelations.gov.au/content/npa/health/_archive/national-agreement.pdf.
- Daouia, A., & Laurent, T. (2013). Frontiles: Partial frontier efficiency analysis. *R package version*, 1, 2. <https://cran.r-project.org/web/packages/frontiles>.
- Daouia, A., & Simar, L. (2007). Nonparametric efficiency analysis: A multivariate conditional quantile approach. *Journal of Econometrics*, 140(2), 375–400. <https://doi.org/10.1016/j.jeconom.2006.07.002>.
- Daraio, C., & Simar, L. (2007). Economies of scale, scope and experience in the Italian motorvehicle sector. In Daraio, C., & Simar, L. (eds.), *Advanced robust and nonparametric methods in efficiency*

- analysis: Methodology and applications* (pp. 135–165). Springer Science & Business Media. https://doi.org/10.1007/978-0-387-35231-2_6.
- Daraio, C., Simar, L., & Wilson, P. W. (2018). Central limit theorems for conditional efficiency measures and tests of the “separability” condition in non-parametric, two-stage models of production. *The Econometrics Journal*, 21(2), 170–191. <https://doi.org/10.1111/ectj.12103>.
- Deprins, D., Simar, L., & Tulkens, H. (1984). Measuring labor efficiency in post offices. In M. G. Marchand, P. Pestieau, & H. Tulkens (Eds.), *The performance of public enterprises: Concepts and measurements* (pp. 243–267). North-Holland: Amsterdam.
- Färe, R., Grosskopf, S., & Logan, J. (1983). The relative efficiency of Illinois electric utilities. *Resources and Energy*, 5(4), 349–367. [https://doi.org/10.1016/0165-0572\(83\)90033-6](https://doi.org/10.1016/0165-0572(83)90033-6).
- Färe, R., & Zelenyuk, V., (2003). On aggregate Farrell efficiencies. *European Journal of Operational Research*, 146(3), 615–620. [https://doi.org/10.1016/S0377-2217\(02\)00259-X](https://doi.org/10.1016/S0377-2217(02)00259-X).
- Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, 120(3), 253–290. <https://doi.org/10.2307/2343100>.
- Grosskopf, S., Nguyen, B. H., Yong, J., & Zelenyuk, V. (2020). Healthcare structural reform and the performance of public hospitals: The case of Queensland, Australia [In Progress].
- Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics*, 17(10), 1107–1128. <https://doi.org/10.1002/hec.1391>.
- Hu, H. H., Qi, Q., & Yang, C. H. (2012). Evaluation of China’s regional hospital efficiency: DEA approach with undesirable output. *Journal of the Operational Research Society*, 63(6), 715–725. <https://doi.org/10.1057/jors.2011.77>.
- Kneip, A., Park, B. U., & Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, 14, 783–793. <https://doi.org/10.1017/S0266466698146042>.
- Kneip, A., Simar, L., & Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econometric Theory*, 24(6), 1663–1697. <https://doi.org/10.1017/S0266466608080651>.
- Kohl, S., Schoenfelder, J., & Fugener, A., & Brunner, J. O., (2019). The use of data envelopment analysis (DEA) in healthcare with a focus on hospitals. *Health Care Management Science*, 22(2), 245–286. <https://doi.org/10.1007/s10729-018-9436-8>.
- Meeusen, W., & van Den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18(2), 435–444. <https://doi.org/10.2307/2525757>.
- Munson, F. C., & Zuckerman, H. S. (1983). The managerial role. In Shortell, S. M., & Kaluzny, A. D. (eds.), *Health care management: A text in organization theory and behavior* (pp. 48–58). Wiley.
- Nguyen, B. H., & Zelenyuk, V. (2021). Aggregate efficiency of industry and its groups: The case of Queensland public hospitals. *Empirical Economics*. forthcoming. <https://doi.org/10.1007/s00181-020-01994-1>.
- O’Neill, L., Rauner, M., Heidenberger, K., & Kraus, M. (2008). A cross-national comparison and taxonomy of DEA-based hospital efficiency studies. *Socio-Economic Planning Sciences*, 42(3), 158–189. <https://doi.org/10.1016/j.seps.2007.03.001>.
- Park, B. U., Jeong, S.-O., & Simar, L. (2010). Asymptotic distribution of conical-hull estimators of directional edges. *The Annals of Statistics*, 38(3), 1320–1340. <https://doi.org/10.1214/09-AOS746>.
- Park, B. U., Simar, L., & Weiner, C. (2000). FDH efficiency scores from a stochastic point of view. *Econometric Theory*, 16, 855–877. <https://doi.org/10.1017/S0266466600166034>.
- Parmeter, C. F., & Zelenyuk, V. (2019). Combining the virtues of stochastic frontier and data envelopment analysis. *Operations Research*, 67(6), 1628–1658. <https://doi.org/10.1287/opre.2018.1831>.
- Paul, C. J. M. (2002). Productive structure and efficiency of public hospitals. In Fox, K. J. (ed.), *Efficiency in the public sector* (pp. 219–248). Boston, MA: Springer. https://doi.org/10.1007/978-1-4757-3592-5_9.

- Productivity Commission. (2010). Public and private hospital: Multivariate analysis (tech. rep. Supplement to Research Report). Productivity Commission. Canberra, ACT66. <https://www.pc.gov.au/inquiries/completed/hospitals/supplement/supplement.pdf>.
- Queensland Health. (2016). Health funding principles and guidelines 2016-17 financial year.
- R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Shephard, R. W. (1953). *Cost and production functions*. Princeton University Press.
- Shephard, R. W. (1970). *Theory of cost and production functions*. Princeton University Press.
- Sickles, R., & Zelenyuk, V. (2019). *Measurement of productivity and efficiency*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781139565981>.
- Simar, L. (2007). How to improve the performances of DEA/FDH estimators in the presence of noise? *Journal of Productivity Analysis*, 28(3), 183–201. <https://doi.org/10.1007/s11123-007-0057-3>.
- Simar, L., Van Keilegom, I., & Zelenyuk, V. (2017). Nonparametric least squares methods for stochastic frontier models. *Journal of Productivity Analysis*, 47(3), 189–204. <https://doi.org/10.1007/s11123-016-0474-2>.
- Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, 136(1), 31–64. <https://doi.org/10.016/j.jeconom.2005.07.009>.
- Simar, L., & Wilson, P.W., (2011). Two-stage DEA: Caveat emptor. *Journal of Productivity Analysis*, 36(2), 205–218. <https://doi.org/10.1007/s11123-011-0230-6>.
- Simar, L., & Wilson, P. W. (2013). Estimation and inference in nonparametric frontier models: Recent developments and perspectives. *Foundations and Trends® in Econometrics*, 5(3–4), 183–337. <https://doi.org/10.1561/08000000020>.
- Simar, L., & Wilson, P. W. (2015). Statistical approaches for non-parametric frontier models: A guided tour. *International Statistical Review*, 83(1), 77–110. <https://doi.org/10.1111/insr.12056>.
- Simar, L., & Wilson, P. W. (2020). Hypothesis testing in nonparametric models of production using multiple sample splits. *Journal of Productivity Analysis*, 53(3), 287–303. <https://doi.org/10.1007/s11123-020-00574-w>.
- Simar, L., & Zelenyuk, V. (2007). Statistical inference for aggregates of Farrell-type efficiencies. *Journal of Applied Econometrics*, 22(7), 1367–1394. <https://doi.org/10.1002/jae.991>.
- Simar, L., & Zelenyuk, V. (2011). Stochastic FDH/DEA estimators for frontier analysis. *Journal of Productivity Analysis*, 36(1), 1–20. <https://doi.org/10.1007/s11123-010-0170-6>.
- Simar, L., & Zelenyuk, V. (2018). Central limit theorems for aggregate efficiency. *Operations Research*, 66(1), 137–149. <https://doi.org/10.1287/opre.2017.1655>.
- Weisgrau, S. (1995). Issues in rural health: Access, hospitals, and reform. *Health care Financing Review*, 17(1), 1–14.
- Zelenyuk, V. (2020). Aggregation of inputs and outputs prior to data envelopment analysis under big data. *European Journal of Operational Research*, 282(1), 172–187. <https://doi.org/10.1016/j.ejor.2019.08.007>.

On the Behavior of Extreme d -dimensional Spatial Quantiles Under Minimal Assumptions



Davy Paindaveine and Joni Virta

Abstract *Spatial* or *geometric* quantiles are among the most celebrated concepts of multivariate quantiles. The spatial quantile $\mu_{\alpha,u}(P)$ of a probability measure P over \mathbb{R}^d is a point in \mathbb{R}^d indexed by an order $\alpha \in [0, 1)$ and a direction u in the unit sphere S^{d-1} of \mathbb{R}^d —or equivalently by a vector αu in the open unit ball of \mathbb{R}^d . Recently, Girard and Stupfler (2017) proved that (i) the extreme quantiles $\mu_{\alpha,u}(P)$ obtained as $\alpha \rightarrow 1$ exit all compact sets of \mathbb{R}^d and that (ii) they do so in a direction converging to u . These results help understanding the nature of these quantiles: the first result is particularly striking as it holds even if P has a bounded support, whereas the second one clarifies the delicate dependence of spatial quantiles on u . However, they were established under assumptions imposing that P is non-atomic, so that it is unclear whether they hold for empirical probability measures. We improve on this by proving these results under much milder conditions, allowing for the sample case. This prevents using gradient condition arguments, which makes the proofs very challenging. We also weaken the well-known sufficient condition for the uniqueness of finite-dimensional spatial quantiles.

1 Introduction

The problem of defining a satisfactory concept of multivariate quantiles in \mathbb{R}^d is a classical one and has generated a huge literature in nonparametric statistics; we refer to Serfling (2002) and the references therein. One of the most famous solutions is

D. Paindaveine (✉)

Université libre de Bruxelles (ECARES and Department of Mathematics) and Université Toulouse 1 Capitole (Toulouse School of Economics), Av. F.D. Roosevelt, 50, CP114/04, 1050, Brussels, Belgium
e-mail: dpaindav@ulb.ac.be

J. Virta

University of Turku (Department of Mathematics and Statistics) and Aalto University School of Science (Department of Mathematics and Systems Analysis), 20014 Turun yliopisto, Finland
e-mail: joni.virta@utu.fi

given by the *spatial* or *geometric* quantiles introduced in Chaudhuri (1996), which are a particular case of the multivariate M-quantiles from Breckling and Chambers (1988); see also Koltchinski (1997). Spatial quantiles are defined as follows.

Definition 1 Let P be a probability measure over \mathbb{R}^d . Fix $\alpha \in [0, 1)$ and $u \in \mathcal{S}^{d-1}$, where $\mathcal{S}^{d-1} := \{z \in \mathbb{R}^d : \|z\|^2 = z'z = 1\}$ is the unit sphere in \mathbb{R}^d . We will say that $\mu_{\alpha,u} = \mu_{\alpha,u}(P)$ is a *spatial quantile of order α in direction u for P* if and only if it minimizes the objective function

$$\mu \mapsto O_{\alpha,u}^P(\mu) := \int_{\mathbb{R}^d} \{\|z - \mu\| - \|z\| - \alpha u'z\} dP(z)$$

over \mathbb{R}^d (the second term in the integrand may look superfluous as it does not depend on μ , but it actually allows avoiding any moment conditions on P).

Existence and uniqueness of $\mu_{\alpha,u}$ will be discussed in the next section. It is easy to check that, for $d = 1$, spatial quantiles reduce to the usual univariate quantiles. The success of spatial quantiles is partly explained by their ability to cope with high-dimensional data and even functional data; see, e.g., Cardot et al. (2017), Cardot et al. (2013), Chakraborty and Chaudhuri (2014) and Chakraborty and Chaudhuri (2014). These quantiles were also used with much success to conduct multiple-output quantile regression, again also in the framework of functional data analysis; we refer to Chaouch and Laïb (2013), Cheng and De Gooijer (2007), and Chowdhury and Chaudhuri (2019). The present work, however, focuses on the finite-dimensional case.

In a slightly different perspective, spatial quantiles allow measuring the centrality of any given location in \mathbb{R}^d with respect to the probability measure P at hand: if the location z in \mathbb{R}^d coincides with the quantile $\mu_{\alpha,u}$, then a centrality measure for z is given by its *spatial depth* $1 - \alpha$; see Gao (2003), Serfling (2002) or Vardi and Zhang (2000). This also leads to a spatial concept of multivariate ranks; see, e.g., Serfling (2010). For recent results on spatial depth and spatial ranks, we refer to Serfling (2021a, b) and to the references therein. The deepest point of P , equivalently its most central quantile, is the quantile $\mu_0 := \mu_{0,u}$ obtained for $\alpha = 0$ (the dependence on u of course vanishes at $\alpha = 0$). This is the celebrated *spatial median*, which is one of the earliest robust location functionals; see, e.g., Brown (1983) or Haldane (1948). For the other quantiles, the larger α is, the less central the quantiles $\mu_{\alpha,u}$ are in each direction u .

The focus of the present work is on the extreme spatial quantiles that are obtained as α converges to one. Recently, Girard and Stupfler (2017) derived striking results on the behavior of such extreme spatial quantiles; see also Girard and Stupfler (2015). In particular, they showed that, under some assumptions on P that do not require that P has a bounded support, these quantiles exit all compact sets of \mathbb{R}^d . Their results, however, require in particular that P is non-atomic, hence remain silent about empirical distributions P_n associated with a random sample of size n from P . Of course, consistency results will imply that the behavior of sample extreme quantiles will mimic the behavior of the corresponding population quantiles as n diverges to

infinity; yet for any fixed n , even for large n , there is no guarantee that the results of Girard and Stupfler (2017) will apply. The goal of the present work is therefore to establish some of these results on extreme spatial quantiles under less stringent assumptions, that will allow for the sample case. Beyond this, we will also weaken the well-known sufficient condition for uniqueness of spatial quantiles. Our results are stated and discussed in Sect. 2, then are proved in Sect. 3.

2 Results

We will say that P is concentrated on a line with direction u_* ($\in \mathcal{S}^{d-1}$) if and only if there exists $z_0 \in \mathbb{R}^d$ such that $P[\{z_0 + \lambda u_* : \lambda \in \mathbb{R}\}] = 1$. Of course, we will say that P is concentrated on a line if and only if there exists $u_* \in \mathcal{S}^{d-1}$ such that P is concentrated on a line with direction u_* . We then have the following existence and uniqueness result.

Theorem 1 *Let P be a probability measure over \mathbb{R}^d . Fix $\alpha \in [0, 1)$ and $u \in \mathcal{S}^{d-1}$. Then, (i) P admits a spatial quantile $\mu_{\alpha,u}$. (ii) If P is not concentrated on a line, then $\mu_{\alpha,u}$ is unique. (iii) If P is not concentrated on a line with direction u , then $\mu_{\alpha,u}$ is unique for any $\alpha > 0$. (iv) If P is concentrated on a line with direction u , say, the line $\mathcal{L} = \{z_0 + \lambda u, \lambda \in \mathbb{R}\}$, then any spatial quantile $\mu_{\alpha,u}$ belongs to \mathcal{L} ; in this case, any such quantile is of the form $\mu_{\alpha,u} = z_0 + \ell_\alpha u$, where ℓ_α is a spatial quantile of order α in direction 1 for $P_{z_0,u}$, with $P_{z_0,u}$ the distribution of $u'(Z - z_0)$ when Z has distribution P .*

The existence result in Theorem 1(i) was established by Kemperman (1987), but, since this paper is not easily accessible, we provide our own proof in Sect. 3. The uniqueness result in Theorem 1(ii) is well-known and can be proved by generalizing to an arbitrary quantile the proof for the median in Milasevic and Ducharme (1987). The result in Theorem 1(iii) is original and shows that the only case where uniqueness of $\mu_{\alpha,u}$, $\alpha > 0$, may fail is the one where P is concentrated on a line with the corresponding direction u . If P is indeed of this form, then uniqueness may fail exactly as for univariate (spatial) quantiles; for instance, if P is the uniform distribution on $\{(-2, 0), (-1, 0), (0, 0), (1, 0), (2, 0)\}$, then any point of the form $(z, 0)$ with $1 \leq z \leq 2$ is a spatial quantile of order $\alpha = .6$ in direction $u = (1, 0)$ (recall that the indexing of the classical univariate quantiles differs from the center-outward indexing used for spatial quantiles). Finally, note that, in case (iii), the spatial quantile $\mu_{\alpha,u}$ may belong to the line on which P is concentrated (an example is given below the proof of Lemma 3).

Our main goal is to establish, under very mild conditions, two results that were recently proved in Girard and Stupfler (2017) under the assumptions that P is nonatomic and is not concentrated on a line. The first result states that, as α converges to one, spatial quantiles with order α will exit all compact sets in \mathbb{R}^d . Our extension of this result is the following.

Theorem 2 *Let P be a probability measure over \mathbb{R}^d . Let (α_n) be a sequence in $[0, 1)$ that converges to one and let (u_n) be a sequence in S^{d-1} . Assume that, for any accumulation point u_* of (u_n) , P is not concentrated on a line with direction u_* or*

$$\int_{\mathbb{R}^d} (\|z\| + u'_*z) dP(z) = \infty. \tag{1}$$

Then, $\|\mu_{\alpha_n, u_n}\| \rightarrow \infty$ as $n \rightarrow \infty$ for any sequence of quantiles (μ_{α_n, u_n}) .

Some comments are in order. First, the result does not require that spatial quantiles are unique, which materializes in the fact that the result is stated “for any sequence of quantiles”. Second, the result allows for distributions that are concentrated on a line, provided that the “moment-type” Condition (1) is satisfied. Clearly, it is necessary that P has infinite first-order moments (hence, an unbounded support) for this condition to be satisfied. It is not sufficient, though, as can be seen by considering the limiting behavior, as $\alpha \rightarrow 1$, of $\mu_{\alpha, u}$ for a probability measure that would be the distribution of the random vector $Z = -|\Lambda|u$, where Λ is Cauchy. Third, note that the result applies as soon as P is not concentrated on (typically, a few) specific lines, namely those with a direction given by an accumulation point of (u_n) . For instance, if $u_n = u$ for any n , then the result applies in particular as soon as P is not concentrated on a line with direction u . But this condition is not even necessary, as the above Cauchy example shows: for instance, in the Cauchy example above, $\|\mu_{\alpha, -u}\| \rightarrow \infty$ as $\alpha \rightarrow 1$. Last but not least, Theorem 2 does not require that P is non-atomic.

We illustrate this result on the basis of the following four examples, in which $P = P_n$ is the empirical measure associated with a sample $z_1, \dots, z_n \in \mathbb{R}^2$. In Example (a), $n = 4$ and the z_i 's were randomly drawn from the uniform distribution over $[-2, 2]^2$. The z_i 's in Example (b) are obtained by projecting those in Example (a) onto the line $\{(\lambda, 0) : \lambda \in \mathbb{R}\}$, whereas those in Example (c) are $z_i = (\cos \theta_i, \sin \theta_i)$, $i = 1, 2, 3$, with $\theta_i = 2\pi i/3$, hence are the vertices of an equilateral triangle. Finally, the four z_i 's in Example (d) are the vertices $(\pm 2, \pm 1)$ of a rectangle. These four settings were chosen since they represent point patterns in general position, along a line, on the vertices of a regular polygon, and on the vertices of a stretched regular polygon, respectively. For each of these examples, Fig. 1 shows the corresponding z_i 's as well as, for four different directions u (namely, $u = (\cos(\pi j/6), \sin(\pi j/6))$, $j = 0, 1, 2, 3$), (linear interpolations of) the spatial quantiles $\mu_{\alpha_m, u}$, $\alpha_m = .001, .002, \dots, .999$. The results are perfectly in line with Theorem 2. Note in particular that, in Example (b), in which P is concentrated on the line with direction $u_* = (1, 0)$, the spatial quantiles $\mu_{\alpha, u}$ exit all compact sets of \mathbb{R}^2 when $u \neq (\pm 1)u_*$, as anticipated by Theorem 2. This fails to happen for $u = u_*$, which is the only case in Fig. 1 for which our theoretical result remains silent.

The second result from Girard and Stupfler (2017) we generalize essentially states that the extreme spatial quantiles $\mu_{\alpha, u}$ are eventually to be found in direction u , which gives a clear interpretation to the direction u in which quantiles are considered (the

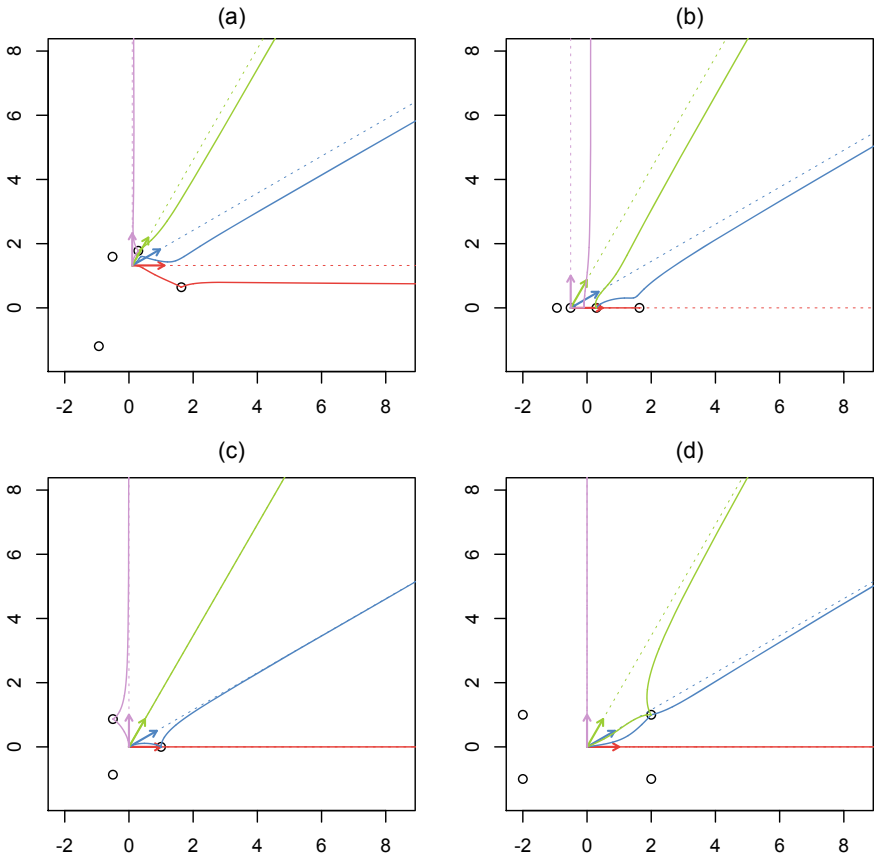


Fig. 1 For $u = (\cos(\pi j/6), \sin(\pi j/6))$, with $j = 0$ (red), 1 (blue), 2 (green), and 3 (purple), the plots show (linear interpolations of) the spatial quantiles $\mu_{\alpha_m, u}$, $\alpha_m = 0.001, 0.002, \dots, 0.999$, in each of the examples **a–d** described in Sect. 2. Dashed lines are showing the halflines with corresponding directions u originating from the spatial median

directions of non-extreme spatial quantiles do not allow for such a clear interpretation). Our version of this result is the following.

Theorem 3 *Let P be a probability measure over \mathbb{R}^d . Let (α_n) be a sequence in $[0, 1)$ that converges to one and let (u_n) be a sequence in \mathcal{S}^{d-1} that converges to u . Assume that P is not concentrated on a line with direction u or that*

$$\int_{\mathbb{R}^d} (\|z\| + u'z) dP(z) = \infty$$

Then, $\mu_{\alpha_n, u_n} / \|\mu_{\alpha_n, u_n}\| \rightarrow u$ as $n \rightarrow \infty$ for any sequence of quantiles (μ_{α_n, u_n}) .

The same comments made below Theorem 2 can be repeated here, but for the fact that the sequence (u_n) here may only have one accumulation point, namely its limit u . Again, the result holds for atomic probability measures, which allows us to illustrate the results in Examples (a)–(d) above. Clearly, Fig. 1 reflects well the conclusion of Theorem 3 in all cases, including those where the probability measure P is concentrated on a line (again, the case associated with $u = (1, 0)$ in Example (b) is the only one for which our result remains silent).

3 Proofs

The proof of Theorem 1 requires the following three lemmas.

Lemma 1 *Let P be a probability measure over \mathbb{R}^d . Fix $\alpha \in [0, 1)$ and $u \in S^{d-1}$. Then, (i) $\mu \mapsto O_{\alpha,u}^P(\mu)$ is convex over \mathbb{R}^d , that is, for $\mu_0, \mu_1 \in \mathbb{R}^d$ ($\mu_0 \neq \mu_1$) and $t \in (0, 1)$, one has $O_{\alpha,u}^P(\mu_t) \leq (1 - t)O_{\alpha,u}^P(\mu_0) + tO_{\alpha,u}^P(\mu_1)$, where we let $\mu_t := (1 - t)\mu_0 + t\mu_1$. (ii) With the same notation, if P is not concentrated on the line containing μ_0 and μ_1 , then $O_{\alpha,u}^P(\mu_t) < (1 - t)O_{\alpha,u}^P(\mu_0) + tO_{\alpha,u}^P(\mu_1)$.*

Proof of Lemma 1. Fix $\mu_0, \mu_1 \in \mathbb{R}^d$ and $t \in (0, 1)$. Then, with $\mu_t = (1 - t)\mu_0 + t\mu_1$, we readily have

$$\begin{aligned} & \|z - \mu_t\| - \|z\| - \alpha u' \mu_t \\ & \leq (1 - t)\{\|z - \mu_0\| - \|z\| - \alpha u' \mu_0\} + t\{\|z - \mu_1\| - \|z\| - \alpha u' \mu_1\}. \end{aligned} \tag{2}$$

Part (i) of the result is then obtained by integrating over \mathbb{R}^d with respect to P . As for Part (ii), it follows from the fact that the inequality in (2) is strict for any z that does not belong to the line containing μ_0 and μ_1 . □

Lemma 2 *Let P be a probability measure over \mathbb{R}^d . Fix $\alpha \in [0, 1)$ and $u \in S^{d-1}$. Then, P admits a spatial quantile $\mu_{\alpha,u}$.*

Proof of Lemma 2. Write $B_R := \{z \in \mathbb{R}^d : \|z\| \leq R\}$ and fix $\lambda > (1 + \alpha)/(1 - \alpha)$. Pick R_0 large enough so that $P[B_{R_0}] \geq \lambda/(\lambda + 1)$. Then,

$$O_{\alpha,u}^P(\mu) = \int_{\mathbb{R}^d} \{\|z - \mu\| - \|z\| - \alpha u' \mu\} dP(z) = O_1(\mu) + O_2(\mu),$$

where we have

$$\begin{aligned}
 O_1(\mu) &:= \int_{B_{R_0}} \{ \|z - \mu\| - \|z\| - \alpha u' \mu \} dP(z) \\
 &\geq \int_{B_{R_0}} \{ \|\mu\| - 2\|z\| - \alpha \|\mu\| \} dP(z) \\
 &\geq \frac{\lambda(1 - \alpha)\|\mu\|}{\lambda + 1} - 2R_0
 \end{aligned}$$

and

$$\begin{aligned}
 O_2(\mu) &:= \int_{\mathbb{R}^d \setminus B_{R_0}} \{ \|z - \mu\| - \|z\| - \alpha u' \mu \} dP(z) \\
 &\geq \int_{\mathbb{R}^d \setminus B_{R_0}} \{ -\|\mu\| - \alpha \|\mu\| \} dP(z) \\
 &\geq -\frac{(1 + \alpha)\|\mu\|}{\lambda + 1}.
 \end{aligned}$$

Therefore, for any μ , we have

$$O_{\alpha,u}^P(\mu) \geq \frac{\lambda(1 - \alpha) - (1 + \alpha)}{\lambda + 1} \|\mu\| - 2R_0 =: c_{\lambda,\alpha} \|\mu\| - 2R_0,$$

where $c_{\lambda,\alpha}$ is strictly positive. To conclude, pick $R > 0$ so that $c_{\lambda,\alpha}R - 2R_0 > O_{\alpha,u}^P(0)$. As a convex function, $\mu \mapsto O_{\alpha,u}^P(\mu)$ is continuous, hence admits a minimum, μ_* say, in the compact set $K := \{\mu \in \mathbb{R}^d : \|\mu\| \leq R\}$. Since any $\mu \notin K$ is such that

$$O_{\alpha,u}^P(\mu) \geq c_{\lambda,\alpha}R - 2R_0 > O_{\alpha,u}^P(0) \geq \min_{\mu \in K} O_{\alpha,u}^P(\mu),$$

we conclude that μ_* also minimizes $\mu \mapsto O_{\alpha,u}^P(\mu)$ over \mathbb{R}^d , which establishes the result. \square

Lemma 3 *Let P be a probability measure over \mathbb{R}^d that is concentrated on a line, \mathcal{L} say, with direction $u_* \in \mathcal{S}^{d-1}$. Fix $\alpha \in (0, 1)$ and $u \in \mathcal{S}^{d-1} \setminus \{\pm u_*\}$. Then, either $\mu_{\alpha,u}$ is unique and belongs to \mathcal{L} , or there exists a quantile $\mu_{\alpha,u}$ that does not belong to \mathcal{L} .*

Proof of Lemma 3. By Lemma 2, there exists at least a quantile $\mu_{\alpha,u}$. Trivially, the same proof also shows that $\mu \mapsto O_{\alpha,u}^P(\mu)$ has a minimizer on \mathcal{L} . Fix then $\mu_*(\in \mathcal{L})$ arbitrarily such that $O_{\alpha,u}^P(\mu_*) \leq O_{\alpha,u}^P(\mu)$ for any $\mu \in \mathcal{L}$.

Let Z be a random d -vector with distribution P . By assumption, $Z = \mu_* + \Lambda u_*$ for some random variable Λ , with distribution P^Λ say. For any $v \in \mathcal{S}^{d-1}$ and any $h > 0$, we then have

$$\begin{aligned} \frac{O_{\alpha,u}^P(\mu_* + hv) - O_{\alpha,u}^P(\mu_*)}{h} &= -\alpha u'v + \int_{\mathbb{R}^d} \frac{\|z - (\mu_* + hv)\| - \|z - \mu_*\|}{h} dP(z) \\ &= -\alpha u'v + \int_{\mathbb{R}} \frac{\|\lambda u_* - hv\| - \|\lambda u_*\|}{h} dP^\Lambda(\lambda), \end{aligned}$$

so that

$$\frac{O_{\alpha,u}^P(\mu_* + hv) - O_{\alpha,u}^P(\mu_*)}{h} - \{P^\Lambda[\{0\}] - s_P u'_*v - \alpha u'v\} = \int_{\mathbb{R}} \ell_h(\lambda) dP^\Lambda(\lambda),$$

where we let $s_P := E[\text{Sign}(\Lambda)]$ and

$$\ell_h(\lambda) := \frac{\|\lambda u_* - hv\| - \|\lambda u_*\|}{h} - \left\{ \mathbb{I}[\lambda = 0] - \text{Sign}(\lambda)(u'_*v)\mathbb{I}[\lambda \neq 0] \right\}.$$

It is easy to check that, for any $\lambda \in \mathbb{R}$, the limit of $\ell_h(\lambda)$ as $h \rightarrow 0$ from above exists and is equal to zero. Moreover, by using the inequality $|\|x\| - \|y\|| \leq \|x - y\|$, it is readily seen that the function $\lambda \mapsto |\ell_h(\lambda)|$ is upper-bounded by the function $\lambda \mapsto 2 + |u'_*v|$ that does not depend on h and is P^Λ -integrable. Therefore, Lebesgue's Dominated Convergence Theorem entails that $\mu \mapsto O_{\alpha,u}^P(\mu)$ admits a directional derivative in direction v at μ_* , and that this directional derivative is given by

$$\frac{\partial O_{\alpha,u}^P}{\partial v}(\mu_*) = P^\Lambda[\{0\}] - v'(s_P u_* + \alpha u). \tag{3}$$

Now, using the fact that u_* and u are linearly independent and that $\alpha > 0$, one has

$$m_{\alpha,u}(\mu_*) := \min_{v \in \mathcal{S}^{d-1}} \frac{\partial O_{\alpha,u}^P}{\partial v}(\mu_*) = P^\Lambda[\{0\}] - \|s_P u_* + \alpha u\|,$$

where the minimum is reached at $v_0 := (s_P u_* + \alpha u) / \|s_P u_* + \alpha u\| (\neq u_*)$ only. We then consider two cases. (i) $m_{\alpha,u}(\mu_*) < 0$: then, there exists $h > 0$ such that $O_{\alpha,u}^P(\mu_* + hv_0) < O_{\alpha,u}^P(\mu_*)$, in which case $O_{\alpha,u}^P(\mu_* + hv_0) < O_{\alpha,u}^P(\mu)$ for any $\mu \in \mathcal{L}$, so that any global minimizer of $\mu \mapsto O_{\alpha,u}^P(\mu)$ does not belong to \mathcal{L} . (ii) $m_{\alpha,u}(\mu_*) \geq 0$: then, any directional derivative in (3) associated with $v \in \mathcal{S}^{d-1} \setminus \{v_0\}$ is strictly positive, so that, for any such v , one has $O_{\alpha,u}^P(\mu_* + hv) > O_{\alpha,u}^P(\mu_*)$ for any h in an interval of the form $(0, \varepsilon_v)$. Pick then, for a fixed $v \in \mathcal{S}^{d-1} \setminus \{v_0\}$ and the corresponding interval $(0, \varepsilon_v)$, an arbitrary $h \in [\varepsilon_v, \infty)$ and any $h_\varepsilon \in (0, \varepsilon_v)$, and write $h_\varepsilon = (1 - \lambda) \times 0 + \lambda h$, for $\lambda := h_\varepsilon / h \in (0, 1)$. The convexity of $O_{\alpha,u}^P$ (Lemma 1(i)) entails that

$$\lambda \{O_{\alpha,u}^P(\mu_* + hv) - O_{\alpha,u}^P(\mu_*)\} \geq O_{\alpha,u}^P(\mu_* + h_\varepsilon v) - O_{\alpha,u}^P(\mu_*) > 0,$$

showing that actually $O_{\alpha,u}^P(\mu_* + hv) > O_{\alpha,u}^P(\mu_*)$ for any $h > 0$. Continuity of $\mu \mapsto O_{\alpha,u}^P(\mu)$ (which also follows from convexity) implies that $f(h) := O_{\alpha,u}^P(\mu_* +$

$h v_0) \geq f(0)$ for any $h > 0$ (would there exist $h > 0$ such that $O_{\alpha,u}^P(\mu_* + h v_0) - O_{\alpha,u}^P(\mu_*) = f(h) - f(0) < 0$, then, from continuity, there would exist $v \in \mathcal{S}^{d-1} \setminus \{v_0\}$ such that $O_{\alpha,u}^P(\mu_* + h v) - O_{\alpha,u}^P(\mu_*) < 0$, a contradiction). It follows that μ_* minimizes $\mu \mapsto O_{\alpha,u}^P(\mu)$ over \mathbb{R}^d . If $f(h) > f(0)$ for any $h > 0$, then this minimizer is unique, whereas if $O_{\alpha,u}^P(\mu_* + h_0 v_0) = f(h_0) = f(0) = O_{\alpha,u}^P(\mu_*)$ for some $h_0 > 0$, then $\mu_* + h_0 v_0 \notin \mathcal{L}$ also minimizes $\mu \mapsto O_{\alpha,u}^P(\mu)$ over \mathbb{R}^d . The result follows. \square

In the framework of Lemma 3, it may indeed happen that $\mu_{\alpha,u}$ is unique and belongs to \mathcal{L} . For instance, if P is the uniform distribution over $\{(-1, 0), (0, 0), (1, 0)\} \subset \mathbb{R}^2$, $\alpha \in (0, \frac{1}{3})$ and $u = (0, 1)$, then P is concentrated on the line $\mathcal{L} = \{\lambda u_* : \lambda \in \mathbb{R}\}$, with $u_* = (1, 0)$, and $\mu_{\alpha,u} = (0, 0) \in \mathcal{L}$ is the unique order- α quantile in direction u for P (this can be checked by proceeding as in the proof of Lemma 3).

We can now prove Theorem 1.

Proof of Theorem 1. (i) The result is an exact restatement of Lemma 2.

(ii) The proof is a straightforward extension of the one in Milasevic and Ducharme (1987). By contradiction, assume that there exist μ_0 and μ_1 , with $\mu_0 \neq \mu_1$, such that $O_{\alpha,u}^P(\mu_0) = O_{\alpha,u}^P(\mu_1)$ is the minimum of $\mu \mapsto O_{\alpha,u}^P(\mu)$ over \mathbb{R}^d . Since, by assumption, P is not concentrated on the line containing μ_0 and μ_1 , Lemma 1(ii) readily yields that, for any $t \in (0, 1)$,

$$O_{\alpha,u}^P((1-t)\mu_0 + t\mu_1) < (1-t)O_{\alpha,u}^P(\mu_0) + tO_{\alpha,u}^P(\mu_1) = O_{\alpha,u}^P(\mu_0),$$

which contradicts the fact that μ_0 minimizes $\mu \mapsto O_{\alpha,u}^P(\mu)$.

(iii) As in the proof of Part (ii), assume by contradiction that $\mu \mapsto O_{\alpha,u}^P(\mu)$ has at least two minimizers in \mathbb{R}^d , now with $\alpha > 0$. In view of Part (ii) of the result, it is enough to consider the case where P would be concentrated on a line \mathcal{L} with direction $u_*(\neq \pm u)$. Lemma 3 thus applies and guarantees that there exists a minimizer of $\mu \mapsto O_{\alpha,u}^P(\mu)$ that does not belong to \mathcal{L} . Thus, it is possible to pick minimizers μ_0 and μ_1 of $\mu \mapsto O_{\alpha,u}^P(\mu)$, with $\mu_0 \notin \mathcal{L}$ and $\mu_0 \neq \mu_1$. Clearly, P is not concentrated on the line containing μ_0 and μ_1 (would it be the case, then P would be the Dirac measure at the intersection, $\{\mu\}$ say, between \mathcal{L} and the line containing μ_0 and μ_1 , hence in particular would be concentrated on the line $\{\mu + \lambda u : \lambda \in \mathbb{R}\}$ that has direction u , a contradiction). Therefore, Lemma 1(ii) again yields that, for any $t \in (0, 1)$,

$$O_{\alpha,u}^P((1-t)\mu_0 + t\mu_1) < (1-t)O_{\alpha,u}^P(\mu_0) + tO_{\alpha,u}^P(\mu_1) = O_{\alpha,u}^P(\mu_0),$$

which contradicts the fact that μ_0 minimizes $\mu \mapsto O_{\alpha,u}^P(\mu)$.

(iv) Assume that P is concentrated on $\mathcal{L} = \{z_0 + \lambda u, \lambda \in \mathbb{R}\}$. Fix $\mu \notin \mathcal{L}$. Let us first show that μ is not a spatial quantile of order α in direction u for P . To do so, write $Z = \mu_{\mathcal{L}} + \Lambda u$, where $\mu_{\mathcal{L}}$ is the orthogonal projection of μ onto \mathcal{L} . Define further $w := (\mu_{\mathcal{L}} - \mu)/c$, with $c := \|\mu_{\mathcal{L}} - \mu\|$. Since $u'w = 0$, we then have

$$\begin{aligned} \frac{O_{\alpha,u}^P(\mu + hw) - O_{\alpha,u}^P(\mu)}{h} &= -\alpha u'w + \int_{\mathbb{R}^d} \frac{\|z - (\mu + hw)\| - \|z - \mu\|}{h} dP(z) \\ &= \int_{\mathbb{R}} \frac{\|(\mu_{\mathcal{L}} + \lambda u) - (\mu + hw)\| - \|(\mu_{\mathcal{L}} + \lambda u) - \mu\|}{h} dP^\Lambda(\lambda) \\ &= \int_{\mathbb{R}} \frac{\|\lambda u + cw - hw\| - \|\lambda u + cw\|}{h} dP^\Lambda(\lambda). \end{aligned}$$

This yields

$$\frac{O_{\alpha,u}^P(\mu + hw) - O_{\alpha,u}^P(\mu)}{h} + \int_{\mathbb{R}} \frac{w'(\lambda u + cw)}{\|\lambda u + cw\|} dP^\Lambda(\lambda) = \int_{\mathbb{R}} g_h(\lambda) dP^\Lambda(\lambda),$$

where

$$\begin{aligned} g_h(\lambda) &:= \frac{\|\lambda u + cw - hw\| - \|\lambda u + cw\|}{h} + \frac{w'(\lambda u + cw)}{\|\lambda u + cw\|} \\ &= \frac{h^2 - 2hw'(\lambda u + cw)}{h(\|\lambda u + cw - hw\| + \|\lambda u + cw\|)} + \frac{w'(\lambda u + cw)}{\|\lambda u + cw\|}. \end{aligned}$$

Clearly, $\lambda \mapsto |g_h(\lambda)|$ is, for $h \in (0, 1)$ say, upper-bounded by the function $\lambda \mapsto (1/\|\lambda u + cw\|) + 3$ that is P^Λ -integrable and does not depend on h (integrability follows from the fact that $\|\lambda u + cw\|^2 = \lambda^2 + c^2 \geq c^2$). Moreover, $g_h(\lambda) \rightarrow 0$ as $h \rightarrow 0$ for any λ . Lebesgue's Dominated Convergence Theorem thus shows that the directional derivative of $O_{\alpha,u}^P$ at μ in direction w exists and is equal to

$$\frac{\partial O_{\alpha,u}^P}{\partial w}(\mu) = - \int_{\mathbb{R}} \frac{w'(\lambda u + cw)}{\|\lambda u + cw\|} dP^\Lambda(\lambda) = - \int_{\mathbb{R}} \frac{c}{\|\lambda u + cw\|} dP^\Lambda(\lambda) < 0.$$

Therefore, μ is not a spatial quantile of order α in direction u for P .

Consequently, all spatial quantiles of order α in direction u for P belong to \mathcal{L} . These can be characterized as follows. Redefine the random variable Λ through $Z = z_0 + \Lambda u$ (in other words, $\Lambda = u'(Z - z_0)$). Spatial quantiles are the minimizers of $\mu \mapsto O_{\alpha,u}^P(\mu)$ over \mathbb{R}^d , which (we just showed it) coincide with the minimizers of the same mapping over \mathcal{L} . These minimizers take the form $z_0 + \ell_\alpha u$, where ℓ_α minimizes

$$\begin{aligned} \lambda \mapsto O_{\alpha,u}^P(z_0 + \lambda u) &= \int_{\mathbb{R}^d} \{ \|z - (z_0 + \lambda u)\| - \|z\| - \alpha u'(z_0 + \lambda u) \} dP(z) \\ &= -\alpha u'z_0 + \int_{\mathbb{R}} \{ |t - \lambda| - \|z_0 + tu\| - \alpha \lambda \} dP^\Lambda(t), \end{aligned}$$

or, equivalently, minimizes

$$\lambda \mapsto \int_{\mathbb{R}} \{|t - \lambda| - |t| - \alpha\lambda\} dP^\Lambda(t)$$

(note that this last (objective) function, hence also the corresponding minimizers, do not depend on u , which a posteriori justifies the notation ℓ_α). In other words, ℓ_α is a spatial quantile of order α in direction 1 for P^Λ . \square

The proof of Theorem 2 requires both following preliminary results.

Lemma 4 *Let P be a probability measure over \mathbb{R}^d . Then, the function*

$$(\alpha, u, \mu) \mapsto O_{\alpha,u}^P(\mu) = \int_{\mathbb{R}^d} \{\|z - \mu\| - \|z\| - \alpha u' \mu\} dP(z) \tag{4}$$

is continuous over $[0, 1] \times \mathcal{S}^{d-1} \times \mathbb{R}^d$.

Proof of Lemma 4. Since

$$\begin{aligned} & |O_{\alpha_2, u_2}^P(\mu_2) - O_{\alpha_1, u_1}^P(\mu_1)| \\ & \leq \int_{\mathbb{R}^d} \left| \|z - \mu_2\| - \|z - \mu_1\| - (\alpha_2 u_2' \mu_2 - \alpha_1 u_1' \mu_1) \right| dP(z) \\ & \leq \|\mu_2 - \mu_1\| + |\alpha_2 u_2' \mu_2 - \alpha_1 u_1' \mu_1| \\ & \leq \|\mu_2\| |\alpha_2 - \alpha_1| + \|\mu_2\| \|u_2 - u_1\| + (1 + \alpha_1) \|\mu_2 - \mu_1\|, \end{aligned}$$

the function in (4) is Lipschitz over any bounded subset of $[0, 1] \times \mathcal{S}^{d-1} \times \mathbb{R}^d$. The result follows. \square

Lemma 5 *Let P be a probability measure over \mathbb{R}^d and fix $u \in \mathcal{S}^{d-1}$. Assume that P is not concentrated on a line with direction u or that*

$$\int_{\mathbb{R}^d} (\|z\| + u'z) dP(z) = \infty. \tag{5}$$

Then the function

$$\mu \mapsto O_{1,u}^P(\mu) := \int_{\mathbb{R}^d} \{\|z - \mu\| - \|z\| - u' \mu\} dP(z)$$

does not have a minimum in \mathbb{R}^d .

Proof of Lemma 5. Since P and u are fixed, we will write $g(\mu) := O_{1,u}^P(\mu)$ throughout the proof. Letting $\mu_n := n\mu$ (with n a positive integer), this allows us to write

$$\begin{aligned}
 g(\mu_n) &= \int_{\mathbb{R}^d} \{ \|z - nu\| - (\|z\| + n) \} dP(z) \\
 &= -2n \int_{\mathbb{R}^d} \frac{\|z\| + u'z}{\|z - nu\| + \|z\| + n} dP(z) \\
 &= g_{<}(\mu_n) + g_{\geq}(\mu_n),
 \end{aligned}$$

where we let

$$g_{<}(\mu_n) := -2n \int_{\mathbb{R}^d} \frac{(\|z\| + u'z)\mathbb{I}[u'z < 0]}{\|z - nu\| + \|z\| + n} dP(z) (\leq 0)$$

and

$$g_{\geq}(\mu_n) := -2n \int_{\mathbb{R}^d} \frac{(\|z\| + u'z)\mathbb{I}[u'z \geq 0]}{\|z - nu\| + \|z\| + n} dP(z) (\leq 0).$$

Now, note that if (5) holds, then

$$\int_{\mathbb{R}^d} \|z\|\mathbb{I}[u'z \geq 0] dP(z) = \infty \quad \text{or} \quad \int_{\mathbb{R}^d} (\|z\| + u'z)\mathbb{I}[u'z < 0] dP(z) = \infty$$

(or both integrals are infinite). This leads to consider three cases.

Case (A): $\int_{\mathbb{R}^d} \|z\|\mathbb{I}[u'z \geq 0] dP(z) = \infty$. Of course, we have

$$-g_{\geq}(\mu_n) \geq 2n \int_{\mathbb{R}^d} \frac{\|z\|\mathbb{I}[u'z \geq 0]}{\|z - nu\| + \|z\| + n} dP(z).$$

Since $(\|z\| + n)^2 - \|z - nu\|^2 = 2n\|z\| + 2nu'z \geq 0$, we also have

$$-g_{\geq}(\mu_n) \geq \int_{\mathbb{R}^d} \frac{n\|z\|\mathbb{I}[u'z \geq 0]}{\|z\| + n} dP(z) =: \int_{\mathbb{R}^d} h_n(z) dP(z). \tag{6}$$

Since $h_n(z) \leq h_{n+1}(z)$ for any z and the pointwise limit of h_n is the function h defined by $h(z) := \|z\|\mathbb{I}[u'z \geq 0]$, the Monotone Convergence Theorem yields

$$\int_{\mathbb{R}^d} h_n(z) dP(z) \rightarrow \int_{\mathbb{R}^d} h(z) dP(z) = \infty,$$

which, jointly with (6), establishes that $g_{\geq}(\mu_n) \rightarrow -\infty$. Since $g(\mu_n) \leq g_{\geq}(\mu_n)$, we conclude that $g(\mu_n) \rightarrow -\infty$, so that g does not have a minimum in Case (A).

Case (B): $\int_{\mathbb{R}^d} (\|z\| + u'z)\mathbb{I}[u'z < 0] dP(z) = \infty$. Using the Monotone Convergence Theorem as in Case (A) readily provides that

$$\begin{aligned}
 -g_{<}(\mu_n) &= 2n \int_{\mathbb{R}^d} \frac{(\|z\| + u'z)\mathbb{I}[u'z < 0]}{\|z - nu\| + \|z\| + n} dP(z) \\
 &= 2 \int_{\mathbb{R}^d} \frac{(\|z\| + u'z)\mathbb{I}[u'z < 0]}{\sqrt{\frac{1}{n^2}\|z\|^2 + 1} + \frac{2}{n}|u'z| + \frac{1}{n}\|z\| + 1} dP(z)
 \end{aligned}$$

converges to

$$\int_{\mathbb{R}^d} (\|z\| + u'z)\mathbb{I}[u'z < 0] dP(z) = \infty$$

as $n \rightarrow \infty$. Since $g(\mu_n) \leq g_{<}(\mu_n)$, this yields $g(\mu_n) \rightarrow -\infty$. It follows that g does not have a minimum in Case (B).

Case (C): $\int_{\mathbb{R}^d} \|z\|\mathbb{I}[u'z \geq 0] dP(z) < \infty$ and $\int_{\mathbb{R}^d} (\|z\| + u'z)\mathbb{I}[u'z < 0] dP(z) < \infty$. Using the finiteness of the first and second integrals, Lebesgue's Dominated Convergence Theorem readily yields

$$g_{\geq}(\mu_n) \rightarrow - \int_{\mathbb{R}^d} (\|z\| + u'z)\mathbb{I}[u'z \geq 0] dP(z)$$

and

$$g_{<}(\mu_n) \rightarrow - \int_{\mathbb{R}^d} (\|z\| + u'z)\mathbb{I}[u'z < 0] dP(z),$$

respectively. Therefore,

$$g(\mu_n) = g_{<}(\mu_n) + g_{\geq}(\mu_n) \rightarrow - \int_{\mathbb{R}^d} (\|z\| + u'z) dP(z) =: i_u^P.$$

In Case (C), P is not concentrated on a line with direction u by assumption, which implies that, for any $\mu \in \mathbb{R}^d$,

$$g(\mu) - i_u^P = \int_{\mathbb{R}^d} \{\|z - \mu\| + u'(z - \mu)\} dP(z) > 0.$$

This shows that the function g does not have a minimum in Case (C) either. The result is thus proved. \square

Theorem 2 then follows from Lemmas 4–5 in the same way as Theorem 2.1(i) in Girard and Stupfler (2017) (but for the fact that we are considering distributions that do not ensure uniqueness of quantiles). We still report the proof for the sake of completeness.

Proof of Theorem 2. Ad absurdum, assume that there exists a sequence of quantiles (μ_{α_n, u_n}) such that $\|\mu_{\alpha_n, u_n}\|$ does not diverge to infinity. Then, $(\mu_{\alpha_n, u_n}, u_n)$ has a subsequence that is bounded, hence from compactness, possesses a further subsequence, $(\mu_{\alpha_{n_\ell}, u_{n_\ell}}, u_{n_\ell})$ say, that converges in $\mathbb{R}^d \times \mathcal{S}^{d-1}$, to (μ_∞, u_∞) , say. By construction, u_∞ is an accumulation point of the sequence (u_n) . For any ℓ , we have

$$O_{\alpha_{n_\ell}, u_{n_\ell}}^P(\mu_{\alpha_{n_\ell}, u_{n_\ell}}) \leq O_{\alpha_{n_\ell}, u_{n_\ell}}^P(\mu)$$

for any $\mu \in \mathbb{R}^d$. In view of Lemma 4, taking limits as $\ell \rightarrow \infty$ then provides

$$O_{1, u_\infty}^P(\mu_\infty) \leq O_{1, u_\infty}^P(\mu)$$

for any $\mu \in \mathbb{R}^d$. Since this contradicts Lemma 5, the result is proved. □

The proof of Theorem 3 requires the following lemma.

Lemma 6 *Let P be a probability measure over \mathbb{R}^d and fix $m \in (0, 2)$. Then,*

$$t_P(r) := \int_{\mathbb{R}^d} \frac{\|z\|}{\sqrt{(\|z\| - r)^2 + mr\|z\|}} dP(z) \rightarrow 0$$

as $r \rightarrow \infty$.

Proof of Lemma 6. Fix $\delta > 0$. For any $r > 0$, let $Y_r := \|Z\|/r$, where Z is a random d -vector with distribution P . Then, with $h := m\delta^2/4$,

$$\begin{aligned} t_P(r) &= \mathbb{E} \left[\frac{\|Z\|}{\sqrt{(\|Z\| - r)^2 + mr\|Z\|}} \right] = \mathbb{E} \left[\frac{Y_r}{\sqrt{(Y_r - 1)^2 + mY_r}} \right] \\ &= \mathbb{E} \left[\frac{Y_r \mathbb{I}[Y_r \leq h]}{\sqrt{(Y_r - 1)^2 + mY_r}} \right] + \mathbb{E} \left[\frac{Y_r \mathbb{I}[Y_r > h]}{\sqrt{(Y_r - 1)^2 + mY_r}} \right]. \end{aligned}$$

Since $y/\sqrt{(y - 1)^2 + my} \leq 2/\sqrt{m(4 - m)}$ for any $y \geq 0$, this provides

$$\begin{aligned} t_P(r) &\leq \mathbb{E} \left[\frac{\sqrt{Y_r} \mathbb{I}[Y_r \leq h]}{\sqrt{m}} \right] + \frac{2}{\sqrt{m(4 - m)}} P[Y_r > h] \\ &\leq \frac{\delta}{2} + \frac{2}{\sqrt{m(4 - m)}} P[\|Z\| > rh] < \delta, \end{aligned}$$

for r large enough. □

Proof of Theorem 3. In this proof, we use the notation

$$\mathcal{S}_{u,c}^{\text{in}} := \mathcal{S}^{d-1} \cap \{z \in \mathbb{R}^d : u'z \geq 1 - c\}$$

and

$$\mathcal{S}_{u,c}^{\text{out}} := \mathcal{S}^{d-1} \cap \{z \in \mathbb{R}^d : u'z \leq 1 - c\}.$$

Ad absurdum, assume that there exists a sequence of quantiles (μ_{α_n, u_n}) such that $(w_n := \mu_{\alpha_n, u_n} / \|\mu_{\alpha_n, u_n}\|)$ does not converge to u . Thus, there exists $\varepsilon > 0$ such that $w_n \in \mathcal{S}_{u, \varepsilon}^{\text{out}}$ for infinitely many n . Upon extraction of a subsequence, we may assume that w_n belongs to $\mathcal{S}_{u, \varepsilon}^{\text{out}}$ for any n . By assumption, we may, still upon extraction of

a subsequence, assume that $u_n \in \mathcal{S}_{u,\varepsilon/2}^{\text{in}}$ for any n . Assume for a moment that there exist $R > 0$ and $\eta \in (0, 1)$ such that

$$O_{\alpha,v}^P(rw) > O_{\alpha,v}^P(rv) \tag{7}$$

for any $\alpha \in [\eta, 1)$, $r \geq R$, $v \in \mathcal{S}_{u,\varepsilon/2}^{\text{in}}$ and $w \in \mathcal{S}_{u,\varepsilon}^{\text{out}}$. Pick then n large enough to have $\alpha_n \geq \eta$ and $\|\mu_{\alpha_n, u_n}\| \geq R$ (existence follows from Theorem 2). By definition, this implies that

$$O_{\alpha_n, u_n}^P(\|\mu_{\alpha_n, u_n}\|w_n) = O_{\alpha_n, u_n}^P(\mu_{\alpha_n, u_n}) \leq O_{\alpha_n, u_n}^P(\|\mu_{\alpha_n, u_n}\|u_n),$$

which contradicts (7).

Therefore, it is sufficient to prove (7). To do so, fix $v \in \mathcal{S}_{u,\varepsilon/2}^{\text{in}}$, $w \in \mathcal{S}_{u,\varepsilon}^{\text{out}}$ and $\eta \in (0, 1)$ (we show that (7) holds, actually, not just for some $\eta \in (0, 1)$ but for any $\eta \in (0, 1)$). Note that one has $\sqrt{2(1 - v'w)} = \|v - w\| \geq u'(v - w) = u'v - u'w \geq (1 - \varepsilon/2) - (1 - \varepsilon) = \varepsilon/2$ so that $2(1 - v'w) \geq \varepsilon^2/4$, hence

$$v'w \leq 1 - \frac{\varepsilon^2}{8}.$$

Write then

$$\begin{aligned} O_{\alpha,v}^P(rw) - O_{\alpha,v}^P(rv) &= \int_{\mathbb{R}^d} \{ \|z - rw\| - \|z - rv\| - \alpha(rv'w - r) \} dP(z) \\ &= r\alpha(1 - v'w) + \int_{\mathbb{R}^d} \frac{\|z - rw\|^2 - \|z - rv\|^2}{\|z - rw\| + \|z - rv\|} dP(z) \\ &\geq \frac{r\eta\varepsilon^2}{8} + \int_{\mathbb{R}^d} \frac{2r(v - w)'z}{\|z - rv\| + \|z - rw\|} dP(z) \\ &\geq r \left[\frac{\eta\varepsilon^2}{8} - 4 \int_{\mathbb{R}^d} \frac{\|z\|}{\|z - rv\| + \|z - rw\|} dP(z) \right]. \end{aligned}$$

Now, using the fact that $\|v + w\|^2 = 2(1 + v'w) \leq 2(2 - \varepsilon^2/8)$, we obtain

$$\begin{aligned} \{ \|z - rv\| + \|z - rw\| \}^2 &\geq \|z - rv\|^2 + \|z - rw\|^2 \\ &= 2\|z\|^2 + 2r^2 - 2r(v + w)'z \geq 2\|z\|^2 + 2r^2 - 2\sqrt{2(2 - \varepsilon^2/8)}r\|z\| \\ &= 2\{(\|z\| - r)^2 + \sqrt{2}(\sqrt{2} - \sqrt{2 - \varepsilon^2/8})r\|z\|\} =: 2\{(\|z\| - r)^2 + m_\varepsilon r\|z\|\}, \end{aligned}$$

which provides

$$O_{\alpha,v}^P(rw) - O_{\alpha,v}^P(rv) \geq r \left[\frac{\eta\varepsilon^2}{8} - 2\sqrt{2} \int_{\mathbb{R}^d} \frac{\|z\|}{\sqrt{(\|z\| - r)^2 + m_\varepsilon r\|z\|}} dP(z) \right].$$

Since $m_\varepsilon \in (0, 2)$, Lemma 6 guarantees that there exists $R > 0$, not depending on the choice of v, w, η and α , such that for any $r \geq R$, $O_{\alpha,v}^P(rw) - O_{\alpha,v}^P(rv) \geq r\eta\varepsilon^2/16 > 0$. This proves (7), hence the result. \square

Acknowledgements Davy Paindaveine’s research is supported by a research fellowship from the Francqui Foundation and by the Program of Concerted Research Actions (ARC) of the Université libre de Bruxelles. The research of Joni Virta was supported by the Academy of Finland (grant 321883).

References

- Breckling, J., & Chambers, R. (1988). M-quantiles. *Biometrika*, 75, 761–771.
- Brown, B. (1983). Statistical uses of the spatial median. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 45, 25–30.
- Cardot, H., Cénac, P., & Godichon-Baggioni, A. (2017). Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls. *Annals of Statistics*, 45, 591–614.
- Cardot, H., Cénac, P., & Zitt, P. A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19, 18–43.
- Chakraborty, A., & Chaudhuri, P. (2014). On data depth in infinite dimensional spaces. *Annals of the Institute of Statistical Mathematics*, 66, 303–324.
- Chakraborty, A., & Chaudhuri, P. (2014). The spatial distribution in infinite dimensional spaces and related quantiles and depths. *Annals of Statistics*, 42, 1203–1231.
- Chaouch, M., & Laïb, N. (2013). Nonparametric multivariate L_1 -median regression estimation with functional covariates. *Electronic Journal of Statistics*, 7, 1553–1586.
- Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91, 862–872.
- Cheng, Y., & De Gooijer, J. (2007). On the u th geometric conditional quantile. *Journal of Statistical Planning and Inference*, 137, 1914–1930.
- Chowdhury, J., & Chaudhuri, P. (2019). Nonparametric depth and quantile regression for functional data. *Bernoulli*, 25, 395–423.
- Gao, Y. (2003). Data depth based on spatial rank. *Statistics & Probability Letters*, 65, 217–225.
- Girard, S., & Stupfler, G. (2015). Extreme geometric quantiles in a multivariate regular variation framework. *Extremes*, 18, 629–663.
- Girard, S., & Stupfler, G. (2017). Intriguing properties of extreme geometric quantiles. *REVSTAT-Statistical Journal*, 15, 107–139.
- Haldane, J. (1948). Note on the median of a multivariate distribution. *Biometrika*, 35, 414–417.
- Kemperman, J. (1978). The median of a finite measure on a Banach space. In *Statistical Data Analysis Based on the L_1 -Norm and Related Methods*, (pp. 217–230). North-Holland, Amsterdam.
- Koltchinski, V. I. (1997). M-estimation, convexity and quantiles. *Annals of Statistics*, 25, 435–477.
- Milasevic, P., & Ducharme, G. (1987). Uniqueness of the spatial median. *Annals of Statistics*, 15, 1332–1333.
- Serfling, R. (2021a) Depth functions on general data spaces, i. Perspectives, with consideration of “density” and “local” depths. Submitted.
- Serfling, R. (2021b) Depth functions on general data spaces, ii. Formulation and maximality, with consideration of the Tukey, projection, spatial, and “contour” depths. Submitted.
- Serfling, R. (2002). A depth function and a scale curve based on spatial quantiles. In *Statistical Data Analysis Based on the L_1 -Norm and Related Methods* (pp. 25–38). Springer.

- Serfling, R. (2002). Quantile functions for multivariate analysis: Approaches and applications. *Statistica Neerlandica*, 56, 214–232.
- Serfling, R. (2010). Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation. *Journal of Nonparametric Statistics*, 22, 915–936.
- Vardi, Y., & Zhang, C. H. (2000). The multivariate L_1 -median and associated data depth. *Proceedings of the National Academy of Sciences*, 97, 1423–1426.

Modelling Flow in Gas Transmission Networks Using Shape-Constrained Expectile Regression



Fabian Otto-Sobotka, Radoslava Mirkov, Benjamin Hofner,
and Thomas Kneib

Abstract The flow of natural gas within a gas transmission network is studied with the aim to model high-demand situations. Knowledge about the latter can be used to optimise such networks. The analysis of data using shape-constrained expectile regression provides deeper insights into the behaviour of gas flow within the network. The models describe dependence of the maximal daily gas flow on the air temperature, including further effects, like day of the week and type of node. Particular attention is given to spatial effects. Geoadditive models offer a combination of such effects and are easily estimated with penalised mean regression. In order to put special emphasis on the highest demands, we use expectile regression, a quantile-like extension of mean regression that offers the same flexibility. Additional assumptions on the influence of the temperature can be added via shape-constraints. The forecast of gas loads for very low temperatures based on this approach and the application of the obtained results is discussed.

F. Otto-Sobotka (✉)

Division of Epidemiology and Biometry, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

e-mail: fabian.otto-sobotka@uni-oldenburg.de

R. Mirkov

Department of Mathematics, Humboldt University Berlin, Berlin, Germany

e-mail: radoslava.mirkov@gmail.com

B. Hofner

Section Biostatistics, Paul-Ehrlich-Institut, Langen, Germany

e-mail: Benjamin.Hofner@pei.de

T. Kneib

Department of Economics, Georg-August-Universität Göttingen, Göttingen, Germany

e-mail: tkneib@uni-goettingen.de

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_14

1 Introduction

Stochastic properties of the flow of gas in transmission networks are the subject of our study, where we aim to support the optimisation of such networks. Statistical models are suitable to describe the gas loads on nodes of the network, and enable viable prediction of the future gas flow. This leads to the reduction of operational costs, as the cost of control energy necessary to satisfy peak demand at low temperatures can be minimised with a good forecast. Additionally, gas transportation operators are obliged to sustain the supply of gas even during very cold days. Since there is not much data available for very low temperatures, a good prediction is crucial for reliable operation. Also, an evaluation of risk measures is important for the estimation of flow limits at a given temperature.

Similar problems are investigated in Friedl et al. (2012), Mirkov and Friedl (2011), where various nonlinear parametric and semiparametric regression models are suggested to tackle the problem. The methodology suggested therein is suitable for assessing the maximal gas loads at a single node but neglects the interplay of nodes and the spatial effects within the network.

In this chapter, we propose a versatile and flexible semiparametric expectile regression model with shape- constraints. Expectile regression was introduced by Newey and Powell (1987) and constructed in analogy to quantile regression. The latter had been proposed by Koenker and Bassett (1978) just a few years earlier. In a regression scenario

$$y_i = \eta_{i,\tau} + \varepsilon_{i,\tau}, \quad i = 1 \dots, n, \quad (1)$$

with quantile-specific predictor $\eta_{i,\tau}$, they relied on the assumption that for the quantile function of the error term we have $Q_{\varepsilon_{i,\tau}}(\tau) = 0$ for some fixed quantile level $\tau \in (0, 1)$, i.e. the τ -quantile of the error distribution is considered to be zero. From this assumption, it follows that the conditional quantile of level τ for the response y_i is given by the predictor $\eta_{i,\tau}$, i.e. $Q_{y_i}(\tau) = \eta_{i,\tau}$. Although the assumption on the error term can only hold for one specific quantile level τ , one can still estimate a series of regression specifications with dense set of quantile levels to allow for the characterisation of the complete conditional distribution of the response variable y instead of only the mean. We basically construct an empirical distribution from the quantile estimates.

Especially in the present case, where extreme scenarios of gas usage and not just the mean are of interest, such methods are preferable. Estimation of quantile-specific predictors relies on minimising the asymmetrically weighted absolute residuals criterion

$$\sum_{i=1}^n w_{i,\tau} |y_i - \eta_{i,\tau}|$$

with weights

$$w_{i,\tau} = w_{i,\tau}(\eta_{i,\tau}, y_i) = \begin{cases} \tau, & \text{for } y_i \geq \eta_{i,\tau} \\ 1 - \tau, & \text{for } y_i < \eta_{i,\tau}. \end{cases} \quad (2)$$

However, it is not at all straightforward to include a flexible semiparametric model structure with nonlinear, parametric, and spatial effects into a quantile regression model. Especially a spatial effect usually relies on a quadratic penalty for smooth results. Those penalties are natural partners to a least-squares estimate. Therefore, we use expectile regression estimates that are obtained by minimising

$$\sum_{i=1}^n w_{i,\tau} (y_i - \eta_{i,\tau})^2.$$

with the underlying assumption that in regression model (1) the τ -expectiles e_τ of the error terms, defined as

$$e_\tau = \arg \min_e E [w_{i,\tau}(e, \varepsilon_{i,\tau})(\varepsilon_{i,\tau} - e)^2]$$

are zero. This definition also allows the estimation of the tails of the response distribution while simultaneously enabling easy smoothing. Hence, we can analyse the properties of the gas flow through the pipelines of the network not only in dependence on the air temperature, and the weekday type, but also consider further covariates like the type of nodes, and the geographic location of the node within the network. Shape-constraints induce realistic behaviour of the estimated gas flow for low air temperatures. We, therefore, use additional asymmetric penalties for monotonicity and introduce additional boundary constraints. The modelling assumptions ensure the estimation of an adequate probability distribution and a good assessment of risk.

Semiparametric M-Quantile regression (Pratesi et al. 2009) or generalised additive models for location, scale and shape (GAMLSS) (Rigby and Stasinopoulos 2005; Mayr et al. 2012) could be alternative approaches for this scenario. However, M-Quantiles incorporate a certain robustness against extreme observations, while we are aiming explicitly at the tails of the response. GAMLSS, on the other hand, still require the selection of an appropriate parametrised distribution for the response. We, therefore, choose the most flexible approach that also allows for the estimation of a risk for scenarios of extreme usage.

This chapter is organised as follows: Section 2 describes the available data and motivates the choice of the studied models. The utilised methods are presented in Sect. 3, whereas Sect. 4 provides details about the application of the methods to model the gas flow and analyse the risk. The obtained results are also compared with previous analyses. Section 5 concludes the chapter.

2 Description of Data and Motivation

2.1 Data

Data for this study was provided by a large German energy company in the context of a larger research project. The data set contains hourly gas flow for 238 network nodes for the period between June 2009 and May 2010. Mean daily temperatures from the corresponding weather stations are also provided. Additionally, we distinguish several types of nodes. Typical nodes in such networks are public utilities, industrial and areal consumers, as well as nodes on border and market crossings. Continuous geographic coordinates, i.e. longitude and latitude for every node are also included.

We study the dependence of gas loads on the air temperature, the type of weekday, the node type and the geographic location of nodes within the network, simultaneously on all nodes along the pipelines of the gas transmission network. Since we want to maximise the transportation capacity through the pipelines, we concentrate on the daily maximum flows $y_{i,k}^{max}$, $k = 1, \dots, 238$, $i = 1, \dots, n$ ($n = 365$), at each node, for all 238 nodes of the network. This results in a sample size of 86,870. The temperature and gas flow data is shown in a scatter plot in Fig. 2. Due to the high number of observations in our data, all scatterplots in this chapter are smoothed according to Eilers and Goeman (2004). The set of network exits is shown on a map in Fig. 3. We note here that in this study we concentrate on the so-called H-network, which denotes the network with high Wobbe Index (Energy Charter Secretariat 2004). The response values have been standardised per node as the range of values was originally very heterogenous.

An important aspect of this modelling approach is the forecast of gas loads on nodes at the so-called design temperature. The design temperature is defined as the lowest temperature at which the gas operator is still obliged to supply gas without failure, and varies within Germany, depending on the climate conditions in different regions. It usually lies between -12°C and -16°C . Such low mean daily temperatures are very uncommon in Germany, and there is no observed gas flow data available at the design temperature. For this reason, gas operators are forced to use predicted gas loads at the design temperature, and we utilise shape-constrained geoadditive expectile regression for this purpose. Shape-constraints are introduced to prevent further increase of the predicted gas loads below the design temperature. As in Friedl et al. (2012), the way the society uses the gas supply system influences the choice of the model. On the one end, there is the maximum possible gas usage at low temperatures. And even on very hot days there will usually be a constant minimum of used gas. In between we expected the gas usage to strongly depend on the daily temperature. Hence, we chose to model the effect of the temperature in a sigmoid shape.

We are also interested in estimating the upper bound of the gas flow at each node, and we are looking for a tail expectation, i.e. the maximum gas loads that will not be exceeded for the given level of risk at the given fixed temperature. In particular, according to Taylor (2008), a dependence between quantiles and expectiles

is convenient for the estimation of a tail expectation as the conditional value-at-risk is also called expected shortfall (ES).

Finally, we would like to estimate the probability distribution of the flow at each node, in order to generate gas network nominations. A nomination describes the balanced in- and outflow of gas at entries and nodes of the network, and needs to be feasible for every temperature, including the design temperature. A nomination is said to be feasible, or validated, if for the given inflow and outflow allocation, a flow of gas through the network can exist, taking into account all technical limitations of the gas network and physical properties of gas. The distribution of the maximal gas inflow on nodes of the network over the whole temperature range necessary for the nomination validation can be estimated using the expectiles.

In what follows, we study the standardised daily maximum flows

$$y_{i,k} = \frac{y_{i,k}^{max} - \bar{y}_k}{\hat{\sigma}(y_k)}, \quad (3)$$

where \bar{y}_k denotes the empirical mean of all maximal daily gas flows and $\hat{\sigma}(y_k)$ is the standard deviation estimated from the data available for node k .

2.2 Previous Models and Advantages of the New Approach

Friedl et al. (2012) and Mirkov and Friedl (2011) provide models for the maximum gas loads for a single node in dependence of the temperature and the weekday using mean regression, whereas additional information like the type of node or the spatial location of the node is neglected. Each node within the network is observed independently, while the interdependence of nodes and their location within the network is ignored. Hence, the previous models are of the form

$$y_{i,k} = \beta_0 + \mathbf{x}'_i \beta_1 + f(\text{temp}_{i,k}) + \varepsilon_{i,k}$$

where β_0 represents the intercept, the weekday covariate information is included in $\mathbf{x}'_i \beta_1$ and $f(\text{temp}_{i,k})$ is a nonlinear temperature effect. Furthermore, these models assume a homogeneous error distribution with zero mean and constant variance. For convenience, the Gaussian distribution is applied, although it is not compatible with the data. Since the forecast includes the level of uncertainty described by the confidence or prediction intervals, the assumption about the distribution of the error terms plays an important role here. Under an appropriate distributional assumption for the error term and the variance homogeneity, the flow limits for the given acceptable level of risk at the given temperature are assessed. Also, the mean value as well as the upper bound of the flow are the basis for the evaluation of several types of temperature-dependent contracts. Here, the adequate probability distribution for the given temperature is of particular interest.

The new approach based on shape-constrained geoadditive expectile models offers a solution to all inconsistencies mentioned above. The geoadditive component in models enables the estimation of the maximum gas loads in dependence of the temperature and the weekday at each node for the whole network simultaneously, as the model embodies the effect of the geographic location of the nodes within the network. Additional information like the type of node is also included.

Also, the models proposed in this chapter do not specify the error distribution, but estimate it from the data. Thus, the appropriate distribution of the gas flow for each node can be obtained from the model. This enables the adequate quantification of the upper limit of gas flow at each node. Also, quantiles of the distribution can be determined and used for the assessment of risk and the evaluation of some contracts, as mentioned above.

Shape-constraints allow for the flexible modelling of gas flow behaviour for very low or very high temperatures. In particular, the shape of the increase of gas loads is easily controlled by the means of these additional conditions. The discussed features of the shape-constrained geoadditive expectile regression models provide us with a tool for more reliable and accurate prediction of gas flow for very low and generally non-observed temperatures, and thus improve the results obtained in Friedl et al. (2012).

3 Methods

3.1 Geoadditive Regression Models

We aim to predict gas usage using categorical, continuous and spatial covariates in a least squares regression setting. We make use of geoadditive models as introduced by Kammann and Wand (2003) and include all available external information like longitude, latitude and type of node. The distinction between working days and weekend/holidays is a sensible covariate. For our analysis we define the following model:

$$y_{i,k} = \beta_0 + \mathbf{x}'_{1,i} \boldsymbol{\beta}_1 + \mathbf{x}'_{2,k} \boldsymbol{\beta}_2 + f(\text{temp}_{i,k}) + g(\text{long}_k, \text{lat}_k) + \varepsilon_{i,k} \quad (4)$$

for $k = 1, \dots, 238$ and $i = 1, \dots, 365$ with the standardised gas flow as response y , working day indicator \mathbf{x}_1 , the category of a node \mathbf{x}_2 , a nonlinear function f modelling the effect of the temperature and a smooth surface g for the effect of the location given by longitude and latitude.

As the functions f and g are unknown, they are approximated in terms of basis function representations

$$f(x) = \sum_{j=1}^J \beta_{j,f} B_j(x)$$

with the number of basis elements J . For the possibly nonlinear effect of the temperature, we construct a cubic B-spline basis on the range between -15°C and 30°C with $J - 6 = 20$ inner knots and a penalty matrix consisting of second-order differences D between neighbouring coefficients as suggested by Eilers and Marx (1996).

The set of 238 different locations for the observations is neither equally spaced nor observed in a rectangular domain. Hence, we choose to model the spatial effect using a Kriging basis instead of a tensor-product P-spline basis. When Kriging is used to approximate an unknown spatial function, we start by selecting the knots for the basis functions as a subset of all observed locations $\{k_1, \dots, k_p\} \subset \{x_1, \dots, x_n\}$. For our purposes, we choose 100 knots that best cover all the observations as suggested by Johnson et al. (1990). As the spatial effects $\beta = (\beta_1, \dots, \beta_p)'$ at the knots should model similarities between nearby observations, we use a Matérn basis function to model the connection between the locations and construct each basis element B_k for node k as

$$B_k(r, \phi) = \exp(-|r/\phi|)(1 + |r/\phi|)$$

with $r = ||k - x||$ and fixed $\phi \propto \max_{i,j} (||k_i - k_j||)$. Smoothness is achieved with a penalty $\beta' K_1 \beta$ based on the proximity of the knots $K_1 = (B_{k_i} (||k_i - k_j||), \phi)_{i,j}$. This allows us to effectively estimate a spatial effect in this scenario containing rather spread out nodes as well as clustered nodes in the Ruhr area.

3.2 Shape-Constrained P-splines

While penalised splines allow for the flexible approximation of any unknown function, further content-driven information about the overall shape of the function might be available. In our case, a sigmoid shape for the effect of the temperature is assumed. A certain shape for the estimated function can be achieved by adding linear constraints to the least-squares estimation or—much simpler—by extending the P -splines with a further penalty. For specific areas where the estimated function deviates from the assumed shape, this penalty can take effect and thus prevent unwanted behaviour.

A sigmoidal function estimate for temperature—in the sense that the estimate is monotonic and has constant boundaries—is anticipated. An asymmetric penalty enforces monotonic function estimates (Eilers 2005). We extend the approach described in Bollaerts et al. (2006) by adding multiple shape penalties to our estimate in order to gain flexibility. This penalty is given as

$$\beta' K_2 \beta = \beta' D'_{(c)} V D_{(c)} \beta = \sum_{j=c+1}^{J-c} v_j (\Delta^c \beta_j)^2, \tag{5}$$

where c is the order of the difference penalty Δ (monotonicity for $c = 1$), the asymmetric weights are given as

$$v_j = \begin{cases} 0 & \text{if } \Delta^c \beta_j > 0 \\ 1 & \text{if } \Delta^c \beta_j \leq 0 \end{cases}$$

and are collected in the diagonal matrix $\mathbf{V} = \text{diag}(\mathbf{v})$. The associated penalty parameter λ_2 is chosen a priori and should be quite large (e.g. 10^6). For more details see Eilers (2005) and Hofner et al. (2011).

Previous approaches also concentrate on achieving monotonicity. However, we also want to include boundary constraints. Constant boundaries are obtained by a strong penalty on the first differences for the outer three spline coefficients at each end. Consequently, extrapolation does not depend on the temperature anymore. We get a constant effect for the minimum and maximum gas usage, respectively. These boundary constraints are enforced by the penalty

$$\beta' \mathbf{K}_3 \beta = \beta' \mathbf{D}'_{(e)} \mathbf{V}^{(3)} \mathbf{D}_{(e)} \beta = \sum_{j=\epsilon+1}^{J-c} v_j^{(3)} (\Delta^e \beta_j)^2, \tag{6}$$

where $v_j^{(3)}$ is one if the corresponding knot is subject to a boundary constraint. Thus, here the first and the last three elements of $\mathbf{v}^{(3)}$ are equal to one and the remaining weights are equal to zero. The weight matrix is constructed as $\mathbf{V}^{(3)} = \text{diag}(\mathbf{v}^{(3)})$. As we use cubic splines, the three outer splines cover the appropriate temperature intervals. The difference order e controls the type of boundary constraint: differences of order one correspond to constant boundaries, differences of order two correspond to linear boundaries, etc. The associated penalty parameter λ_3 is chosen quite large (as is λ_2).

If we combine P -splines with the two penalties introduced above, we obtain monotonic, smooth effect estimates with boundary constraints. Setting $\lambda_2 = 0$ results in P -splines with boundary constraints only, i.e. without monotonicity assumptions.

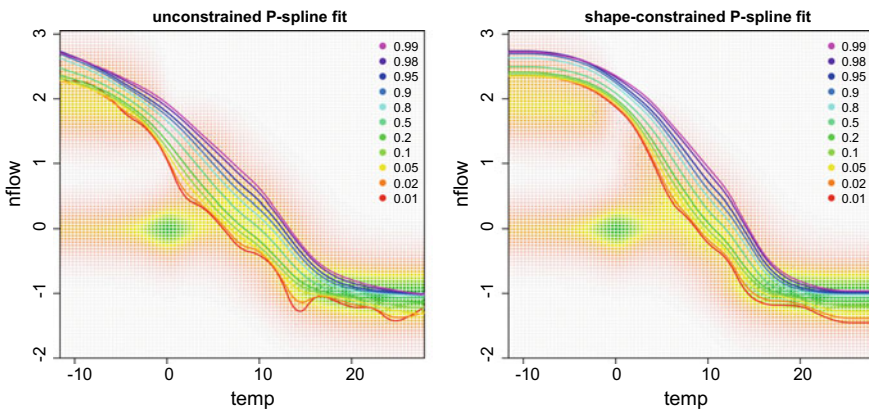


Fig. 1 Example of estimated gas flow for one node without and with shape-constraints

Setting $\lambda_3 = 0$ leads to P -splines without boundary constraints. Note that if the monotonicity constraint is active, one needs to *iteratively* solve the penalised least squares equation as \mathbf{V} depends on $\boldsymbol{\beta}$, and $\boldsymbol{\beta}$ depends on \mathbf{V} . An example of the effects of shape-constraints is given in Fig. 1.

3.3 Semiparametric Expectile Regression

We consider two types of model fitting in this study. The first method directly minimises the sum of the asymmetrically weighted squared residuals via iteratively weighted least squares. The second method is based on a boosting approach and relies on a functional gradient descent approach based on the asymmetrically weighted squares loss function.

3.3.1 Least Asymmetrically Weighted Squares

For the estimation, all the bases of the semiparametric model (4) are combined in a complete design matrix $\mathbf{B} = (\mathbf{1}, \mathbf{X}, \mathbf{B}_{\text{spline}}, \mathbf{B}_{\text{spat}})$ and the penalties are ordered in block-diagonal form $\mathcal{K}_k = \text{diag}(0, \mathbf{0}_X, \mathbf{K}_{\text{spline}}, \mathbf{K}_{\text{spat}})$, $k = 1, \dots, 3$. The vector of regression coefficients is $\boldsymbol{\beta} = (\beta_0, \beta_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_f, \boldsymbol{\beta}_g)$. However, to maintain full rank in the design matrix a reparameterisation has to be applied to the basis matrices (except the shape-constrained P -splines) as presented in Fahrmeir et al. (2004). An expectile regression estimate is obtained by minimising the asymmetrically weighted squared residuals

$$\hat{\boldsymbol{\beta}}_\tau = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n w_{i,\tau} (y_i - \mathbf{b}_i \boldsymbol{\beta})^2 + \lambda_1 \boldsymbol{\beta}' \mathcal{K}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}' \mathcal{K}_2 \boldsymbol{\beta} + \lambda_3 \boldsymbol{\beta}' \mathcal{K}_3 \boldsymbol{\beta}$$

with \mathbf{b}_i as the i th row of the combined basis matrix \mathbf{B} from the semiparametric model and penalty matrices \mathcal{K}_1 for the smoothness, and \mathcal{K}_2 and \mathcal{K}_3 for the shape-constraints. In practice, iteratively weighted penalised least squares updates are performed. The estimation starts with weights of $w_{i,\tau} = 0.5$ irrespective of τ and the estimate

$$\hat{\boldsymbol{\beta}}_\tau^{[k]} = (\mathbf{B}' \mathbf{W}_\tau^{[k-1]} \mathbf{B} + \lambda_1 \mathcal{K}_1 + \lambda_2 \mathcal{K}_2 + \lambda_3 \mathcal{K}_3)^{-1} \mathbf{B} \mathbf{W}_\tau^{[k-1]} \mathbf{y}$$

with weight matrix $\mathbf{W}_\tau^{[k]} = \text{diag}(w_{1,\tau}^{[k]}, \dots, w_{n,\tau}^{[k]})$ is calculated. In the following steps, the weights $w_{i,\tau}^{[k]}$ are determined according to Eq. (2) using $\hat{\boldsymbol{\beta}}_\tau^{[k-1]}$ and a new estimate $\hat{\boldsymbol{\beta}}_\tau^{[k]}$ is obtained. This is repeated until convergence, i.e. the weights do not change anymore. Afterwards, the shape penalties λ_2 or λ_3 are increased if the final estimate does not meet the assumed shape-constraints. The previous procedure has to be repeated once. This estimation works for a fixed vector of smoothing parameters

λ_1 , one for each penalised covariate. The optimal values of λ_1 can be estimated by minimising the generalised cross-validation criterion

$$V_g^w = \frac{n \sum_{i=1}^n w_{i,\tau} (y_i - \mathbf{B}_i \hat{\boldsymbol{\beta}}_\tau)^2}{[\text{tr}(\mathbf{I} - \mathbf{H}^{(\tau)})]^2}$$

numerically, given the generalised hat matrix

$$\mathbf{H}^{(\tau)} = (\mathbf{W}_\tau)^{1/2} \mathbf{B} (\mathbf{B}' \mathbf{W}_\tau \mathbf{B} + \lambda_1 \mathcal{K}_1 + \lambda_2 \mathcal{K}_2 + \lambda_3 \mathcal{K}_3)^{-1} \mathbf{B}' (\mathbf{W}_\tau)^{1/2}. \quad (7)$$

This has been introduced to expectile regression by Schnabel and Eilers (2009) and adapted to additive models in Sobotka and Kneib (2012).

3.3.2 Boosting

As an alternative to directly solving the multidimensional optimisation problem, component-wise functional gradient boosting (Bühlmann and Hothorn 2007) has proven to be a valuable method: Boosting algorithms are especially attractive due to their intrinsic variable selection properties and the ease of combining a wide range of modelling alternatives (such as linear effects, smooth effects, constrained smooth effects and spatial effects) in a single model specification (Kneib et al. 2009; Hofner et al. 2011). Furthermore, smoothing parameters are also optimised within the algorithm, based on the data.

In short, one begins with a constant model $\hat{f} \equiv 0$ and computes the negative gradient of the loss function ρ evaluated at the fit of the previous iteration $\hat{f}_i^{[m-1]}$

$$\mathbf{u} = (u_1, \dots, u_n)' := \left(-\frac{\partial}{\partial f} \rho \left(y_i, \hat{f}_i^{[m-1]} \right) \right)_{i=1, \dots, n}$$

(See Bühlmann and Hothorn 2007; Hofner et al. 2014). Here, the negative gradient is given by the weighted residuals

$$u_i = \begin{cases} 2\tau \cdot (y_i - \hat{f}_i^{[m-1]}) & \text{if } (y_i - \hat{f}_i^{[m-1]}) > 0 \\ 2(1 - \tau) \cdot (y_i - \hat{f}_i^{[m-1]}) & \text{if } (y_i - \hat{f}_i^{[m-1]}) < 0. \end{cases}$$

For each component in our model (4) we specify a so-called “base-learner”. A model component usually represents a single variable (or a group of variables in the case of spatial effects). Thus we get separate base-learners for each linear effect (i.e. one base-learner per variable), one base-learner for the smooth effect of temperature f_{spline} and one base-learner for the spatial effect f_{spat} . Each base-learner is then fitted separately to the weighted residuals \mathbf{u} by penalised least squares, and only the model component that describes these residuals best is updated by adding a small proportion of the fit (e.g. 10%) to the *current* model fit. Subsequently, the residuals are updated

and the whole procedure is iterated until a fixed number of iterations is reached. The final model $\hat{\eta}$ is defined as the sum of all models fitted in this process. As we update only *one* base-learner in each boosting iteration and as each base-learner usually only depends on one or very few variables, variable selection can be obtained by stopping the boosting procedure after an appropriate number of iterations. This is the major tuning parameter of boosting and is usually optimised using cross-validation techniques.

Base-learners

To achieve the desired model structure, we only need to define appropriate base-learners for each predictor: For example, linear effects are fitted using linear base-learners (which are just simple regression models with the negative gradient as outcome and one predictor). Smooth effects can be fitted using P-spline base-learners (Schmid and Hothorn 2008), and spatial effects can be fitted using a radial basis function base-learners, i.e. Kriging base-learners (Hofner 2011). Additionally, constrained effect estimates can be specified. Monotonicity-constrained base-learners are studied in Hofner et al. (2011, 2014). Here, we expand monotonicity constrained base-learners to monotonic effects with additional boundary constraints. For the negative gradient vector $\mathbf{u} = (u_1, \dots, u_n)'$, i.e. the (continuous) vector of weighted residuals, we can estimate a smooth monotonic function using P-splines with additional asymmetric difference penalty (5) and an additional penalty for the boundary effects (6) via the penalised least squares criterion

$$(\mathbf{u} - \mathbf{B}\boldsymbol{\beta})'(\mathbf{u} - \mathbf{B}\boldsymbol{\beta}) + \lambda_1 \boldsymbol{\beta}'\mathcal{K}_1\boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}'\mathcal{K}_2\boldsymbol{\beta} + \lambda_3 \boldsymbol{\beta}'\mathcal{K}_3\boldsymbol{\beta}, \tag{8}$$

where \mathbf{B} denotes the B-spline basis matrix, and $\boldsymbol{\beta}$ the regression coefficients of a single nonlinear function. Let furthermore $\mathbf{K}_{\text{spline}} = \mathbf{D}'_d \mathbf{D}_d$ be the standard P-spline penalty with difference order d , and let λ_1 be the associated smoothing parameter.

Note that the base-learner (8) does not contain the weights $w_{i,\tau}$ but is a simple unweighted penalised least squares criterion. The weights $w_{i,\tau}$ are only used in the derivation of the negative gradient vector \mathbf{u} . Still, the resulting model can be interpreted as an expectile regression model.

3.3.3 Interpretation of Expectiles as Risk Measure

While the interpretation of the mean or a quantile is relatively straightforward, there is in general no intuitive interpretation for a single expectile. This becomes obvious when from the definition of a univariate τ -expectile μ_τ for a random variable Y , which is only available in the implicit form

$$\tau = \frac{\int_{-\infty}^{\mu_\tau} |y - \mu_\tau| f_Y(y) dy}{\int_{-\infty}^{\infty} |y - \mu_\tau| f_Y(y) dy} = \frac{G(\mu_\tau) - \mu_\tau F(\mu_\tau)}{2(G(\mu_\tau) - \mu_\tau F(\mu_\tau)) + (\mu_\tau - \mu_{0.5})},$$

where $G(e) = \int_{-\infty}^e y f_Y(y) dy$ is the partial moment function, F is the cdf, and $G(\infty) = \mu_{0.5}$ is the expectation of Y . In most cases, expectile regression will be performed in order to have a complete overview over the whole conditional distribution of the response variable. However, as it is the case in this chapter, one can be interested in only the upper tail of this distribution, for example. We make use of the connection between expectiles and the risk measure *expected shortfall* (ES) to facilitate an interpretation. The expected shortfall is originally defined as

$$ES_p(t) = E(Y(x)|Y(x) > \tilde{y}_p(x))$$

denoting the mean beyond a pre-specified p -quantile \tilde{y}_p . This equation has been rewritten, for example, by Taylor (2008), as a function of a τ -expectile μ_τ and its value $p = F(\mu_\tau(x))$ of the unknown distribution function F , resulting in

$$ES_p(x) = \left(1 + \frac{1 - \tau}{(1 - 2\tau)p}\right) \mu_\tau(x) - \frac{1 - \tau}{(1 - 2\tau)p} \mu_{0.5}(x). \tag{9}$$

For the estimation of F , Taylor (2008) suggested to start by estimating a dense set of expectiles for asymmetries $0 < \tau_1 < \dots < \tau_T < 1$ and then construct an empirical distribution function from this set. Given the fixed quantile value p , an expectile has to be selected that best fits this quantile as suggested by Efron (1991). A more advanced solution to this problem is given by Schulze Waltrup et al. (2015) and will be used here. We calculated a smooth density based on the dense set of expectiles, where quantiles for every $p \in [0, 1]$ can be extracted. The fixed p , the appropriate τ and the estimated expectile μ_τ are then entered into Eq. (9). Then the expected shortfall can be estimated. With that we can supply more results than the distributional estimate of the response, we can construct an estimate for an upper tail expectation.

3.3.4 Confidence Intervals

While parametric confidence intervals for boosting estimates cannot be derived, an asymptotic result for a least asymmetrically weighted least squares estimate has been constructed in Sobotka et al. (2013). The resulting vector of regression coefficients is asymptotically normal

$$\hat{\beta}_\tau \stackrel{a}{\sim} N(\beta_\tau^0, \text{Cov}(\beta_\tau^0))$$

with

$$\text{Cov}(\beta_\tau^0) = (\mathbf{B}'\mathbf{W}_\tau\mathbf{B} + \mathcal{K})^{-1}\mathbf{B}'\mathbf{W}_\tau^2 \text{diag}(\text{Var}(y - \mathbf{B}\beta_\tau^0))\mathbf{B}(\mathbf{B}'\mathbf{W}_\tau\mathbf{B} + \mathcal{K})^{-1},$$

where the heteroscedastic residual variance is estimated by

$$\widehat{\text{Var}}(y_i - \mathbf{B}\beta_\tau^0) = \frac{(y_i - \mathbf{B}\hat{\beta}_\tau)^2}{1 - \mathbf{H}_{ii}^{(\tau)}}$$

given the generalised hat matrix (7). Confidence intervals can then be constructed point-wise for a fixed covariate value. For boosting estimates, confidence intervals can be derived using bootstrap techniques.

4 Estimating and Forecasting Gas Flow

The data, as presented in Sect. 2, contains a few, very important possible covariates that should allow for an accurate prediction of the gas usage. For each observed amount of gas flow, we have the mean daily temperature at the node, the binary distinction between workday and weekend/holiday, the type of node in five categories and the location of the node.

The response needs to be standardised according to Eq. (3) in order to allow the inclusion of all nodes into the same model. For this regression problem we have constructed a model covering the whole data set instead of single nodes, and two estimation procedures, iteratively reweighted least squares and boosting, each with different strengths.

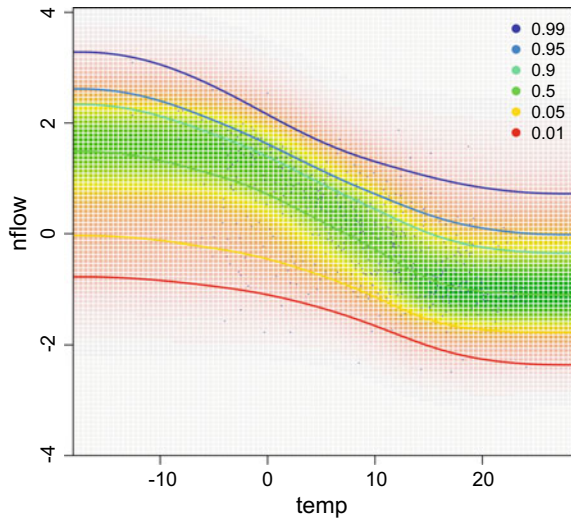
4.1 Results

The categorical effects in Table 1 show a significant difference in gas usage between weekdays and the weekend throughout the whole distribution of the response. Further, we mainly find a difference between industry nodes and municipalities. Interestingly, this distinction is not significant for the mean regression ($\tau = 0.5$) but only

Table 1 Categorical effects for the upper half of the conditional distribution of the response. Significant effects are set in bold

τ	0.5	0.9	0.95	0.99
Weekday	0.289	0.328	0.348	0.454
Type	<i>Reference: "industry"</i>			
Municipal	-0.040	-0.226	-0.292	-0.376
Areal	0.019	-0.018	-0.042	-0.067
Border	-0.035	0.052	0.059	0.031
Market	-0.058	-0.069	-0.064	0.029

Fig. 2 Standardised flow for predefined temperature interval $[-15; 30]$. Combined results for 238 knots. 6 expectiles calculated



in the upper tail of the conditional distribution of the response. The nonlinear effect of the temperature in Fig. 2 shows the modelled constant effects for very high and very low temperatures which state a minimum amount of gas usage independent from a further increase in temperature, for example, in households for cooking and in factories, and also a gas flow that does not exceed a practical limit. Otherwise, we see a decrease in gas usage with increasing temperatures, but also a change in variance and skewness of the conditional distribution of the response. This means that the effect of the temperature for the upper tail is different from the mean effect. The contrast is even stronger in the spatial effect depicted in Fig. 3. For the mean regression, we observe almost no dependence between the mean gas flow and the location of the nodes. However, for expectiles from the upper tail we observe a specific region with increased gas flow, in the south-west of Germany. This model shows that for the prediction of extreme scenarios of gas usage a mean regression is just not sufficient and results from the upper extremes of the response strongly differ from the mean. An extreme scenario can now be constructed by choosing a prediction from one node and a high expectile and revert the standardisation of the response (3). As expected, we found stronger change throughout Germany from west to east while, for example, the estimated gas usage was almost constant throughout the Ruhr area where housing is dense and the industries are very similar.

The results based on boosting are very similar and are provided in the appendix. The application of boosting includes possible variable selection and automatic selection of all smoothing parameters by an overall cross-validation. Hence, we observe that all small parametric effects are close to zero, especially for the areal, border and market dummies. The differences in the selection of smoothing parameters also explain small differences in the spatial effects, but the overall results do not change.

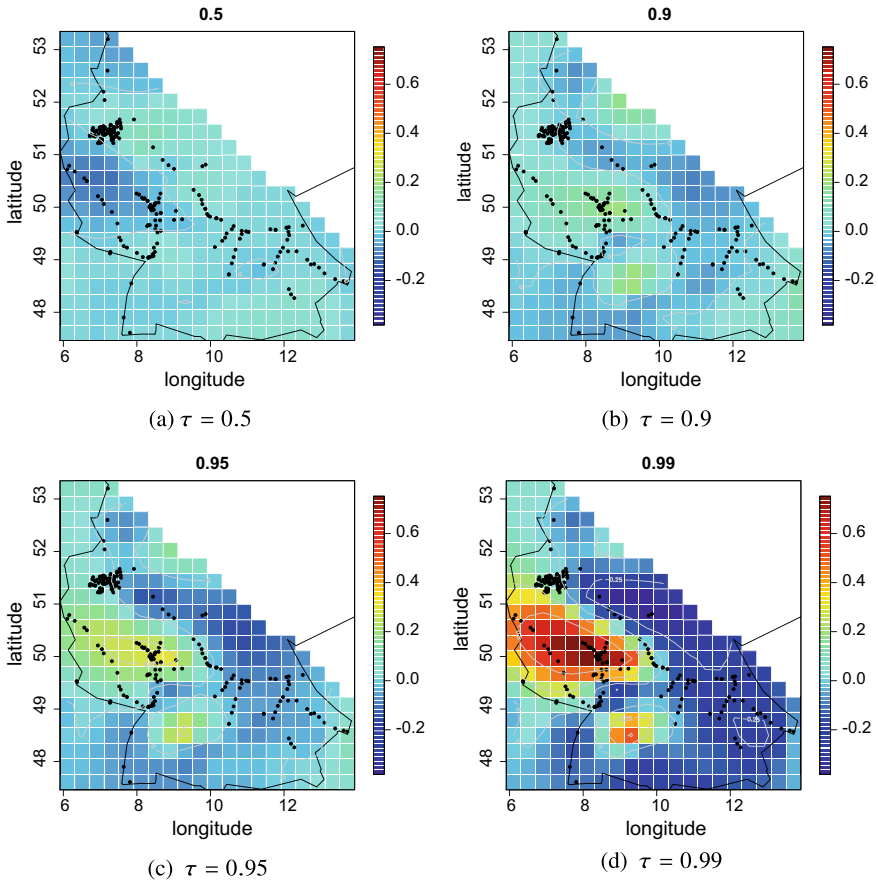


Fig. 3 Partial spatial effect of standardised flow for four expectiles. Lighter colours indicate higher values of gas flow, i.e. white represents a partial effect of 0.5 and dark red stands for -0.5 . Observed locations are included as black points

4.2 Risk Analysis

In order to estimate the expected shortfall of the gas flow in the geoaddivitive model we start by estimating 99 expectiles from $\tau = 0.01$ to $\tau = 0.99$. We then obtain the expectile from $\tau = 0.98$ which best corresponds to the properties of a 0.95-quantile such that 5% of the observations are above the estimated expectile. This expectile is then transformed as described in Sect. 3.3.3. The resulting risk measure is then depicted in Fig. 4. Especially the estimated temperature curve predicts the desired extreme scenario of gas flow and delivers a smooth estimate in a region of very few observations.

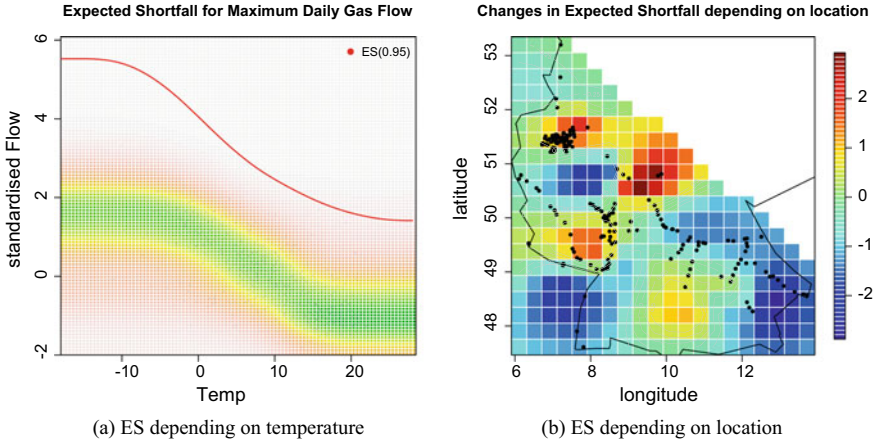


Fig. 4 Expected shortfall for a level of 0.95 estimated from a dense set of 99 expectiles. Estimated effects of the temperature and the location, where black dots again denote the nodes of the network

5 Conclusion

In this chapter, we present a novel modelling approach useful for forecasting gas flow on nodes of gas transmission networks. This approach extends the previous work on the same topic introduced in Friedl et al. (2012) and offers solutions to the inconsistencies mentioned there. In particular, we show that the modelling of high-demand scenarios in a gas transmission network can be improved in several ways. Geoadditive regression models considering the whole network simultaneously instead of a single node make use of similarities between different nodes, either on a spatial or on a behavioural level. Differences in gas usage can be uncovered depending on the coordinates of the nodes, however, those dependencies mainly arise in high-demand situations and not in the previously applied mean regression.

The use of expectiles improves our knowledge about extreme scenarios of gas usage from the previous mean regression attempts. We find strong effects of the covariates in the upper tail of the distribution of gas usage while there is little to find in the mean. However, we still retain the simplicity of a least-squares estimate with its asymptotic normal confidence and the flexibility of spatial and shape-constrained modelling.

The inclusion of spatial information into the model also brings up further possibilities for future research. Since we could use the information of the gas loads over the course of one year also as a time series, spatio-temporal models as introduced to expectiles by Spiegel et al. (2019) could be applied. By introducing an interaction term between the temperature or time information and the coordinate information, we could have one gas usage curve per node while still maintaining spatial similarities. Further, the results from our geoadditive model have shown an artefact effect in areas where the network is not present. This could possibly be improved by restricting the

spatial smoothing to those areas where the network is observed. Soap film smoothing Wood et al. (2008) could be applied within an expectile regression regarding this problem.

Acknowledgements Financial support from the German Research Foundation (DFG) grants KN 922/4-1 and KN 922/9-1 is gratefully acknowledged. We would also like to thank Werner Roemisch and Paul Eilers for their valuable comments during the development of this research. We acknowledge the detailed feedback of two anonymous referees. Finally, we would like to thank the editors of this festschrift, Abdelaati Daouia and Anne Ruiz-Gazen, for inviting us and our contribution.

Appendix

For comparison we also provide boosting estimates for the geoaddivitive model. With such a large data set we do not expect a lot of variable selection taking place, however, the selection of smoothness should be easier, especially for the spatial effect. We estimate the expectiles using 10000 initial boosting iterations and a tenfold cross-validation to find the optimal stopping iteration.

The temperature curves in Fig. 5 also show the changes in skewness of the response distribution and the deviation from normality in the tails. The categorical effects in Table 2 also support the previous results. In the mean regression model ($\tau = 0.5$) effects for “municipal” and “market” node types have been selected to enter the boosting model. For the other expectiles only an effect for the municipal node was selected in the boosting model. This effect increases with increasing expectile. The

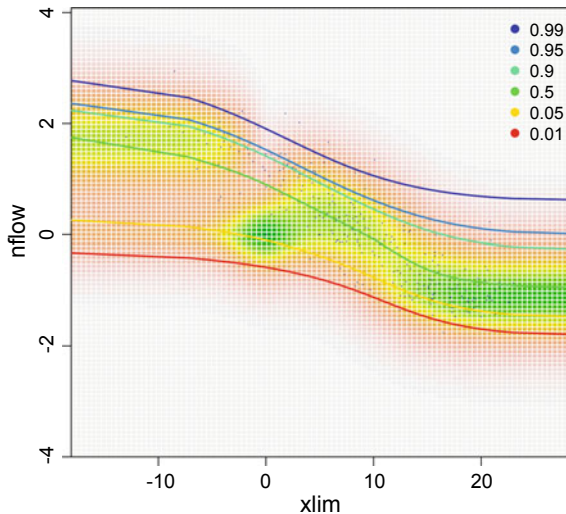


Fig. 5 Standardised flow for temperature interval. Overall results for 238 knots. 6 expectiles calculated

Table 2 Estimated categorical effects for the upper half of the conditional distribution of the response. Unselected effects are set to zero

τ	0.5	0.9	0.95	0.99
Weekday	0.288	0.288	0.293	0.338
Type	<i>Reference: "industry"</i>			
Municipal	-0.010	-0.178	-0.241	-0.324
Areal	0	0	0	0
Border	0	0	0	0
Market	0.006	0	0	0

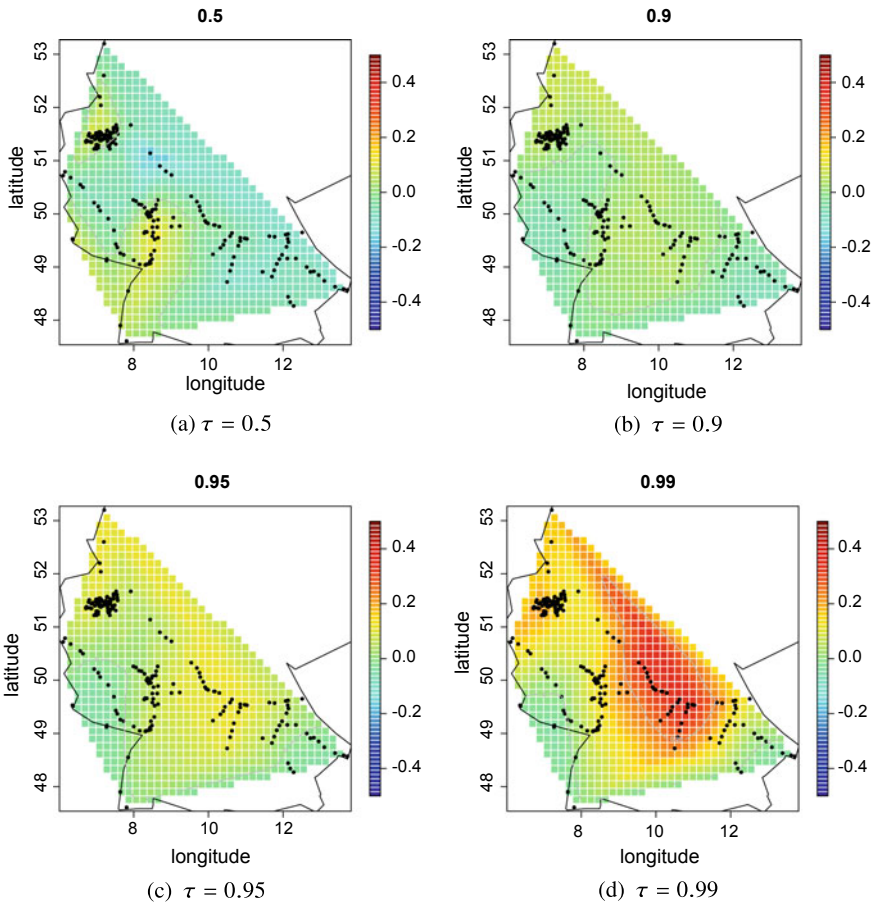


Fig. 6 Partial spatial effect of standardised flow for four expectiles. Lighter colours indicate higher values. Observed locations are included as black points

distinction between working day and holiday is again strong for all expectiles of the response.

In the spatial effect shown in Fig. 6 we again see a rather small change for the mean. In extreme scenarios, on the other hand, it becomes more clear with boosting that there is a difference in gas usage between the south-west and the centre of Germany.

Overall, these results support the use of expectile regression and the importance of effects in the upper tail of the conditional distribution of the response for forecasting gas flow.

References

- Bollaerts, K., Eilers, P. H. C., & Van Mechelen, I. (2006). Simple and multiple p-splines regression with shape constraints. *British Journal of Mathematical & Statistical Psychology*, 59, 451–469.
- Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22, 477–505.
- Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1, 93–125.
- Eilers, P., & Goeman, J. (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics*, 20(5), 623–628.
- Eilers, P. H. C. (2005). Unimodal smoothing. *Journal of Chemometrics*, 19, 317–328.
- Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–121.
- Energy Charter Secretariat (2004). LNG and natural gas quality standards. *Occasional Papers 2*.
- Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized structured additive regression: A Bayesian perspective. *Statistica Sinica*, 14, 731–761.
- Friedl, H., Mirkov, R., & Steinkamp, A. (2012). Modeling and Forecasting Gas Flow on Exits of Gas Transmission Networks. *International Statistical Review Special Issue on Energy Statistics*, 80(1), 24–39.
- Hofner, B. (2011). *Boosting in Structured Additive Models*. Ph. D. thesis, LMU München. Verlag Dr. Hut, München.
- Hofner, B., Hothorn, T., Kneib, T., & Schmid, M. (2011). A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, 20, 956–971.
- Hofner, B., Kneib, T., Hothorn, T. (2014). A unified framework of constrained regression. [arXiv:1403.7118](https://arxiv.org/abs/1403.7118).
- Hofner, B., Mayr, A., Robinzonov, N., & Schmid, M. (2014). Model-based boosting in R - A hands-on tutorial using the R package mboost. *Computational Statistics*, 29, 3–35.
- Hofner, B., Müller, J., & Hothorn, T. (2011). Monotonicity-constrained species distribution models. *Ecology*, 92, 1895–1901.
- Johnson, M., Moore, L., & Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26, 131–148.
- Kammann, E. E., & Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1), 1–18.
- Kneib, T., Hothorn, T., & Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics*, 65, 626–634.
- Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.
- Mayr, A., Fenske, N., Hofner, B., Kneib, T., & Schmid, M. (2012). Generalized additive models for location, scale and shape for high-dimensional data - a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61, 403–427.

- Mirkov, R., & Friedl, H. (2011). Nonlinear and Spline Regression Models for Forecasting Gas Flow on Exits of Gas Transmission Networks. In Conesa, D., Forte, A., Lopez-Quilez, and A., Munoz, F. (eds.), *Proceedings of the 26th International Workshop on Statistical Modelling* (pp. 394–399). ADEIT, The University of Valencia.
- Newey, W. K., & Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55, 819–847.
- Pratesi, M., Ranalli, M. G., & Salvati, N. (2009). Nonparametric m-quantile regression using penalised splines. *Journal of Nonparametric Statistics*, 21(3), 287–304.
- Rigby, R., & Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54, 507–554.
- Schmid, M., & Hothorn, T. (2008). Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis*, 53, 298–311.
- Schnabel, S., & Eilers, P. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, 53, 4168–4177.
- Schulze Waltrup, L., Sobotka, F., Kneib, T., & Kauermann, G. (2015). Expectile and Quantile Regression—David and Goliath? *Statistical Modelling*, 15(5), 433–456.
- Sobotka, F., Kauermann, G., Schulze Waltrup, L., & Kneib, T. (2013). On confidence intervals for semiparametric expectile regression. *Statistics and Computing*, 23(2), 135–148.
- Sobotka, F., & Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56, 755–767.
- Spiegel, E., Kneib, T., & Otto-Sobotka, F. (2019). Spatio-temporal expectile regression models. *Statistical Modelling*.
- Taylor, J. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, 6(2), 231–252.
- Wood, S. N., M. V. Bravington, and S. L. Hedley (2008). Soap film smoothing. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 70(5), 931–955.

Spatial Statistics and Econometrics

Asymptotic Analysis of Maximum Likelihood Estimation of Covariance Parameters for Gaussian Processes: An Introduction with Proofs



François Bachoc

Abstract This article provides an introduction to the asymptotic analysis of covariance parameter estimation for Gaussian processes. Maximum likelihood estimation is considered. The aim of this introduction is to be accessible to a wide audience and to present some existing results and proof techniques from the literature. The increasing-domain and fixed-domain asymptotic settings are considered. Under increasing-domain asymptotics, it is shown that in general all the components of the covariance parameter can be estimated consistently by maximum likelihood and that asymptotic normality holds. In contrast, under fixed-domain asymptotics, only some components of the covariance parameter, constituting the microergodic parameter, can be estimated consistently. Under fixed-domain asymptotics, the special case of the family of isotropic Matérn covariance functions is considered. It is shown that only a combination of the variance and spatial scale parameter is microergodic. A consistency and asymptotic normality proof is sketched for maximum likelihood estimators.

1 Introduction

Kriging Stein (1999), Rasmussen and Williams (2006) consists of inferring the values of a (Gaussian) process given observations at a finite set of points. It has become a popular method for a large range of applications such as geostatistics Matheron (1970), numerical code approximation Sacks et al. (1989), Santner et al. (2003), Bachoc et al. (2016), calibration Paulo et al. (2012), Bachoc et al. (2014), Kennedy

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-73249-3_15) contains supplementary material, which is available to authorized users.

F. Bachoc (✉)
Institut de Mathématique, UMR5219; Université de Toulouse; CNRS, UPS IMT, 31062 Toulouse
Cedex 9, France
e-mail: francois.bachoc@math.univ-toulouse.fr

and O'Hagan (2001), global optimization Jones et al. (1998), and machine learning Rasmussen and Williams (2006).

If the mean and covariance function of the Gaussian process are known, then the unknown values of the Gaussian process can be predicted based on Gaussian conditioning Rasmussen and Williams (2006), Santner et al. (2003). Confidence intervals are associated with the predictions. In addition, in the case where the observation points of the Gaussian process can be selected, efficient goal-oriented sequential sampling techniques are available, for instance, for optimization Jones et al. (1998) or estimation of failure domains Bect et al. (2012).

Nevertheless, the mean and covariance functions are typically unknown, so that the above methods are typically carried out based on a mean and covariance function selected by the user, that differs from the true ones. Here we shall consider the case where the mean function is known to be equal to zero and the covariance function is known to belong to a parametric set of covariance functions. In this case, selecting a covariance function amounts to estimating the covariance parameter. Large estimation errors of the covariance parameter can be harmful to the quality of the above methods based on Gaussian processes. Hence, one may hope to obtain theoretical guarantees that estimators of the covariance parameters converge to the true ones.

Here we will review some of such guarantees in the case of maximum likelihood estimation Rasmussen and Williams (2006), Stein (1999), which is the most standard estimation method of covariance parameters. The two main settings for these guarantees are the increasing and fixed-domain asymptotic frameworks. Under increasing-domain asymptotics, we will show that, generally speaking, the covariance parameter is fully estimable consistently and asymptotic normality holds. Under fixed-domain asymptotics, only a subcomponent of the covariance parameter, called the microergodic parameter, can be estimated consistently. We will show that the microergodic parameter is estimated consistently by maximum likelihood in the case of the family of isotropic Matérn covariance functions, with asymptotic normality. In both asymptotic settings, we will provide sketches of the proofs. We will also highlight the technical differences between the proofs in the two settings.

The rest of the article is organized as follows. Gaussian processes, estimation of covariance parameters and maximum likelihood are introduced in Sect. 2. Increasing-domain asymptotics is studied in Sect. 3. Fixed-domain asymptotics is studied in Sect. 4. Concluding remarks and pointers to additional references are provided in Sect. 5. A supplementary material contains the asymptotic normality results for the Matérn model and the expressions of means and covariances of quadratic forms of a Gaussian vector.

2 Framework and Notations

2.1 Gaussian Processes and Covariance Functions

We consider a Gaussian process $\xi : \mathbb{R}^d \rightarrow \mathbb{R}$. We recall that ξ is a stochastic process such that for any $m \in \mathbb{N}$ and for any $u_1, \dots, u_m \in \mathbb{R}^d$, the random vector $(\xi(u_1), \dots, \xi(u_m))$ is a Gaussian vector Rasmussen and Williams (2006). Here and in the rest of the paper, \mathbb{N} is the set of positive integers.

We assume throughout that ξ has mean function zero, that is $\mathbb{E}(\xi(u)) = 0$ for $u \in \mathbb{R}^d$. Thus, the distribution of ξ is characterized by its covariance function

$$(u, v) \in \mathbb{R}^{2d} \mapsto \text{cov}(\xi(u), \xi(v)).$$

We assume in all the paper that the covariance function of ξ is stationary, that is, there exists a function $k^* : \mathbb{R}^d \rightarrow \mathbb{R}$ such that for $u, v \in \mathbb{R}^d$,

$$\text{cov}(\xi(u), \xi(v)) = k^*(u - v).$$

In a slight abuse of language, we will also refer to k^* as the (stationary) covariance function of ξ . The function k^* is symmetric because for $u \in \mathbb{R}^d$, $k^*(u) = \text{cov}(\xi(u), \xi(0)) = \text{cov}(\xi(0), \xi(u)) = k^*(-u)$. This function is positive definite in the sense of the following definition.

Definition 1 A function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is positive definite if for any $m \in \mathbb{N}$ and for any $u_1, \dots, u_m \in \mathbb{R}^d$, the $m \times m$ matrix $[\phi(u_i - u_j)]_{i,j=1,\dots,m}$ is positive semi-definite.

The function k^* is positive definite because the matrices $[k^*(u_i - u_j)]_{i,j=1,\dots,m}$ of the form of Definition 1 are covariance matrices (of Gaussian vectors).

We then consider a set of functions $\{k_\theta; \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^p$ and where for $\theta \in \Theta$, k_θ is a function from $\mathbb{R}^d \rightarrow \mathbb{R}$ that is symmetric and positive definite. We also call k_θ a covariance function and θ a covariance parameter for $\theta \in \Theta$.

The set $\{k_\theta; \theta \in \Theta\}$ is a set of candidate covariance functions for ξ , that is, this set is known to the statistician who aims at selecting an appropriate parameter θ such that k_θ is as close as possible to k^* . In the rest of the paper, we will consider that k^* belongs to $\{k_\theta; \theta \in \Theta\}$. Hence, there exists $\theta_0 \in \Theta$ such that $k^* = k_{\theta_0}$. This setting is called the well-specified case in Bachoc (2013a, b, 2018). Under this setting, we have a classical parametric statistical estimation problem, where the goal is to estimate the true covariance parameter θ_0 .

2.2 Classical Families of Covariance Functions

For $q \in \mathbb{N}$ and for a vector x in \mathbb{R}^q , we let $\|x\|$ be the Euclidean norm of x . A first classical family of covariance functions is composed by the isotropic exponential ones with $\Theta \subset (0, \infty)^2$ and

$$k_\theta(x) = \sigma^2 e^{-\alpha \|x\|},$$

for $\theta = (\sigma^2, \alpha)$ and $x \in \mathbb{R}^d$. A second classical family is composed by the isotropic Gaussian covariance functions, with $\Theta \subset (0, \infty)^2$ and

$$k_\theta(x) = \sigma^2 e^{-\alpha^2 \|x\|^2},$$

for $\theta = (\sigma^2, \alpha)$ and $x \in \mathbb{R}^d$.

Finally, a third classical family is composed by the isotropic Matérn covariance functions, with $\Theta \subset (0, \infty)^3$ and

$$k_\theta(x) = \frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)} (\alpha \|x\|)^\nu \mathcal{K}_\nu(\alpha \|x\|), \tag{1}$$

where Γ is the gamma function, \mathcal{K}_ν is the modified Bessel function of the second kind, for $\theta = (\sigma^2, \alpha, \nu)$ and $x \in \mathbb{R}^d$. These families of covariance functions, and other ones, can be found, for instance, in Bevilacqua et al. (2019), Genton and Kleiber (2015), Gneiting and Schlather (2004), Rasmussen and Williams (2006), Santner et al. (2003), Stein (1999). We remark that the isotropic exponential covariance functions are special cases of the isotropic Matérn covariance functions with $\nu = 1/2$ Stein (1999).

For these three families of covariance functions, one can check that $k_\theta(0) = \sigma^2$ (in the Matérn case the function is extended at zero by continuity). Hence σ^2 is called the variance parameter, because if ξ has covariance function k_θ we have $\text{var}(\xi(u)) = \sigma^2$ for $u \in \mathbb{R}^d$. In these three families of covariance functions, for $u, v \in \mathbb{R}^d$, if ξ has covariance function k_θ we have that $\text{cov}(\xi(u), \xi(v))$ depends on $\alpha \|u - v\|$. Hence α is called the spatial scale parameter because changing α can be interpreted as changing the spatial scale when measuring differences between input locations of ξ . In the three examples, $k_\theta(x)$ is a decreasing function of $\|x\|$, thus a large α makes the covariance decrease more quickly with $\|x\|$ and provides a small spatial scale of variation of ξ . Conversely, a small α makes the covariance decrease more slowly and provides a large spatial scale of variation of ξ .

Finally, for the family of Matérn covariance functions, ν is called the smoothness parameter. To interpret this, for $\theta \in \Theta$, let us call spectral density the function $\hat{k}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}$ such that for $u \in \mathbb{R}^d$

$$k_\theta(u) = \int_{\mathbb{R}^d} \hat{k}_\theta(\omega) e^{i\omega^\top u} d\omega,$$

with $i^2 = -1$. Under mild regularity assumptions, that hold for the three families above, the function \hat{k}_θ is the Fourier transform of k_θ . When k_θ is a Matérn covariance function, we have

$$\hat{k}_\theta(\omega) = \sigma^2 \frac{\Gamma(\nu + d/2) \alpha^{2\nu}}{\Gamma(\nu) \pi^{d/2}} \frac{1}{(\alpha^2 + \|\omega\|^2)^{\nu + d/2}}, \tag{2}$$

for $\omega \in \mathbb{R}^d$ Gneiting et al. (2010). Hence, we see that for larger ν , the Fourier transform $\hat{k}_\theta(\omega)$ converges to zero faster as $\|\omega\| \rightarrow \infty$, which implies that the function k_θ is smoother at zero (this function is already infinitely differentiable on $\mathbb{R}^d \setminus \{0\}$). This is why ν is called the smoothness parameter.

There is an important body of literature on the interplay between the smoothness of the covariance function of ξ and the smoothness of ξ Adler (1981), Adler (1990), Azaïs and Wschebor (2009). In our case, if ξ has an exponential covariance function, then it is continuous and not differentiable (almost surely and in quadratic mean). If ξ has a Gaussian covariance function, then it is infinitely differentiable (almost surely and in quadratic mean). The Matérn covariance functions provide, so to speak, a continuum of smoothness in between these two cases. Indeed, consider ξ with Matérn covariance function with smoothness parameter $\nu > 0$. Then ξ is m times differentiable (almost surely and in quadratic mean) if $\nu > m$.

2.3 Maximum Likelihood

Consider a sequence $(s_i)_{i \in \mathbb{N}}$ of spatial locations at which we observe ξ , with $s_i \in \mathbb{R}^d$. Assume from now on that the locations $(s_i)_{i \in \mathbb{N}}$ are two-by-two distinct. Then, for $n \in \mathbb{N}$, we consider the Gaussian observation vector $y = (y_1, \dots, y_n)^\top = (\xi(s_1), \dots, \xi(s_n))^\top$.

We consider a family of covariance functions $\{k_\theta; \theta \in \Theta\}$ and assume further that for $n \in \mathbb{N}$, the covariance matrix $R_\theta := [k_\theta(s_i - s_j)]_{i,j=1,\dots,n}$ is invertible. Then, when ξ has covariance function k_θ , the Gaussian density of y is

$$\mathcal{L}_n(\theta) = \frac{1}{\sqrt{|R_\theta|}(2\pi)^{n/2}} e^{-\frac{1}{2}y^\top R_\theta^{-1}y},$$

with $|R_\theta|$ the determinant of R_θ . The focus of this paper will be on maximum likelihood estimation. A maximum likelihood estimator is a (measurable) estimator of θ_0 that satisfies

$$\hat{\theta}_{\text{ML}} \in \operatorname{argmax}_{\theta \in \Theta} \mathcal{L}_n(\theta). \quad (3)$$

We remark that, in general, there may not be a unique estimator $\hat{\theta}_{\text{ML}}$ satisfying (3). Furthermore, the existence of measurable estimators satisfying (3) is not a trivial problem. We refer, for instance, to Giné and Nickl (2016), Molchanov (2005) on this point.

In this paper, we assume that there exists at least one measurable estimator satisfying (3) and the results hold for any choice of such an estimator. A notable particular case is when $\Theta = (0, \infty)$, $\theta = \sigma^2$ and $k_\theta = \sigma^2 k^*$. In this case, there is a unique

estimator satisfying (3) (see also (21) in Sect. 4). In this special case, we can call $\hat{\theta}_{\text{ML}}$ the maximum likelihood estimator. In general, one may rather call it a maximum likelihood estimator.

It is convenient to consider the following decreasing transformation of the logarithm of the likelihood:

$$L_n(\theta) = \frac{1}{n} \log(|R_\theta|) + \frac{1}{n} y^\top R_\theta^{-1} y, \quad (4)$$

for $\theta \in \Theta$. We have

$$\hat{\theta}_{\text{ML}} \in \underset{\theta \in \Theta}{\operatorname{argmin}} L_n(\theta).$$

The problem of studying the asymptotic properties of $\hat{\theta}_{\text{ML}}$ as $n \rightarrow \infty$ presents several differences compared to the most standard parametric estimation setting where the observations are independent and identically distributed Van der Vaart (2000). Indeed, in our case the components of the observation vector y are dependent, so the logarithm of the likelihood is not a sum of independent random variables. Furthermore, the likelihood function involves the quantities $|R_\theta|$ and R_θ^{-1} for which, often, no explicit expressions exist. Finally, for asymptotic statistics with independent and identically distributed data, there is a single asymptotic setting as $n \rightarrow \infty$. Here there exist several possible asymptotic settings, depending on how the spatial locations s_1, \dots, s_n behave as $n \rightarrow \infty$. The proof techniques and the results obtained strongly depend on the asymptotic setting. We will now review some results under the two main existing asymptotic frameworks: increasing-domain and fixed-domain asymptotics.

3 Increasing-Domain Asymptotics

In Section 3, we assume that there exists a fixed $\Delta > 0$ such that

$$\inf_{\substack{i, j \in \mathbb{N} \\ i \neq j}} \|s_i - s_j\| \geq \Delta. \quad (5)$$

This assumption is the main assumption considered in the literature for increasing-domain asymptotics (see Bachoc 2014, for instance, and see also Bachoc 2018 for one of the few exceptions). This assumption implies that the spatial locations $(s_i)_{i \in \mathbb{N}}$ are not restricted to a bounded set. The results and proofs that will be presented in Sect. 3 can mainly be found in Bachoc (2014).

3.1 Consistency

Here the aim is to show that $\hat{\theta}_{\text{ML}}$ converges to θ_0 , weakly. We consider a general family of covariance functions $\{k_\theta; \theta \in \Theta\}$, where Θ is compact, that satisfies

$$\sup_{\theta \in \Theta} |k_\theta(x)| \leq \frac{C_{\text{sup}}}{1 + \|x\|^{d+C_{\text{inf}}}} \quad (6)$$

and

$$\max_{s=1,2,3} \max_{\substack{i_1, \dots, i_s = \\ 1, \dots, p}} \sup_{\theta \in \Theta} \left| \frac{\partial^s}{\partial \theta_{i_1} \dots \partial \theta_{i_s}} k_\theta(x) \right| \leq \frac{C_{\text{sup}}}{1 + \|x\|^{d+C_{\text{inf}}}}, \quad (7)$$

where $0 < C_{\text{inf}}$ and $C_{\text{sup}} < \infty$ are fixed constants and for $x \in \mathbb{R}^d$.

We also assume that

$$(\theta, \omega) \in \Theta \times \mathbb{R}^d \mapsto \hat{k}_\theta(\omega) \text{ is continuous and strictly positive.} \quad (8)$$

The families of isotropic exponential, Gaussian, and Matérn covariance functions do satisfy (6) and (7), when Θ is compact, and ν is fixed for Matérn. Indeed, these functions and their partial derivatives, with respect to σ^2 and α , are exponentially decaying as $\|x\| \rightarrow \infty$, where x is their input. For the exponential and Gaussian covariance functions, this can be seen simply and for the Matérn covariance function, this follows from the properties of the modified Bessel functions of the second kind Abramowitz and Stegun (1964). Also, when Θ is compact, exponentially decaying functions bounding the covariance functions and their partial derivatives can be chosen uniformly over $\theta \in \Theta$ (see again Abramowitz and Stegun 1964 for the Matérn covariance functions).

These three families of covariance functions also satisfy (8). The expressions of the Fourier transforms of these covariance functions can be found, for instance, in Gneiting et al. (2010) and Stein (1999).

Then the next lemma enables to control the term R_θ^{-1} in (4). We let $\lambda_{\text{inf}}(M)$ be the smallest eigenvalue of a symmetric matrix M .

Lemma 1 (Proposition D.4 in Bachoc 2014, Theorem 5 in Bachoc and Furrer 2016)
Assume that (5), (6) and (8) hold. We have

$$\inf_{n \in \mathbb{N}} \inf_{\theta \in \Theta} \lambda_{\text{inf}}(R_\theta) > 0.$$

Proof (sketch) We have, for $n \in \mathbb{N}$ and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$,

$$\begin{aligned}
\sum_{i,j=1}^n \lambda_i \lambda_j (R_\theta)_{i,j} &= \sum_{i,j=1}^n \lambda_i \lambda_j k_\theta(s_i - s_j) \\
&= \sum_{i,j=1}^n \lambda_i \lambda_j \int_{\mathbb{R}^d} \hat{k}_\theta(\omega) e^{i\omega^\top (s_i - s_j)} d\omega \\
&= \int_{\mathbb{R}^d} \hat{k}_\theta(\omega) \left(\sum_{i,j=1}^n \lambda_i \lambda_j e^{i\omega^\top s_i} e^{-i\omega^\top s_j} \right) d\omega \\
&= \int_{\mathbb{R}^d} \hat{k}_\theta(\omega) \left| \sum_{i=1}^n \lambda_i e^{i\omega^\top s_i} \right|^2 d\omega, \tag{9}
\end{aligned}$$

where $|z|$ is the modulus of a complex number z . In (9), $\hat{k}_\theta(\omega)$ is strictly positive. Furthermore, because s_1, \dots, s_n are two-by-two distinct, the family of functions $(\omega \mapsto e^{i\omega^\top s_i})_{i=1, \dots, n}$ is linearly independent. Hence, $\sum_{i,j=1}^n \lambda_i \lambda_j (R_\theta)_{i,j} > 0$ for $(\lambda_1, \dots, \lambda_n) \neq 0$. This shows that $\lambda_{\inf}(R_\theta) > 0$ for $n \in \mathbb{N}$ and $\theta \in \Theta$. Proving that the infimum in the lemma is also strictly positive is also based on (9). We refer to the proofs of Proposition D.4 in Bachoc (2014) or of Theorem 5 in Bachoc and Furrer (2016). \square

The next lemma will enable to control the variance of the likelihood criterion and the order of magnitude of its derivatives.

Lemma 2 *Assume that (5)–(8) hold. For any $\theta \in \Theta$, as $n \rightarrow \infty$,*

$$\text{var}(L_n(\theta)) = o(1).$$

Furthermore

$$\max_{i=1, \dots, p} \sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_i} L_n(\theta) \right| = O_p(1).$$

Proof (sketch) Using that y is a centered Gaussian vector, we have, with $\text{cov}(z)$ the covariance matrix of a random vector z , from Appendix B in the supplementary material

$$\text{var}(L_n(\theta)) = \frac{1}{n^2} \text{var}(y^\top R_\theta^{-1} y) = \frac{2}{n^2} \text{tr} \left(R_\theta^{-1} \text{cov}(y) R_\theta^{-1} \text{cov}(y) \right) = \frac{2}{n^2} \text{tr} \left(R_\theta^{-1} R_{\theta_0} R_\theta^{-1} R_{\theta_0} \right).$$

Let $\lambda_{\sup}(M)$ be the largest eigenvalue of a symmetric matrix M . From Gershgorin circle theorem, we have

$$\begin{aligned} \lambda_{\text{sup}}(R_{\theta_0}) &\leq \max_{i=1, \dots, n} \sum_{j=1}^n |(R_{\theta_0})_{i,j}| \\ &= \max_{i=1, \dots, n} \sum_{j=1}^n |k_{\theta_0}(s_i - s_j)| \\ \text{(from (6) :)} &\leq \max_{i=1, \dots, n} \sum_{j=1}^n \frac{C_{\text{sup}}}{1 + \|s_i - s_j\|^{d+C_{\text{inf}}}}. \end{aligned}$$

It is shown in Bachoc (2014) that (5) implies that

$$\max_{i=1, \dots, \infty} \sum_{j=1}^{\infty} \frac{C_{\text{sup}}}{1 + \|s_i - s_j\|^{d+C_{\text{inf}}}} < \infty.$$

Hence there is a constant $A_1 < \infty$ such that $\lambda_{\text{sup}}(R_{\theta_0}) \leq A_1$. Also, from Lemma 1, there is a constant $A_2 < \infty$ such that $\sup_{\theta \in \Theta} \lambda_{\text{sup}}(R_{\theta}^{-1}) \leq A_2$. Hence, we have $\text{var}(L_n(\theta)) \leq 2A_1^2 A_2^2/n$ which proves the first part of the lemma.

For the second part of the lemma, let $\rho_{\text{sup}}(M)$ be the largest singular value of a matrix M . Using Gershgorin circle theorem again, together with (7), we show that there is a constant $A_3 < \infty$ such that

$$\max_{i=1, \dots, p} \sup_{\theta \in \Theta} \rho_{\text{sup}} \left(\frac{\partial R_{\theta}}{\partial \theta_i} \right) \leq A_3.$$

With this, we have

$$\begin{aligned} \max_{i=1, \dots, p} \sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_i} L_n(\theta) \right| &= \max_{i=1, \dots, p} \sup_{\theta \in \Theta} \left| \frac{1}{n} \text{tr} \left(R_{\theta}^{-1} \frac{\partial R_{\theta}}{\partial \theta_i} \right) - \frac{1}{n} y^{\top} R_{\theta}^{-1} \frac{\partial R_{\theta}}{\partial \theta_i} R_{\theta}^{-1} y \right| \\ &\leq A_2 A_3 + A_2^2 A_3 \frac{\|y\|^2}{n}. \end{aligned}$$

This last quantity is a $O_p(1)$ because $\|y\|^2/n$ is non-negative with (bounded) expectation $\text{var}(\xi(0))$. □

The consistency result will rely on the following asymptotic identifiability assumption. We assume that for all $\epsilon > 0$,

$$\liminf_{n \rightarrow \infty} \inf_{\substack{\theta \in \Theta \\ \|\theta - \theta_0\| \geq \epsilon}} \frac{1}{n} \sum_{i,j=1}^n (k_{\theta}(s_i - s_j) - k_{\theta_0}(s_i - s_j))^2 > 0. \tag{10}$$

This assumption means that for θ bounded away from θ_0 , there is sufficient information in the spatial locations s_1, \dots, s_n to distinguish between the two covariance functions k_{θ} and k_{θ_0} . In Bachoc (2014), an explicit example is provided for which (10) holds.

We remark that, even though there are n^2 terms in the sum in (10), this sum can be shown to be a $O(n)$ for any fixed $\theta \in \Theta$, because of (6) (by proceeding as in the proof of Lemma 2). The intuition is that, asymptotically, for many pairs $i, j \in \{1, \dots, n\}$, $k_\theta(s_i - s_j)$ and $k_{\theta_0}(s_i - s_j)$ are small. This is why the normalization factor is $1/n$ rather than $1/n^2$ in (10).

With the assumption (10), we can now state the consistency result.

Theorem 1 (Bachoc (2014)) *Assume that (5)–(8) and (10) hold. As $n \rightarrow \infty$*

$$\hat{\theta}_{ML} \rightarrow^p \theta_0.$$

Proof (sketch) From Lemma 2 we have, for any $\theta \in \Theta$,

$$L_n(\theta) - \mathbb{E}(L_n(\theta)) \rightarrow_{n \rightarrow \infty}^p 0.$$

Furthermore one can show, similarly as in Lemma 2,

$$\max_{i=1, \dots, p} \sup_{\theta \in \Theta} \left| \frac{\partial}{\partial \theta_i} \mathbb{E}(L_n(\theta)) \right| = O(1).$$

Hence, using Lemma 2, we obtain

$$\sup_{\theta \in \Theta} |L_n(\theta) - \mathbb{E}(L_n(\theta))| = o_p(1). \tag{11}$$

Next, it is shown in Bachoc (2014) that there exists a constant $A_4 > 0$ such that for $\theta \in \Theta$

$$\mathbb{E}(L_n(\theta)) - \mathbb{E}(L_n(\theta_0)) \geq A_4 \frac{1}{n} \sum_{i,j=1}^n (k_\theta(s_i - s_j) - k_{\theta_0}(s_i - s_j))^2. \tag{12}$$

From (12) and (10), we then obtain, for $\epsilon > 0$, with a strictly positive constant A_5 , for n large enough,

$$\inf_{\substack{\theta \in \Theta \\ \|\theta - \theta_0\| \geq \epsilon}} (\mathbb{E}(L_n(\theta)) - \mathbb{E}(L_n(\theta_0))) \geq A_5. \tag{13}$$

Combining (11) and (13) enables to conclude the proof with a standard M-estimator argument (for instance, as in the proof of Theorem 5.7 in Van der Vaart 2000). \square

3.2 Asymptotic Normality

For $i \in \{1, \dots, p\}$, we have seen in the proof of Lemma 2 that the i th partial derivative of L_n at θ_0 is

$$\frac{\partial}{\partial \theta_i} L_n(\theta_0) = \frac{1}{n} \text{tr} \left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} \right) - \frac{1}{n} y^\top R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} y.$$

Since y is a centered Gaussian vector and using Appendix B in the supplementary material, the element i, j of the covariance matrix of the gradient of L_n at θ is thus, for $i, j = 1, \dots, p$,

$$\begin{aligned} \text{cov} \left(\frac{\partial}{\partial \theta_i} L_n(\theta_0), \frac{\partial}{\partial \theta_j} L_n(\theta_0) \right) &= \frac{2}{n^2} \text{tr} \left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} R_{\theta_0} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} R_{\theta_0}^{-1} R_{\theta_0} \right) \\ &= \frac{2}{n^2} \text{tr} \left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} \right). \end{aligned} \quad (14)$$

It is shown in Bachoc (2014) that for $i, j \in \{1, \dots, p\}$,

$$\mathbb{E} \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} L_n(\theta_0) \right) = \frac{1}{n} \text{tr} \left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} \right).$$

We will thus need to ensure that the $p \times p$ matrix with element i, j equal to

$$\frac{1}{n} \text{tr} \left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} \right)$$

is asymptotically invertible. For this, we assume that for all $(\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p \setminus \{0\}$,

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i,j=1}^n \left(\sum_{m=1}^p \lambda_m \frac{\partial k_{\theta_0}(s_i - s_j)}{\partial \theta_m} \right)^2 > 0. \quad (15)$$

This assumption is interpreted as a local identifiability condition around θ_0 . In Bachoc (2014), an explicit example is provided for which (15) holds.

We can now state the asymptotic normality result for maximum likelihood estimators.

Theorem 2 Assume that (5)–(8), (10) and (15) hold. Let Σ_{θ_0} be the $p \times p$ matrix with element i, j equal to

$$\frac{1}{2} \frac{1}{n} \text{tr} \left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} \right).$$

Then

$$0 < \liminf_{n \rightarrow \infty} \lambda_{\inf}(\Sigma_{\theta_0}) \leq \limsup_{n \rightarrow \infty} \lambda_{\sup}(\Sigma_{\theta_0}) < \infty. \quad (16)$$

Furthermore, with $M^{-1/2}$ the unique symmetric matrix square root of M^{-1} for a symmetric strictly positive definite M , we have

$$\sqrt{n} (\Sigma_{\theta_0}^{-1})^{-1/2} (\hat{\theta}_{ML} - \theta_0) \xrightarrow{d}_{n \rightarrow \infty} \mathcal{N}(0, I_p). \tag{17}$$

We remark that in Theorem 2, $\Sigma_{\theta_0}^{-1}$ is the asymptotic covariance matrix, but this matrix is not necessarily assumed to converge as $n \rightarrow \infty$. This matrix has its eigenvalues bounded away from zero and infinity asymptotically, so that the rate of convergence is \sqrt{n} in Theorem 2.

Remark 1 Here the element i, j of $n\Sigma_{\theta_0}$ is $n^2/4$ times the covariance between the elements i and j of the gradient of L_n , from (14). Note that L_n is $-2/n$ times the log-likelihood (up to a constant not depending on y or θ). Consider now the score vector that is equal to the gradient of the log-likelihood. Then, we obtain that the covariance between the elements i and j of the score is $n^2/4$ times $4/n^2$ times the element i, j of $n\Sigma_{\theta_0}$.

In other words, $n\Sigma_{\theta_0}$ is the (theoretical) Fisher information matrix. In agreement with this, remark that from Theorem 2 the inverse of $n\Sigma_{\theta_0}$ provides the asymptotic covariance matrix of maximum likelihood estimators as $n \rightarrow \infty$.

Proof (sketch) In Bachoc (2014), it is shown that there exists a strictly positive constant A_6 such that for any $\lambda_1, \dots, \lambda_p$ with $\lambda_1^2 + \dots + \lambda_p^2 = 1$, we have

$$\sum_{i,j=1}^p \lambda_i \lambda_j \frac{1}{2} \frac{1}{n} \text{tr} \left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} \right) \geq A_6 \frac{1}{n} \sum_{i,j=1}^n \left(\sum_{m=1}^p \lambda_m \frac{\partial k_{\theta_0}(s_i - s_j)}{\partial \theta_m} \right)^2.$$

Hence, from (15),

$$0 < \liminf_{n \rightarrow \infty} \lambda_{\inf}(\Sigma_{\theta_0}).$$

Hence Σ_{θ_0} is invertible for n large enough. Let n be large enough so that this is the case in the rest of the proof.

One can show as in the proof of Lemma 2 (see also Bachoc 2014) that

$$\limsup_{n \rightarrow \infty} \lambda_{\sup}(\Sigma_{\theta_0}) < \infty.$$

Hence (16) is proved. Let us now prove (17).

It is shown in Bachoc (2014) (see also Bachoc et al. 2020), using a standard M-estimator argument together with techniques similar as above, that

$$\begin{aligned} \sqrt{n}(\hat{\theta}_{ML} - \theta_0) &= - \left(\left[\mathbb{E} \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} L_n(\theta_0) \right) \right]_{i,j=1,\dots,p} \right)^{-1} \sqrt{n} \left(\frac{\partial}{\partial \theta_i} L_n(\theta_0) \right)_{i=1,\dots,p} + o_p(1) \\ &= - \frac{1}{2} \Sigma_{\theta_0}^{-1} \sqrt{n} \left(\frac{\partial}{\partial \theta_i} L_n(\theta_0) \right)_{i=1,\dots,p} + o_p(1). \end{aligned}$$

Hence to conclude the proof, it is sufficient to show that

$$(4\Sigma_{\theta_0})^{-1/2} \sqrt{n} \left(\frac{\partial}{\partial \theta_i} L_n(\theta_0) \right)_{i=1, \dots, p} \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, I_p).$$

Let us show this using linear combinations. Let us write the $p \times 1$ gradient vector

$$\frac{\partial}{\partial \theta} L_n(\theta_0) = \left(\frac{\partial}{\partial \theta_i} L_n(\theta_0) \right)_{i=1, \dots, p}.$$

Let $\lambda = (\lambda_1, \dots, \lambda_p)^\top \in \mathbb{R}^p$ be fixed with $\lambda_1^2 + \dots + \lambda_p^2 = 1$. We have

$$\sum_{i=1}^p \lambda_i \left((4\Sigma_{\theta_0})^{-1/2} \sqrt{n} \frac{\partial}{\partial \theta} L_n(\theta_0) \right)_i = \sum_{i=1}^p \left((4\Sigma_{\theta_0})^{-1/2} \lambda \right)_i \sqrt{n} \frac{\partial}{\partial \theta_i} L_n(\theta_0).$$

Let us now write $\beta_i = \left((4\Sigma_{\theta_0})^{-1/2} \lambda \right)_i$. We have

$$\begin{aligned} & \sum_{i=1}^p \lambda_i \left((4\Sigma_{\theta_0})^{-1/2} \sqrt{n} \frac{\partial}{\partial \theta} L_n(\theta_0) \right)_i \\ &= \sum_{i=1}^p \beta_i \sqrt{n} \frac{\partial}{\partial \theta_i} L_n(\theta_0) \\ &= -\sqrt{n} \left(y^\top \left(\frac{1}{n} \sum_{i=1}^p \beta_i R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \right) y - \mathbb{E} \left(y^\top \left(\frac{1}{n} \sum_{i=1}^p \beta_i R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \right) y \right) \right), \end{aligned}$$

using for the last equality that the gradient of the logarithm of the likelihood at θ_0 has mean zero. Letting $z = (z_1, \dots, z_n)^\top = R_{\theta_0}^{-1/2} y$, the negative of the above quantity is equal to

$$\sqrt{n} \left(z^\top \left(\frac{1}{n} \sum_{i=1}^p \beta_i R_{\theta_0}^{-1/2} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1/2} \right) z - \mathbb{E} \left(z^\top \left(\frac{1}{n} \sum_{i=1}^p \beta_i R_{\theta_0}^{-1/2} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1/2} \right) z \right) \right). \quad (18)$$

Letting ρ_1, \dots, ρ_n be the eigenvalues of $(1/n) \sum_{i=1}^p \beta_i R_{\theta_0}^{-1/2} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1/2}$ and letting $w = (w_1, \dots, w_n) \sim \mathcal{N}(0, I_n)$, (18) is equal, in distribution, to

$$\sqrt{n} \sum_{i=1}^n (w_i^2 - 1) \rho_i. \quad (19)$$

Let us show that (19) converges to a standard Gaussian distribution. We have

$$\begin{aligned}
\text{var} \left(\sqrt{n} \sum_{i=1}^n (w_i^2 - 1) \rho_i \right) &= 2n \sum_{i=1}^n \rho_i^2 \\
&= \text{var} \left(\sum_{i=1}^p \lambda_i \left((4\Sigma_{\theta_0})^{-1/2} \sqrt{n} \frac{\partial}{\partial \theta} L_n(\theta_0) \right)_i \right) \\
&= \lambda^\top \text{cov} \left((4\Sigma_{\theta_0})^{-1/2} \sqrt{n} \frac{\partial}{\partial \theta} L_n(\theta_0) \right) \lambda \\
(\text{from (14) :}) &= \lambda^\top I_p \lambda \\
&= 1.
\end{aligned}$$

One can show as in the proof of Lemma 2 (see also Bachoc 2014) that $\max_{i=1}^n |\rho_i| = O(1/n)$. Hence, the classical Lindeberg–Feller central limit theorem enables to conclude that (19) converges to a standard Gaussian distribution (see also Istas and Lang 1997). This concludes the proof. \square

To conclude Sect. 3, the consistency and asymptotic normality results given here are quite generally applicable to families of stationary covariance functions and to Gaussian processes with zero mean functions. Some extensions to non-zero constant mean functions are discussed in Bachoc et al. (2020). It would be interesting to provide extensions to non-stationary covariance functions or to unknown non-constant mean functions, with a parametric family of mean functions. It is possible that some of the proof techniques and intermediary results presented in Sect. 3 and in Bachoc (2014) would be relevant for these extensions. Nevertheless, new arguments would also need to be developed, and appropriate assumptions, on the non-stationary covariance functions and non-constant mean functions, would need to be considered.

4 Fixed-Domain Asymptotics

4.1 What Changes

Under fixed-domain asymptotics, the spatial locations s_1, \dots, s_n are restricted to a compact set $D \subset \mathbb{R}^d$. In this case, almost none of the proof techniques above for increasing-domain asymptotics can be applied. Indeed, they are based on the fact that for a given $i \in \{1, \dots, n\}$, $\xi(s_i)$ has a very small covariance with $\xi(s_j)$ for most s_j , $j = 1, \dots, n$. On the contrary, under fixed-domain asymptotics, for instance if k_{θ_0} is non-zero on \mathbb{R}^d , $\xi(s_i)$ has a non negligible covariance with all the $\xi(s_j)$, $j = 1, \dots, n$.

In particular, contrary to Lemma 1, if $\theta \in \Theta$ is such that k_θ is continuous at zero, then the smallest eigenvalue of R_θ goes to zero as $n \rightarrow \infty$. This is seen by considering a sequence of 2×2 submatrices based on s_{i_n}, s_{j_n} with $\|s_{i_n} - s_{j_n}\| \rightarrow 0$ as $n \rightarrow \infty$. Similarly, the largest eigenvalue of R_θ goes to infinity as $n \rightarrow \infty$ for any $\theta \in \Theta$ if k_θ is, for instance, non-zero on \mathbb{R}^d .

4.2 Microergodic and Non-microergodic Parameters

The conclusion of Sect. 3 on increasing-domain asymptotics is that the family of stationary covariance functions $\{k_\theta; \theta \in \Theta\}$ can be fairly general to prove the consistency and asymptotic normality of maximum likelihood estimators of θ_0 . In particular, under the reasonable conditions (10) and (15), θ_0 can be entirely consistently estimable.

We will now see that, in contrast, for a family $\{k_\theta; \theta \in \Theta\}$ of covariance functions, under fixed-domain asymptotics, it can regularly be the case that θ_0 is not entirely consistently estimable.

The notion that makes this more precise is that of the equivalence of Gaussian measures Ibragimov and Rozanov (1978), Stein (1999). Consider two covariance parameters $\theta_1, \theta_2 \in \Theta, \theta_1 \neq \theta_2$. If ξ has covariance function k_{θ_1} , ξ yields a measure \mathcal{M}_{θ_1} on the set of functions from D to \mathbb{R} , with respect to the cylindrical sigma-algebra.¹ Similarly, if ξ has covariance function k_{θ_2} , ξ yields a measure \mathcal{M}_{θ_2} . When D is compact, these two measures can be equivalent (for a set A of functions, $\mathcal{M}_{\theta_1}(A) = 0$ if and only if $\mathcal{M}_{\theta_2}(A) = 0$) even when the covariance functions k_{θ_1} and k_{θ_2} are different.

The notion of equivalence of Gaussian measures enables to definition non-microergodic parameters.

Definition 2 Let Φ be a function from Θ to \mathbb{R}^q for $q \in \mathbb{N}$. We say that $\Phi(\theta_0)$ is non-microergodic if there exists $\theta_1 \in \Theta$ such that $\Phi(\theta_1) \neq \Phi(\theta_0)$ and the measures \mathcal{M}_{θ_1} and \mathcal{M}_{θ_0} are equivalent.

If a covariance parameter is non-microergodic, it cannot be estimated consistently.

Lemma 3 Let $(s_i)_{i \in \mathbb{N}}$ be any sequence of points in D . If $\Phi(\theta_0)$ is non-microergodic, there does not exist a sequence of functions $\hat{\Phi}_n : \mathbb{R}^n \rightarrow \mathbb{R}^q$ such that, for any $\theta \in \Theta$, if ξ has covariance function k_θ then $\hat{\Phi}_n(\xi(s_1), \dots, \xi(s_n))$ goes to $\Phi(\theta)$ in probability as $n \rightarrow \infty$.

Proof Let $\Phi(\theta_0)$ be non-microergodic. Then fix $\theta_1 \in \Theta$ such that $\Phi(\theta_1) \neq \Phi(\theta_0)$ and the measures \mathcal{M}_{θ_1} and \mathcal{M}_{θ_0} are equivalent.

Assume that an estimator sequence $\hat{\Phi}_n$ as described in the lemma exists. Then, when ξ has covariance function k_{θ_0} , as $n \rightarrow \infty$,

$$\hat{\Phi}_n(\xi(s_1), \dots, \xi(s_n)) \rightarrow^p \Phi(\theta_0).$$

Hence there exists a subsequence n' such that as $n' \rightarrow \infty$, almost surely,

$$\hat{\Phi}_{n'}(\xi(s_1), \dots, \xi(s_{n'})) \rightarrow \Phi(\theta_0).$$

¹If Gaussian processes with continuous realizations on compact sets are considered, one can also define Gaussian measures over the Banach space of continuous functions (on a compact set) endowed with the supremum norm and the corresponding Borel sigma-algebra.

This can be written in the form

$$\mathcal{M}_{\theta_0} \left(\left\{ f \text{ function from } D \text{ to } \mathbb{R} \text{ such that } \hat{\Phi}_{n'}(f(s_1), \dots, f(s_{n'})) \rightarrow_{n' \rightarrow \infty} \Phi(\theta_0) \right\} \right) = 1.$$

Then since the measures \mathcal{M}_{θ_1} and \mathcal{M}_{θ_0} are equivalent

$$\mathcal{M}_{\theta_1} \left(\left\{ f \text{ function from } D \text{ to } \mathbb{R} \text{ such that } \hat{\Phi}_{n'}(f(s_1), \dots, f(s_{n'})) \rightarrow_{n' \rightarrow \infty} \Phi(\theta_0) \right\} \right) = 1.$$

This means that, when ξ has covariance function k_{θ_1} , the sequence $\hat{\Phi}_{n'}(\xi(s_1), \dots, \xi(s_{n'}))$ goes almost surely to $\Phi(\theta_0) \neq \Phi(\theta_1)$. Hence the sequence $\hat{\Phi}_n(\xi(s_1), \dots, \xi(s_n))$ does not go to $\Phi(\theta_1)$ in probability as $n \rightarrow \infty$. This is a contradiction which concludes the proof. \square

Hence, one should not expect to have accurate estimators of non-microergodic parameters under fixed-domain asymptotics. The interpretation of non-microergodic parameters is that, even if $\Phi(\theta_0)$ and $\Phi(\theta_1)$ are different, there is not enough information in a single realization of the random function $\{\xi(s); s \in D\}$ (even if this realization was observed continuously) to distinguish between $\Phi(\theta_0)$ and $\Phi(\theta_1)$. This lack of information stems from the boundedness of D .

It is important to remark that there exist results showing that non-microergodic parameters have an asymptotically negligible impact on prediction of unknown values of ξ Stein (1988), Stein (1990a), Stein (1990c), Zhang (2004). In Stein (1999), this situation is interpreted as an instance of the following principle, called Jeffreys’s law: “things we shall never find much out about cannot be very important for prediction”.

Finally, we can define microergodic parameters.

Definition 3 Let Φ be a function from Θ to \mathbb{R}^q for $q \in \mathbb{N}$. We say that $\Phi(\theta_0)$ is microergodic if for any $\theta_1 \in \Theta$ such that $\Phi(\theta_1) \neq \Phi(\theta_0)$, the measures \mathcal{M}_{θ_1} and \mathcal{M}_{θ_0} are orthogonal (i.e., there exists a set of functions A such that $\mathcal{M}_{\theta_1}(A) = 0$ and $\mathcal{M}_{\theta_0}(A) = 1$).

4.3 Consistent Estimation of the Microergodic Parameter of the Isotropic Matérn Model

Let us now focus on the family of isotropic Matérn covariance functions (1), in the case where the smoothness parameter ν is known. We thus consider $\theta = (\sigma^2, \alpha) \in \Theta = (0, \infty) \times [\alpha_{\text{inf}}, \alpha_{\text{sup}}]$ with $0 < \alpha_{\text{inf}} < \alpha_{\text{sup}} < \infty$ fixed. We thus have

$$k_{\theta}(x) = \frac{\sigma^2 2^{1-\nu}}{\Gamma(\nu)} (\alpha \|x\|)^{\nu} \mathcal{K}_{\nu}(\alpha \|x\|), \tag{20}$$

for $x \in \mathbb{R}^d$ where $0 < \nu < \infty$ is fixed and known. We let $\theta_0 = (\sigma_0^2, \alpha_0)$. In the rest of Sect. 4, we set the dimension as $d \in \{1, 2, 3\}$.

Then the parameters σ_0^2 and α_0 are non-microergodic, while the parameter $\sigma_0^2 \alpha_0^{2\nu}$ is microergodic.

Proposition 1 (Zhang 2004) *With the family of covariance functions given by (20), the measures \mathcal{M}_{θ_1} and \mathcal{M}_{θ_0} are equivalent if $\sigma_1^2 \alpha_1^{2\nu} = \sigma_0^2 \alpha_0^{2\nu}$ and are orthogonal if $\sigma_1^2 \alpha_1^{2\nu} \neq \sigma_0^2 \alpha_0^{2\nu}$. Hence, $\sigma_0^2 \alpha_0^{2\nu}$ is microergodic, and in particular σ_0^2 and α_0 are non-microergodic.*

We remark that Proposition 1 holds for $d \in \{1, 2, 3\}$, which is the ambient assumption in Sect. 4.3. When $d \geq 5$, Anderes (2010) proved that the full parameter (σ_0^2, α_0) is microergodic (thus in particular σ_0^2 and α_0 are microergodic). At the time of Anderes (2010), it was mentioned there that the case $d = 4$ was open, that is, it was not known if σ_0^2 and α_0 are microergodic in this case. Currently, this case is still open, to the best of our knowledge.

Then, Zhang (2004) finds a consistent estimator of $\sigma_0^2 \alpha_0^{2\nu}$ by fixing α to an arbitrary value and by maximizing the likelihood with respect to σ^2 only. Hence, for $\alpha \in [\alpha_{\text{inf}}, \alpha_{\text{sup}}]$, let

$$\hat{\sigma}^2(\alpha) = \underset{\sigma^2 \in (0, \infty)}{\operatorname{argmin}} L_n(\sigma^2, \alpha).$$

We remark that the argmin is unique from (22) in the proof of Theorem 3. By canceling the derivative of $L_n(\sigma^2, \alpha)$ with respect to σ^2 , we find

$$\hat{\sigma}^2(\alpha) = \frac{1}{n} y^\top \Sigma_\alpha^{-1} y, \tag{21}$$

with $\Sigma_\alpha = R_{\sigma^2, \alpha} / \sigma^2$, based on (22) in the proof of Theorem 3.

Theorem 3 (Zhang 2004) *Let α_1 be any fixed element of $[\alpha_{\text{inf}}, \alpha_{\text{sup}}]$. As $n \rightarrow \infty$, almost surely,*

$$\hat{\sigma}^2(\alpha_1) \alpha_1^{2\nu} \rightarrow \sigma_0^2 \alpha_0^{2\nu}.$$

Proof (sketch) Let

$$\sigma_1^2 = \frac{\sigma_0^2 \alpha_0^{2\nu}}{\alpha_1^{2\nu}}.$$

Let $\epsilon > 0$. From Proposition 1, the measures $\mathcal{M}_{\sigma_0^2, \alpha_0}$ and $\mathcal{M}_{\sigma_1^2, \alpha_1}$ are equivalent and the measures $\mathcal{M}_{\sigma_0^2, \alpha_0}$ and $\mathcal{M}_{\sigma_1^2 + \epsilon, \alpha_1}$ are orthogonal. Hence, Zhang (2004), based on Gikhman and Skorokhod (2004), obtains that, almost surely,

$$nL_n(\sigma_1^2 + \epsilon, \alpha_1) - nL_n(\sigma_1^2, \alpha_1) \rightarrow \infty.$$

Similarly, we can show that, almost surely,

$$nL_n(\sigma_1^2 - \epsilon, \alpha_1) - nL_n(\sigma_1^2, \alpha_1) \rightarrow \infty.$$

Let $\Sigma_{\alpha_1} = R_{\sigma^2, \alpha_1} / \sigma^2$. Then

$$L_n(\sigma^2, \alpha_1) = \log(\sigma^2) + \frac{1}{n} \log(|\Sigma_{\alpha_1}|) + \frac{1}{\sigma^2} \frac{1}{n} y^\top \Sigma_{\alpha_1}^{-1} y. \quad (22)$$

Hence, $1/\sigma^2 \mapsto nL_n(\sigma^2, \alpha_1)$ is convex, and thus by convexity we obtain, as $n \rightarrow \infty$,

$$\left(\inf_{\substack{\sigma^2 \in (0, \infty) \\ |\sigma^2 - \sigma_1^2| \geq \epsilon}} nL_n(\sigma^2, \alpha_1) \right) - nL_n(\sigma_1^2, \alpha_1) \rightarrow \infty$$

almost surely. This implies that $\hat{\sigma}^2(\alpha_1) \rightarrow \sigma_1^2$ almost surely as $n \rightarrow \infty$ which concludes the proof. \square

In the supplementary material, still for the Matérn covariance functions, we also provide asymptotic normality results for the estimator $\hat{\sigma}^2(\alpha_1) \alpha_1^{2\nu}$ and for the “full” maximum likelihood estimator, where the likelihood is maximized with respect to both σ^2 and α .

We remark that, in general and outside of the Matérn case, consistency results for maximum likelihood under fixed-domain asymptotics are quite scarce. We mention a few such other consistency results at the end of Appendix A in the supplementary material and in Sect. 5.

5 Conclusion

We have presented some asymptotic results on covariance parameter estimation under increasing and fixed-domain asymptotics. The presentation highlights the strong differences between the two settings. Under increasing-domain asymptotics, with mild identifiability conditions, all the components of the covariance parameter can be estimated consistently, and with asymptotic normality. The proof techniques hold for general families of stationary covariance functions. They are based on the asymptotic independence between most pairs of observations, as $n \rightarrow \infty$, that enables to control the logarithm of the likelihood and its gradient and to apply general methods for M-estimators.

In contrast, under fixed-domain asymptotics, typically all pairs of observations have a covariance that is not small. As a consequence, some components of the covariance parameter cannot be estimated consistently, even if changing the component changes the covariance function. The notion of equivalence of Gaussian measures, yielding the notion of microergodicity, is central. The results and proofs are not general in the current state of the literature. Here we have presented results and proofs related to the family of isotropic Matérn covariance functions in dimension $d = 1, 2, 3$. The presented proofs rely on the Fourier transforms of these covariance functions (through the results taken from the cited references) and also on the explicit expression of the logarithm of the likelihood as a function of the variance parameter σ^2 .

There are many other existing contributions in the literature that we have not presented here. Under increasing-domain asymptotics, earlier results on maximum likelihood were provided by Mardia and Marshall (1984), using general results from Sweeting (1980) (the latter not necessarily considering Gaussian processes). Restricted maximum likelihood was then studied in Cressie and Lahiri (1993). Cross-validation was considered in Bachoc (2014), Bachoc (2018). Extensions to transformed Gaussian processes were studied in Bachoc et al. (2020). Pairwise likelihood was studied in Bevilacqua and Gaetan (2015). Multivariate processes were considered in Furrer et al. (2016), Shaby and Ruppert (2012). Finally, more generally, the increasing-domain asymptotic framework is investigated in spatial statistics, for instance, in Hallin et al. (2009), Lahiri and Robinson (2016), Lahiri (2003), Lahiri and Mukherjee (2004).

Under fixed-domain asymptotics, earlier results for the estimation of the microergodic parameter in the family of exponential covariance functions in dimension one were obtained in Ying (1991). The estimation of parameters for the Brownian motion is addressed in Stein (1990b). Various additional results on maximum likelihood are obtained in Loh (2005), Loh and Lam (2000), Van der Vaart (1996), Ying (1993). Variation-based estimators are studied in Anderes (2010), Blanke and Vial (2014), Istas and Lang (1997), Loh (2015). Composite likelihood is addressed in Bachoc et al. (2019). The case of covariance parameter estimation for constrained Gaussian processes is addressed in Bachoc et al. (2019), López-Lopera et al. (2018). Cross-validation is addressed in Bachoc et al. (2017). Finally, extensions of the fixed-domain asymptotic results presented here to the family of isotropic Wendland covariance functions are provided in Bevilacqua et al. (2019).

Acknowledgements We are grateful to Abdelaati Daouia and Anne Ruiz-Gazen for suggesting to write this manuscript. We are also grateful to two anonymous referees, for their constructive comments, which led to an improvement of this manuscript.

References

- Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York, ninth Dover printing, tenth GPO printing edition: Dover.
- Adler, R. (1981). *The geometry of random fields*. New York: Wiley.
- Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes*. IMS.
- Anderes, E. (2010). On the consistent separation of scale and variance for Gaussian random fields. *Annals of Statistics*, 38, 870–893.
- Azaïs, J.-M., & Wschebor, M. (2009). *Level sets and extrema of random processes and fields*. Wiley.
- Bachoc, F. (2013a). Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics and Data Analysis*, 66, 55–69.
- Bachoc, F. (2013b) *Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments*. Ph.D. thesis, Université Paris-Diderot - Paris VII. <https://tel.archives-ouvertes.fr/tel-00881002/document>.

- Bachoc, F. (2014). Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *Journal of Multivariate Analysis*, 125, 1–35.
- Bachoc, F. (2018). Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case. *Bernoulli*, 24(2), 1531–1575.
- Bachoc, F., Ammar, K., & Martinez, J. (2016). Improvement of code behavior in a design of experiments by metamodeling. *Nuclear Science and Engineering*, 183(3), 387–406.
- Bachoc, F., Bétancourt, J., Furrer, R., & Klein, T. (2020). Asymptotic properties of the maximum likelihood and cross validation estimators for transformed Gaussian processes. *Electronic Journal of Statistics*, 14(1), 1962–2008.
- Bachoc, F., Bevilacqua, M., & Velandia, D. (2019). Composite likelihood estimation for a Gaussian process under fixed domain asymptotics. *Journal of Multivariate Analysis*, 174.
- Bachoc, F., Bois, G., Garnier, J., & Martinez, J.-M. (2014). Calibration and improved prediction of computer models by universal Kriging. *Nuclear Science and Engineering*, 176(1), 81–97.
- Bachoc, F., & Furrer, R. (2016). On the smallest eigenvalues of covariance matrices of multivariate spatial processes. *Stat*, 5(1), 102–107.
- Bachoc, F., Lagnoux, A., & López-Lopera, A. F. (2019). Maximum likelihood estimation for Gaussian processes under inequality constraints. *Electronic Journal of Statistics*, 13(2), 2921–2969.
- Bachoc, F., Lagnoux, A., & Nguyen, T. M. N. (2017). Cross-validation estimation of covariance parameters under fixed-domain asymptotics. *Journal of Multivariate Analysis*, 160, 42–67.
- Bect, J., Ginsbourger, D., Li, L., Picheny, V., & Vazquez, E. (2012). Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3), 773–793.
- Bevilacqua, M., Faouzi, T., Furrer, R., & Porcu, E. (2019). Estimation and prediction using generalized Wendland covariance functions under fixed domain asymptotics. *The Annals of Statistics*, 47(2), 828–856.
- Bevilacqua, M., & Gaetan, C. (2015). Comparing composite likelihood methods based on pairs for spatial Gaussian random fields. *Statistics and Computing*, 25(5), 877–892.
- Blanke, D., & Vial, C. (2014). Global smoothness estimation of a Gaussian process from general sequence designs. *Electronic Journal of Statistics*, 8(1), 1152–1187.
- Cressie, N., & Lahiri, S. (1993). The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis*, 45, 217–233.
- Furrer, R., Bachoc, F., & Du, J. (2016). Asymptotic properties of multivariate tapering for estimation and prediction. *Journal of Multivariate Analysis*, 149, 177–191.
- Genton, M. G., & Kleiber, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statistical Science*, 30(2), 147–163.
- Gikhman, I., & Skorokhod, A. (2004). *The theory of stochastic processes II*. Springer Science & Business Media.
- Giné, E., & Nickl, R. (2016). *Mathematical foundations of infinite-dimensional statistical models* (Vol. 40). Cambridge University Press.
- Gneiting, T., Kleiber, W., & Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491), 1167–1177.
- Gneiting, T., & Schlather, M. (2004). Stochastic models that separate fractal dimension and the hurst effect. *SIAM Review*, 46(2), 269–282.
- Hallin, M., Lu, Z., & Yu, K. (2009). Local linear spatial quantile regression. *Bernoulli*, 22(1), 659–686.
- Ibragimov, I., & Rozanov, Y. (1978). *Gaussian random processes*. New York: Springer.
- Istas, J., & Lang, G. (1997). Quadratic variations and estimation of the local Hölder index of a Gaussian process. *Annales Institut Henri Poincaré Probability Statistics*, 33(4), 407–436.
- Jones, D., Schonlau, M., & Welch, W. (1998). Efficient global optimization of expensive black box functions. *Journal of Global Optimization*, 13, 455–492.
- Kennedy, M. C., & O’Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3), 425–464.

- Lahiri, S., & Robinson, P. (2016). Central limit theorems for long range dependent spatial linear processes. *Bernoulli*, 22(1), 345–375.
- Lahiri, S. N. (2003). Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs. *Sankhyā: The Indian Journal of Statistics*, 65, 356–388.
- Lahiri, S. N., & Mukherjee, K. (2004). Asymptotic distributions of M-estimators in a spatial regression model under some fixed and stochastic spatial sampling designs. *Annals of the Institute of Statistical Mathematics*, 56, 225–250.
- Loh, W. (2005). Fixed domain asymptotics for a subclass of Matérn type Gaussian random fields. *Annals of Statistics*, 33, 2344–2394.
- Loh, W., & Lam, T. (2000). Estimating structured correlation matrices in smooth Gaussian random field models. *Annals of Statistics*, 28, 880–904.
- Loh, W.-L. (2015). Estimating the smoothness of a Gaussian random field from irregularly spaced data via higher-order quadratic variations. *The Annals of Statistics*, 43(6), 2766–2794.
- López-Lopera, A. F., Bachoc, F., Durrande, N., & Roustant, O. (2018). Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3), 1224–1255.
- Mardia, K., & Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71, 135–146.
- Matheron, G. (1970). *La Théorie des Variables Régionalisées et ses Applications*. Fascicule 5 in Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau. Ecole Nationale Supérieure des Mines de Paris.
- Molchanov, I. (2005). *Theory of random sets* (Vol. 19). Springer.
- Paulo, R., Garcia-Donato, G., & Palomo, J. (2012). Calibration of computer models with multivariate output. *Computational Statistics and Data Analysis*, 56, 3959–3974.
- Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning*. Cambridge: The MIT Press.
- Sacks, J., Welch, W., Mitchell, T., & Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4, 409–423.
- Santner, T., Williams, B., & Notz, W. (2003). *The design and analysis of computer experiments*. New York: Springer.
- Shaby, B. A., & Ruppert, D. (2012). Tapered covariance: Bayesian estimation and asymptotics. *Journal of Computational and Graphical Statistics*, 21(2), 433–452.
- Stein, M. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Annals of Statistics*, 16, 55–63.
- Stein, M. (1990). Bounds on the efficiency of linear predictions using an incorrect covariance function. *Annals of Statistics*, 18, 1116–1138.
- Stein, M. (1990). A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. *Annals of Statistics*, 18, 1139–1157.
- Stein, M. (1990). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *Annals of Statistics*, 18, 850–872.
- Stein, M. (1999). *Interpolation of spatial data: some theory for kriging*. New York: Springer.
- Sweeting, T. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Annals of Statistics*, 8, 1375–1381.
- Van der Vaart, A. W. (1996). Maximum likelihood estimation under a spatial sampling scheme. *Annals of Statistics*, 24(5), 2049–2057.
- Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge University Press.
- Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36, 280–296.
- Ying, Z. (1993). Maximum likelihood estimation of parameters under a spatial sampling scheme. *Annals of Statistics*, 21, 1567–1590.
- Zhang, H. (2004). Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99, 250–261.

Global Scan Methods for Comparing Two Spatial Point Processes



Florent Bonneu and Lionel Cucala

Abstract In many scientific areas such as forestry, ecology, or epidemiology, deciding whether two spatial point patterns are equally distributed is an important issue. This work proposes an adaptation of spatial scan methods, originally designed for local cluster detection, in order to test for the global similarity between two spatial point patterns. We design two spatial global scan statistics based on likelihood ratio on the one hand and on moments on the other, and explain how to compute their significance. A simulation procedure is conducted to compare these global scan methods to others based on kernel density estimation or second-order summary statistics. We also apply them to a dataset of wildfires registered in France.

1 Introduction

In the past years, the analysis of spatial point patterns has received much attention (Baddeley et al. 2015; Diggle 2003; Møller and Waagepetersen 2003). Thanks to the daily use of sensors, many datasets consist nowadays of spatial locations of random events: earthquakes epicentres, tree species in a forest, addresses of people affected by a certain disease... These spatial locations are often completed by a mark, that can be either binary, continuous, or any other type. In many cases, a comparison is needed between the spatial distribution of two point patterns: for example, Diggle et al. (1990) compare the spatial distribution of larynx cancers with the underlying population. The main concern is often to know whether events from the first dataset

This work is dedicated to Christine Thomas-Agnan who directed our both Phd theses. We thank her for the constant support and help. Thanks to her human skills of listening and understanding, working with her is always a delight.

F. Bonneu
Avignon Université, Avignon, France
e-mail: florent.bonneu@univ-avignon.fr

L. Cucala (✉)
Université de Montpellier, Montpellier, France
e-mail: lionel.cucala@umontpellier.fr

tend to cluster around one or more hotspots compared to events from the second dataset.

In this case-control framework, tests of spatial clustering can be split into three types (Lawson and Denison 2002; Gelfand et al. 2010). Focussed tests are designed to check whether events are abnormally clustered around a specific point, such as a pollution source (Diggle 2003). Global tests measure how events tend to cluster in the whole observation domain: some are based on the comparison of kernel intensity estimators (Fuentes-Santos et al. 2017) or similarly analyse the kernel estimation of the relative risk (Kelsall and Diggle 1995), others are based on the partitioning of the observation domain in cells (Alba-Fernández et al. 2016; Allard and Fraley 1997) whereas distance-based methods compare second-order characteristics such as their K functions (Diggle and Chetwynd 1991; Bonneu and Thomas-Agnan 2015). A number of papers also investigate the distribution properties of the marks, either by testing the mark independence (Grabarnik et al. 2011) or by computing the sample cross K-function (Illian et al. 2008). Finally, local tests identify the specific area where events are most clustered, for example, with LISA functions (Moraga and Montes 2011). Among these local tests, variable-window scan statistics play an essential role since the initial role by Kulldorff and Nagarwalla (1995). A scan statistic is the maximum value of a clustering index measured on a set of candidate clusters.

The idea we investigate in this paper is the adaptation of scan statistics from local tests to global tests by considering all the clustering indices measured on the potential clusters instead of only retaining the maximum one. Section 2 describes the global scan statistics and their computational aspects. These scan statistics are then applied to real and simulated datasets in Sect. 3. The paper concludes with a discussion.

2 Methodology

Let $X = \{x_1, \dots, x_n\}$ be a realization of a spatial point process observed in a bounded region $W \subset \mathbb{R}^2$. Suppose that a bivariate mark is associated to each location x_i : $\forall i = 1, \dots, n, m_i \in \{0, 1\}$ without loss of generality. Thus, the initial point pattern can be split up:

$$X = X_0 \cup X_1,$$

where $X_0 = \{x_{0,1}, \dots, x_{0,n_0}\}$ and $X_1 = \{x_{1,1}, \dots, x_{1,n_1}\}$ are, respectively, the spatial patterns of type 0 and 1, and $n_0 + n_1 = n$.

We would like to test for H_0 : “Conditionally to the events locations X_1, \dots, X_n , the marks M_1, \dots, M_n are independent and identically distributed”. This corresponds to the so-called random labeling hypothesis (events of a single point process are independently assigned to the two groups) rather than the so-called random superposition hypothesis (superposition of two independent populations of events) (Diggle 2010).

A direct consequence of this hypothesis is that the marginal spatial distributions X_0 and X_1 correspond to random thinned versions of the same spatial point process model. This assumption is the starting point of spatial scan methods, whose objective

is to detect the most significant cluster, i.e. the area of W in which the “distance” between what we observe and what should be observed under H_0 is the greatest. Remark that this null hypothesis does not assume anything about the spatial distribution of X but just considers that the distribution of the mark M_i does not depend on the location X_i . Contrarily to our test, many other procedures for comparing point patterns rely on a Poisson assumption (Fuentes-Santos et al. 2017).

2.1 Spatial Scan Statistics for Bivariate Data

As said before, a scan statistic, as originally defined by Cressie (1977), is the maximum value of a clustering index on a set of candidate clusters. Let a Borel set $Z \subset W$ be any candidate cluster. The clustering index, denoted by $I(Z)$, measures the clustering of abnormal marks in Z with respect to H_0 . We will describe later on the clustering indices that can be considered.

The set of candidate clusters can be chosen in different ways. In this article, we will focus on circular clusters, such as Kulldorff (1997), but all our definitions in the sequel are valid for any set of candidate clusters. The set of potential clusters, denoted by \mathcal{D} , is the set of discs centred on one location and passing through another one:

$$\mathcal{D} = \{D_{i,j}, 1 \leq i \leq n, 1 \leq j \leq n\},$$

where $D_{i,j} := \{s \in W : \|s - x_i\| \leq \|x_j - x_i\|\}$ is the closed disc centred on x_i and passing through x_j . Since the disc may have null radius (if $i = j$), the number of potential clusters is n^2 .

The original idea for comparing all these candidate clusters has been introduced by Kulldorff (1997). It relies on the likelihood ratio between two parametric hypotheses. Under H_0 , the marks M_1, \dots, M_n are independent and identically distributed. Since they are bivariate, they follow a Bernoulli distribution with parameter p , $\mathcal{B}(p)$. The likelihood ratio relating to the null hypothesis is thus

$$LR_0(m_1, \dots, m_n; p) = \prod_{i=1}^n p^{m_i} (1 - p)^{1-m_i}.$$

For any candidate cluster $Z \subset W$, an alternative hypothesis $H_{1,Z}$ is introduced, stating that the marks are still independent but follow a Bernoulli distribution with different parameters in Z and in Z^c , the complementary set of Z in W :

$$\begin{cases} M_i \sim \mathcal{B}(p_Z) & \text{if } x_i \in Z, \\ M_i \sim \mathcal{B}(p_{Z^c}) & \text{if } x_i \in Z^c, \end{cases}$$

and the likelihood ratio relating to this alternative hypothesis is thus

$$LR_{1,Z}(m_1, \dots, m_n; p_Z, p_{Z^c}) = \prod_{i: x_i \in Z} (p_Z)^{m_i} (1 - p_Z)^{1 - m_i} \prod_{i: x_i \in Z^c} (p_{Z^c})^{m_i} (1 - p_{Z^c})^{1 - m_i}.$$

Let p^* , p_Z^* and $p_{Z^c}^*$ be the maximum likelihood estimators of p , p_Z and p_{Z^c} , respectively. Their computation is straightforward:

$$p^* = \arg \max_{p \in [0,1]} LR_0(m_1, \dots, m_n; p) = \frac{n_1}{n}$$

and

$$(p_Z^*, p_{Z^c}^*) = \arg \max_{(p_Z, p_{Z^c}) \in [0,1]^2} LR_{1,Z}(m_1, \dots, m_n; p_Z, p_{Z^c}) = \left(\frac{n_1(Z)}{n(Z)}, \frac{n_1(Z^c)}{n(Z^c)} \right),$$

where $n(Z) = \sum_{i=1}^n \mathbb{1}(x_i \in Z)$ and $n_1(Z) = \sum_{i=1}^n m_i \mathbb{1}(x_i \in Z)$ are, respectively, the total number of events and the number of type 1 events in Z . The log-likelihood ratio (LLR) between hypotheses H_0 and $H_{1,Z}$ can thus be computed:

$$\begin{aligned} LLR(Z) &= \log(LR_{1,Z}(m_1, \dots, m_n; p_Z^*, p_{Z^c}^*)) - \log(LR_0(m_1, \dots, m_n; p^*)) \\ &= n_1(Z) \log\left(\frac{n_1(Z)}{n(Z)}\right) + (n(Z) - n_1(Z)) \log\left(1 - \frac{n_1(Z)}{n(Z)}\right) \\ &\quad + n_1(Z^c) \log\left(\frac{n_1(Z^c)}{n(Z^c)}\right) + (n(Z^c) - n_1(Z^c)) \log\left(1 - \frac{n_1(Z^c)}{n(Z^c)}\right) \\ &\quad - n_1 \log\left(\frac{n_1}{n}\right) - (n - n_1) \log\left(1 - \frac{n_1}{n}\right). \end{aligned}$$

This log-likelihood ratio $LLR(Z)$, which has already been used by Allard and Fraley (1997) to estimate the support domain of a bounded point process in presence of background noise, is a relevant clustering index for abnormal marks in Z since it increases when the proportions of type 1 events differ between Z and Z^c . However, it does not indicate whether the proportion of type 1 events in Z is higher or lower than in Z^c . Therefore, in order to collect this information, we propose to use the following LLR-based clustering index:

$$I_{LLR(Z)} = LLR(Z) \left[\mathbb{1}\left(\frac{n_1(Z)}{n(Z)} > \frac{n_1(Z^c)}{n(Z^c)}\right) - \mathbb{1}\left(\frac{n_1(Z)}{n(Z)} < \frac{n_1(Z^c)}{n(Z^c)}\right) \right].$$

Another clustering index for comparing candidate clusters has been introduced by Cucala (2014). This one is not restricted to bivariate marks but can be applied to any numerical marks since it does not rely on any distribution assumption. Under H_0 , the marks M_1, \dots, M_n are all independent and identically distributed, so that their means and variances are equal. Therefore, the number of type 1 events in Z , $N_1(Z)$, follows the binomial distribution $Bin(n(Z), \frac{n_1}{n})$. Since $\mathbb{E}(N_1(Z)) = n(Z) \frac{n_1}{n}$ and $Var(N_1(Z)) = n(Z) \frac{n_0 n_1}{n^2}$, the moment-based clustering index

$$I_M(Z) = \frac{N_1(Z) - n(Z)\frac{n_1}{n}}{\sqrt{n(Z)\frac{n_0n_1}{n^2}}}$$

has null mean and unit variance for every $Z \subset W$. Since it increases when the proportion of type 1 events in Z is higher than in Z^c and decreases when it is lower in Z than in Z^c , its behaviour is very similar to $I_{LLR}(Z)$.

Both indices have been designed for computing local spatial scan statistics and detecting the area where the hypothesis H_0 is most violated. If one looks for either excess or default of type 1 events, the relevant local spatial scan statistic is

$$\Psi^{max} = \max_{Z \in \mathcal{D}} |I(Z)|$$

and the most likely cluster is

$$\hat{C} = \arg \max_{Z \in \mathcal{D}} |I(Z)|.$$

From now on, Ψ_{LLR}^{max} and Ψ_M^{max} will denote the local spatial scan statistics computed using respectively $I_{LLR}(Z)$ and $I_M(Z)$. Remark that, if one looks only for excess of type 1 events or only for default of type 1 events, $\max_{Z \in \mathcal{C}} I(Z)$ or $\min_{Z \in \mathcal{C}} I(Z)$ should be, respectively, investigated.

These local spatial scan statistics are not designed for global investigation of the differences of spatial distribution between X_0 and X_1 since they only retain the maximum (in absolute value) of the abnormality index. A consequence of X_0 and X_1 being differently spatially distributed would be the presence of many candidate clusters with high index $I(Z)$ (i.e. excess of type 1 events), and also many candidate clusters with low index $I(Z)$ (i.e. default of type 1 events). Thus, it is necessary to introduce summary statistics measuring the empirical distribution of these indices. To this end, we propose the following global spatial scan statistics

$$\Psi_{LLR}^{var} = \text{Var}(\{I_{LLR}(Z), Z \in \mathcal{D}\}) :$$

and

$$\Psi_M^{var} = \text{Var}(\{I_M(Z), Z \in \mathcal{D}\}),$$

where $\text{Var}()$ stands for the empirical variance of a real sample.

2.2 Significance Issues

Once a scan statistic (either local or global) is computed, we need to evaluate its significance. Unfortunately, its null distribution is untractable due to the dependence between $I(Z)$ and $I(Z')$ if $n(Z \cap Z') \neq 0$. Another solution, chosen by Kulldorff

(1997) or Kulldorff et al. (2009), would be to simulate random datasets under the null hypothesis but this is only possible when one can simulate random locations similarly to the ones in X . Here, since we do not want to assume anything about the spatial distribution of the events, we decided to run a technique called random labelling: a simulated dataset is obtained by randomly associating the marks to the spatial locations. When doing this, the overall spatial structure of locations is preserved so that the simulated dataset satisfies the H_0 hypothesis of a random distribution of marks conditionally to the point locations. Notice that the random labelling technique allows to satisfy the requirements of a good Monte Carlo procedure mentioned by Baddeley et al. (2015) and in particular the exchangeability condition (Mrkvička et al. 2020).

Let T denote the number of randomized datasets and $\Psi^{(1)}, \dots, \Psi^{(T)}$ be the observations of the spatial scan statistic associated with these datasets. Throughout this paper, for the simulation study or the application to forest fire occurrences, T is always fixed to 99 simulated patterns under H_0 . According to Dwass (1957), the Monte-Carlo-based p-value of the scan statistic Ψ , observed on the initial sample, is $\frac{R}{T+1}$, where R is the rank of Ψ in the $(T + 1)$ -sample $(\Psi^{(1)}, \dots, \Psi^{(T)}, \Psi)$. Note that this p-value is unbiased in the sense that under the null hypothesis, the probability of observing a p-value less than or equal to p is exactly p . According to the classical test theory, the most likely cluster \hat{C} is said to be significant if the associated p-value is less than the type I error α .

3 Applications

We illustrate our global clustering tests on a simulation study comparing the results of different methods and on a real dataset of forest fire occurrences in France. All the implementation is made in R code,¹ using sometimes existing functions available in R packages.

3.1 Simulation Study

In order to evaluate the performance of our tests based on Ψ_M^{var} , Ψ_M^{max} , Ψ_{LLR}^{var} and Ψ_{LLR}^{max} , we design three simulated scenarios of bivariate data with different spatial structures: absence of spatial variation in risk (model ASVR), presence of spatial variation in risk (model PSVR) and stationary spatial clustering (model SSC) and we compare them with existing tests constructed with Ψ_F and Ψ_{DC} statistics.

The statistic Ψ_F is a modified version of the statistic introduced by Fuentes-Santos et al. (2017) in the spatial point process framework corresponding to the following square discrepancy measure

¹Available on demand at the authors' email addresses.

$$\Psi_F = \int_W (\hat{f}_0(x) - \hat{f}_1(x))^2 dx,$$

where \hat{f}_0 and \hat{f}_1 are, respectively, the estimated densities of event locations of type 0 and 1. Our statistic Ψ_F differs from that in Fuentes-Santos et al. (2017) in the estimation procedure of the case and control densities. We use here a Gaussian kernel and the bandwidth selection in Cronie and Van Lieshout (2018) instead of a radial kernel and the plug-in bandwidth selector described in Fuentes-Santos et al. (2017). The test based on Ψ_F is realized with the `kde.test` function of the R package `kde`. The statistic Ψ_{DC} is the statistic introduced by Diggle and Chetwynd (1991) defined by

$$\Psi_{DC} = \frac{\sum_{k=1}^m \hat{D}(r_k)}{\sqrt{\text{Var}(\{\hat{D}(r_k), k = 1, \dots, m\})}}$$

where $\hat{D}(r_k) = \hat{K}_0(r_k) - \hat{K}_1(r_k)$ is the difference of the estimated Ripley K functions of cases and controls and r_k are equally spaced distances for $k = 1, \dots, m$.

The significances of the six statistics are estimated via the random labelling approach described in the previous section, and using the same permuted samples in order to avoid overdispersion. The type I error is set to $\alpha = 5\%$.

We now describe the bivariate point process models considered in our simulation study.

- ASVR model: spatial Poisson point process with intensity function $\lambda(x, y) = N \exp(-3y)$ and types 0 and 1 allocated by random labelling.
- PSVR model: superimposed Poisson point processes with intensity functions $\lambda_0(x, y) = N_1 \exp(-3y)$ for type 0 and $\lambda_1(x, y) = N_2(x^2 + y^2)$ for type 1.
- SSC model: superimposed 50 points uniformly distributed for type 0 and a Thomas cluster point pattern for type 1 with intensity of the Poisson process of cluster centres equal to 5, a standard deviation of random displacement corresponding to 0.1 and a mean number of points per cluster equal to 10.

The constants N , N_1 and N_2 are chosen so that the mean number of points per marginal point process is equal to 50.

Table 1 shows the rejection rate of the null hypothesis obtained with 100 simulations of each scenario for the six tests considered. The results are relatively equivalent between the different test procedures except for the Diggle–Chetwynd statistic which is based on the second-order structure of the spatial point patterns and is not adapted to detect differences of concentration at the first order. Remark that the rejection rates for scenario ASVR are very close to the nominal type I error since this satisfies the null hypothesis.

Because the chosen scenarios PSVR and SSC are far from the null hypothesis of no spatial variation in risk, we construct several other examples with a rate ϵ indicating the departure from H_0 . For $\epsilon = 0$, we are under the ASVR model and for $\epsilon = 1$ we are further than ever from the null hypothesis H_0 . Considering ϵ values between 0 and 1 allows us to detect the range where the departure from the null hypothesis is

Table 1 Rejection rate of the null hypothesis for the six tests computed on 100 simulations of each model

Model	Ψ_M^{max}	Ψ_M^{var}	Ψ_{LLR}^{max}	Ψ_{LLR}^{var}	Ψ_{DC}	Ψ_F
ASVR	0.04	0.03	0.07	0.05	0.04	0.04
PSVR	1.00	1.00	1.00	1.00	0.03	0.98
SSC	1.00	1.00	1.00	0.95	1.00	1.00

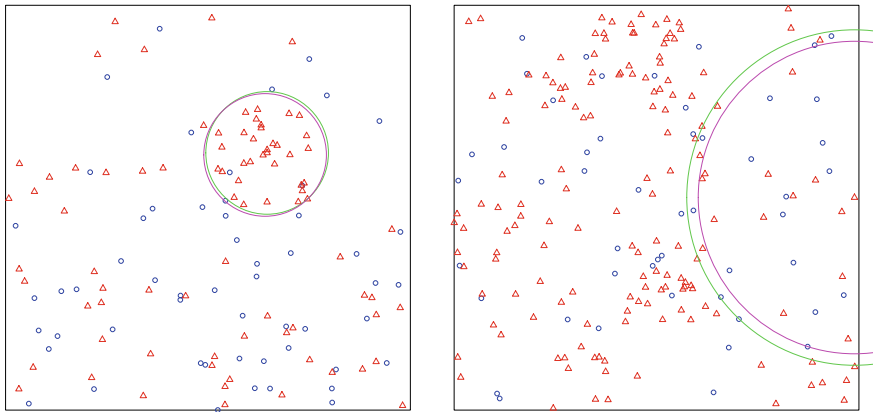


Fig. 1 Simulated patterns of bivariate data of type 0 in blue circles and of type 1 in red triangles with an epsilon parameter equal to 0.5 for model PSVReps (left) and for model SSCeps (right)

detected and thus to have an indication of the power of the test. We present below the two new models named PSVReps and SSCeps and show in Fig. 1 simulated patterns for each one with $\epsilon = 0.5$ and the most likely clusters, i.e. the discs maximizing $|I_M(Z)|$ and $|I_{LLR}(Z)|$, respectively, coloured in green and magenta.

- PSVReps model: superimposed Poisson point processes with intensity functions $\lambda_0(x, y) = N_1 \exp(-2y)$ for type 0 and $\lambda_1(x, y) = N_1 \exp(-2y) + \epsilon \frac{50}{N_1} \frac{1}{4}^{-2} \mathbb{1}\{0.5 < x < 0.75; 0.5 < y < 0.75\}$ for type 1.
- SSCeps model: superimposed 50 points uniformly distributed for type 0 and a Thomas cluster point pattern for type 1 with intensity of the Poisson process of cluster centres equal to $50(1 - 0.9\epsilon)$, a standard deviation of random displacement corresponding to $0.1 - 0.09\epsilon$ and a mean number of points per cluster equal to $1 + 9\epsilon$.

Tables 2 and 3 present the rejection rates of the null hypothesis for the two models PSVReps and SSCeps for different values of ϵ from 0 to 1 by step 0.2. For the PSVReps model, we notice that the test based on Ψ_F appears to be the most sensitive and gives the best global results. The local scan statistics Ψ_M^{max} and Ψ_{LLR}^{max} give good results, similar to Ψ_F , when ϵ is low and are less sensitive for average values of ϵ . On the opposite, the statistic Ψ_{DC} is not adequate in this situation. For the SSCeps model,

Table 2 Rejection rate of the null hypothesis for the six tests computed on 100 simulations for each epsilon in {0, 0.2, 0.4, 0.6, 0.8, 1} with model PSVReps

Model	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.6$	$\epsilon = 0.8$	$\epsilon = 1$
Ψ_M^{max}	0.02	0.24	0.66	0.86	0.90	0.96
Ψ_M^{var}	0.05	0.18	0.66	0.91	0.99	0.99
Ψ_{LLR}^{max}	0.02	0.28	0.74	0.94	0.99	0.99
Ψ_{LLR}^{var}	0.04	0.18	0.65	0.92	0.99	0.99
Ψ_{DC}	0.04	0.11	0.42	0.75	0.98	0.97
Ψ_F	0.04	0.26	0.82	0.99	1.00	1.00

Table 3 Rejection rate of the null hypothesis for the six tests computed on 100 simulations for each epsilon in {0, 0.2, 0.4, 0.6, 0.8, 1} with model SSCeps

Model	$\epsilon = 0$	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.6$	$\epsilon = 0.8$	$\epsilon = 1$
Ψ_M^{max}	0.04	0.25	0.45	0.81	0.92	1.00
Ψ_M^{var}	0.05	0.21	0.31	0.43	0.76	1.00
Ψ_{LLR}^{max}	0.03	0.30	0.45	0.53	0.95	1.00
Ψ_{LLR}^{var}	0.05	0.22	0.29	0.41	0.89	0.95
Ψ_{DC}	0.04	0.28	0.61	0.85	0.99	1.00
Ψ_F	0.03	0.18	0.49	0.92	1.00	1.00

we observe approximately the same results: Ψ_M^{max} and Ψ_{LLR}^{max} are very sensitive to detect departure from H_0 with low ϵ but not necessary for ϵ values around 0.6. The test based on Ψ_{DC} gives here very good results due to the fact that the second-order distribution of points is very different between type 0 and 1.

In conclusion, the tests based on local scan statistics Ψ_M^{max} and Ψ_{LLR}^{max} give relatively good results near from those of Ψ_F and are competitive in the case where few points are far from the null hypothesis which it is the case when ϵ is low. The tests based on global scan statistics Ψ_M^{var} and Ψ_{LLR}^{var} are slightly less powerful than the local ones but this might be improved by the choice of another set of candidate clusters. Finally, the choice of the clustering index does not seem to be crucial since the results are similar between methods based on LLR index and moment-based index.

3.2 Forest Fire Occurrences

Forest fires have substantial worldwide impacts on human societies causing notably health issues, environmental disasters, and economic losses. To limit wildfire occurrences and their negative effects, fire risk prevention begins with describing and understanding the stochastic mechanisms governing the spatio-temporal distributions of locations and the propagation dynamics of forest fires. To reach this goal,

many statistical analyses and modelling techniques have been used in the spatio-temporal point process framework, for example, in Opitz et al. (2020) and references therein. In our study, we focus on the French Mediterranean basin which is very exposed to wildfires due to its climate with hot/dry summers and cool/wet winters, its vegetation with highly inflammable species and its important human activity. Forest fires are the primary cause of forest destruction in this region.

Since 1973 in the French Mediterranean region, Fire Departments recorded the locations and characteristics of forest fire occurrences in the Prométhée² database in order to analyse them by statistical tools for improving the knowledge of the spatio-temporal distribution of wildfires and their causes. Historically, the spatial location of wildfires is given by their DFCI coordinates spanning a grid with quadratic cells of 4 km², causing a loss of information due to positional uncertainties. For this reason and because of the growth of GPS³ resources and powerful statistical techniques for spatio-temporal point patterns, the authorities started to record the GPS locations for some forest fires. We consider the overall database of GPS locations of wildfire occurrences registered between 2013 and 2019 in the Var and Haute-Corse departments, two french administrative units with, respectively, 5973 km² and 4666 km².

In forest fire analysis, it is of particular interest to detect and measure spatial distribution differences between two or more sub-samples of forest fire point patterns. For example, we want to know whether the spatial distribution for one year is different from that of the previous year. Another issue consists in assessing whether the spatial distribution for arsons is different for involuntary forest fires (natural, due to negligence...). We have applied the statistical tests described previously to detect overall spatial distribution differences for two types of recorded wildfires for the two issues described above.

For our first application, we focus on the testing of spatial distribution differences between all wildfire locations recorded by GPS over two successive years. These GPS datasets represent samples of all the forest fires recorded at the DFCI scale. Our testing procedure is equivalent to applying first-order separability tests in time. Pairwise comparisons between all the couples of spatial point patterns at the annual scale are feasible but we decide to be concise and restrict our attention to the three most relevant examples. Figure 2 shows the spatial distribution of forest fire occurrences in Var and Haute-Corse departments and the cluster of points with the maximum index value defined, respectively, by $|I_M(Z)|$ and $|I_{LLR}(Z)|$, without taking into account if they are significant for the moment. We can observe that the cluster with maximum $|I_{LLR}(Z)|$ has often a larger radius than the one with maximum $|I_M(Z)|$; this phenomenon has already been noticed by Cucala (2017) when doing cluster detection.

Table 4 gives the p-values of the separability tests for the three cases: *Var1516*, *Var1617* and *HauteCorse1314*, corresponding, respectively, to the spatial point patterns of wildfires in left, middle and right panels in Fig. 2. For *Var1516*, all the

²<http://www.promethee.com/>.

³Global Positioning System.

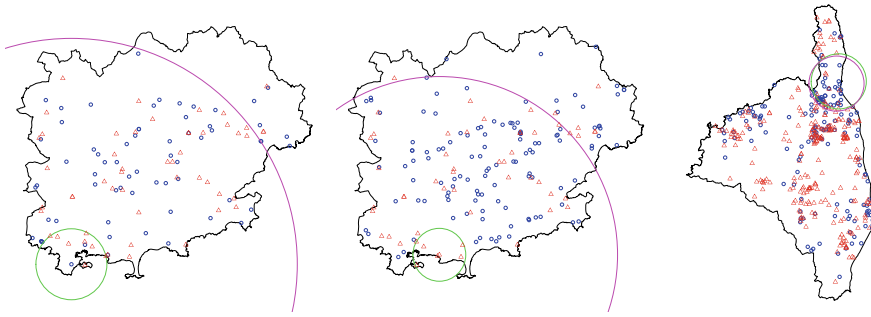


Fig. 2 Locations of forest fire occurrences recorded by GPS in the Var (left-middle) and Haute-Corse (right) departments. Left-middle: 54 locations in 2016 are in red triangles whereas the blue circles represent 66 locations in 2015 (left) and 123 locations in 2017 (middle). Right: 165 locations in 2013 in blue circles and 266 locations in 2014 in red triangles. Green circles represent the clusters with maximum $|I_M(Z)|$ value and the magenta circles represent the clusters with maximum $|I_{LLR}(Z)|$ value. Geographical scales are different for the two administrative units: Var and Haute-Corse

Table 4 P-values of the separability tests of the spatial distributions of forest fires recorded during two successive years in the Var and Haute-Corse departments

Samples	Ψ_M^{max}	Ψ_M^{var}	Ψ_{LLR}^{max}	Ψ_{LLR}^{max}	Ψ_{DC}	Ψ_F
Var1516	0.99	0.9	0.43	0.76	0.25	0.93
Var1617	0.02	0.1	0.02	0.02	0.54	0.4
HauteCorse1314	0.01	0.01	0.01	0.01	0.99	0.01

tests conclude to non-significant differences between the distributions of locations in 2015 and 2016. For the years 2016 and 2017, the tests based on scan statistics Ψ_M^{max} , Ψ_{LLR}^{max} , Ψ_M^{var} and Ψ_{LLR}^{var} indicate significant differences whereas the tests based on Ψ_{DC} and Ψ_F do not. The Ψ_F statistic seems to ignore the differences between the two spatial point patterns when they are spatially distributed in nearby areas, because the main difficulty of this approach is the global bandwidth selection, when the number of points is low as it is suggested in Fuentes-Santos et al. (2017). The test based on Ψ_{DC} indicates that the two spatial point patterns have the same second-order structure. These results are consistent with our knowledge on forest fires occurrences because when the conditions are approximately similar (climate, vegetation...) the second-order structure is unchanged whereas the first-order structure is modified by previous occurrences of large and/or many wildfires that burnt the vegetation locally and avoids new forest fires in the future. In example *HauteCorse1314*, all the tests conclude to spatial distribution difference except the one based on Ψ_{DC} that suggests no-difference in the second-order spatial structure.

For our second application, we group wildfires into two types depending on their nature (arson or not) or their burnt area (less or greater than 1 hectare). Figure 3

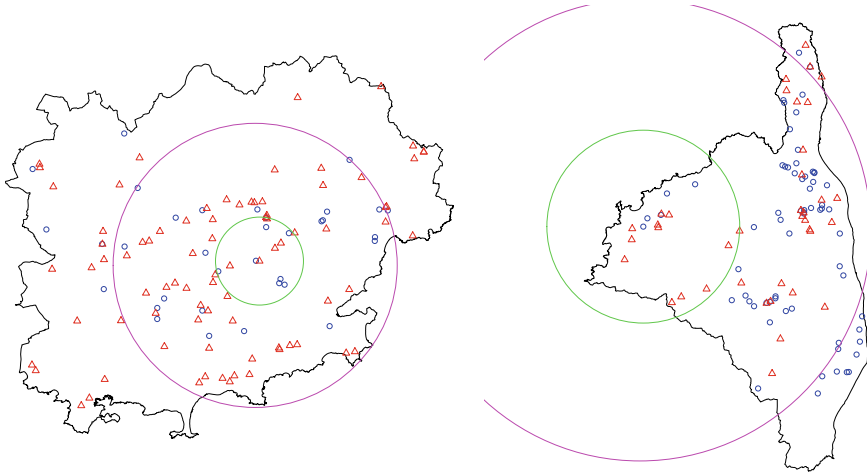


Fig. 3 Left: GPS locations of arsons in red triangles and non-arsons in blue circles in the Var department during the year 2017. Right: GPS locations of forest fire occurrences with a burnt area less than 1 hectare in red triangles and greater than 1 hectare in blue circles in the Haute-Corse department during the year 2016

Table 5 P-values of the comparison tests between spatial distributions of wildfires with burnt area less or greater than 1 hectare in 2017 in the Var department, and arsons and others types (natural, due to negligence...) in 2016 in the Haute-Corse department

Categories	Ψ_M^{max}	Ψ_M^{max}	Ψ_{LLR}^{max}	Ψ_{LLR}^{var}	Ψ_{DC}	Π_F
Arsons/non-Arsons	0.53	0.43	0.4	0.33	0.89	0.35
< 1 ha / > 1 ha	0.35	0.13	0.34	0.11	0.78	0.06

presents the considered samples and the cluster with the largest value for the indices $|I_M(Z)|$ and $|I_{LLR}(Z)|$.

The p-values of all the tests are larger than 5% in Table 5, so we do not reject the null hypothesis of spatial differences between the two point patterns in each case. We notice that the test based on Ψ_F is very close to 5% and the rejection of H_0 , maybe because it does not depend on a fixed cluster family based on discs and so it can detect difference notably in the South East of the Haute-Corse department.

4 Discussion

The scan methods introduced in this article are parameter-free competitive techniques for detecting significant differences between two spatial point patterns. The application to forest fire occurrences has shown that they can be more efficient than

others on a certain type of data. As already mentioned, the global scan methods are more sensitive to the choice of the set of candidate clusters so it could be worth analyzing whether the performances of these methods vary when this set is modified. For example, candidate clusters with a radius higher than a given threshold could be eliminated, such as done by Kulldorff (1997) for cluster detection.

In the simulation study, we compared the local and global scan statistics to a density-based statistic and another comparing the second-order structure of the point pattern. Notice that this study could be completed by computing also distance-based statistics (Hahn 2012) or area-based statistics (Andresen 2009), and by analysing the influence of the number of events of each type.

The global scan statistics we introduced in this paper are only based on the different values of a clustering index computed on candidate clusters. However, the additional information concerning these candidate clusters (centre and radius) could be used by a post-processing technique to obtain the most relevant cluster areas, using for example a density-based method.

Finally, the scan tests that have been designed here for comparing two spatial point patterns could be extended to compare as many point patterns as wanted by using the concentration index introduced by Jung et al. (2010) for multinomial data.

References

- Alba-Fernández, M. V., Ariza-López, F. J., Jiménez-Gamero, M. D., & Rodríguez-Avi, J. (2016). On the similarity analysis of spatial patterns. *Spatial Statistics*, 18, 352–362.
- Allard, D., & Fraley, C. (1997). Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoï tessellation. *Journal of the American Statistical Association*, 92, 1485–1493.
- Andresen, M. (2009). Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. *Applied Geography*, 29, 333–345.
- Baddeley, A., Rubak, E. H., & Turner, R. (2015). *Spatial point patterns: Methodology and applications with R*. Boca Raton: CRC Press, Chapman and Hall.
- Bonneu, F., & Thomas-Agnan, C. (2015). Measuring and testing spatial mass concentration with micro-geographic data. *Spatial Economic Analysis*, 10, 289–316.
- Cressie, N. (1977). On some properties of the scan statistic on the circle and the line. *Journal of Applied Probability*, 14, 272–283.
- Cronie, O., & Van Lieshout, M. N. M. (2018). A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. *Biometrika*, 105, 455–462.
- Cucala, L. (2014). A distribution-free spatial scan statistic for marked point processes. *Spatial Statistics*, 10, 117–125.
- Cucala, L. (2017). Variable window scan statistics: Alternatives to generalized likelihood ratio tests. In J. Glaz & M. Koutras (Eds.), *Handbook of scan statistics*. New York: Springer.
- Diggle, P. J. (2003). *Statistical analysis of spatial and spatio-temporal point patterns*. Boca Raton: CRC Press.
- Diggle, P. J. (2010). Nonparametric methods. In A. E. Gelfand, P. J. Diggle, M. Fuentes, P. Guttorp (Eds.), *Handbook of spatial statistics* (1st ed.). Handbooks of modern statistical methods. Boca Raton: CRC Press.

- Diggle, P. J., Gatrell, A., & Lovett, A. (1990). Modelling the prevalence of cancer of the larynx in part of Lancashire: A new methodology for spatial epidemiology. In R. W. Thomas (Ed.), *Spatial epidemiology*. New York: Wiley.
- Diggle, P. J., & Chetwynd, A. G. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, *47*, 1155–1163.
- Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, *28*, 181–187.
- Fuentes-Santos, I., González-Manteiga, W., & Mateu, J. (2017). A nonparametric test for the comparison of first-order structures of spatial point processes. *Spatial Statistics*, *22*, 240–260.
- Gelfand, A. E., Diggle, P. J., Fuentes, M., & Guttorp, P. (2010). *Handbook of spatial statistics*. Boca Raton: CRC Press.
- Grabarnik, P., Myllymäki, M., & Stoyan, D. (2011). Correct testing of mark independence for marked point patterns. *Ecological Modelling*, *222*, 3888–3894.
- Hahn, U. (2012). A studentized permutation test for the comparison of spatial point patterns. *Journal of the American Statistical Association*, *107*, 754–764.
- Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*. New York: Wiley.
- Jung, I., Kulldorff, M., & Richard, O. (2010). A spatial scan statistic for multinomial data. *Statistics in Medicine*, *29*, 1910–1918.
- Kelsall, J., & Diggle, P. J. (1995). Kernel estimation of relative risk. *Bernoulli*, *1*, 3–16.
- Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics. Theory and Methods*, *26*, 1481–1496.
- Kulldorff, M., Huang, L., & Konty, K. (2009). A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics*, *8*, 58.
- Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, *14*, 799–810.
- Lawson, A., & Denison, D. (2002). *Spatial cluster modelling*. Boca Raton: Chapman and Hall/CRC.
- Møller, J., & Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. Boca Raton: Chapman and Hall/CRC.
- Moraga, P., & Montes, F. (2011). Detection of spatial disease clusters with LISA functions. *Statistics in Medicine*, *30*, 1057–1071.
- Mrkvíčka, T., Dvořák, J., González, J. A., & Mateu, J. (2020). Revisiting the random shift approach for testing in spatial statistics. *Spatial Statistics*. <https://doi.org/10.1016/j.spasta.2020.100430>
- Opitz, T., Bonneu, F., & Gabriel, E. (2020). Point-process based Bayesian modeling of space–time structures of forest fire occurrences in Mediterranean France. *Spatial Statistics*. <https://doi.org/10.1016/j.spasta.2020.100429>

Assessing Spillover Effects of Spatial Policies with Semiparametric Zero-Inflated Models and Random Forests



Hervé Cardot and Antonio Musolesi

Abstract The aim of this work is to estimate the variation over time of the spatial spillover effects of a public policy that was devoted to boost rural development in France over the period 1993–2002. At a micro data level, it is often observed that the dependent variable, such as local employment in a municipality, does not vary along time, so that we face a kind of zero inflated phenomenon that cannot be dealt with a classical continuous response model or propensity score approaches. We consider two recent non parametric techniques that are able to deal with that estimation issue. The first approach consists in fitting two generalized additive models to estimate both the probability of no variation as well as the variation along time of the continuous part of the response. The second approach is based on the use of random forests which can naturally handle the observation of a mixture of a discrete response as well as a continuous one. Instead of estimating average treatment effects, we take advantage of the flexibility of the non parametric approaches to estimate what would have been the potential outcome under treatment, as well as treatment of the neighboring municipalities, on some particular municipalities chosen as being representative or as being of particular interest. The results indicate the evidence of interesting patterns of temporal spatially-mediated spillover effects of the policy with relevant nonlinear effects. Policy spillovers matter, even if they are generally not high in magnitude, for some municipalities with specific demographic and economic characteristics.

Dedicated to Pr. Christine Thomas-Agnan who initiated both of us to non-parametrics and spline fitting and their infinite ways of playing with data.

H. Cardot (✉)
Université de Bourgogne, Institut de Mathématiques de Bourgogne, Dijon, France
e-mail: herve.cardot@u-bourgogne.fr

A. Musolesi
Department of Economics and Management and SEEDS, Ferrara University, Ferrara, Italy
e-mail: mslntn@unife.it

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_17

319

1 Introduction

This work is motivated by the evaluation of the variation along time of the spatial spillover effects of a public policy that was devoted to boost rural development in France over the period 1993–2002. Economic and demographic characteristics are measured at the municipality level for a sample of more than 25000 municipalities and the aim is to estimate the effect of the rural policy on the variation of local employment, taking also account of the treatment of the neighboring municipalities. More generally, the main goal of this paper is to propose statistical non parametric methods that are suitable to assess spillover effects of spatial policies at a micro level.

The notion of spillovers is related to the idea that the effect of a variable on another variable may spill over to other statistical units of the sample. Such an idea has a long tradition in economics and there exists a huge literature, theoretical as well empirical on such a relevant issue at all levels of aggregation. A typical field of study of spillover effects is economics and econometrics of innovation, research and development and productivity/economic growth (Griliches 1998; Ertur and Koch 2007; Ertur and Musolesi 2017; Charlot et al. 2015). Also note that spillovers may arise not only through geographical proximity and that alternative channels of spillovers can be effective, such as trade, foreign direct investment, bilateral technological proximity, patent citations between countries, language skills and genetic proximity (Coe and Helpman 1995; Potterie van Pottelsberghe and Lichtenberg 2001; Spolaore and Wacziarg 2009).

However, while the issue of policy spillover effects is extremely relevant for properly designing public policies, such a topic is still at an early stage of development and, to the best of our knowledge, only few studies exist (see, e.g. Angelucci and Di Maro 2015; Clarke 2017).

From a methodological viewpoint, a common practice with panel data, when the model contains individual random effects that may be correlated with the explanatory variables and in particular with the treatment variable, is to assume, for identification purposes, that the conditional independence assumption holds for the difference of the outcome after and before the beginning of the policy. At the same time, at a micro data level, it is often observed that the dependent variable, such as local employment in a municipality, does not vary along time, so that when considering the distribution of the individual differences over time we face a kind of zero inflated phenomenon that cannot be simply handled with a classical continuous response model or propensity score approaches.

When dealing with large samples, non parametric approaches allow modeling complex nonlinear relations, threshold effects and interactions and can be preferred to more rigid parametric statistical models (see e.g. Wood 2017 for a discussion). Furthermore, as being more flexible, non parametric models may be employed successfully to estimate treatment effects for particular configurations of the conditioning variables.

We consider in this work two recent non parametric techniques that are able to deal with the zero inflation phenomenon and that allow a relevant degree of flexibility

and permit to estimate heterogeneous policy effects. We first extend the approach developed in Cardot and Musolesi (2020) in order to take into account eventual spillover effects of the policy. This first approach consists in fitting two generalized additive models to estimate both the probability of no variation as well as the variation along time of the continuous part of the response. As far as the second approach is concerned, it is not unusual anymore to have to deal with large databases and there is a growing interest in considering modern machine learning tools for evaluating treatment effects in the literature (see for example Zhao et al. 2016; Belloni et al. 2017; Goller et al. 2019). Random forests (see Breiman 2001 for a seminal paper), which are built by aggregating regression trees fitted on subsamples of the initial sample, are particularly interesting in our micro data context since they can naturally deal with a zero inflated continuous response. Random forests are known to be highly efficient in terms of prediction in many situations (see Hastie et al. 2009 for a general presentation) and freely available libraries, such as `ranger` in R (see Wright and Ziegler 2017), allow for fast computation on large samples of high dimensional data. Recently, some asymptotic convergence results for random forests have been obtained by Scornet et al. (2015) whereas a guided tour on recent development of random forests is proposed in Biau and Scornet (2016).

Most of the policy evaluation studies considering random forests are based on propensity score matching (see for example Zhao et al. 2016; Belloni et al. 2017; Goller et al. 2019). The originality of our work is to deal with selection bias by non-parametrically estimating the counterfactual with random forests, taking into account many covariates that can be related with the potential outcome and the selection variables.

The paper is organized as follows. First, we introduce in Sect. 2 notations as well as the fundamental conditional independence assumption that ensures that the effects of the policy as well as the spill over effects can be estimated. Section 3 presents the two non parametric econometric models considered in this study to estimate counterfactuals expected responses, given the set of confounding variables. We illustrate in Sect. 4 our methodology on the evaluation of a local development policy based on a large sample of more than 25000 municipalities in France. We consider the example of four municipalities chosen by a clustering method based on k -medoids (see Kaufman and Rousseeuw 1990). These municipalities can be seen as a representative unit, in terms of being the most central municipality, of each corresponding cluster. This allows us to estimate localized effects of the policy for central statistical units and to avoid the non common support classical issue in treatment effect evaluation.

2 Conditional Average Treatment Effect, Identification and Model Specification

Let i denote a statistical unit (a municipality in our framework) which is assigned to some treatment. We denote by $Y_i^r(t)$ the potential employment level for municipality i at time t under treatment (incentive) r , for $r \in \{0, 1\}$, with the convention that $r = 0$ corresponds to no treatment. Time t is discrete, taking values in $t_0 < t_1 < \dots < t_m$. We assume that the incentives are allocated after t_0 and that they may produce an effect from period k , with $t_k > t_0$. All the counterfactuals are assumed to be equal before the treatment begins, that is to say $Y_i^1(t) = Y_i^0(t)$ for $t_0 \leq t < t_k$.

To account for spatial spillover effects of the policy we also introduce a binary indicator of treatment, denoted by W , taking the value 1 (some) when the considered unit has some neighboring municipalities that also receive the treatment and the value 0 (none) when all its neighboring municipalities do not receive any treatment. We introduce the counterfactual response $Y^{r,w}(t)$ which represents the response that would have been observed at time t under treatment $r \in \{0, 1\}$ and neighbors values $w \in \{0, 1\}$.

For each municipality $i \in \{1, \dots, n\}$, we denote by $X_i = (X_{i1}, \dots, X_{ip})$ a set of characteristics observed during the first period of time t_0 , which are the *initial conditions*. We denote by D_i , with $D_i \in \{0, 1\}$, the treatment status of municipality i , that is supposed to be a binary random variable.

Our aim is to estimate, for $t \geq t_k$, the expected treatment effect given $X = x$, according to the treatment status $(w_0, w_1) \in \{0, 1\} \times \{0, 1\}$, of the neighboring municipalities,

$$\tau(t, x, w_0, w_1) = \mathbb{E}[Y^{1,w_1}(t) - Y^{0,w_0}(t) | X = x] \quad (1)$$

based on the observation of $(Y_i^0(t_0), \dots, Y_i^0(t_{k-1}), Y_i^{D_i, W_i}(t_k), \dots, Y_i^{D_i, W_i}(t_m), X_i, D_i, W_i)$, for $i = 1, \dots, n$. Spillover effects of the policy can be captured by considering different configuration for w_0 and w_1 . For instance, if $w_0 = 1$ and $w_1 = 1$, $\tau(t, x, 1, 1)$ is the expected treatment effect at time t when some of the neighboring municipalities have been treated. If $w_0 = 0$ and $w_1 = 1$, $\tau(t, x, 0, 1)$ is the expected effect of the policy combined with the fact that some neighbors have received the treatment compared to no policy and no treated neighbors.

2.1 Identification Issues and Conditional Independence Assumption

In order to be able to identify and to estimate the conditional treatment effect (1) and taking advantage of the fact that we have panel data, we consider a *before-after* approach and thus assume that the conditional independence assumption holds for the difference of the outcome after and before the beginning of the policy,

$$Y^{r,w}(t) - Y^0(t_0) \perp\!\!\!\perp (D, W) \mid X, \quad r \in \{0, 1\}, \quad w \in \{0, 1\}, \quad t \in \{t_k, \dots, t_m\}. \quad (2)$$

The above condition is more general than the standard conditional independence assumption, unconfoundedness or selection on observables, which is typically employed with cross-sectional data. Indeed, since selection bias may not be completely eliminated after controlling for the observables, it is important to note that exploiting the longitudinal structure of the data may help to address the issue of selection on unobservables. The conditional independence assumption (2) holds for example when the unobservable part of the model contains correlated random (individual) time invariant effects, thus allowing for any kind of dependence between selection into the treatment and time-invariant individual characteristics. Since our aim is to estimate the treatment effect given $X = x$, we do not consider the propensity score as an interesting minimal conditioning variable as it is often advised when one is interested in estimating average effects (see Rosenbaum and Rubin 1983).

When the conditional independence assumption (2) holds, we can expand $\tau(t, x, w_0, w_1)$ as follows to obtain a difference-in-differences expansion, as discussed in Abadie (2005),

$$\tau(t, x, w_0, w_1) = \mathbb{E} [Y^{1,w_1}(t) - Y^0(t_0) \mid X = x] - \mathbb{E} [Y^{0,w_0}(t) - Y^0(t_0) \mid X = x] \quad (3)$$

and estimate each term at the right-hand side of (3) separately.

Note that (2) is a non parametric identification condition. However, adopting a parametric specification for difference-in-differences expansion (3) implies a specific parametric form and any deviation from such a form may invalidate the estimates (Abadie 2005). Theoretical results about identification indicate that the validity of the common trend assumption implicit in difference-in-differences models is functional form dependent (Lechner 2011, 2015). These reasons suggests that adopting a flexible non parametric model is useful for credible identification.

As far as the set of confounding variables X is concerned, we gather demographic, education and work’s qualification information aggregated at the municipality level. We also have at hand information on land use, obtained thanks to satellite images. These variables are indicated as relevant by the related literature on local employment growth (Carlino and Mills 1987). We consider pre-treatment covariates, to ensure that D (and eventually also W) causes X and Y causes X do not occur (Lee 2016). This is likely to be relevant in our economic context where it could be expected that the covariates prior the introduction of the policy, such as the share of qualified workers or the existing stock of infrastructure, cause both the inclusion in the program D , and the potential local employment Y . After the introduction of the policy, the level of such covariates is likely to be affected by their past values, by the treatments D and finally also by the response variable Y . In such a causal framework, pre-treatment covariates should be controlled for whereas post-treatment covariates should not (see Lee 2016; Lechner 2011). Another relevant variable included in the model which is worth mentioning is the initial level of employment. Including the initial outcome as a regressor implies assuming unconfoundedness given lagged outcome. This inclusion

avoids having an omitted variable bias which would be particularly relevant if the average outcome of the treated and control groups differs substantially at the first period (Imbens and Wooldridge 2009), as in this case.

2.2 Zero Inflation and Conditional Mixtures

When dealing with micro count data a non negligible fraction of the dependent variable Y may not vary at all over a period of time. This means that, for $t \geq t_k$,

$$\mathbb{P}[Y^{r,w}(t) - Y^0(t_0) = 0 \mid X = x] > 0,$$

inducing a zero inflation phenomenon. A conditional mixture model was introduced in Cardot and Musolesi (2020) to describe the probability law of $Y^r(t) - Y^0(t_0)$ with a mixture of a Dirac at 0 and a continuous distribution, with mixture weights depending on the treatment status D and the conditioning variables X . Taking account of the neighbor's treatment indicator, we slightly modify the mixture model studied in Cardot and Musolesi (2020) and we assume that, in distribution, for $W = w$ and given $X = x$, we have

$$Y^{r,w}(t) - Y^0(t_0) \mid X = x \sim \pi^r(t, x, w)\delta_0 + (1 - \pi^r(t, x, w))f^r(t, x, w) \quad (4)$$

where \sim denotes equality in distribution, $\pi^r(t, x, w) = \mathbb{P}[Y^{r,w}(t) - Y^0(t_0) = 0 \mid X = x]$ is the probability of no variation, δ_0 is the Dirac mass at zero and $f^r(t, x, w)$ is a continuous density that varies over time t .

Combining (3) and the conditional mixture assumption (4), we can write

$$\begin{aligned} \tau(t, x, w_0, w_1) &= \mathbb{E}\left[Y^{1,w_1}(t) - Y^0(t_0) \mid X = x, \Delta^{1,w_1}(t) \neq 0\right] \times \left(1 - \pi^1(t, x, w_1)\right) \\ &\quad - \mathbb{E}\left[Y^{0,w_0}(t) - Y^0(t_0) \mid X = x, \Delta^{0,w_0}(t) \neq 0\right] \times \left(1 - \pi^0(t, x, w_0)\right), \end{aligned} \quad (5)$$

where $\Delta^{r,w}(t) = Y^{r,w}(t) - Y^0(t_0)$, $r \in \{0, 1\}$. The decomposition given in (5) takes explicitly account of the zero inflation feature of the counterfactual outcome variations and allows for direct econometric estimation.

3 Econometric Modeling and Estimation Procedures

For $t \in \{t_1, \dots, t_m\}$ and $i = 1, \dots, n$, we denote by

$$\Delta_i^{D_i, W_i}(t) = Y_i^{D_i, W_i}(t) - Y_i^0(t_0) \quad (6)$$

the outcome variation for municipality i between t and t_0 , with the convention that $Y_i^{D_i, W_i}(t) = Y_i^0(t)$ before the treatment begins, that is to say when $t \in \{t_1, \dots, t_{k-1}\}$.

We consider two different approaches for estimating $\tau(t, x, w_0, w_1)$. The first one is directly based on decomposition (5) and relies on the use of additive and generalized additive models for estimating the conditional expectation for the continuous parts and the conditional probability of no variation for the discrete part of $\Delta_i^{D_i, W_i}(t)$ given X_i . The second one is more simple in some sense and is based on the decomposition (3) and the ability of random forests to naturally deal with the zero inflation phenomenon and a relatively large number p of potential covariates.

3.1 A Flexible Semi-parametric Modeling Approach Based on Additive Models and Conditional Mixtures

Additive and generalized additive models are generally assumed to be an interesting compromise between parametric models, which may not be flexible enough and may not let the data a chance to speak, and purely non parametric models based on kernel smoothers or splines which suffer from the curse of dimensionality and have very poor rates of convergence even if the number of covariates is moderate (see Stone 1985 for convergence properties of non parametric estimators for additive models and Wood (2017) for efficient implementations in R). We suppose that the following additive model holds for the continuous part, for $r \in \{0, 1\}$ and $w \in \{0, 1\}$,

$$\mathbb{E} [\Delta^{r,w}(t)|X = (x_1, \dots, x_p), \Delta^{r,w}(t) \neq 0] = \alpha_0^{r,w}(t) + \sum_{j=1}^p \alpha_j^{r,w}(x_j, t), \quad (7)$$

where, for each $t \in \{t_1, \dots, t_m\}$, $\alpha_0^{r,w}(t)$ is an unknown coefficient and $\alpha_j^{r,w}(x_j, t)$ are unknown smooth functions of x_j , $j = 1, \dots, p$. As far as the discrete part is concerned, we assume that, for $r \in \{0, 1\}$ and $w \in \{0, 1\}$,

$$\text{logit} (\mathbb{P} [\Delta^{r,w}(t) = 0|X = (x_1, \dots, x_p)]) = \beta_0^{r,w}(t) + \sum_{j=1}^p \beta_j^{r,w}(x_j, t), \quad (8)$$

where, for each $t \in \{t_1, \dots, t_m\}$, $\beta_0^{r,w}(t)$ is an unknown coefficient and $\beta_j^{r,w}(x_j, t)$ are unknown smooth functions of x_j , $j = 1, \dots, p$. Note that as usual with additive models (see Wood 2017 for example), identifiability constraints must be added to get a unique representation in (7) and (8).

The terms $\beta_j^{r,w}(x_j, t)$ and $\alpha_j^{r,w}(x_j, t)$ which should not depend on the values of r and w when $t \in \{t_1, \dots, t_{k-1}\}$ can be useful to perform pre-program tests (Heckman and Hotz 1989).

Note that a simple extension of (7) and (8) consists in introducing interactions between covariates instead of additive effects. For example, we could consider a model with a bivariate function $\alpha_{j,j'}^{r,w}(x_j, x_{j'}, t)$ to replace the additive terms

$\alpha_j^{r,w}(x_j, t) + \alpha_j^{r,w}(x_j, t)$. This permits to have a gain in flexibility at the expense of more difficult interpretations and slower rates of convergence.

3.1.1 Estimation with the `mgcv` Library in `R`

For estimation of the unknown parameters and regression functions, we split the initial sample into $t_m - t_0$ samples consisting in the n following observations, $(\Delta_i^{D_i, W_i}(t), X_i, D_i, W_i)_{i=1, \dots, n}$, where t belongs to $\{t_1, \dots, t_m\}$. The fact that the considered mixture is a mixture of a continuous variable and a discrete variable makes the computation rather simple compared to mixtures of continuous variables or mixtures of discrete variables (see McLachlan and Peel 2000). Indeed, as far as the continuous part is concerned, the probability of no variation is equal to zero and we can fit the two underlying distributions separately. The unknown smooth functions $\alpha_j^{r,w}(x_j, t)$ are expanded in spline basis and fitted on the subsamples satisfying $\Delta_i^{D_i, W_i}(t) \neq 0$ with the `bam` function of the `mgcv` library. For the discrete part of the distribution, we create the indicator variable $T_i^{D_i, W_i}(t) = 1$ when $\Delta_i^{D_i, W_i}(t) = 0$ and $T_i^{D_i, W_i}(t) = 0$ if $\Delta_i^{D_i, W_i}(t) \neq 0$. The unknown functions $\beta_j^{r,w}(x_j, t)$ are expanded in spline basis and fitted on the subsamples $(T_i^{D_i, W_i}(t), X_i, D_i, W_i), i = 1, \dots, n$, with the `bam` function with the `binomial` family and the usual `logit` link function.

In order to be able to deal with large datasets and to select effective values of the smoothing parameters in a reasonable time (less than 20s on a personal computer), each estimation is performed with the fast REML approach (see Wood 2017 for details).

Adding a “hat” on the estimated unknown parameters and functions, the estimated value of $\tau(t, x, w_0, w_1)$ is given by

$$\begin{aligned} \widehat{\tau}(t, (x_1, \dots, x_p), w_0, w_1) &= \left(\widehat{\alpha}_1^{1, w_1}(t) + \sum_{j=1}^p \widehat{\alpha}_j^{1, w_1}(x_j, t) \right) \left(1 - \widehat{\pi}^1(t, (x_1, \dots, x_p), w_1) \right) \\ &\quad - \left(\widehat{\alpha}_0^{0, w_0}(t) + \sum_{j=1}^p \widehat{\alpha}_j^{0, w_0}(x_j, t) \right) \left(1 - \widehat{\pi}^0(t, (x_1, \dots, x_p), w_0) \right) \end{aligned} \quad (9)$$

where $\text{logit}(\widehat{\pi}^r(t, (x_1, \dots, x_p), w)) = \widehat{\beta}_0^{r,w}(t) + \sum_{j=1}^p \widehat{\beta}_j^{r,w}(x_j, t)$.

3.2 Estimation of the Conditional Treatment Effect with Random Forests

Random Forests are now widely used in machine learning and data science when large samples with many covariates are available. They are non parametric ensemble techniques which consist in aggregating the prediction obtained by regression trees

(when the target variable is quantitative) or classification trees (when it is qualitative) built on subsamples of the initial sample. Each tree is grown sufficiently deep to have relatively small bias and the aggregation (averaging) procedure of the trees allows for variance reduction. A way to obtain a small variance consists in selecting randomly, at each node of the tree, a subset of the p available covariates so that the correlation between the different trees of the forest is as small as possible. The interested reader can find much more details on the intuition behind random forests and their statistical properties in Breiman (2001); Hastie et al. (2009); Biau and Scornet (2016).

One major advantage of random forests compared to other non parametric competing approaches is that they can naturally deal with qualitative and quantitative explanatory variables and perform variable selection at each node of the trees.

Another advantage compared to other classical non parametric estimation procedures such as regression splines or kernel smoothers is that random forests seem to be much less sensitive to the choice of the hyperparameters, which are the number of trees in the forest, the size of the terminal nodes and the number of selected variables at each node.

Random Forests allow for a direct estimation of complex functions of the initial condition under the treatment of the considered municipality and its neighboring municipalities. As for additive models, we split the initial sample into $t_m - t_0$ samples consisting in the n following observations, $(\Delta_i^{D_i, W_i}(t), X_i, D_i, W_i)_{i=1, \dots, n}$, where t belongs to $\{t_1, \dots, t_m\}$ and fit random forests on each subsample to get estimates of $\mathbb{E}[Y^{1, w_1}(t) - Y^0(t_0) | X = x]$ and $\mathbb{E}[Y^{0, w_0}(t) - Y^0(t_0) | X = x]$, denoted by $\hat{\mu}^{1, w_1}(t, x)$ and $\hat{\mu}^{0, w_0}(t, x)$. Estimation is performed with the `ranger` library in R (see Wright and Ziegler 2017). Then, the expected treatment effect, given $X = x$ is estimated by

$$\hat{\tau}(t, x, w_0, w_1) = \hat{\mu}^{1, w_1}(t, x) - \hat{\mu}^{0, w_0}(t, x). \quad (10)$$

Pseudo confidence intervals are built by using a non parametric bootstrap approach.

4 An Illustration on the Estimation of the Effect of Local Development Policies in France

4.1 Description of the Policy and Data

We illustrate the proposed methods by assessing possible spatial spillover effects of the ZRR (Zones de Revitalisation Rurale) program in France. Specifically designed to boost employment of rural areas, this geographically targeted program, which is based on tax exemptions for new hires to firms located in deprived areas, started the 1st September 1996 and covered the period 1996–2004. A noticeable feature of the program is that the selection of ZRR was clearly not random. A rather complex algorithm was used to determine the eligibility, according to some observable—demographic, economic and institutional—criteria. To be eligible to ZRR, a municipi-

Table 1 Treatment and neighbors treatment contingency table

		Treated neighbors	
		None	Some
ZRR	0	12737	5120
	1	282	7454

pality should be a part of a canton with population density lower than 31 inhabitants per square km (1990 Population Census).¹ The population or the labor force must also have diminished or the share of the agricultural labor employment must be at least twice the French average. Finally, to be included into the program, the municipality should belong to a pre-existing zoning scheme set up by the European Union, which is called TRDP (*Territoire Rural de Développement Prioritaire*). However, due to political tempering, it is also likely that, beyond such observed criteria, other sources of selection on unobservables could affect the process.

Ex ante, spillovers from ZRR may be either positive arising directly through a higher labor demand and/or indirectly from agglomeration economies or negative if some substitution effects occur, that is if geographical shifts in jobs from non-treated to treated areas occur. This is why it is interesting to evaluate both the magnitude and the sign of spatial spillover effects of ZRR.

We exploit the sample used in Cardot and Musolesi (2020). The municipalities, which correspond to the finest available spatial level, are the statistical units of the analysis and the dependent variable is the number of employees. The data were obtained over a period of ten years, 1993–2002. As explanatory variables, we dispose of ZRR zoning during the period. Other explanatory variables come from the 1990 CENSUS and from satellite images that were also taken in 1990 (a brief description of the variables is given in the Appendix). The sample is composed of $n = 25593$ municipalities.

The identification of spillovers is an intricate empirical matter, requiring the definition of the neighborhood and the choice of an adequate channel of transmission. We focus here on purely geographic spillovers and adopt a notion of neighborhood by considering the spillovers arising from the municipalities sharing a common border. Among the 25593 municipalities under study, 13019 municipalities have all of their neighboring municipalities that do not receive the ZRR incentives while the remaining 12574 municipalities have all or some of their neighboring municipalities that are under ZRR (see Table 1).

We can finally note that our identification strategy is affected by the fact that the ZRR program was introduced in 1996 and covered the entire period under investigation and more precisely it covered the period 1996–2004. During this period, firms in ZRR areas benefit from a tax exemption on the new hires during the first 12 months of their contracts. For these reasons our treatment variable does not vary between 1996

¹A canton with a population density less than 5 inhabitants per square km is automatically labelled as ZRR without any other requirement.

to 2002. Assuming a time invariant treatment status within a programming period is a common practice in the literature (Behaghel et al. 2015). We also use the years 1994 and 1995 to conduct ‘pre-program’ tests along the lines depicted by Heckman and Hotz (1989).

4.2 Estimation Results and Counterfactual Analysis at the Municipality Level

We provide a counterfactual analysis, both for the semi-parametric modeling approach (GAM) and random forests (RF), for few representative municipalities for which the evolution of the potential outcomes are estimated and compared under the different possible treatments.

4.2.1 Variable Selection

A backward selection procedure of the variables has been performed to fit more effective additive and generalized additive models. Since the main interest lies in assessing possible heterogeneous treatment effects and heterogeneous spillovers, we examine how such effects may interact with some economic or demographic characteristics of the municipalities. For that purpose, the model selection procedure led us to retain only significant interactions with the initial level of employment (variable `SIZE`) of the municipality and with its population density (variable `DENSITY`). A more detailed description of the fitted additive and generalized additive models is given in the Appendix

4.2.2 Selection of the Municipalities Under Study

We focus the estimation of policy and spillover effects on some municipalities that have been chosen with a clustering partition around medoids procedure (see Kaufman and Rousseeuw 1990 for a description of that particular clustering procedure). The distance between the municipalities has been computed according to the variables `DENSITY`, `INCOME` and `SIZE` (in reduced form) with the Euclidean distance. We have retained four clusters whose centers are the representative, that is to say most central, municipalities within each group. Descriptive statistics are given in Table 2 and mean estimated values of the counterfactuals as well as Pointwise 90% confidence intervals (obtained by non parametric bootstrap) are drawn for the four selected municipalities in Figs. 1, 2, 3 and 4.

Table 2 Descriptive statistics for the municipalities selected for the counterfactual analysis

Municipality	DENSITY	SIZE	INCOME	OLD	FACT	BTS	AGRIH	URB
1	218.85	105	5772	0.11	0.19	0.016	0.08	0.23
2	61.26	48	4324	0.30	0.06	0.037	0.19	0.032
3	41.87	25	6300	0.20	0.13	0.038	0.03	0.028
4	22.74	10	3724	0.14	0.16	0.007	0.22	0.015

4.2.3 Assessing the Econometric Modeling with “Pre-program” Tests

Pre-program tests are based on the idea that a valid estimator would correctly adjust for differences in pre-program outcomes between future participants and non-participants, otherwise the estimator is rejected. The ‘pre-program’ test is generally implemented by considering $t < t_k$, where t_k is here the time of introduction of the policy, and by testing the significance of the treatment effect. If such an effect is significantly different from 0 then the underlying model fails to pass the test. However, even if the logic is compelling, if a shock or an anticipation effect close to the time of the treatment affects only one group but not the other, the results from such a test are potentially misleading. This problem has also been summarized under the heading “fallacy of alignment” (Heckman et al. 1999). In our case, treated firms could (shortly) postpone hiring in order to obtain the fiscal incentives, so that using quite longer lags can be useful in order to obtain an effective test and avoiding to overestimate the treatment effect. Accordingly, use all the available information in the data and use the most distant data before the introduction of the policy to set t_0 and propose two tests by setting $(t_0 = 1993, t = 1994)$, $(t_0 = 1993, t = 1995)$. Note that the year 1996 coincides with the introduction of the policy and some effects may occur. The results indicate that when performing the counterfactual estimation with the proposed nonparametric approaches, for $t = 1994, 1995$, a very good alignment is obtained. Moreover, non reported results indicate that adopting a flexible nonparametric approach instead of considering a linear specification greatly improves the alignment. This is an important result supporting the use of flexible models instead of a linear one not only to improve results’ interpretation but also in terms of credible identification.

4.2.4 Policy and Spillover Effects: Estimation Results

We now present to the main estimation results. The first selected municipality is an extremely dense and urbanized municipality, with values of DENSITY (population density) and URB (urban surface/total surface) greater than the 95th percentile. It is also very rich in terms of INCOME (per capita income) and big in terms of SIZE (initial level of employment) with values of these variables around the 80th percentile. As seen in Fig. 1, GAM and RF provide an almost identical positive estimation of the evolution of employment in the absence of a treatment for that municipality and

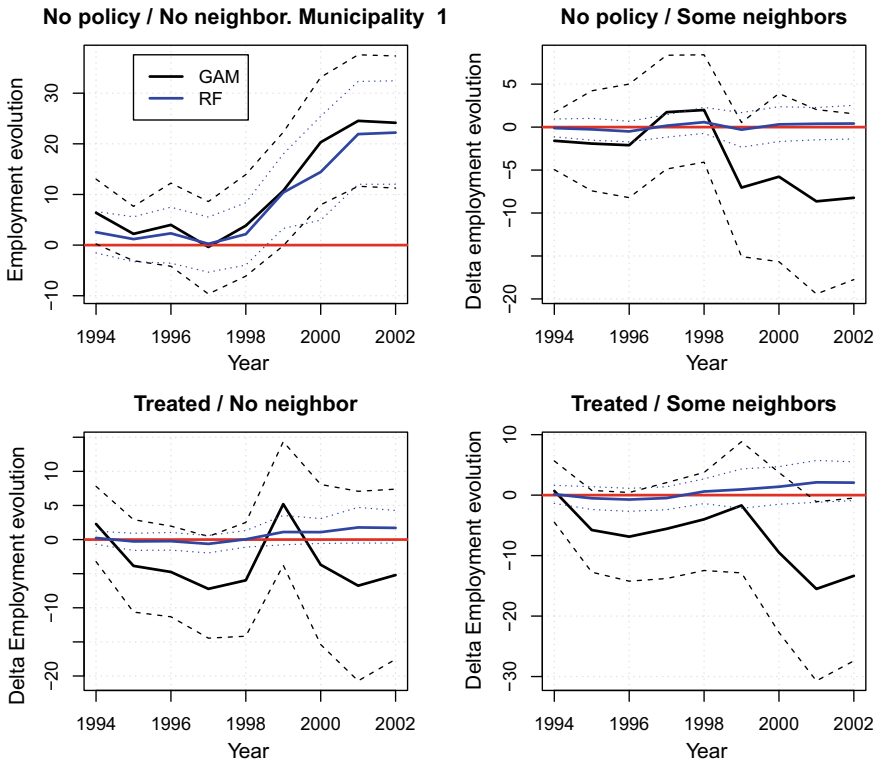


Fig. 1 Counterfactual analysis for municipality 1. Left/top: Estimation of the employment evolution $\hat{\mu}^{0,0}(t, x)$ (see Eq. 10) under no treatment and no treated neighbors, with pointwise 90% bootstrap confidence intervals. Right/top: estimation of the neighbors treatment effect, $\hat{\mu}^{0,1}(t) - \hat{\mu}^{0,0}(t)$, under no policy. Left/bottom: estimation of the treatment effect, $\hat{\mu}^{1,0}(t) - \hat{\mu}^{0,0}(t)$, when no neighbors receive the treatment. Right/bottom: estimation of the treatment effect, $\hat{\mu}^{1,1}(t) - \hat{\mu}^{0,0}(t)$, when some neighbors receive the treatment

in the neighboring municipalities. Moreover, according to both approaches, ZRR would have no significant effect on the evolution of employment for the considered period and, more specifically, policy spillover effects are never significant. This is not surprising given the design of ZRR.

The second municipality is rather dense, urbanized and big, with values of *DENSITY*, *URB* and *SIZE* about the 75th percentile of our sample. The value of *INCOME* is close to the median. For this municipality, see Fig. 2, ZRR alone produces a small positive effect having an inverted U time pattern with the peak reached for $t = 1999$ for both GAM and RF, the latter providing a more gradual decreasing pattern, compared to the former. Very importantly, spatial spillovers matters. Spatial spillover are indeed positive and provide an additional effect to ZRR. This result is obtained using both GAM and RF, despite for the former the additional effect is higher in magnitude.

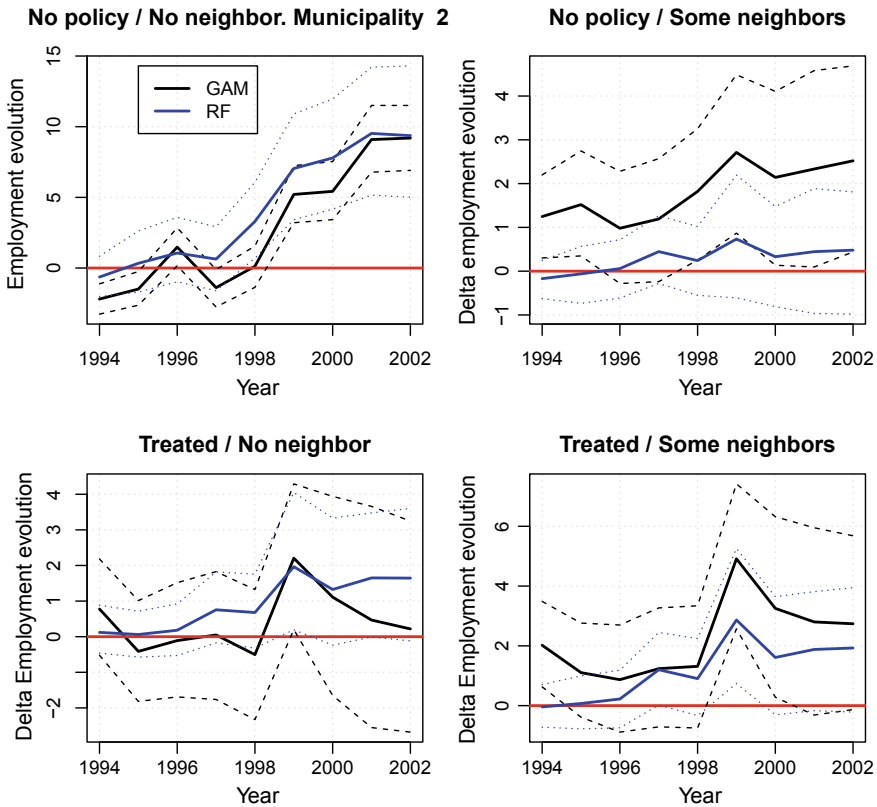


Fig. 2 Counterfactual analysis for municipality 2. Left/top: Estimation of the employment evolution $\hat{\mu}^{0,0}(t, x)$ (see Eq. 10) under no treatment and no treated neighbors, with pointwise 90% bootstrap confidence intervals. Right/top: estimation of the neighbors treatment effect, $\hat{\mu}^{0,1}(t) - \hat{\mu}^{0,0}(t)$, under no policy. Left/bottom: estimation of the treatment effect, $\hat{\mu}^{1,0}(t) - \hat{\mu}^{0,0}(t)$, when no neighbors receive the treatment. Right/bottom: estimation of the treatment effect, $\hat{\mu}^{1,1}(t) - \hat{\mu}^{0,0}(t)$, when some neighbors receive the treatment

This is an important result and it will be of great interest in comparison with the remaining selected municipalities.

The third municipality is quite close to the median values in terms of *DENSITY* and *SIZE*. We note in Fig. 3 that the direct effect of ZRR (no treated neighboring municipalities) is positive and significant for some time periods t for both GAM and RF, despite the estimated time pattern of the effect is somehow different: GAM predicts a rather abrupt but transitory effect, slowly decreasing after 1999, while according to RF, ZRR appears to have a more persistent effect over time on employment. Very interestingly, both approaches suggest the existence of positive spillover effects. According to both approaches, this positive effect increases smoothly over time. Spatial spillover effects matter and are relatively high in magnitude.

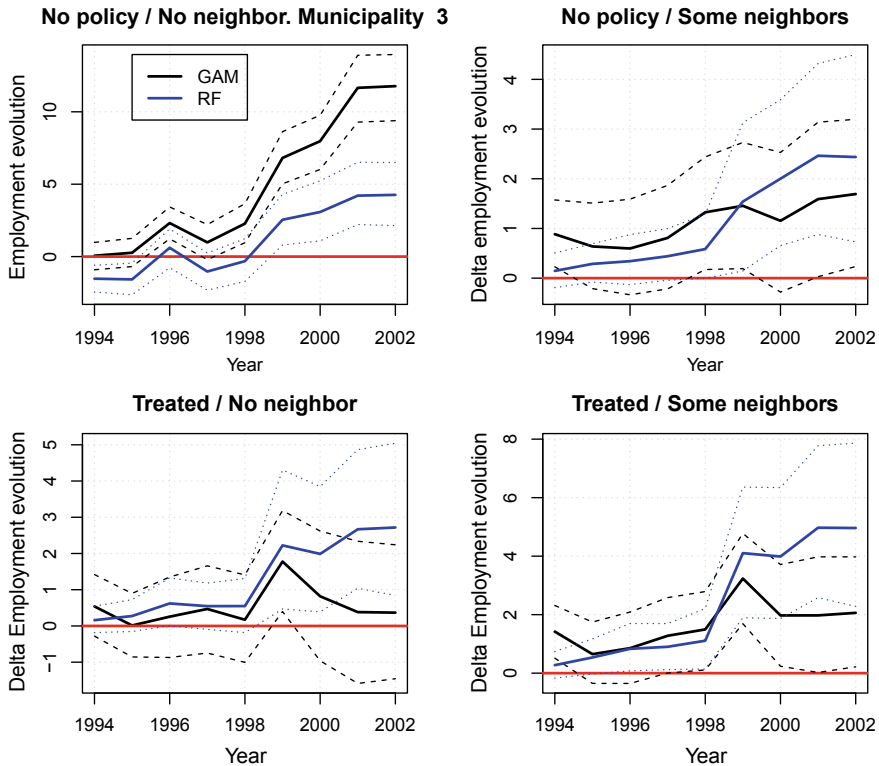


Fig. 3 Counterfactual analysis for municipality 3. Left/top: Estimation of the employment evolution $\hat{\mu}^{0,0}(t, x)$ (see Eq. 10) under no treatment and no treated neighbors, with pointwise 90% bootstrap confidence intervals. Right/top: estimation of the neighbors treatment effect, $\hat{\mu}^{0,1}(t) - \hat{\mu}^{0,0}(t)$, under no policy. Left/bottom: estimation of the treatment effect, $\hat{\mu}^{1,0}(t) - \hat{\mu}^{0,0}(t)$, when no neighbors receive the treatment. Right/bottom: estimation of the treatment effect, $\hat{\mu}^{1,1}(t) - \hat{\mu}^{0,0}(t)$, when some neighbors receive the treatment

Finally, we focus attention on the fourth municipality, which is small, poor and low density municipality and for that reasons it is of particular interest here as being a typical distressed municipality that needs public subsidies to boost its socio-economic development. As seen in Fig. 4, ZRR alone appears to have an abrupt but transitory effect on employment according to both GAM and RF, although it is low in magnitude and much lower than the estimated effect for the second and third municipalities, which are bigger and more dense. Spatial spillovers produce a very small additional positive effect in comparison with the direct ZRR effect.

In summary the proposed flexible non parametric approaches provide a robust picture indicating the existence of positive spatial spillover effects for municipalities with certain economic and demographic characteristics. Two main results emerge. The first one is that for small, poor and low density municipalities, those being the primary target of geographically targeted policies, both direct and spatially-mediated

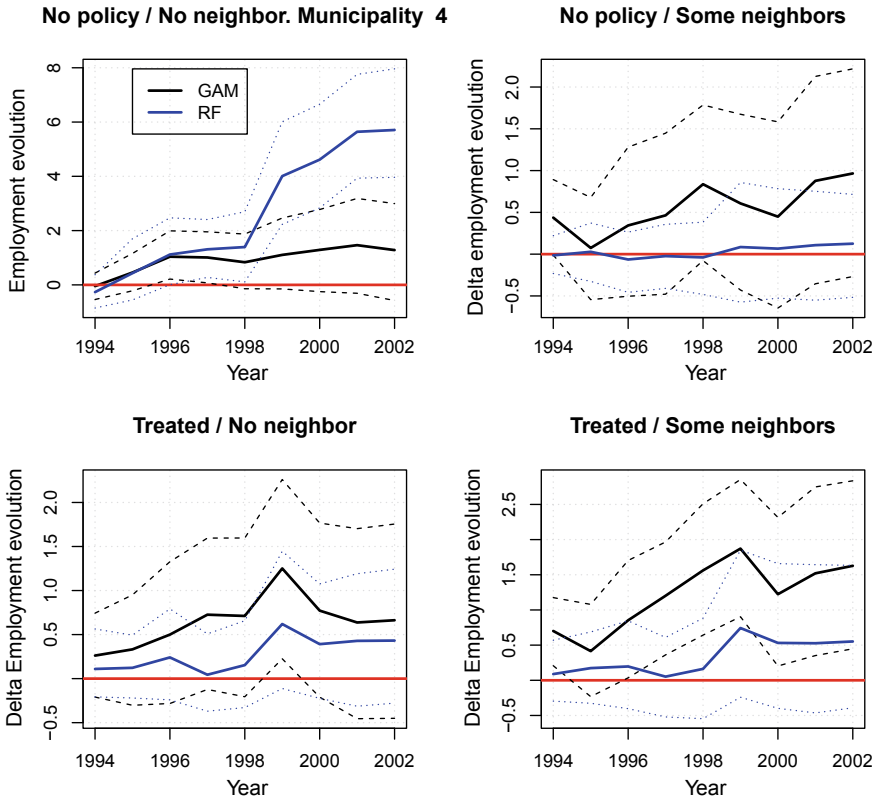


Fig. 4 Counterfactual analysis for municipality 4. Left/top: Estimation of the employment evolution $\hat{\mu}^{0,0}(t, x)$ (see Eq. 10) under no treatment and no treated neighbors, with pointwise 90% bootstrap confidence intervals. Right/top: estimation of the neighbors treatment effect, $\hat{\mu}^{0,1}(t) - \hat{\mu}^{0,0}(t)$, under no policy. Left/bottom: estimation of the treatment effect, $\hat{\mu}^{1,0}(t) - \hat{\mu}^{0,0}(t)$, when no neighbors receive the treatment. Right/bottom: estimation of the treatment effect, $\hat{\mu}^{1,1}(t) - \hat{\mu}^{0,0}(t)$, when some neighbors receive the treatment

spillover effects are very low in magnitude and are effective only over a very short time period. For bigger and more dense/urbanized areas, both effects are higher in magnitude and more persistent over time. This result is consistent with the idea that agglomeration externalities and an adequate size of the local market are essential in order to make such fiscal incentives and their spillover effects effective. The second relevant result is that nonlinear interaction effects appear to be relevant: direct policy effect and indirect effects arising from spatial spillovers vary non-linearly with some demographic and economic characteristics, such as the size and the density of the municipality. Finally note that when considering a standard parametric linear model, we do not find any evidence of significant spillover effects, thus evidencing the importance of adopting flexible models to highlight somehow complex nonlinear

spillover effects, which otherwise will be missed when using a standard parametric model. Detailed results are available upon request.

5 Conclusion

We have considered in this paper two non parametric approaches to assess spillover effects of a spatially targeted policy that was introduced in France to boost rural development. Both approaches are able to handle the zero inflated phenomenon that may arise at a micro level and that cannot be dealt with properly with classical continuous distribution models.

These two approaches provide a more credible identification and more focused analysis in comparison with standard parametric models. They indeed provide a very good alignment when conducting placebo tests and suggest the existence of interesting patterns of temporal spatially-mediated spillover effects of the policy with relevant nonlinear effects. Policy spillovers matter, although they are generally not high in magnitude, for municipalities with some specific demographic and economic characteristics.

Acknowledgements Calculations were performed using HPC resources from PSIUN CCUB (Centre de Calcul de l'Université de Bourgogne, France). The Institut de Mathématiques de Bourgogne is supported by the EIPHI Graduate School (contract ANR-17-EURE-0002).

Appendix

We present here the variables that were considered to adjust the generalized additive models and random forests as well as functions used in R to perform estimation. A detailed description of the definition of these variables as well as some descriptive statistics can be found in the Appendix of Cardot and Musolesi (2020).

Description of the Variables

The dependent variable Y_{it} corresponds to the number of employees at time t for municipality i . The socio-economic and demographic variables come from standard INSEE sources while the variables measuring land use have been obtained from the "Corine Land Cover" base. For each municipality, we have

- $SIZE \equiv Y_{t_0}$ is the initial outcome, i.e. the level of employment at t_0 , with t_0 equals to 1993.
- $DENSITY \equiv (\text{total population}) / (\text{total surface in terms of } km^2)$;
- $INCOME \equiv (\text{net taxable income}) / (\text{total population})$;

- OLD \equiv (population over 65) / (total population) ;
- FACT \equiv (number offactory workers) / (total population);
- EXE \equiv (number ofexecutive workers) / (total population);
- FARM \equiv (number of farmers) / (total population);
- UNIV \equiv $\frac{\text{(number of people with a master level degree called "Maîtrise universitaire")}}{\text{(total population)}}$;
- BTS \equiv $\frac{\text{(number of people with a technicaldegree called "Brevet de Technicien Supérieur")}}{\text{(total population)}}$;
- NOEDU \equiv (number of people without adegree) / (total population);
- AGRI \equiv (farmlandsurface) / (total surface);
- CULT \equiv (cultivated landsurface) / (total surface);
- URB \equiv (urban surface) / (total surface);
- IND \equiv (industrial surface) / (total surface);
- ARA \equiv (arable surface) / (total surface);
- GRA \equiv (grassland surface) / (total surface);

where the total surface and the total population should be understood within the considered municipality.

Fitting the Statistical Models with R

For the random forests we have introduced 13 important demographic and economic variables, plus the indicator of treatment of the neighbors (w_{zrr}) to fit the variation of employment, separately for the municipalities that were treated and those which were not treated,

```
ranger(Delta_Employment ~
wzrr+DENSITY+UNIV+SIZE+OLD+INCOME+FARM+EXE+FACT+BTS+CULT+AGRIH+URB+IND)
```

Putting the two subsamples together, we can also compute the importance of each variable for different moment of time. It appears that w_{zrr} and z_{rr} are the least important variables. This agrees with our conclusions that both the policy and the spillover effects are small or non significant, depending on the economic and demographic characteristics of the municipality.

For the estimation approaches based on mixtures and additive models, we used the `bam` function of the `mgcv` library to perform faster estimation. Since the main interest lies in assessing possible heterogeneous treatment effects and heterogeneous spillovers, we examine how such effects may interact with some economic or demographic characteristics of the municipalities. We first selected the variables to introduce into the two regression functions below by adopting a backward algorithm and then follow again a general-to-specific procedure to select significant interactions. This procedure led us to consider the following model for the continuous part of the response,

```
bam(Continuous_Delta_Employment ~
  ZRR+wzrr+s(SIZE,by=wzrr)+s(DENSITY,by=wzrr)+s(SIZE,by=ZRR)+s(DENSITY,by=ZRR)
  +s(OLD)+s(INCOME)+s(FACT)+s(BTS)+s(CULT)+s(AGRIH)+s(URB)+s(IND), method="fREML")
```

The probability of no variation is fitted with a generalized additive model and the `logit` link function,

```
bam(Zero_Delta_Employment ~
  ZRR+wzrr+s(SIZE,by=wzrr)+s(DENSITY,by=wzrr)+s(SIZE,by=ZRR)+s(DENSITY,by=ZRR)
  +s(UNIV)+s(INCOME)+s(FACT)+s(EXE)+s(FARM)+s(BTS)+s(NOEDU)+s(URB)+s(IND)+s(ARA)
  +s(GRA), method="fREML", family=binomial(link = "logit"))
```

References

- Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72, 1–19.
- Angelucci, M., & Di Maro, V. (2015). Program evaluation and spillover effects. *The World Bank*, 1–19.
- Behaghel, L., Lorenceau, A., & Quantin, S. (2015). Replacing churches and mason lodges? Tax exemptions and rural development. *Journal of Public Economics*, 1–15.
- Belloni, A., Chernozhukov, V., Fernández-Val, I., & Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85, 233–298.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, 25, 197–227.
- Cardot, H., & Musolesi, A. (2020). Modeling temporal treatment effects with zero inflated semi-parametric regression models: The case of local development policies in France. *Econometric Reviews*, 39, 135–157.
- Carlino, G. A., & Mills, E. S. (1987). The determinants of county growth. *Journal of Regional Science*, 27, 39–54.
- Clarke, D. (2017). Estimating difference-in-differences in the presence of spillovers. MPRA Paper 81604, University Library of Munich, Germany.
- Charlot, S., Crescenzi, R., & Musolesi, A. (2015). Econometric modelling of the regional knowledge production function in Europe. *Journal of Economic Geography*, 15, 1227–1259.
- Coe, D., & Helpman, E. (1995). International RD spillovers. *European Economic Review*, 39, 859–887.
- Efron, B., & Tibshirani, R.J. (1993). An introduction to the bootstrap (Vol. 57). Monographs on statistics and applied probability. New York: Chapman and Hall.
- Ertur, C., & Koch, W. (2007). Growth, technological interdependence and spatial externalities: Theory and evidence. *Journal of Applied Econometrics*, 22, 1033–1062.
- Ertur, C., & Musolesi, A. (2017). Weak and strong cross-sectional dependence: A panel data analysis of international technology diffusion. *Journal of Applied Econometrics*, 32, 477–503.
- Frölich, M. (2004). Programme evaluation with multiple treatments. *Journal of Economic Surveys*, 18, 181–224.
- Goller, D., Lechner, M., Andreas Moczall, A., & Wolff, J. (2019). Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. Discussion Paper no. 2019–10, Department of Economics, University of St. Gallen.
- Griliches, Z. (1998). The search for RD spillovers. In *RD and productivity. The econometric evidence* (pp. 251–268). National Bureau of Economic Research Inc.
- Härdle, W., Huet, S., Mammen, E., & Sperlich, S. (2004). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory*, 20, 265–300.

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning - data mining, inference, and prediction* (2nd ed.). New York: Springer.
- Heckman, J., & Hotz, V. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association*, *84*, 862–874.
- Heckman, J., Lalonde, R., & Smith, J. (1999). The economics and econometrics of active labor market programs. *Handbook of Labor Economics*, *3*, 1865–2097.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley Inc.
- Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*, 5–86.
- Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, *4*(3), 165–224.
- Lechner, M. (2015). Treatment effects and panel data. In B. Baltagi (ed.), *The Oxford handbook of panel data*. Oxford: Oxford University Press.
- Lee, M. J. (2016). *Matching, regression discontinuity, difference in differences, and beyond*. New York: Oxford University Press.
- Potterie van Pottelsberghe, B., Lichtenberg, F. (2001). Does foreign direct investment transfer technology across borders? *Review of Economics and Statistics*, *83*, 490–497.
- McLachlan, G. J., & Peel, D. (2000). *Finite Mixture Models*. Wiley series in probability and statistics. New York: Wiley.
- R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
- Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*, 1716–1741.
- Spolaore, E., & Wacziarg, R. (2009). The diffusion of development. *The Quarterly Journal of Economics*, *124*, 469–529.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, *13*, 689–705.
- Wood, S. N. (2017). *Generalized additive models: an introduction with R* (2nd ed.). Texts in statistical science series. Boca Raton: Chapman & Hall/CRC.
- Wooldridge, J.M. (2005). Fixed-effects and related estimators for correlated random coefficient and treatment-effect panel data models. *The Review of Economics and Statistics*, *87*, 395–390.
- Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*, 1–17.
- Zhao, P., Su, X., Ge, T., & Fan, J. (2016). Propensity score and proximity matching using random forest. *Contemporary Clinical Trials*, *47*, 85–92.

Spatial Autocorrelation in Econometric Land Use Models: An Overview



Raja Chakir and Julie Le Gallo

Abstract This chapter provides an overview of the literature on econometric land use models including spatial autocorrelation. These models are useful to analyze the determinants of land use changes and to study their implications for the environment (carbon stocks, water quality, biodiversity, ecosystem services). Recent methodological advances in spatial econometrics have improved the quality of econometric models allowing them to identify more precisely the determinants of land use changes and make more accurate land use predictions. We review the current state of the literature on studies which account explicitly for spatial autocorrelation in econometric land use models or in the environmental impacts of land use.

1 Introduction

Land use plays a vital role in many major societal issues: food security (Verburg et al. 2013), preservation of biodiversity and ecosystem services (Foley 2005), climate change mitigation (Lal 2004) and the achievement of many Sustainable Development Goals (Gao and Bryan 2017). Land use choices are the result of complex decision-making processes related to the local and global biophysical and socioeconomic drivers. The researcher faces two central and related questions: “what drives land use change?” and “what are the (environmental and socioeconomic) impacts of land use change on stakeholders and the whole society?”. The answers to these questions are crucial for the design of public policies related to how to feed the growing world population and avoid unwanted land use effects on the environment.

R. Chakir (✉)

Université Paris-Saclay, INRAE, AgroParisTech, Economie Publique, 78850 Thiverval-Grignon, France

e-mail: raja.chakir@inrae.fr

J. Le Gallo

CESAER UMR1041, Agrosup Dijon, INRAE, Université de Bourgogne Franche-Comté, 21000 Dijon, France

e-mail: julie.le-gallo@agrosupdijon.fr

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_18

339

Various disciplines (economics, statistics, geography, land use science) have developed a range of empirical land use modeling approaches, using either aggregate or individual data. However, most of this work pays little attention to spatial autocorrelation (SA) in modeling land use although spatial interdependence is prevalent in all economic decisions in general and in land use decisions in particular. As a result, “standard” statistical and econometric methods, which assume independent observations, are inappropriate. More generally, taking account of the spatial dimension in econometric models involves two effects: spatial heterogeneity and SA. Spatial heterogeneity is the spatial differentiation of variables and behaviors in space and usually does not require specific econometric methods. Switching models, semi-parametric modeling of coordinates or clustered robust inference can handle this effect appropriately. Conversely, SA refers to the lack of independence among geographic observations. It measures the degree of similarity between an attribute in one location and the same attribute in neighboring locations (Anselin 1988). Unlike temporal autocorrelation, SA is multidimensional requiring a specialized set of techniques, which are not simple extensions of two-dimensional time series methods. In this chapter, we focus on SA in econometric land use models.

There is a growing body of work on econometric modeling of land use. These studies address the determinants of land use and land use change and their impacts on water quality (Bockstael 1996), deforestation (Chomitz and Gray 1996), carbon sequestration costs (Lubowski et al. 2006), and habitat fragmentation (Lewis and Plantinga 2007). Before the 1990s, econometric land use studies that explicitly introduced SA of observations were relatively rare as the presence of SA makes discrete choice models analytically intractable and requires use of computationally expensive Bayesian techniques or simulation estimation methods (Fleming 2004). Thus, most land use studies and especially those based on individual data avoid thorough treatment of spatial effects or use ad hoc procedures aimed at reducing the negative consequences of ignoring them.¹

Although land use studies taking explicit account of SA have increased (Brady and Irwin 2011), they remain relatively scarce (Ay et al. 2017; Chakir and Le Gallo 2013; Li et al. 2013; Sidharthan and Bhat 2012; Ferdous and Bhat 2012; Chakir and Parent 2009). Most econometric land use models in papers published in high quality journals still tend either to ignore SA, or use ad hoc methods to deal with it (e.g., Irwin et al. 2003; Carrion-Flores and Irwin 2004; Lubowski et al. 2008; Fezzi and Bateman 2011). This is because SA raises several issues related to econometric estimation, hypothesis testing, and prediction—especially in the case of discrete choice models (Billé and Arbia 2019).

Then, the aim of this chapter is to present the state of the art in the literature on econometric land use models and to show how methodological developments in spatial econometrics have been introduced into these models. We point out that this is not an exhaustive review; rather the objective is to highlight the main contributions to econometric land use models and their methodological advances. Our literature reviews depart from those provided by Brady and Irwin (2011), which summarize

¹Ignoring spatial effects can result in biased and/or inefficient parameter estimates or assessment of statistical significance (Anselin 1988).

the econometric challenges of spatial models in land use and hedonic model context, Plantinga (2015), who focuses on methods for integrating economic land use and biophysical models and Chakir (2015), who reviews methodological developments in spatial econometrics that have been introduced into land use models. The main goal of this literature review is to summarize the studies which include SA explicitly in land use models or in models of the environmental impacts of land use.²

The remainder of the chapter is organized as follows. First, we provide some general considerations related to the econometric modeling of land use (Sect. 2). Then we focus, respectively, on linear (Sect. 3) and discrete choice econometric land use models (Sect. 4) models. Section 5 shows how SA enhances models that focus on the impact of land use on various environmental outcomes. Section 6 concludes and highlights some directions for further research.

2 Econometric Land Use Models

Most econometric land use studies are based on the classical theory which considers that land use activities are chosen to maximize land rents and that rents vary with land characteristics, in particular soil fertility (Ricardo 1817) and location (von Thunen 1875). Yet, other factors might influence land use decisions for a given land parcel: socioeconomic factors (input and output prices) and policy variables (taxes and subsidies). The extent and significance of these determinants are analyzed in two broad categories of models: aggregate land use models which use aggregate (county level, state level, etc.) data, and individual land use models which are based on parcel-level or sample plot data. Table 1 presents a summary of some papers that provide econometric modeling of land use employing both aggregate and individual data.

Aggregate and individual land use models are complementary and provide different insights into the determinants of land use and land use changes, and their environmental effects. The choice between an aggregate and individual land use model often depends on data availability and the objective of the study. If the objective is to make land use predictions at the scale of one or a group of countries (such as European countries), an aggregate data model is required. If the objective is to study the effects of land use on biodiversity or water quality, a model based on individual data is more relevant. Both approaches have drawbacks.

On the one hand, aggregate data limits the capacity to explain the effects of heterogeneous physical characteristics such as soil quality on land use choices. Because the data are aggregated to units such as the county, intra-county variations in soil quality are ignored. Moreover, while aggregate data can be useful to study global issues (changes in land use shares within a region), the results are of limited use for policy making related to the spatial organization of land use in a region, or local issues related to biodiversity, water quality, or urbanization.

²We did a literature search for articles adopting an explicit spatial econometric approach to land use issues. Then, among these articles, we chose those that we considered the most important either from a methodological point of view or from the point of view of the environmental impacts of land uses.

Table 1 Example of econometric individual and aggregate, spatial and aspatial land use studies. (AER: American Economic Review, AJAE: American Journal of Agricultural Economics, ARER: Agriculture and Resource Econ. Review, EE: Ecological economics, GA: Geographical Analysis, FS: Forest Science, JARE: J. of Agri and Resource Econ. JEEM: Journal of Environmental Economics and Management, JGS: Journal of Geographical Systems, JRS: Journal of Regional Science, LE: Land Economics, LUP: Land Use Policy, PIRS: Papers in regional Science, RSUE: Regional Science and Urban Economics, SEA: Spatial Economic Analysis, WBER: World Bank Economic Review)

Paper	Land use categories	Model	Spatial	Journal
Aggregate land use share studies				
Aliğ (1986)	Crops, 3 types de forest, pasture and urban	Land-use share	No	FS
Lichtenberg (1989)	7 crops	Land-use share	No	AJAE
Stavins and Jaffe (1990)	Crops, forest	Land-use share	No	AER
Wu and Segerson (1995)	6 cultures	Land-use share	No	AJAE
Plantinga (1996)	Agriculture to forest	Land-use share	No	AJAE
Plantinga et al. (1999)	Agriculture, forest and urban/other land use	Land-use share	No	AJAE
Hardie and Parks (1997)	Agriculture, forest, urban/other use	Land-use share	No	AJAE
Plantinga and Ahn (2002)	Crops, forest	Land-use share	No	JARE
Chakir and Le Gallo (2013)	Agriculture, forest, urban and other use	Land-use share	Yes	EE
Marcos-Martinez et al. (2017)	19 land use categories	Land-use share	Yes	LUP
Chakir and Lungarska (2017)	Agriculture, forest, urban and other use	Land-use share	Yes	SEA
Marcos-Martinez et al. (2017)	Extensive grazing, pastures, cereals, annuals, perennials	Land-use share	Yes	LUP
Amin et al. (2019)	Deforestation	Deforestation area	Yes	JEEM

(continued)

Table 1 (continued)

Paper	Land use categories	Model	Spatial	Journal
Individual discrete choice studies				
McMillen (1989)	Farm, residential	Multinomial logit	No	LE
Bockstael (1996)	Urbanization	Probit	No	AJAE
Chomitz and Gray (1996)	Deforestation	Multinomial logit	No	WBER
Claassen and Tegene (1999)	Culture, pasture	Probit	No	ARER
Carrion-Flores and Irwin (2004)	Urbanization	Probit	No	AJAE
Lubowski et al. (2006)	Crops, pasture, forest, urban, <i>range</i> and CRP	Nested logit	No	JEEM
Chakir and Parent (2009)	Agriculture, forest, urban and other uses	Multinomial probit	Yes	PIRS
Wang and Kockelman (2009)	4 levels of urbanization	Ordered probit	Yes	PIRS
Ferdous and Bhat (2012)	4 levels of urbanization	Ordered probit	Yes	JGS
Sidharthan and Bhat (2012)	Urban, commercial, industrial and non-developed	Multinomial probit	Yes	GA
Li et al. (2013)	Farm, forest, grass, water, urban, unused	Multinomial probit	Yes	LE
Bhat et al. (2015)	Commercial, industrial, residential, underdeveloped	Multiple discrete-continuous probit	Yes	JRS
Carrión-Flores et al. (2018)	Commercial, industrial, residential, parks, agriculture	Multinomial logit	Yes	RSUE

On the other hand, one of the frequent difficulties related to modeling land use at the individual level is the lack of “good” explanatory variables or their scale incompatibilities. Although geophysical explanatory variables such as slope, altitude and soil quality are increasingly available at very fine resolution, economic variables

(rents, conversion costs, and prices) are either not available or observable only at aggregate scales. To compensate for this lack of data, empirical models often use proxies for rents at more or less aggregated scales. Another difficulty of individual-level land use models is related to the complexity involved in estimating discrete choice models in the multinomial case. This difficulty is accentuated if SA is included in the specification.

In relation to this latter issue, SA in land use choices tends not to be included in theoretical frameworks but added *ex post* in the empirical specification. In land use modeling, SA can stem from two sources. First, it can arise from spillovers among the error terms due to omitted spatial variables affecting land use decisions such as weather or soil quality. A spatial error model or spatial robust inference allows to control for these omitted variables provided that they are not correlated with the observables. Second, it can arise from spillovers among land use decisions or spatial interaction relationships in the land use choices. This might be due, for example, to the neighboring plots being owned by the same landowner, or to shared information which induces forest or agricultural clustering and landowners adopting the same technology based on shared learning. In this case, a spatial autoregressive model would account for these spatial interactions.

In the case of aggregate data, logarithmic transformation on land use shares implies linear equations that can easily be estimated. Therefore, SA in the case of land use models can be estimated using spatial models in the linear case (Sect. 3). Conversely, in most cases of individual data (Sect. 4), the presence of SA tends to make discrete choice models analytically intractable and requires use of simulation estimation methods or Bayesian techniques (Smith and LeSage 2004). Other estimation procedures have been proposed in the literature: the expectation-maximization method (McMillen 1992), the generalized method of moments (GMM) (Pinkse and Slade 1998), and the composite maximum likelihood method (Sidharthan and Bhat 2012; Ferdous and Bhat 2012). For a detailed review of the inclusion of SA in discrete choice models see Fleming (2004), Smirnov (2010), Billé and Arbia (2019).

Considering SA also sheds new light on the issue of prediction. Comparing individual and aggregate models with respect to their predictive accuracy is an ongoing and still open issue with mixed evidence. The seminal paper by Grunfeld and Griliches (1960) examined the relative power of individual (micro) and aggregate (macro) models for explaining aggregate outcomes and found that an aggregate model often performs better. In the context of land use models, Wu and Adams (2002) show that even in the case of linear models, the choice between the micro- and macro-scales to make aggregate predictions cannot generally be resolved by a priori reasoning. Ay et al. (2017) show that introducing SA in aggregate land use models provides better predictions than using individual aspatial models with higher numbers of observations. This suggests that there might be little to be gained from using individual land use data if the sole objective is to predict land use at the aggregate spatial resolution.

Some studies choose none of these modeling approaches and resort instead to ad hoc methods to circumvent the problems related to estimating discrete choice models in the presence of SA (De Pinto and Nelson 2007). These models are summarized below:

- **Spatial sampling:** Most early studies in the land use literature simply purge the data of SA using a spatial sampling technique which allows construction of a data sample without neighbors. This is a fairly widespread practice: Nelson and Hellerstein (1997), Carrion-Flores and Irwin (2004), Irwin et al. (2003), Irwin and Bockstael (2004), Lewis and Plantinga (2007), Lubowski et al. (2008), De Pinto and Nelson (2009), Fezzi et al. (2015)
- **Introduction of latitude and longitude as explanatory variables:** Nelson et al. (2001), Muller and Zeller (2002) claim to account for SA by using two additional explanatory variables representing the latitude and longitude of each observation. While this type of correction is likely to be useful if the spatial effect is caused by an unobserved variable which varies linearly between regions, it captures spatial heterogeneity rather than capturing the SA as claimed by the authors;
- **The introduction of spatially shifted geophysical variables:** Nelson et al. (2001), Munroe et al. (2002) use spatial shifts (i.e., weighted averages of values in neighboring locations) of geophysical variables such as soil type, slope, and vegetation index as exogenous variables. A possible justification for using these types of variables is that they account for the direct influence of the environment on land use decisions in a particular location.

While useful, these methods cannot control for substantive SA, an issue to which we turn in the next two sections.

3 Linear Land Use Models

The objective of most studies using aggregate data is to identify the determinants of land use shares. Most U.S. econometric studies use the county scale and land use data derived generally from federal sources such as agricultural census. The most common method is to specify county land use shares as a logistic function. Examples of studies that use this method include Lichtenberg (1989), Plantinga (1996), Hardie and Parks (1997). While the authors attempt to explain the factors that influence the share of land allocated to a particular land use, other aggregate data studies try to explain changes in land use shares in an area (Stavins and Jaffe 1990; Plantinga and Ahn 2002). All these studies ignore SA. More recent studies taking explicit account of SA have been conducted at the French level by Chakir and Le Gallo (2013), Ay et al. (2017), Chakir and Lungarska (2017) who estimate aggregate land use share models at the department level, 12×12 km and 8×8 km grid cells, respectively.

3.1 Land Use Share Models

Although all econometric studies are based on the same economic theory, several variants of theoretical land allocation models have been proposed (Lichtenberg 1989; Stavins and Jaffe 1990; Plantinga 1996; Hardie and Parks 1997). We present here a

fairly simple version of these models based on Wu and Segerson (1995)'s static model where the landowner n_i ($n_i = 1, \dots, N$) in the region i ($i = 1, \dots, I$) is assumed to be risk neutral and maximizes his expected profit from the use k ($k = 1, \dots, K$) on quality land j ($j = 1, \dots, J$), at time t ($t = 1, \dots, T$), denoted $\pi_{jk}(x(t, n_i), a_{jk}(t, n_i), n_i)$, where $x(t, n_i)$ is a vector of the exogenous variables such as prices, costs and other economic variables and $a_{jk}(t, n_i)$ is the area of land of quality j allocated to use k . For each quality of land, the landowner chooses the area $a_{jk}(t, n_i) \geq 0$ that maximizes his total profit:

$$\sum_{k=0}^K \pi_{jk}(x(t, n_i), a_{jk}(t, n_i), n_i) \quad \text{subject to} \quad \sum_{k=0}^K a_{jk}(t, n_i) = A_j(t, n_i) \quad (1)$$

where $A_j(t, n_i)$ is the total surface of available quality land j . The resolution of the optimization program (1) gives the optimal area $a_{jk}^*(x(t, n_i), A_j(t, n_i), n_i)$ allocated to each use k for each quality of the land j at time t . The optimal share of land allocated to the use k is

$$s_k(x(t, n_i), t, n_i) = \frac{1}{A_j(t, n_i)} \sum_j a_{jk}^*(t, n_i) \quad (2)$$

The optimal uses derived from the theoretical model for each owner should be aggregated to match the scale of the observed data. In practice, the available data are the shares of land uses at an aggregate resolution (county, region, municipality). The land use share k ($k = 1, \dots, K$) in the region i at time t is written as

$$s_{ikt} = p_{ikt} + \varepsilon_{ikt} = \frac{e^{\beta'_k X_{it}}}{\sum_{j=1}^K e^{\beta'_j X_{it}}} + \varepsilon_{ikt} \quad \forall i = 1, \dots, I, \forall k = 1, \dots, K \text{ and } \forall t = 1, \dots, T \quad (3)$$

where p_{ikt} is the expected share of land allocated for use k in the i region at time t . The observed land use share at time t , s_{ikt} may differ from the optimal land use share due to possible hazards such as climate or policy shocks. These elements, of zero average, are captured by the error term ε_{ikt} . X_{it} are the explanatory variables and β'_k are the associated coefficients.

As in Wu and Segerson (1995), Plantinga et al. (1999), most aggregate studies specify land use shares in the logistic functional form for three reasons: first, this functional form allows predicted land use shares to stay between 0 and 1, second, this specification is parsimonious in terms of parameters, and third, logarithmic transformation allows use of linear equations which are easily estimated. This transformation³ has been proposed by Zellner and Lee (1965) and, applied to land use choices, it allows to write the logarithm of each use share normalized by a given share as follows:

³This transformation corresponds to the additive log ratio (ALR) transformation in the literature on composition data in statistics, see Aitchison (1986) for more details.

$$\tilde{y}_{ikt} = \ln(s_{ikt}/s_{ikt}) = \beta'_k X_{it} + u_{ikt} \quad \forall i = 1, \dots, I, \forall k = 1, \dots, K \text{ et } \forall t = 1, \dots, T \quad (4)$$

where u_{ikt} is the transformed error term. Model (4) has $K - 1$ equations that are seemingly unrelated regressions (SUR) and can be estimated by methods accounting for correlations between the error terms associated with each equation.

3.2 *Spatial Autocorrelation in Linear Models*

In linear specification models, SA is handled by the inclusion of spatially lagged variables, that is, weighted averages of the observations of “neighbors” of a given location. These spatially lagged variables can be used as the dependent variable (spatial autoregressive SAR models), explanatory variables (spatial cross regressive SLX models), or the error terms (SEM) or any combination of these options which results in a range of spatial models (Elhorst 2010). For instance, the spatial autoregressive combined (SARAR) model accounts simultaneously for autocorrelation in the error term and for spatial associations of the dependent variable. The spatial Durbin model (SDM) is a combination of SAR and SLX and can be reduced to SEM (LeSage and Pace 2009), while the spatial Durbin error model (SDEM) integrates all the elements of the SLX and the SEM. Finally, the general nesting spatial (GNS) model combines the SARAR and the SLX models (see Table 2). Until the early 2000s, most empirical spatial econometric studies were interested mainly in two specifications: SAR and SEM. Specifications accounting for richer and combined forms of SA are now more commonly estimated. For more details on the taxonomy of linear SA models for cross-sectional data see Elhorst (2014).

The choice of the best spatial specification can be made based on theory or by applying statistical tests to different models. The literature proposes several strategies, the most common being either the so-called classical strategy starting from the simplest “specific to general” model, the most general model going from “general to specific”. Florax et al. (2003) compare these strategies and show that the classical approach gives the best results in terms of identifying the best specification and most precisely estimated parameters but LeSage and Pace (2009) argue that the choice of the best specification should start with the SDM. Elhorst (2010) proposes a mix of these two strategies.

3.3 *Example of Spatial Land Studies with Linear Models*

This section provides some examples of aggregate land use studies which take account of SA.

Some works include SA in order to improve the specification and understanding of what drives land use change. For instance, Meyfroidt and Lambin (2008) analyze the causes of reforestation in Vietnam during the 1990s on a national scale, and test emerging forest transition theories on the same scale. They build a reforestation spatial lag regression model using census and geographic data at a fine level of

Table 2 Summary table of the estimated linear land use (LU) spatial model specifications (Chakir and Lungarska 2017). ρ is the spatial autoregressive coefficient, λ the SA coefficient, γ and β represent a vector of unknown parameters to be estimated. W is a nonnegative $n \times n$ matrix describing the spatial configuration or arrangement of the units in the sample

Model	Model	Interpretation
SEM	$\tilde{y} = X\beta + \varepsilon$ and $\varepsilon = \lambda W\varepsilon + u$	Unobserved omitted variables follow a spatial pattern, data measurement errors
SAR	$\tilde{y} = \rho W\tilde{y} + X\beta + \varepsilon$	LU for one location is determined jointly with that of neighbors
SLX	$\tilde{y} = X\beta + WX\gamma + \varepsilon$	LU for one location is determined by the explanatory variables of neighbors
SDM	$\tilde{y} = \rho W\tilde{y} + X\beta + WX\gamma + \varepsilon$	A combination of SLX and SAR and can be reduced to SEM
SARAR	$\tilde{y} = \rho W\tilde{y} + X\beta + \varepsilon$ and $\varepsilon = \lambda W\varepsilon + u$	A combination of SEM and SAR
SDEM	$\tilde{y} = X\beta + WX\gamma + \varepsilon$ and $\varepsilon = \lambda W\varepsilon + u$	A combination of SEM and SLX
GNS	$\tilde{y} = \rho W\tilde{y} + X\beta + WX\gamma + \varepsilon$ and $\varepsilon = \lambda W\varepsilon + u$	A combination of SLX and SARAR

aggregation for the whole country. Their results show that forest land distribution affects forests not just in the focal district but also in neighboring districts. This observation can be interpreted in terms of a diffusion process: early and successful implementation of the policy in some districts may have facilitated its rapid adoption by neighboring districts.

Marcos-Martinez et al. (2017) estimate the determinants of land use in Australia’s intensive agricultural region during the period 1992–2010. They estimate land use shares with spatial error and random effects combined with variance decomposition analysis to identify the statistical significance, direction and magnitude of the observed associations between land-uses and their drivers. Their results show that improved transportation infrastructure, zoning regulations and mechanisms to reduce exposure to farm debt and climate variability risks have significant impacts on the configuration of the Australian agricultural landscape.

Amin et al. (2019) analyze whether protected areas are efficient instruments to fight deforestation in Brazilian Amazonia. They estimate a dynamic SDM and assess the impact of different types of protected areas (integral protected areas, sustainable protected areas, indigenous lands) on deforestation. The results differ according to the type of protected area: (i) integral protected areas and indigenous lands reduce deforestation; (ii) sustainable use areas do not contribute to reducing deforestation;

and (iii) the spillover effects generated by integral protected areas and indigenous lands lead to a reduction in the deforestation in their vicinities.

Two studies focus on prediction in spatial land use share models. Chakir and Le Gallo (2013) make a methodological contribution to the literature by controlling for both unobservable individual heterogeneity and SA in an aggregate land use model. Their study was conducted on a panel of land use data at the French departments NUTS3 scale, observed between 1992 and 2003. The authors were interested in the relationship between four land uses (agriculture, forest, urban, and other) and their potential economic and demographic determinants. The econometric model consists of a system of three equations with a panel dimension and SA in the errors associated to each equation. Thus, their econometric model is a SUR model with random individual effects and autoregressive spatial structure of the error term. The model was estimated using the feasible generalized least square (FGLS) estimation method proposed by Baltagi and Pirotte (2011) for SUR-SEM-RE (Seemingly Unrelated Regressions-Spatial Error Model-Random Effects) model estimations. Their results are of three orders: first, controlling for both unobservable individual heterogeneity and SA yields the best predictions relative to any other specification in which SA and/or individual heterogeneity are ignored. Second, taking into account the correlations between the error terms in the different equations does not seem to improve prediction performance. Third, ignoring individual heterogeneity introduces substantial loss of prediction accuracy.

Chakir and Lungarska (2017) estimate land use share models for France at a homogeneous (8×8 km) grid scale for five land use classes—agriculture, pasture, forest, urban, and other. They investigate the determinants of land use shares using economic, physical and demographic explanatory variables. They model SA between grid cells and compare prediction accuracy and estimated elasticities for the different spatial model specifications (ordinary least square (OLS), SLX, SEM, SAR, SDM, SDEM, SARAR, GNS). They compare these spatial specifications using three rent proxies: farmers' revenues, land prices, and shadow land prices derived from a mathematical programming model. Their comparison is based on several criteria: quality of economic explanation (significance of agricultural rents and their marginal impacts), prediction quality (NRMSE), specification tests (LM tests), and goodness of fit (log-likelihood, R2, AIC). The test results show that the SDM, SDEM, SARAR, and GNS models should be considered. According to the goodness of fit (pseudo-R2, log-likelihood and AIC) and prediction quality criteria, GNS is the specification that best fits their data. In a context of aggregate land use, the existence of autocorrelation is due mainly to spatially correlated errors—essentially a data measurement problem. This applies especially to their case since they use artificially constructed grids, and different scales for the explanatory variables and land use data. Their results show also that including SA in land use share models improves the quality of the predictions which confirms the results in the previous aggregate land use literature.

4 Discrete Choice Land Use Models

When using individual (parcel or plot) data, the land use variable is generally a categorical variable so that estimating land use patterns on individual data usually requires a discrete choice framework. Discrete choice models are based on McFadden (1974)'s random utility theory which states that the landowner decides to switch from one use to another if the expected net revenues exceed the revenues from the original use.

4.1 Individual Choice Land Use Model

This section presents the theoretical land use model based on individual data as in Lubowski et al. (2008). We assume that the landowner chooses the land use of a plot based on the costs and benefits associated with each possible use. For example, the landowner chooses land use k at time t if:

$$R_{kt} - rC_{jkt} > R_{jt} \quad \forall j, k = 1, \dots, K \quad \text{and} \quad \forall t = 1, \dots, T \quad (5)$$

where R_{jt} and R_{kt} represent the discounted expected net benefits at time t of a unit of land for uses j and k , respectively, C_{jkt} is the marginal cost of converting a unit of land from use j to use k at time t ($C_{jjt} = 0$) and r is the discount rate.

In order to estimate the determinants of land use econometrically, the theoretical model suggests comparing the benefits and costs of converting land from one use to another at each date. To move to the econometric specification, land use conversion revenues and costs are rewritten as functions of the observed and unobserved variables. Thus, the utility U_{ikt} of the owner of parcel i with land use k at time t is written as follows:

$$U_{ikt} = \beta x_{ikt} + \epsilon_{ikt} \quad \forall i = 1, \dots, N, \quad \forall k = 1, \dots, K \quad \text{and} \quad \forall t = 1, \dots, T, \quad (6)$$

where x_{ikt} are the observed explanatory variables, β the vector of parameters to be estimated and ϵ_{ikt} are the error terms which take account of the unobserved variables that might affect the landowner's utility.

We assume that the owner has a choice between K land use categories for each parcel at each date. The landowner chooses the optimal land use for his or her plot by comparing the utilities associated to each land use category. If $y_{it} = 1, 2, \dots, K$; is the landowner's land use choice for the parcel i at time t , we obtain

$$y_{it} = k, \text{ if } U_{ikt} \geq \max U_{ijt} \quad \forall i = 1, \dots, N, \quad \forall j, k = 1, \dots, K \quad \text{and} \quad \forall t = 1, \dots, T, \quad (7)$$

Thus, the probability that the parcel i is allocated to the use k at the time t is written as

$$P(y_{ikt} = 1) = Pr[U_{ikt} \geq \max U_{ijt}] \quad (8)$$

for all $j = 1, \dots, K$ with $y_{ikt} = 1$ if k is the observed use and 0 otherwise; U_{ikt} is the utility associated with land use k .

Since estimation of discrete choice models in the multinomial case is dimensionally constrained, some studies are limited to two use categories and use a probit model in the binary case (Bockstael 1996; Kline and Alig 1999; Irwin and Bockstael 2002). Other studies estimate a multinomial logit model because of its computational simplicity (Chomitz and Gray 1996; Nelson and Hellerstein 1997; Nelson et al. 2001) which involves the questionable assumption of independence of irrelevant alternatives (IIA). Finally, a nested logit model could be a good alternative if the alternatives can be partitioned into several subsets.

4.2 Spatial Autocorrelation in Discrete Choice Models

SA can be accounted for in the discrete choice land use model (Eqs. (6)–(8)), by including the spatially lagged variables or the error terms. Simplifying the notations and removing the subscripts, the general nonlinear nesting model (GNNM) can be written as follows:

$$U = \rho WU + X\beta + WX\gamma + \varepsilon, \text{ and } \varepsilon = \lambda W\varepsilon + u \quad (9)$$

where WU is the shifted utility function for the weight matrix W , ρ is the autoregressive spatial parameter which indicates the magnitude of the interaction between the latent variables U , γ , like β , is a vector of the unknown parameters to be estimated, λ is the parameter of the intensity of the SA between the residuals, and u is a classical error term such as $u \sim iid(0, \sigma^2 I)$. The GNNM model presented in Eq. (9) becomes a SEM model if $\rho = 0$ and $\gamma = 0$, and becomes a SAR model if $\lambda = 0$ and $\gamma = 0$. In contrast to the linear case, the spatially lagged variable in the SAR model is not observable. For example, in the case of a land use model, it is the utility associated to the profitability of neighboring plots and not the observed land use which should define the utility function of the landowner (Anselin and Cho 2002).

In the case where the error terms ε follow a normal distribution, estimation of the probit-SAR model raises two problems. On the one hand, heteroskedasticity makes the classical estimators inconsistent. On the other hand, estimation of a probit-SAR requires computation of a likelihood function with $N - 1$ (where N is the number of observations) integrals which makes maximum likelihood estimation impossible. This second difficulty applies also to the logit model case (Anselin and Cho 2002). Several approaches have been proposed in the literature to deal with these estimation problems, including simulation estimation (Geweke et al. 1994) or Bayesian (LeSage 2000) methods able to deal with the computation of multidimensional integrals of the likelihood function. Other estimation procedures have been proposed to cope with the problems associated to the introduction of SA in the case of discrete choice models. These include the expectation-maximization method (McMillen 1992), the GMM

Table 3 Summary table of the estimated spatial discrete choice models

Model	Estimation method	Example
Spatial autoregressive logit	Bayesian	Blackman et al. (2008)
Ordered probit	Bayesian	Wang and Kockelman (2009)
Multinomial probit	Bayesian	Chakir and Parent (2009)
Random parameter logit	Max simulated likelihood	Lewis et al. (2011)
Multinomial probit	Max approximate CML	Sidharthan and Bhat (2012)
Ordered probit	Max CML	Ferdous and Bhat (2012)
Multinomial logit	GMM	Li et al. (2013)
Multiple discrete-continuous probit	Max CML	Bhat et al. (2015)
Conditional parametric probit	Max Locally Weighted log-Likelihood estimator	McMillen and Soppelsa (2015)
Multinomial logit	GMM	Carrión-Flores et al. (2018)

(Pinkse and Slade 1998), the maximum pseudo-likelihood method (Smirnov 2010), and finally the method of maximum approximate composite marginal likelihood (CML) (Sidharthan and Bhat 2012). For detailed reviews of SA in discrete choice models see Fleming (2004), Smirnov (2010). Simulation estimation and Bayesian methods have been employed only recently to deal with the computational problems associated to considering SA in discrete choice models. Because these methods are still relatively expensive to implement, their use in the land use literature remains limited. Table 3 provides an overview of these studies.

4.3 *Examples of Spatial Land Use Studies with Discrete Choice Models*

To tackle the complexities induced by SA in discrete choice models for land use, some papers resort to Bayesian methods, for example, Wang and Kockelman (2009) who estimate an ordered probit spatial dynamic model using satellite land cover data. Chakir and Parent (2009) also use a Bayesian approach to estimate land use determinants in a multinomial probit econometric model which accounts for both unobservable individual heterogeneity and SA in errors. They analyze the determinants of land based on a panel of 3,130 points in the Rhône department in France between 1992 and 2003. It appears that land use changes are indeed influenced by unobserved factors in neighboring plots. Finally, Blackman et al. (2008) estimate a bayesian heteroskedastic SAR logit model of land cover for a shade-grown coffee region in southern Mexico. Their results show that all other things being equal plots close to large cities are less likely to be cleared which contrasts to the pattern usually observed in natural forests. They also find that belonging to a coffee-marketing cooperative, farm size, and certain soil types are associated to tree cover while prox-

imity to a small town center is associated to forest clearing. This study is extended in Blackman et al. (2012) who estimate a SAR probit model.

Other papers use variants of maximum simulated likelihood. Lewis et al. (2011)⁴ estimate a random parameter logit model to take account of the non-observed space-time components of the willingness to pay. This specification makes it possible to take account of spatial heterogeneity rather than SA and also allows consideration of heteroskedasticity via a block variance-covariance matrix with individual effects which depend on space. It is a kind of SA but with no spatial structure and with a matrix of weights as in spatial models. In the spatial econometrics literature, CML has become a popular approach for estimating spatial probit models and has been used to model land use. For instance, Ferdous and Bhat (2012) analyze changes in the intensity of urban land use taking account of both the spatial dimension and temporal dynamics. Their econometric model is an ordered probit estimated using CML. The results show that ignoring the presence of spatial autocorrelation and spatial heterogeneity introduces important bias and that ignoring spatial heterogeneity is more serious than ignoring lagged spatial dynamics. Sidharthan and Bhat (2012) use maximum approximate CML (MACML) to estimate a multinomial probit-type land use model with SA between plots and spatial heterogeneity.

Finally, rather than tackling the spatial autoregressive coefficient directly as in the previous papers, McMillen and Soppelsa (2015) estimate a conditional parametric spatial probit model imposing far less structure on the data than conventional parametric models. They illustrate the approach using data on 474,170 individual lots in the City of Chicago. Their results suggest that simple functional forms are not appropriate for explaining the spatial variation in residential land use across the entire city. Similarly, Carrión-Flores et al. (2018) propose a GMM spatial estimator for a multinomial logit model with spatial lag dependence. The model is linearized to avoid the repeated matrix inversion required for the full GMM estimation. The linearization breaks up the estimation procedure into two simple steps: a standard multinomial logit model with no SA followed by a two-stage least squares (TSLS) estimation of the linearized model which accounts for SA. This model is applied to estimate land use conversion in the rural-urban fringe for four different land uses (agricultural, residential, industrial and commercial). The results show a positive SA of about 0.36—a result consistent with the widely-accepted idea that land use conversion is a spatial process.

5 Land Use and Its Impacts on the Environment

Land use is considered as one of the main drivers of global changes to nature, which endanger numerous species or cause their extinction and compromise the supply of ecosystem services (ES) which are important for humans (Millennium Ecosystem Assessment 2005). The protection of ES is emerging as a major concern alongside

⁴Several attempts in the literature introduce spatial dependence in multinomial models but, except for Lewis et al. (2011) to the best of our knowledge, they have not been used in the land use literature.

climate change issues (IPCC 2019) and biodiversity conservation (IPBES 2019). This has resulted in land use becoming a growing concern for policy-makers as means of protecting ecosystems (Bateman et al. 2013). There is a large literature estimating the effects of land use on various ES: water quality (Fezzi et al. 2015), carbon sequestration (Lubowski et al. 2006), and biodiversity (Polasky et al. 2008).

In this context, accounting for SA when studying the impacts of land use on ES is a major issue. Research shows that including SA in species distribution models improves model fit and prediction accuracy (Record et al. 2013) and that ignoring SA can produce inaccurate results (Kühn 2007). Below, we review a selection of those studies that model SA explicitly to estimate the impacts of land use on the environment.

5.1 Land Use and ES

The relationships between land use and ES is complex. For instance, some land uses such as intensive agriculture could have negative impacts on ecosystems while others could contribute to the provision of many ES. For example, tropical forests are an example of a supplier of ES at various scales. At the local scale, these services include wood, secondary forest products, pollination, etc. More generally, they sequester large amounts of carbon which regulates the global climate (IPCC 2019). In addition, the productivity of some land uses such as agriculture is dependent on ecosystems such as biological pest control, soil fertility, and pollination. Thus, degradation of these ecosystems constitutes a serious threat to the long-term agricultural productivity growth. Below, we provide two examples of spatial studies dealing with this link.

Chen et al. (2020) employ an integrated spatial panel approach to examine the geographic variations and spatial determinants of the ES balance in the middle reaches of the Yangtze River urban agglomerations (MRYRUA) in China. They analyze the spatio-temporal evolution features of landscape patterns and the supply of demand for and balance among ES and landscape pattern metrics for the period 1995–2015. The results indicate that construction land in the MRYRUA has increased continuously, while farmland has decreased. Counties with higher ES supply and balance indices are concentrated primarily in mountainous areas, while the indices of ES demand in the three smaller urban agglomerations, plains areas, counties surrounding major cities, and along major traffic routes are higher. SA and spatial spillover effects of the ES balance index are observed in the MRYRUA. Population density and road density are negatively associated to an ES balance. Landscape pattern metrics are also statistically significant, either positive or negative. The findings suggest that both drivers and spillover effects should be accounted for when considering integrative ecosystem management and land use sustainability measures in urban agglomerations. Both have important implications for urban planning and decision-making related to development and ES.

Klemick (2011) uses cross-sectional farm survey data to estimate the value of fallow ES in shifting cultivation in one region in the Brazilian Amazon. The objective is to test whether it provides economically significant local externalities which

might justify forest conservation from a local perspective. The author estimates a production function to determine the contributions to agricultural income of on-farm and off-farm forest fallow. Soil quality controls, instrumental variables and spatial econometric approaches help address issues of endogeneity and variation in unobservable factors over space. The results suggest that Bragantina farmers generally allocate land between cultivation and fallow efficiently taking account of beneficial spillovers. This finding does not necessarily imply that farmers intentionally internalize the value of these services but might suggest that private land tenure plays a role in promoting sustainable land management given the different findings from other studies of shifting cultivation in common property tenure regimes which identify overexploitation of fallow biomass.

5.2 *Land Use and Water Quality*

There is a large literature on the effects of land use on water quality and freshwater biodiversity. Most of these papers ignore SA. Here, we provide some examples of studies that model SA explicitly in a study of land use and water quality.

Most studies show that forest areas have a positive impact on water quality compared to intensive agriculture, livestock, and urban areas. For example, Abildtrup et al. (2015) analyze the economic impacts of land use on the cost of drinking water supply, taking account of both the organizational choice of water supply and spatial factors in the same model. They estimate a model for the choice of management type and for the price of water, accounting for the potential dependence of the error terms between equations, as well as between neighboring water services. They estimate a sample selection model adapted to a spatial context, that is, allowing for spatial lags and spatial error processes. The model is applied to data from the French Vosges department. The results show spatial interactions related to the characteristics of neighboring water services but no SA of the error terms in the management choice equation, or in prices. They show that forest land cover significantly reduces water supply costs at the large but not the local scale.

Induced land use adaptation on freshwater biodiversity is analyzed by Bayramoglu et al. (2020). They study the links between land use (agriculture, pasture, forest, and urban environment) and the fish-based index (FBI) an indicator of the ecological state of surface water measured for various French rivers observed between 2001 and 2013. They estimate two models: a spatial econometric model of land use and a spatial panel statistical model of the FBI. Their results indicate that adapting land use to climate change is reducing the biodiversity of freshwater in France. Furthermore, rivers located in regions with intensive agriculture and pastures are associated to lower freshwater biodiversity than those in forest regions. Simulations show that climate change will exacerbate these negative impacts through changes to land use. They show how two policies for regulating the level of fertilizers in agriculture and carrying capacity in grasslands could help improve freshwater biodiversity and cope with the adverse effects of land use and climate change.

5.3 *Land Use and Climate Change*

The interactions between land use and climate are complex (IPCC 2019). First, land use and land practices affect the global concentration of greenhouse gases (Houghton 2003). Second, while land use change is an important driver of climate change, a changing climate can lead to changes in land use. For example, farmers might convert pasture to crop land which has higher economic returns under changing climatic conditions. Third, spatially heterogeneous land use activities have important impacts on local weather (Feddema et al. 2005). Fourth, land use changes could play an important role in mitigating climate change either by increasing carbon sequestration or by reducing greenhouse gas emissions. This could be achieved by adopting land uses such as afforestation or preservation of permanent pasture (Pielke 2005).

Land use adaptation to climate change could exacerbate the adverse impacts of land use on the environment. For example, Lungarska and Chakir (2018) show that in France, climate change will reduce forest areas which could increase greenhouse gas emissions. They estimate a spatial econometric land use model and simulate the impacts of two IPCC climate change scenarios (A2 and B1, horizon 2100) and a mitigation policy in the form of a tax on greenhouse gas emissions (0–200 euros/tCO₂) aimed at reducing agricultural greenhouse gas emissions. They show that both climate change scenarios lead to an increase in agricultural area at the expense of forests. Greenhouse gas mitigation policies reduce expansion of agriculture, and therefore could counteract the consequences of climate change on land use. Taking account of land use adaptations to climate change makes it possible to reduce abatement costs in the agricultural sector.

6 Conclusion

The objective of this review was to summarize the literature on econometric land use modeling and show how SA can be accounted for in these models. Despite the recent advances in econometric land use models, several research directions remain to be explored and several issues need to be addressed concerning data, theories, and empirical models.

First, there is a frequent lack of data to construct relevant explanatory variables implied by theoretical models. In particular, land rents are described in the theoretical model as among the main decision variables related to land use or land use change but are unobservable in the case of agricultural or urban use. In the case of forestry use, these rents are even more difficult to calculate. More research is needed along these lines, and especially to investigate the question of the links between land price and land rent, drawing on the work of Randall and Castle (1985), Goodwin et al. (2003).

Second, more investigation is needed into scale issues in land use studies. For example, most economic variables refer to administrative units rather than grids which makes it easier to estimate econometric models at the same administrative

scale (such as department, municipality, or small agricultural regions in the French case). However, a land use model with aggregate spatial resolution is less relevant for assessing the local ecological effects of land uses. Ecological issues such as habitat quality or dispersion of species operate on fine scales. Ecological conditions vary considerably within each administrative unit, introducing additional uncertainty for ecological assessments.

Third, in addition to the spatial dimension, it would be interesting to incorporate the dynamic dimension explicitly in econometric land use models (Epanchin-Niell et al. 2017). Methodological advances in the specification and estimation of spatio-temporal panel models are one of the difficulties related to spatial econometrics as noted in Arbia (2011). The estimation methods developed by Ferdous and Bhat (2012), Sidharthan and Bhat (2012) seem promising as alternatives to the computationally intensive Bayesian or simulation methods.

Fourth, all the models presented here assume implicitly that the spatial weight matrix is exogenous. If spatial units refer to individual landowners making land use choices, these choices might be influenced substantially by the choices of peers with whom they choose to be linked in which case the weight matrix becomes endogenous. Identification of endogenous peer effects and how to disentangle them from exogenous effects and correlated effects in networks has been studied extensively.⁵ The way landowners form networks and how these affect land use decisions are of considerable interest to understand the drivers of these decisions.

Finally and related to this issue, structural models should be further developed to study the links between land use and land use changes, and their effects on the environment for example on GHG emissions and biodiversity. The advantage of a structural approach is that it makes more explicit assumptions about observable and unobservable variables. The structural approach also makes it possible to unambiguously account for the endogeneity of prices and the feedbacks that determine the market equilibrium (Timmins and Schlenker 2009). The aim is to propose a theoretical economic model which includes the farmer's decisions about crop rotations, choice of inputs (fertilizers), land allocation between agricultural and grassland uses, and herd size and composition. This would be quite challenging and would force a limited focus on a subset of these decisions (Kaminski et al. 2013).

Addressing these issues would help to improve the quality of econometric land use models. Developing accurate models is important for policy making to allow for more accurate predictions about land use and future changes and more accurate measurement of the effects of these changes on natural resources (biodiversity, water quality, soil quality, and air quality).

⁵See Hsieh et al. (2019) for a recent paper on the specification and estimation of network formation and network interaction and applications of this literature to land use issues can be found in Isaac and Matous (2017), Baird et al. (2016).

Acknowledgements We dedicate this contribution to Christine Thomas-Agnan, a friend and colleague for many years. We wish her a nice, sweet, and pleasant retirement. We are grateful for helpful comments from the editors and the two anonymous reviewers. Raja Chakir acknowledges the support from the French state aid managed by the Agence Nationale de la Recherche as part of the “Investments d’Avenir” Programme within STIMUL (Scenarios Towards integrating multi-scale land use tools) flagship project (LabEx BASC; ANR-11-LABX-0034) and Cland Institut de convergence (ANR-16-CONV-0003).

References

- Abildtrup, J., Garcia, S., & Kere, E. N. (2015). Land use and drinking water supply: A sample selection model with spatial dependence. *Revue d’Économie Régionale et Urbaine*, 2015(1–2), 321–342.
- Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.
- Alig, R. (1986). Econometric analysis of the factors influencing forest acreage trends in the south-east. *Forest Science*, 32(1), 119–134.
- Amin, A., Choumert-Nkolo, J., Combes, J.-L., Motel, P. C., Kéré, E. N., Ongono-Olinga, J.-G., et al. (2019). Neighborhood effects in the brazilian amazônia: Protected areas and deforestation. *Journal of Environmental Economics and Management*, 93, 272–288.
- Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht: Kluwer Academic Publishers.
- Anselin, L., & Cho, W. (2002). Spatial effect and ecological inference. *Political Analysis*, 10, 276–297.
- Arbia, G. (2011). A lustrum of sea: Recent research trends following the creation of the spatial econometrics association (2007–2011). *Spatial Economic Analysis*, 6(4), 377–395.
- Ay, J.-S., Chakir, R., & Gallo, J. L. (2017). Aggregated versus individual land-use models: Modeling spatial autocorrelation to increase predictive accuracy. *Environmental Modeling & Assessment*, 22, 129–145.
- Baird, J., Jollineau, M., Plummer, R., & Valenti, J. (2016). Exploring agricultural advice networks, beneficial management practices and water quality on the landscape: A geospatial social-ecological systems analysis. *Land Use Policy*, 51, 236–243.
- Baltagi, B. H., & Pirotte, A. (2011). Seemingly unrelated regressions with spatial error components. *Empirical Economics*, 40, 5–49.
- Bateman, I. J., Harwood, A. R., Mace, G. M., Watson, R. T., Abson, D. J., Andrews, B., et al. (2013). Bringing ecosystem services into economic decision-making: Land use in the United Kingdom. *Science*, 341, 45–50.
- Bayramoglu, B., Chakir, R., & Lungarska, A. (2020). Impacts of land use and climate change on freshwater ecosystems in France. *Environmental Modeling & Assessment*, 25(2), 147–172.
- Bhat, C. R., Dubey, S. K., Alam, M. J. B., & Khushefati, W. H. (2015). A new spatial multiple discrete-continuous modeling approach to land use change analysis. *Journal of Regional Science*, 55(5), 801–841.
- Billé, A. G., & Arbia, G. (2019). Spatial limited dependent variable models: A review focused on specification, estimation, and health economics applications. *Journal of Economic Surveys*, 33(5), 1531–1554.
- Blackman, A., Albers, H. J., Avalos-Sartorio, B., & Murphy, L. C. (2008). Land cover in a managed forest ecosystem: Mexican shade coffee. *American Journal of Agricultural Economics*, 90(1), 216–231.
- Blackman, A., Avalos-Sartorio, B., & Chow, J. (2012). Land cover change in agroforestry: Shade coffee in El Salvador. *Land Economics*, 88(1), 75–101.

- Bockstael, N. E. (1996). Modeling economics and ecology: The importance of a spatial perspective. *American Journal of Agricultural Economics*, 78(5), 1168–1180.
- Brady, M., & Irwin, E. (2011). Accounting for spatial effects in economic models of land use: Recent developments and challenges ahead. *Environmental & Resource Economics*, 48(3), 487–509.
- Carrion-Flores, C., & Irwin, E. G. (2004). Determinants of residential land-use conversion and sprawl at the rural-urban fringe. *American Journal of Agricultural Economics*, 86(4), 889–904.
- Carrión-Flores, C. E., Flores-Lagunes, A., & Guci, L. (2018). An estimator for discrete-choice models with spatial lag dependence using large samples, with an application to land-use conversions. *Regional Science and Urban Economics*, 69, 77–93.
- Chakir, R. (2015). L'espace dans les modèles économétriques d'utilisation des sols: enjeux méthodologiques et applications empiriques. *Revue d'Economie Regionale Urbaine*, 1/2(1), 59–82.
- Chakir, R., & Le Gallo, J. (2013). Predicting land use allocation in France: A spatial panel data analysis. *Ecological Economics*, 92, 114–125.
- Chakir, R., & Lungarska, A. (2017). Agricultural rent in land-use models: Comparison of frequently used proxies. *Spatial Economic Analysis*, 12(2–3), 279–303.
- Chakir, R., Parent, O. (2009). Determinants of land use changes: A spatial multinomial probit approach. *Papers in Regional Science*, 88(2), 327–344, 06.
- Chen, W., Chi, G., & Li, J. (2020). The spatial aspect of ecosystem services balance and its determinants. *Land Use Policy*, 90, 104263 (2020)
- Chomitz, K. M., & Gray, D. A. (1996). Roads, land use, and deforestation: A spatial model applied to Belize. *World Bank Economic Review*, 10(3), 487–512.
- Claassen, R., & Tegene, A. (1999). Agricultural land use choice: A discrete choice approach. *Agricultural and Resource Economics Review*, 28(1), 26–36.
- De Pinto, A., & Nelson, G. C. (2007). Modelling deforestation and land-use change: Sparse data environments. *Journal of Agricultural Economics*, 58(3), 502–516.
- De Pinto, A., & Nelson, G. C. (2009). Land use change with spatially explicit data: A dynamic approach. *Environmental and Resource Economics*, 43, 209–229.
- Elhorst, J. P. (2010). Applied spatial econometrics: Raising the bar. *Spatial Economic Analysis*, 5(1), 9–28.
- Elhorst, J. P. (2014). *Spatial econometrics: From cross-sectional data to spatial panels*. Berlin: Springer.
- Epanchin-Niell, R., Kuwayama, Y., & Walls, M. (2017). Spatial-dynamic complexities of climate challenge for rural areas: Integrating resource and regional economic insights. *American Journal of Agricultural Economics*, 99(2), 447–463.
- Feddema, J. J., Oleson, K. W., Bonan, G. B., Mearns, L. O., Buja, L. E., Meehl, G. A., et al. (2005). The importance of land-cover change in simulating future climates. *Science*, 310(5754), 1674–1678.
- Ferdous, N., & Bhat, C. R. (2012). A spatial panel ordered-response model with application to the analysis of urban land-use development intensity patterns. *Journal of Geographical Systems*, 15, 1–29.
- Fezzi, C., & Bateman, I. J. (2011). Structural agricultural land use modeling for spatial agro-environmental policy analysis. *American Journal of Agricultural Economics*, 93(4), 1168–1188.
- Fezzi, C., Harwood, A. R., Lovett, A. A., & Bateman, I. J. (2015). The environmental impact of climate change adaptation on land use and water quality. *Nature Climate Change*, 5(3), 255–260.
- Fleming, M. M. (2004). Techniques for estimating spatially dependent discrete choice models. In R. Florax & L. Anselin (Eds.), *Advances in spatial econometrics: Methodology, tools, and applications* (pp. 145–167). Berlin: Springer.
- Florax, R. J. G. M., Folmer, H., & Rey, S. J. (2003). Specification searches in spatial econometrics: The relevance of Hendry's methodology. *Regional Science and Urban Economics*, 33, 557–579.
- Foley, J. A. (2005). Global consequences of land use. *Science*, 309(5734), 570–574.
- Gao, L., & Bryan, B. A. (2017). Finding pathways to national-scale land-sector sustainability. *Nature*, 544(7649), 217–222.

- Geweke, J., Keane, M., & Runkle, D. (1994). Alternative computational approaches to inference in the multinomial probit model. *The Review of Economics and Statistics*, 76(4), 609–632.
- Goodwin, B. K., Mishra, A. K., & Ortalo-Magne, F. N. (2003). What's wrong with our models of agricultural land values? *American Journal of Agricultural Economics*, 85(3), 744–752.
- Grunfeld, Y., & Griliches, Z. (1960). Is aggregation necessarily bad? *The Review of Economics and Statistics*, 42(1), 1–13.
- Hardie, I. W., & Parks, P. J. (1997). Land use with heterogeneous land quality: An application of an area base model. *American Journal of Agricultural Economics*, 79(2), 299–310.
- Houghton, R. A. (2003). Revised estimates of the annual net flux of carbon to the atmosphere from changes in land use and land management 1850–2000. *Tellus B*, 55(2), 378–390.
- Hsieh, C.-S., Lee, L.-F., Boucher, V. (2019). Specification and estimation of network formation and network interaction models with exponential probability distribution. *SSRN*.
- IPBES. (2019). Global assessment report on biodiversity and ecosystem services. Technical report, IPBES.
- IPCC. (2019). IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems. Technical report, IPCC.
- Irwin, E., & Bockstael, N. (2002). Interacting agents, spatial externalities and the evolution of residential land use patterns. *Journal of Economic Geography*, 2, 331–54.
- Irwin, E. G., Bell, K. P., & Geoghegan, J. (2003). Modeling and managing urban growth at the rural-urban fringe: A parcel-level model of residential land use change. *Agricultural and Resource Economics Review*, 32(1), 83–102.
- Irwin, E. G., & Bockstael, N. E. (2004). Land use externalities, open space preservation, and urban sprawl. *Regional Science and Urban Economics*, 34(6), 705–725.
- Isaac, M., & Matous, P. (2017). Social network ties predict land use diversity and land use change: A case study in Ghana. *Regional Environmental Change*, 17, 1823–1833.
- Kaminski, J., Kan, I., & Fleischer, A. (2013). A structural land-use analysis of agricultural adaptation to climate change: A proactive approach. *American Journal of Agricultural Economics*, 95(1), 70–93.
- Klemick, H. (2011). Shifting cultivation, forest fallow, and externalities in ecosystem services: Evidence from the eastern amazon. *Journal of Environmental Economics and Management*, 61(1), 95–106.
- Kline, J. D., & Alig, R. J. (1999). Does land use planning slow the conversion of forest and farm lands? *Growth and Change*, 30, 3–22.
- Kühn, I. (2007). Incorporating spatial autocorrelation may invert observed patterns. *Biodiversity Letter*, 13, 66–69.
- Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, 304(5677), 1623–1627.
- LeSage, J., & Pace, R. (2009). *Introduction to spatial econometrics*. Boca Raton: CRC Press.
- LeSage, J. P. (2000). Bayesian estimation of limited dependent variable spatial autoregressive model. *Geographical Analysis*, 32, 19–35.
- Lewis, D. J., Plantinga, A. J., Nelson, E., & Polasky, S. (2011). The efficiency of voluntary incentive policies for preventing biodiversity loss. *Resource and Energy Economics*, 33(01), 192–211.
- Lewis, D. J., & Plantinga, A. J. (2007). Policies for habitat fragmentation: Combining econometrics with GIS-based landscape simulations. *Land Economics*, 83(19), 109–127.
- Li, M., Wu, J., & Deng, X. (2013). Identifying drivers of land use change in China: A spatial multinomial logit model analysis. *Land Economics*, 89, 632–654.
- Lichtenberg, E. (1989). Land quality, irrigation development, and cropping patterns in the northern high plains. *American Journal of Agricultural Economics*, 71(1), 187–194.
- Lubowski, R., Plantinga, A., & Stavins, R. (2008). What drives land-use change in the united states? A national analysis of landowner decisions. *Land Economics*, 84(4), 551–572.

- Lubowski, R. N., Plantinga, A. J., & Stavins, R. N. (2006). Land-use change and carbon sinks: Econometric estimation of the carbon sequestration supply function. *Journal of Environmental Economics and Management*, 51, 135–152.
- Lungarska, A., & Chakir, R. (2018). Climate-induced land use change in France: Impacts of agricultural adaptation and climate change mitigation. *Ecological Economics*, 147, 134–154.
- Marcos-Martinez, R., Bryan, B. A., Connor, J. D., & King, D. (2017). Agricultural land-use dynamics: Assessing the relative importance of socioeconomic and biophysical drivers for more targeted policy. *Land Use Policy*, 63, 53–66.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, Chap. 2. New York: Academic.
- McMillen, D., & Soppelsa, M. E. (2015). A conditionally parametric probit model of microdata land use in Chicago. *Journal of Regional Science*, 55(3), 391–415.
- McMillen, D. P. (1989). An empirical model of urban fringe land use. *Land Economics*, 65(2), 138–145.
- McMillen, D. P. (1992). Probit with spatial autocorrelation. *Journal of Regional Science*, 32(3), 335–348.
- Meyfroidt, P., & Lambin, E. F. (2008). The causes of the reforestation in Vietnam. *Land Use Policy*, 25(2), 182–197.
- Millenium Ecosystem Assessment. (2005). Ecosystems and human well-being: Synthesis. Technical report, World Resources Institute. Washington, DC: Island Press.
- Muller, D., & Zeller, M. (2002). Land use dynamics in the central highlands of Vietnam: A spatial model combining village survey data with satellite imagery interpretation. *Agricultural Economics*, 27(3), 333–354.
- Munroe, D. K., Southworth, J., & Tucker, C. M. (2002). The dynamics of land-cover change in western Honduras: Exploring spatial and temporal complexity. *Agricultural Economics*, 27(3), 355–369.
- Nelson, G. C., Harris, V., & Stone, S. W. (2001). Deforestation, land use, and property rights: Empirical evidence from Darien, Panama. *Land Economics*, 77(2), 187–205.
- Nelson, G. C., & Hellerstein, D. (1997). Do roads cause deforestation? Using satellite images in econometric analysis of land use. *American Journal of Agricultural Economics*, 79(1), 80–88. Feb.
- Pielke, R. A. (2005). Land use and climate change. *Science*, 310(5754), 1625–1626.
- Pinkse, J., & Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics*, 85(1), 125–154.
- Plantinga, A., Mauldin, T., & Miller, D. (1999). An econometric analysis of the costs of sequestering carbon in forests. *American Journal of Agricultural Economics*, 81, 812–24.
- Plantinga, A. J. (1996). The effect of agricultural policies on land use and environmental quality. *American Journal of Agricultural Economics*, 78(4), 1082–1091.
- Plantinga, A. J. (2015). Integrating economic land-use and biophysical models. *The Annual Review of Resource Economics*, 7(1), 233–249.
- Plantinga, A. J., & Ahn, S. (2002). Efficient policies for environmental protection: An econometric analysis of incentives for land conversion and retention. *Journal of Agricultural and Resource Economics*, 27(1), 128–145.
- Polasky, S., Nelson, E., Camm, J., Csuti, B., Fackler, P., Lonsdorf, E., et al. (2008). Where to put things? Spatial land management to sustain biodiversity and economic returns. *Biological Conservation*, 141, 1505–1524.
- Randall, A., & Castle, E. N. (1985). Land resources and land markets. In A. V. Kneese & J. L. Sweeney (Eds.), *Handbook of natural resource and energy economics* (Vol. 2, Chap. 13, pp. 571–620). Amsterdam: Elsevier.
- Record, S., Fitzpatrick, M. C., Finley, A. O., Veloz, S., & Ellison, A. M. (2013). Should species distribution models account for spatial autocorrelation? A test of model projections across eight millennia of climate change. *Global Ecology and Biogeography*, 22, 760–771.

- Ricardo, D. (1817). *The principles of political economy and taxation*. London: John Murray, Albemarle-Street.
- Sidharthan, R., & Bhat, C. R. (2012). Incorporating spatial dynamics and temporal dependency in land use change models. *Geographical Analysis*, 44(4), 321–349.
- Smirnov, O. A. (2010). Modeling spatial discrete choice. *Regional Science and Urban Economics*, 40, 292–298.
- Smith, T. E., & LeSage, J. P. (2004). A Bayesian probit model with spatial dependencies. *Advances in Econometrics*, 18.
- Stavins, R. N., & Jaffe, A. B. (1990). Unintended impacts of public investments on private decisions: The depletion of forested wetlands. *American Economic Review*, 80(3), 337–352.
- Timmins, C., & Schlenker, W. (2009). Reduced-form versus structural modeling in environmental and resource economics. *Annual Review of Resource Economics*, 1(1), 351–380.
- Verburg, P. H., Mertz, O., Erb, K.-H., Haberl, H., & Wu, W. (2013). Land system change and food security: Towards multi-scale land system solutions. *Current Opinion in Environmental Sustainability*, 5(5), 494–502.
- von Thunen, J. H. (1875). *Der isolierte staat in beziehung auf landwirthschaft und nationalokonomie*; hrsg. von h. schumacher-zarchlin.
- Wang, X., & Kockelman, K. M. (2009). Application of the dynamic spatial ordered probit model: Patterns of land development change in Austin, Texas. *Papers in Regional Science*, 88(2), 345–365.
- Wu, J., & Adams, R.-M. (2002). Micro versus macro acreage response models: Does site-specific information matter? *Journal-of-Agricultural-and-Resource-Economics*, 27(1), 40–60.
- Wu, J., & Segerson, K. (1995). The impact of policies and land characteristics on potential groundwater pollution in Wisconsin. *American Journal of Agricultural Economics*, 77(4), 1033–1047.
- Zellner, A., & Lee, T. (1965). Joint estimation of relationships involving discrete random variables. *Econometrica*, 33, 382–94.

Modeling Dependence in Spatio-Temporal Econometrics



Noel Cressie and Christopher K. Wikle

Abstract This chapter is concerned with lattice data that have a temporal label as well as a spatial label, where these spatio-temporal data appear in the “space-time cube” as a time series of spatial lattice (regular or irregular) processes. The spatio-temporal autoregressive (STAR) models have traditionally been used to model such data but, importantly, one should include a component of variation that models instantaneous spatial dependence as well. That is, the STAR model should include the spatial autoregressive (SAR) model as a subcomponent, for which we give a generic form. Perhaps more importantly, we illustrate how noisy and missing data can be accounted for by using the STAR-like models as process models, alongside a data model and potentially a parameter model, in a hierarchical statistical model (HM).

1 Introduction

Spatial Econometrics has its origins in the statistical modeling of data that are labeled with a spatial (regular or irregular) lattice and, hence, they fall under Tobler’s first law of geography (everything is related to everything else, but near things are more related than distant things; Tobler 1970). Spatial-econometric models were inspired by the autoregressive (AR) statistical models found in time series analysis, where the data are temporally labeled and things in the recent past are more related than things in the distant past. The area of study known as Econometrics has these AR (combined with moving average) models at its core. Spatial Econometrics has mimicked Econometrics with spatial autoregressive (SAR) models at its core (e.g., Anselin 1988; Arbia 2006). In this chapter, we consider the spatial and the temporal aspects together and

N. Cressie (✉)
University of Wollongong, Wollongong, Australia
e-mail: ncressie@uow.edu.au

C. K. Wikle
University of Missouri, Columbia, USA
e-mail: wiklec@missouri.edu

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_19

give spatio-temporal-econometric models based on spatio-temporal autoregressive moving average (STARMA) models that can be fitted to spatio-temporal data.

One might think that SAR models have very similar statistical properties to those of AR models. However, the time dimension is ordered, whereas the spatial dimension is not (unless one-dimensional space provides the spatial labels or a partial order is imposed on a high-dimensional space; see, e.g., Tjøstheim 1978; Cressie and Davidson 1998). While the SAR models of Spatial Econometrics can be defined analogously to the AR (temporal) models of Econometrics, some of their spatial-statistical dependence properties are quite different from those of their temporal counterparts. Furthermore, the notion of filtering out noise due to measurement error, which is common in signal processing, has not been given the emphasis it deserves in Spatial Econometrics. These and other issues will be discussed and extended to spatio-temporal-econometric models.

Consider now data with both a spatial label and a temporal label, where these spatio-temporal data appear in the “space-time cube,” as a time series of spatial lattice processes. The spatio-temporal autoregressive (STAR) models have traditionally been used but, importantly, one should include a component of variation that models instantaneous spatial dependence as well. That is, the STAR model should include the SAR model as a subcomponent. In this chapter, we give the generic form for such a spatio-temporal model. We also consider the fundamental problem of how to handle measurement error in the data as well as missing data, by introducing a data model along with the STAR model, which defines a hierarchical statistical model (HM).

This chapter is organized as follows. Section 2 motivates why space and time are important factors in any scientific investigation and why modeling statistical dependence is key when making inferences from spatio-temporal data. Section 3 develops the core statistical models of Spatio-Temporal Econometrics. Section 4 looks back at the evolution of Spatial Econometrics and notes how some key advances in spatial statistical modeling have been slow to take hold. Section 5 returns to the core spatio-temporal econometric models presented in Sect. 3 and gives a modern, HM approach to modeling spatio-temporal data on regular or irregular spatial lattices. Some general remarks are given in Sect. 6, and a brief technical appendix concludes the chapter.

2 Spatio-Temporal Statistics

Spatio-temporal data were essential to the nomadic tribes of early civilizations, who used them to return to seasonal hunting grounds. On a bigger scale, data sets on weather, geology, plants, animals, and indigenous people were collected by early explorers seeking to map and exploit new lands. In a sense, we are all analyzers of spatial and temporal data. As we plan our futures (economically, socially, educationally, etc.), we must take into account the present and seek guidance from the past. As we look at a map to plan a trip, we are letting its spatial abstraction guide us.

There is an important statistical characteristic of spatio-temporal data that is almost ubiquitous, namely that nearby (in space and time) observations tend to be more alike than those that are far apart. A simple, often-effective forecast of tomorrow's weather is to use today's observed weather. This "persistence" forecast is based on observing large autocorrelations between successive days. Such dependence behavior in "nearby" temporal data is also seen in "nearby" spatial data, such as in studies of the environment. Statistics for spatio-temporal data is challenging due to this dependence in time and space. One fundamental scientific problem that arises is understanding the evolution of spatial processes over time (e.g., the evolution of sea-ice coverage in the Arctic; sea surface temperature and the El Niño phenomenon; and time trends of precipitation in agricultural regions). Proper inference to determine if evolutionary components (natural or anthropogenic) are real requires a spatio-temporal *statistical* methodology.

The scientific method involves observation, inspiration, hypothesis-generation, experimentation (to support or refute the current scientific hypothesis), inference, more inspiration, more hypothesis-generation, and so forth. In a sense, everything begins with observation, but it is quickly apparent to a scientist that unless data are obtained in a more-or-less controlled manner (i.e., according to an experimental design), scientifically defensible inference can be challenging. Understanding the role of dependencies when the data are spatial or temporal or both, provides an important perspective when working with experimental data versus observational data.

It is our belief that statistical models used for describing temporal variability in space should represent the variability dynamically. Models used in Physics, Chemistry, Biology, Economics, etc., do this all the time with difference equations and differential equations to express the dynamical evolutionary mechanisms. Why should this change when the models become statistical? Perhaps it is because there is often an alternative framework, for example, a model based on correlations, that describes the spatio-temporal dependence. However, this descriptive approach does not directly involve evolutionary mechanisms and, as a consequence, it can push scientific understanding of the Physics/Chemistry/Biology/Economics/etc., into the background. There is in fact a way to have both, in the form of a scientific-statistical model that recognizes the dynamical scientific aspects of the phenomenon, with its uncertainties expressed through statistical models. Obviously, descriptive (correlational) statistical models have a role to play when little is known about the etiology of the phenomenon; however, when possible, we believe that one should use a dynamical statistical approach to model spatio-temporal data, such as the models given in Sect. 3.2.

2.1 *Uncertainty and Data*

Central to the observation, summarization, and inference (including prediction) of spatio-temporal processes are *data*. All data come bundled with error. In particular,

along with the obvious errors associated with measuring, manipulating, and archiving, there are other errors, such as discrete spatial and temporal sampling of an inherently continuous system. Consequently, there are always scales of variability that are unresolvable and that will further “contaminate” the data. For example, in Atmospheric Science, this is considered as a form of “turbulence,” and it corresponds to the well known aliasing problem in time series analysis (e.g., Chatfield 1989, p. 126) and the micro-scale component of the “nugget effect” in geostatistics (e.g., Cressie 1993, p. 59).

Furthermore, spatio-temporal data are rarely sampled at spatial or temporal locations that are optimal for the analysis of a specific scientific problem. For instance, in environmental studies there is often a bias in data-coverage toward areas where population density is large and, within a given area, the coverage is often limited by cost. Thus, the location of a measuring site and its temporal sampling frequency may have very little to do with the underlying scientific mechanisms. A scientific study should include the *design* of data locations and sampling frequencies when framing questions, when choosing statistical-analysis techniques, and when interpreting results. This task is complicated since the data are nearly always statistically dependent in space and time, and hence most of the traditional statistical methods taught in introductory statistics courses (which assume independent and identically distributed, or *iid*, errors) do not apply or have to be modified.

2.2 *Uncertainty and Models*

Science attempts to *explain* the world in which we live, but that world is very complex. A model is a simplification of some well chosen aspects of the world, where the level of complexity often depends on the question being asked. Pragmatically, the goal of a model is to predict and, at the same time, scientists want to incorporate their understanding of how the world works, into their models. For example, the motion of a pendulum can be modeled using Newton’s second law and the simple gravity pendulum that ignores the effect of friction and air resistance. The model predicts future locations of the pendulum quite well, with smaller-order modifications needed when the pendulum is used for precise time-keeping. Models that are scientifically meaningful, that predict well, and that are conceptually simple are generally preferred. However, an injudicious application of Occam’s razor (or “the law of parsimony”) might elevate simplicity over the other two criteria. For example, a statistical model based on correlational associations might be simpler than a model based on scientific theory.

The way to bridge this divide is to focus on what is more-or-less-certain in the scientific theory, and use *scientific-statistical* relationships to characterize it. In other words, we suggest that the uncertainties in the models be expressed probabilistically. As the data become more expansive, it is natural that they might suggest a more complex model. Clearly, there is a balance to be struck between too much simplicity, so failing to recognize an important signal in the data, and too much complexity,

which results in a non-existent signal being “discovered.” The research area known as *model choice* uses various criteria (e.g., AIC, DIC) to achieve this balance (e.g., Wikle et al. 2019, pp. 284–287).

2.3 Conditional Probabilities in a Hierarchical Statistical Model (HM)

There is a very general way to express uncertainties coming from different sources, through an approach known as hierarchical statistical modeling. There are data Z that measure Y (with measurement uncertainty), there is the scientific process Y (with less or more uncertainty), and there are parameters θ (unknown, not certain) that control the conditional probability distribution of Z given Y , and the probability distribution of Y . In this chapter, the quantities in which we are interested are random vectors and random variables.

The following conditional probabilities are the basic building blocks of a hierarchical statistical model (HM):

Data model : $[Z|Y, \theta]$

Process model : $[Y|\theta]$

where, using generic random quantities A and B , $[A]$ denotes the marginal distribution of A , $[A, B]$ denotes the joint distribution of A and B , and $[A|B]$ denotes the conditional distribution of A given B . Now the joint distribution of Z and Y can be decomposed as follows. From the equation $[A, B] = [A|B][B]$, we have

$$[Z, Y|\theta] = [Z|Y, \theta][Y|\theta], \tag{1}$$

which is simply a product of the data model and the process model.

In the HM above, it is assumed that θ is fixed (not random), and that all probability distributions are conditional on the fixed values of the parameters. Inference on Y depends on the following distribution (sometimes called the *predictive distribution*), obtained from Bayes’ Theorem:

$$[Y|Z, \theta] = \frac{[Z|Y, \theta][Y|\theta]}{[Z|\theta]}, \tag{2}$$

where the normalizing “constant,” $[Z|\theta] = \int [Z|Y, \theta][Y|\theta]dY$, ensures that the total probability of the predictive distribution is 1.

What can be done about θ ? A Bayesian approach would augment the HM with a *parameter model*, $[\theta]$, which is usually called the *prior*. In this chapter, we want to make an important point, that Spatial (and Spatio-Temporal) Econometrics does *not* need to adopt a Bayesian approach to use Bayes’ Theorem, given by (2), and to exploit the power of an HM. Henceforth in this chapter, we adopt a non-Bayesian approach and assume that θ is fixed but unknown, with some closing remarks about this given in Sect. 6.

In practice, θ is often specified using an estimate, in which case (2) is replaced with $[Y|Z, \hat{\theta}]$, where $\hat{\theta}$ is an estimate of θ (i.e., depends on the data Z). It is also possible that θ is estimated from an independent study or is simply an educated guess. It is this “empirical” step of “plugging in $\hat{\theta}$ ” that we shall adopt in this chapter. A fully Bayesian HM can be found in, for example, Wikle et al. (2019, pp. 168–170).

2.4 “Classical” Statistical Modeling

Here we use “classical” as an adjective for both frequentist and Bayesian modeling. The HM introduces data Z , process Y , and parameters θ ; however, the “classical” model found in the work of Fisher (e.g., Fisher 1935) has only data Z and parameters θ , as does the “classical” model of Bayes and many who followed him (e.g., Press 1989). Classical frequentists base their inferences on the *likelihood*, $[Z|\theta]$. Classical Bayesians base their inferences on the *posterior distribution*, $[\theta|Z]$, which requires both a likelihood $[Z|\theta]$ and a prior $[\theta]$ to be specified. Both classical approaches miss the fundamental importance of modeling the *latent process* Y , where the Physics/Chemistry/Biology/Economics/etc., typically resides.

To be sure, Statistics has played and continues to play an important role in Science, but often using simple, introductory-textbook approaches based on correlation and regression. Without Y being made explicit in statistical models, Science has often chosen its own path to statistical inference. Scientists know that parameters θ are important; these might be starting values, or boundary conditions, or diffusion constants, and so forth. In what follows, we give a deliberately simplistic description of how a traditional scientist might use Statistics in her/his research, although we note that in some disciplines this is changing fast. It is our hope that a more modern way of building statistical-dependence models will happen in Spatial Econometrics and in Spatio-Temporal Econometrics (presented in Sects. 4 and 5 of this chapter).

Scientific experiments produce data Z , and variability in the data is generally recognized by scientists. One approach to support, refine, or refute a scientific theory has been to “smooth” the data first. Consider the smoother f , and write

$$\tilde{Y} = f(Z).$$

The scientist might then assume that any (random) variability has been removed and that \tilde{Y} can now be treated as the true process with no uncertainty. A less extreme viewpoint would be to consider that \tilde{Y} is “close to” the true process Y . In that case, the scientist might fit a model for Y using the “data” \tilde{Y} . If the model for Y is $[Y|\theta_P]$, namely a process model with parameters θ_P that are a subset of θ , the scientist might use classical Statistics to fit $[Y|\theta_P]$ to \tilde{Y} . While the approach just described can be effective when the “signal” is strong, it also has the potential to declare the presence of a signal when it may simply be the result of chance fluctuations.

Given the data are to be smoothed, it should be recognized that they are often a combination of raw observations and algorithmic manipulations. The statistical scientist might write instead,

$$\tilde{Z} = f(Z), \tag{3}$$

where the notation \tilde{Z} in (3) is deliberate and suggests an important difference between the two ways to think about $f(Z)$.

An HM can be fitted using the data \tilde{Z} , where the data model, $[\tilde{Z}|Y, \theta]$, recognizes any remaining uncertainty in \tilde{Z} after smoothing. Inference on the process Y is based on the predictive distribution obtained from (2):

$$[Y|\tilde{Z}, \theta] \propto [\tilde{Z}|Y, \theta][Y|\theta], \tag{4}$$

where “ \propto ” means “is proportional to.” By writing the data manipulation and pre-processing according to (3), we have a coherent way to decompose the variability in \tilde{Z} through (4). (Bayesian statisticians would then specify a prior distribution $[\theta]$, but the ultimate goal of inference on Y and θ remains unchanged.)

While the picture painted above is simplistic, it does illustrate that scientific interest is in Y . If a classical frequentist statistician were to include the scientific model $[Y|\theta]$ in the analysis, it should be done in the calculation of the marginal model,

$$[\tilde{Z}|\theta] = \int [\tilde{Z}|Y, \theta][Y|\theta]dY.$$

That is, the classical frequentist who bases inference on the likelihood should recognize Y and then integrate it out. However, if there is no such recognition in the first place, the model chosen to be fitted, $[\tilde{Z}|\theta]$, may be difficult to interpret scientifically or, worse yet, may be inappropriately interpreted.

The classical Bayesian is also compromised; inclusion of the scientific model $[Y|\theta]$ yields the posterior distribution of θ ,

$$[\theta|\tilde{Z}] \propto \int [\tilde{Z}|Y, \theta][Y|\theta]dY \times [\theta].$$

This has the same potential for misinterpretation, if the Bayesian modeler tries to model directly $[\tilde{Z}|\theta]$ and uses it in $[\tilde{Z}|\theta] \times [\theta]$ to obtain the posterior distribution.

Spatial Econometrics has a tradition of fitting data directly to process models, and hence from the HM perspective it leaves the data model out of its formalism. As a result, variability due to measurement error is confounded with process-model error. That is, Spatial Econometrics has traditionally taken the classical frequentist approach to inference. In the next section, we concentrate on *process models* for processes indexed by both space and time and, in Sects. 4 and 5, we return to the HM where the data model is formulated along with the process model (and θ is estimated).

3 Spatio-Temporal-Econometric Modeling

There are a number of ways to express statistically that “things” nearby (in space and time) are more related than distant “things.” In this section, we illustrate the fundamental difference between space and time with a simple example, and then we show how dynamical spatio-temporal-econometric models can be built that capture the best features of Spatial Econometrics and multivariate time series analysis. In what follows, we let $Y_t(\mathbf{s})$ denote a random variable at spatial location \mathbf{s} and time t , and then we allow \mathbf{s} and t to vary over a spatio-temporal domain of interest.

3.1 Spatial Description and Temporal Dynamics: A Simple Example

The best way to compare space and time in our statistical context is to consider a simple example, where the spatial domain $D_s \equiv \{s_0, s_0 + \Delta, \dots, s_0 + 99\Delta\}$ is defined in one dimension, and the temporal domain $D_t \equiv \{0, 1, 2, \dots\}$ is defined on the nonnegative integers. Then let $\{Y_t(s) : s \in D_s, t \in D_t\}$ be a spatio-temporal process of interest; recall that in the space-time cube, fixing $t = t_0$ yields a spatial process and fixing $s = s_0$ yields a time series.

Define the spatial process at the fixed time point t_0 to be the 100-dimensional vector,

$$\mathbf{Y}_{t_0} \equiv (Y_{t_0}(s_0), \dots, Y_{t_0}(s_0 + 99\Delta))',$$

and define the time series at fixed spatial location s_0 to be the (different) 100-dimensional vector,

$$\mathbf{Y}(s_0) \equiv (Y_{t_0}(s_0), \dots, Y_{t_0+99}(s_0))'.$$

For illustrative purposes, the dimension of these vectors were arbitrarily chosen to be 100. By comparing spatial statistical models for \mathbf{Y}_{t_0} and time series models for $\mathbf{Y}(s_0)$, we can see to what extent space is modeled differently from time. Note that we deliberately chose the dimensions of the vectors to be the same to make the comparison easier, but they need not be.

Let us consider the vector \mathbf{Y}_{t_0} . A simple departure from independence for a *spatial process* is nearest-neighbor dependence expressed through conditional distributions. Let $\text{Gau}(\mu, \sigma^2)$ denote a Gaussian distribution with mean μ and variance σ^2 . Assume, for $i = 1, \dots, 98$, the Gaussian (conditional) distribution,

$$\begin{aligned} & Y_{t_0}(s_i) | \{Y_{t_0}(s_j) : j = 0, \dots, 99 \text{ and } j \neq i\} \\ & \sim \text{Gau}((\phi_{t_0}/(1 + \phi_{t_0}^2))\{Y_{t_0}(s_{i-1}) + Y_{t_0}(s_{i+1})\}, \sigma_{t_0}^2/(1 + \phi_{t_0}^2)), \end{aligned} \quad (5)$$

where $s_i \equiv s_0 + i\Delta$; $i = 0, \dots, 99$. On the edges of the transect, assume

$$\begin{aligned}
 Y_{t_0}(s_0) | \{Y_{t_0}(s_j) : j = 1, \dots, 99\} &\sim \text{Gau}(\phi_{t_0} Y_{t_0}(s_1), \sigma_{t_0}^2), \\
 Y_{t_0}(s_{99}) | \{Y_{t_0}(s_j) : j = 0, \dots, 98\} &\sim \text{Gau}(\phi_{t_0} Y_{t_0}(s_{98}), \sigma_{t_0}^2).
 \end{aligned}$$

In (5), assume that the *spatial-dependence parameter*, ϕ_{t_0} , satisfies $|\phi_{t_0}| \leq 1$. Based on these assumptions, it can be shown that $E(\mathbf{Y}_{t_0}) = \mathbf{0}$, and the correlation between nearest neighbors is

$$\text{corr}(Y_{t_0}(s_i), Y_{t_0}(s_{i-1})) = \phi_{t_0}; \quad i = 1, \dots, 99. \tag{6}$$

The process given by (6) is *descriptive* in that it is given simply in terms of correlation.

Let us now consider the vector $\mathbf{Y}(s_0)$. A simple departure from independence for a *time series* is a first-order autoregressive process. Assume that

$$Y_t(s_0) = \phi(s_0) Y_{t-1}(s_0) + \delta_t; \quad t = t_0 + 1, \dots, t_0 + 99, \tag{7}$$

where δ_t is independent of $Y_{t-1}(s_0)$, and the elements of $\{\delta_t\}$ are *iid* as $\text{Gau}(0, \sigma_\delta^2(s_0))$, for $t = t_0, t_0 + 1, \dots, t_0 + 99$. To initialize the process, assume

$$Y_{t_0}(s_0) \sim \text{Gau}(0, \sigma_\delta^2(s_0)/(1 - \phi(s_0)^2)),$$

which is a deliberate choice, as is assuming that the *temporal-dependence parameter* $\phi(s_0)$ satisfies $|\phi(s_0)| < 1$. Based on these assumptions, it can be shown that $E(\mathbf{Y}(s_0)) = \mathbf{0}$, $\text{var}(Y_t(s_0))$ does not depend on t , and the correlation between two adjacent time points is:

$$\text{corr}(Y_{t-1}(s_0), Y_t(s_0)) = \phi(s_0); \quad t = t_0 + 1, \dots, t_0 + 99. \tag{8}$$

The dependence in the process given by (7) is *dynamical* in that it shows how current values are related mechanistically to past values. More generally, the dependence of current values on past values can be expressed probabilistically, and (7) has an equivalent probabilistic expression in terms of the conditional probability of $Y_t(s_0)$ given past values:

$$Y_t(s_0) | Y_{t-1}(s_0), \dots, Y_{t_0}(s_0) \sim \text{Gau}(\phi(s_0) Y_{t-1}(s_0), \sigma_\delta^2(s_0)).$$

Such time series models are sometimes referred to as causal.

Let us compare and contrast the spatial process (5) and the time series (7). Both are Gaussian with mean zero. From (6) and (8), we see that if $\phi_{t_0} = \phi(s_0)$, they imply the *same* correlation between adjacent random variables. In fact, because of the Gaussian assumption, if the temporal-dependence and the spatial-dependence variance-covariance parameters are equal, the processes are probabilistically identical! However, the spatial process (5) looks east and west for dependence, in contrast to the time series (7), which is causal and looks to the past. This example has a cautionary aspect. Clearly, a description of the properties of spatial or temporal statistical

dependence of the model through just moments or even through joint probability distributions can completely miss the genesis of the statistical dependence, such as the dynamical structure given by (7).

Now, when it comes to considering space and time together in $\{Y_t(\mathbf{s})\}$, we believe that (whenever possible) the temporal dependence should be expressed dynamically, based on Physical/Chemical/Biological/Economic/etc., considerations, since here the etiology of the phenomenon is clearest. In a contribution to the Statistics literature that was well ahead of its time, Hotelling (1927) gave various statistical analyses based on dynamical models from stochastic differential equations (albeit only for the temporal dimension).

This *dynamical* approach to spatio-temporal statistical modeling contrasts to that of some others, where time is treated as an extra (although different) dimension. In that case, *descriptive* expressions of spatial dependencies through covariance functions are modified to account for the additional temporal dimension. We call this expression descriptive because usually it is not accompanied by an explanation of why the temporal dependence is present.

3.2 Time Series of Spatial Processes

In Spatio-Temporal Econometrics, a generic spatio-temporal process Y is

$$\{Y_t(\mathbf{s}_i) : i = 1, \dots, n; t = 0, 1, \dots\}$$

and, for the moment, we can imagine that $Y_t(\cdot)$ is observed at every one of the n spatial locations for all t . We write the spatial process at time t as the vector,

$$\mathbf{Y}_t \equiv (Y_t(\mathbf{s}_1), \dots, Y_t(\mathbf{s}_n))'; \quad t = 0, 1, \dots$$

Hence, the original spatio-temporal process can be written as the *multivariate time series*,

$$\mathbf{Y}_0, \mathbf{Y}_1, \dots$$

In Spatial Econometrics, the spatial statistical modeling of an individual \mathbf{Y}_t has been largely based on SAR models (see below), although CAR models are equally appropriate (e.g., Allcroft and Glasbey 2003).

The vector notation enables us to express the Markov property for $\{\mathbf{Y}_t\}$ succinctly as,

$$[\mathbf{Y}_t | \mathbf{Y}_0, \dots, \mathbf{Y}_{t-1}] = [\mathbf{Y}_t | \mathbf{Y}_{t-1}]; \quad t = 1, 2, \dots$$

An example of a process satisfying the Markov property is the VAR(1) model of dimension n :

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t \tag{9}$$

where, in its full generality, \mathbf{M} has n^2 parameters, and $\Sigma_\eta \equiv \text{var}(\boldsymbol{\eta}_t)$ has $O(n^2)$ parameters. However, the spatial context can be used to reduce the number of parameters drastically.

For example, suppose we assume that the (i, j) th entry of \mathbf{M} equals 0, unless $\|s_i - s_j\| \leq h$, for a given $h > 0$. Then the current value at s_i is related to those immediate-past values at s_i and nearby values at s_j (within a radius of h). Thus, rather than \mathbf{M} being made up of n^2 parameters, the parameter space can be made $O(n)$ by making \mathbf{M} sparse through spatial proximities of the n locations. A similar modeling strategy that allows further reduction in the size of the parameter space would choose Σ_η to be sparse (a geostatistical-type spatial model) or Σ_η^{-1} to be sparse (a lattice-type spatial model).

The VAR(1) model is a special case of the spatio-temporal autoregressive moving average (STARMA) models. It is generally true that for these and other multivariate time series, the number of parameters can be enormous, and an important skill of the modeler is to reduce drastically the size of the parameter space. We believe that this is best achieved through recognizing and preserving any known spatio-temporal interactions in the underlying process $\{Y_t(\mathbf{s})\}$.

3.3 Space-Time Autoregressive Moving Average (STARMA) Models

We could look for even more generality than a VAR(1) model in the temporal domain, by assuming higher orders of autoregression as well as a moving average type of dependence. Define the *spatio-temporal autoregressive moving average (STARMA)* models (Ali 1979; Pfeifer and Deutch 1980; Cressie 1993, p. 450) as

$$\mathbf{Y}_t = \sum_{k=0}^p \left(\sum_{j=1}^{\lambda_k} f_{kj} \mathbf{U}_{kj} \right) \mathbf{Y}_{t-k} + \sum_{l=0}^q \left(\sum_{j=1}^{\mu_l} g_{lj} \mathbf{V}_{lj} \right) \boldsymbol{\omega}_{t-l}; \quad t = 0, 1, \dots,$$

where $\{\mathbf{U}_{kj}\}$ and $\{\mathbf{V}_{lj}\}$ are known weight matrices; p and q are the orders of the autoregressive part and the moving average part, respectively; $\{f_{kj}\}$ and $\{g_{lj}\}$ are parameters of the model; $\{\boldsymbol{\omega}_t\}$ are iid random vectors with mean $\mathbf{0}$ and covariance matrix Σ_ω ; and the index j is used to denote substructures. These are core models in Spatio-Temporal Econometrics.

Under reparameterization, we obtain

$$\mathbf{Y}_t = \sum_{k=0}^p \mathbf{B}_k \mathbf{Y}_{t-k} + \sum_{l=0}^q \mathbf{E}_l \boldsymbol{\omega}_{t-l} \tag{10}$$

where, without loss of generality, we henceforth put $\Sigma_\omega = \sigma_\omega^2 \mathbf{I}$ and, for identifiability reasons, \mathbf{B}_0 has zero entries down the diagonal. It is important to note that the index

k in (10) starts at $k = 0$; the matrix \mathbf{B}_0 models instantaneous spatial dependence in the same way that spatial dependence is modeled in a SAR model. As for the SAR model, we assume that $(\mathbf{I} - \mathbf{B}_0)$ is invertible.

The number of parameters in (10) is still very large. Consider several simple cases. First, $p = 0$ and $q = 0$ results in a time series of purely spatial processes without any temporal dependence linking them:

$$\mathbf{Y}_t = \mathbf{B}_0 \mathbf{Y}_t + \mathbf{E}_0 \boldsymbol{\omega}_t; \quad t = 0, 1, \dots$$

To see this clearly, rewrite the expression above as

$$\mathbf{Y}_t = (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{E}_0 \boldsymbol{\omega}_t; \quad t = 0, 1, \dots$$

and, since $\{\boldsymbol{\omega}_0, \boldsymbol{\omega}_1, \dots\}$ are mutually independent, we see that the time series $\{\mathbf{Y}_t\}$ defined just above has no temporal dependence. When $\mathbf{E}_0 = \mathbf{I}$, the multivariate time series consists of *iid* mean-zero SARs.

The second case is $p = 1$ and $q = 0$, and recall that \mathbf{B}_0 has all-zero diagonal entries. Then,

$$\mathbf{Y}_t = \mathbf{B}_0 \mathbf{Y}_t + \mathbf{B}_1 \mathbf{Y}_{t-1} + \mathbf{E}_0 \boldsymbol{\omega}_t; \quad t = 0, 1, \dots$$

Given \mathbf{Y}_{t-1} , the vector \mathbf{Y}_t has spatial statistical dependence that is expressed in the form of a SAR model. From Cressie (1993, p. 409), a SAR can be written as a CAR, which is a Markov random field with simple *conditional* probability dependencies. The equation just above can be written equivalently as

$$\mathbf{Y}_t = (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{B}_1 \mathbf{Y}_{t-1} + (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{E}_0 \boldsymbol{\omega}_t \equiv \mathbf{M} \mathbf{Y}_{t-1} + \boldsymbol{\eta}_t,$$

where $\mathbf{M} \equiv (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{B}_1$ and $\{\boldsymbol{\eta}_t\}$ are *iid* with mean zero and $\text{var}(\boldsymbol{\eta}_t) = \boldsymbol{\Sigma}_\eta = \sigma_\omega^2 (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{E}_0 \mathbf{E}_0' (\mathbf{I} - \mathbf{B}_0')^{-1}$. This is a VAR(1) model, and recall that the matrix \mathbf{B}_0 represents “instantaneous” spatial dependence. Notice that if we multiply out $(\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{B}_1$, where $(\mathbf{I} - \mathbf{B}_0)$ is sparse, we obtain a propagator matrix $\mathbf{M} = (\mathbf{I} - \mathbf{B}_0)^{-1} \mathbf{B}_1$ that is generally *not* sparse.

Another way to achieve a VAR(1) model is the third case, $p = 1, q = 0$, and $\mathbf{B}_0 \equiv \mathbf{0}$. Then,

$$\mathbf{Y}_t = \mathbf{B}_1 \mathbf{Y}_{t-1} + \mathbf{E}_0 \boldsymbol{\omega}_t; \quad t = 0, 1, \dots,$$

which is equivalent to

$$\mathbf{Y}_t = \mathbf{M} \mathbf{Y}_{t-1} + \boldsymbol{\eta}_t,$$

where now $\mathbf{M} \equiv \mathbf{B}_1$, and $\{\boldsymbol{\eta}_t\}$ are *iid* with mean zero and $\text{var}(\boldsymbol{\eta}_t) = \boldsymbol{\Sigma}_\eta = \sigma_\omega^2 \mathbf{E}_0 \mathbf{E}_0'$.

There are clearly a number of different ways to arrive at the same type of model. The difference between them lies in their parameterizations. One way to think of \mathbf{B}_0 is that it captures the variability at time steps much smaller than the unit of time specified for the autoregression. Small-temporal-scale dynamics, which may be important and unwise to ignore, are collected together into the matrix \mathbf{B}_0 that models

instantaneous spatial dependence (Cressie 1993, p. 450; LeSage and Pace 2009, Sect. 2.1). Thus, this instantaneous spatial dependence is in fact an approximation of dynamical structure running at time scales much shorter than the unit of time in the autoregression.

4 Spatial-Econometric Modeling

We saw in Sect. 3.1 that a spatial Gaussian process in one-dimensional space that is described through its covariance function can be probabilistically equivalent to a corresponding temporal process (i.e., a time series) that is modeled dynamically through an autoregressive mechanism. Then in Sect. 3.3, we generalized the autoregressive model by collecting all the spatial-process values into a vector, resulting in a very flexible class of multivariate dynamical models for spatio-temporal processes.

Spatial Econometrics grew out of seeing how dependence was modeled in time in Econometrics. This was achieved through Box–Jenkins ARIMA modeling (Box and Jenkins 1970) and the use of “backshift” operators, and then by applying the same idea with “spatial-shift” matrices to generate dependence in space (Paelinck and Klaasen 1979). For example, the mean-zero AR(1) model for the time series $\{Y_t\}$ is defined as $Y_t = \phi Y_{t-1} + \delta_t$, where Y_{t-1} is independent of δ_t , and $\{\delta_1, \delta_2, \dots\}$ are *iid* with $E(\delta_t) = 0$ and $\text{var}(\delta_t) = \sigma_\delta^2$ (see Eq. (7)). This equation can be written equivalently in terms of the backshift operator B as

$$Y_t = \phi B Y_t + \delta_t. \tag{11}$$

At the core of Spatial Econometrics are models for $\mathbf{Y} \equiv (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))'$ that mechanistically connect $Y(\mathbf{s}_i)$ to its “neighbors”: Replace ϕB in (11) with the square-matrix operator \mathbf{B}_0 whose diagonal elements are defined to be zero, and any off-diagonal element that is zero indicates a lack of spatial “connection” between the two corresponding locations. The resulting SAR model is,

$$\mathbf{Y} = \mathbf{B}_0 \mathbf{Y} + \boldsymbol{\omega}, \tag{12}$$

where $E(\boldsymbol{\omega}) = \mathbf{0}$ and $\text{var}(\boldsymbol{\omega}) = \sigma_\omega^2 \mathbf{I}$, which was introduced in Sect. 3.3 as a way to capture instantaneous spatial dependence in a mean-zero spatio-temporal process.

If we write $\mathbf{B}_0 = \phi \mathbf{B}$, where \mathbf{B} is the square matrix (b_{ij}) , then the generalization from “time” in (11) to “space” in (12) looks beguilingly straightforward. However, these are mathematical relationships, and nothing has been said yet about the statistical dependence between $\mathbf{B}_0 \mathbf{Y}$ and $\boldsymbol{\omega}$ in (12). Recall that in the AR(1) process given by (11), $Y_{t-1} (= B Y_t)$ and δ_t are independent. In the SAR process given by (12), $\mathbf{B}_0 \mathbf{Y} = \mathbf{B}_0 (\mathbf{I} - \mathbf{B}_0)^{-1} \boldsymbol{\omega}$, and hence $\text{cov}(\mathbf{B}_0 \mathbf{Y}, \boldsymbol{\omega}) = \mathbf{B}_0 (\mathbf{I} - \mathbf{B}_0)^{-1} \text{var}(\boldsymbol{\omega}) = \sigma_\omega^2 \mathbf{B}_0 (\mathbf{I} - \mathbf{B}_0)^{-1}$, which shows that $\mathbf{B}_0 \mathbf{Y}$ and $\boldsymbol{\omega}$ are statistically dependent.

This latter property means one has to be very careful when interpreting the SAR model. It has been misinterpreted as being causal in Spatial Econometrics; Gibbons

and Overman (2012) address this mistake directly, and the presence of non-zero covariances between the autoregressive part, $\mathbf{B}_0\mathbf{Y}$, and the error, $\boldsymbol{\omega}$, is a manifestation of the fundamentally different structure of the SAR model and the AR model, which is causal.

For $\mathbf{B}_0 = \phi\mathbf{B}$, (12) can be written as

$$Y(\mathbf{s}_i) = \phi \sum_{j=1}^n b_{ij} Y(\mathbf{s}_j) + \omega(\mathbf{s}_i); \quad i = 1, \dots, n,$$

where recall $b_{ii} = 0$. In a naive cross-validation exercise, $Y(\mathbf{s}_i)$ would be deleted and then predicted with $\hat{Y}(\mathbf{s}_i) \equiv \phi \sum_{j=1}^n b_{ij} Y(\mathbf{s}_j)$; then $\hat{Y}(\mathbf{s}_i)$ would be compared to $Y(\mathbf{s}_i)$ via, say, $(\hat{Y}(\mathbf{s}_i) - Y(\mathbf{s}_i))^2$. However, this $\hat{Y}(\mathbf{s}_i)$ is an inferior predictor of $Y(\mathbf{s}_i)$, since the optimal cross-validation predictor of $Y(\mathbf{s}_i)$ is,

$$Y^*(\mathbf{s}_i) \equiv E(Y(\mathbf{s}_i)|\mathbf{Y}_{-i}),$$

for \mathbf{Y}_{-i} the $(n - 1)$ -dimensional vector with $Y(\mathbf{s}_i)$ removed from \mathbf{Y} . From the Lemma given in the Appendix, $Y^*(\mathbf{s}_i)$ can be derived analytically from the full $n \times n$ covariance matrix, $\text{var}(\mathbf{Y}) = \sigma_\omega^2\{(\mathbf{I} - \phi\mathbf{B})(\mathbf{I} - \phi\mathbf{B}')\}^{-1}$, and it is different from $\hat{Y}(\mathbf{s}_i)$.

Note that while a derivation of $Y^*(\mathbf{s}_i)$, albeit straightforward and resulting in a closed-form expression, is necessary for the SAR model, $Y^*(\mathbf{s}_i)$ is immediately available from the conditional autoregressive (CAR) model, although this model is used much less frequently in Spatial Econometrics. (For readers interested in the relationships between SAR and CAR models, see Cressie 1993, p. 408–410, Ver Hoef et al. 2018.)

Another caution with the use of SAR models in Spatial Econometrics comes with how they are specified when the spatial process \mathbf{Y} does *not* have mean zero. One should take guidance from how the time series model (11) would be modified to handle, say, the regression, $E(Y_t) = \mathbf{x}'_t\boldsymbol{\beta}$. The time series model,

$$Y_t - \mathbf{x}'_t\boldsymbol{\beta} = \phi \cdot (Y_{t-1} - \mathbf{x}'_{t-1}\boldsymbol{\beta}) + \delta_t, \tag{13}$$

is an AR(1) process that preserves the mean structure, $E(Y_t) = \mathbf{x}'_t\boldsymbol{\beta}$. For reasons that are not clear, the Spatial-Econometrics literature (e.g., Anselin 1988) shows a preference to include the regression term, $\mathbf{X}\boldsymbol{\beta}$, and the spatial-dependence operator \mathbf{B}_0 in its core model as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}_0\mathbf{Y} + \boldsymbol{\omega}, \tag{14}$$

where $\boldsymbol{\omega} \equiv (\omega(\mathbf{s}_1), \dots, \omega(\mathbf{s}_n))'$ represents model error with $E(\boldsymbol{\omega}) = \mathbf{0}$.

As a consequence of (14), $E(\mathbf{Y}) = (\mathbf{I} - \mathbf{B}_0)^{-1}\mathbf{X}\boldsymbol{\beta}$, which results in the confounding of large-scale regression effects $\boldsymbol{\beta}$ with small-scale spatial-dependence effects \mathbf{B}_0 . This can be avoided by taking a cue from the time series model (13). That is, to generalize the SAR model to include regression, we write

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{B}_0(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\omega}.$$

Now $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{var}(\mathbf{Y}) = \sigma_\omega^2\{\mathbf{I} - \mathbf{B}_0\}^{-1}$, and hence $\boldsymbol{\beta}$ appears only in $E(\mathbf{Y})$, and \mathbf{B}_0 appears only in $\text{var}(\mathbf{Y})$. There is an equivalent way to write this model, namely

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}, \text{ and } \mathbf{U} = \mathbf{B}_0\mathbf{U} + \boldsymbol{\omega}, \tag{15}$$

which does appear in the more recent Spatial-Econometrics literature and is called a spatial error model (e.g., LeSage and Pace 2009, Sect. 2.3). Our point is that software based on the model (14) should not be used when fitting spatial statistical models to spatial data with covariates \mathbf{X} , due to the confounding of large-scale and small-scale effects and the consequent misinterpretation of a fitted model (14).

More generally, confounding between fixed effects and spatial random effects has become an important topic in the spatial-statistics literature (e.g., Reich et al. 2006; Paciorek 2010; Hodges and Reich 2010; Hughes and Haran 2013; Hanks et al. 2015). There is still some uncertainty as to what extent these models are able to account for confounding; appropriate mitigation approaches depend on the underlying dependence structure of the random effects, the extent to which covariates are known, and the spatial support (Hanks et al. 2015). Geographers and spatial econometricians have been aware of spatial confounding for some time in the context of areal data, and they have provided ‘‘Moran’s I’’ eigenvector approaches that make the spatial random effects orthogonal to the fixed effects (e.g., Griffith 2000, 2003). Spatial statisticians have also considered Moran’s I basis functions and extensions in this context (Hughes and Haran 2013; Bradley et al. 2015). However, it is unclear how to force random effects to be in the space orthogonal to the fixed effects if the fixed effects have continuous support as they do in geostatistical models (Hanks et al. 2015). More recently, Bradley et al. (2020) considered confounding between the spatial process and the error process and showed that accounting for dependence between these two processes can improve prediction accuracy.

In Sect. 2.3, we made the point that observations (Z) on a process are different from the values of the process itself (Y). This is typically due to measurement error (‘‘noisiness’’), and it can also be due to gaps in the observations (‘‘missingness’’). This can be captured in a spatial statistical model by writing,

$$Z(\mathbf{s}_i) = Y(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i); \quad \mathbf{s}_i \in D^* \subset \{\mathbf{s}_1, \dots, \mathbf{s}_n\}. \tag{16}$$

In (16), $Z(\mathbf{s}_i)$ is an observation at spatial location \mathbf{s}_i in D^* ; data at locations not in D^* are considered as missing; and $\varepsilon(\cdot)$ is an independent measurement-error process with $\text{var}(\varepsilon(\mathbf{s}_i)) = \sigma_\varepsilon^2 > 0$. Goulaud et al. (2017) consider spatial-econometric models for missing data, but they do not recognize that the measurement-error component of variation $\varepsilon(\cdot)$ is different from the model-error component of variation $\omega(\cdot)$.

In the general case of non-zero mean due to regression effects, (15) is the *process model* that represents all components of $\mathbf{Y} \equiv (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))'$, even though some might not be observed, and (16) is the *data model* for data $\{Z(\mathbf{s}_i) : \mathbf{s}_i \in D^*\}$ that are observed. That is, in terms of the HM presented in Sect. 2.3, (16) defines $[Z|Y]$

and (15) defines $[Y]$, where dependence of these two models on parameters $\theta \equiv \{\boldsymbol{\beta}, \sigma_\omega^2, \sigma_\varepsilon^2\}$ has been dropped from the notation for ease of exposition. Specifically, the HM is:

Data model: $Z(\mathbf{s}_i)|Y(\mathbf{s}_i) \sim \text{Gau}(Y(\mathbf{s}_i), \sigma_\varepsilon^2)$, and define $\mathbf{Z} \equiv (Z(\mathbf{s}) : \mathbf{s} \in D^*)'$.

Process model: $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{B}_0(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\omega}$, where $\boldsymbol{\omega} \sim \text{Gau}(\mathbf{0}, \sigma_\omega^2\mathbf{I})$.

The data model and the process model together allow calculation of the predictive distribution: Since $[\mathbf{Z}|\mathbf{Y}]$ is Gaussian and $[\mathbf{Y}]$ is Gaussian, so too is the joint distribution $[\mathbf{Y}, \mathbf{Z}]$ the marginal distribution $[\mathbf{Z}]$, and most importantly the predictive distribution $[\mathbf{Y}|\mathbf{Z}]$. Hence, the key calculations for inference on \mathbf{Y} from the “imperfect” data \mathbf{Z} are the conditional moments,

$$E(Y(\mathbf{s}_i)|\mathbf{Z}), \text{ var}(Y(\mathbf{s}_i)|\mathbf{Z}), \text{ and } \text{cov}(Y(\mathbf{s}_i), Y(\mathbf{s}_j)|\mathbf{Z}), \text{ for } i, j = 1, \dots, n,$$

and recall that there are locations $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \setminus D^*$ at which there are no observations. These conditional moments are known in closed form from Bayes’ Theorem given by (2), where the distributions in the numerator of (2) are obtained from (15) and (16), and likelihood-based estimates of θ are used in place of θ . No time-consuming iterative algorithms are needed to calculate them; see the Lemma in the Appendix. The one bottleneck may be fast computation of $\text{var}(\mathbf{Z})^{-1}$ when the data set \mathbf{Z} is very large; see Burden et al. (2015) for a reduced-rank approach to this problem and a comparison to the Spatial-Econometrics literature where fast computation of $\text{var}(\mathbf{Y})^{-1}$ is the focus.

The lessons learned from this section are first to *de-trend* the spatio-temporal data using covariates and then to *use HMs* to capture the imperfections of noisy and missing data. The next section will apply these lessons to the spatio-temporal setting given in Sect. 3.

5 Modern Spatio-Temporal-Econometric Hierarchical Models

All the ideas and methodology that are needed have been presented in the preceding sections. It is simply a matter of tying them together now in a series of steps that bears a resemblance to pseudocode for algorithmic development.

Recall that the generic spatio-temporal data are Z , the generic underlying process being measured is Y , which represents the whole process $\{Y_t(\mathbf{s})\}$, and the generic parameters are θ . Due to incomplete data (“missingness”), Z will be of smaller dimension than Y , and the presence of measurement error (“noise”) results in the conditional distribution,

$$Z_t(\mathbf{s})|Y, \sigma_\varepsilon^2 \sim \text{Gau}(Y_t(\mathbf{s}), \sigma_\varepsilon^2),$$

provided an observation occurs at location \mathbf{s} and time t in the spatio-temporal domain of interest $\{\mathbf{s}_1, \dots, \mathbf{s}_n\} \times \{0, 1, \dots, T\}$.

The building blocks of dynamical models in Spatio-Temporal Econometrics are given below in a sequence of eight steps:

1. $[Z|Y, \theta] = \prod_{D^*} [Z_t(\mathbf{s})|Y, \sigma_\varepsilon^2]$, for D^* the set of all *spatio-temporal data locations*, is Gaussian.
2. $[Y|\theta]$ is a (high-dimensional) Gaussian distribution; see, for example, (10) or its modification that includes regression:

$$(\mathbf{Y}_t - \mathbf{X}_t\boldsymbol{\beta}) = \sum_{k=0}^p \mathbf{B}_k(\mathbf{Y}_{t-k} - \mathbf{X}_{t-k}\boldsymbol{\beta}) + \sum_{l=0}^q \mathbf{E}_l\boldsymbol{\omega}_{t-l},$$

for $t = p, p + 1, \dots, T$.

3. $[Z, Y|\theta] = [Z|Y, \theta][Y|\theta]$ is Gaussian (since 1. and 2. are Gaussian).
4. $L(\theta) \equiv [Z|\theta] = \int [Z, Y|\theta] dY$; θ includes σ_ε^2 , $\boldsymbol{\beta}$, the spatio-temporal-variation parameters in $\{\mathbf{B}_k\}$ and $\{\mathbf{E}_l\}$, and $\{\text{var}(\boldsymbol{\omega}_{t-l})\}$. Recall that $[Z|\theta]$ is Gaussian.
5. Estimate θ with $\hat{\theta} = \arg \sup_{\theta} L(\theta)$, the maximum likelihood estimator.
6. $[Y|Z, \hat{\theta}] = [Z|Y, \hat{\theta}][Y|\hat{\theta}]/[Z|\hat{\theta}]$ is a Gaussian distribution called the (empirical) predictive distribution.
7. $E(Y|Z, \hat{\theta})$ and $\text{var}(Y|Z, \hat{\theta})$ characterize the predictive distribution; both can be calculated straightforwardly in closed form, using the Lemma in the Appendix.
8. Estimation and prediction: Report and interpret $\hat{\theta}$ and its uncertainties (estimation). Make a choropleth map of $E(Y|Z, \hat{\theta})$, which is the HM’s spatio-temporal predictor of Y (prediction). Make a second choropleth map of the diagonal elements $(\text{diag}(\text{var}(Y|Z, \hat{\theta})))^{1/2}$, which uses the HM to quantify the uncertainty in the first map.

These are the basic steps taken to fit the dynamical spatio-temporal models given in Chap. 5 of Wikle et al. (2019): There, Sects. 5.2 and 5.3 are the most relevant to the development given in this chapter.

6 Concluding Remarks

We would like to expression our best wishes to Christine (Thomas-Agnan) on the occasion of her 65th birthday. She has been a gracious host and an engaging co-author during several long-terms visits by the first author to Université Toulouse 1 Capitole.

Our approach to the problem of “scientific understanding in the presence of uncertainty” takes a probabilistic viewpoint, which allows us to build useful spatio-temporal statistical models and make scientific inferences for various spatial and temporal scales. Accounting for the uncertainty enables us to look for possible associations within and between variables in the underlying scientific process, with the

potential for finding mechanisms that extend, modify, or even disprove a scientific theory. The dynamical spatio-temporal-econometric models described in this chapter are an important subset of a much larger class of dynamical HMs for the twenty-first century (Wikle et al. 2019). We have concentrated on HMs where the parameters θ are estimated from the data, which are called empirical HMs. Bayesian HMs arise when a prior, $[\theta]$, is assigned to the unknown parameters θ . In many cases, the predictive moments, $E(Y|Z)$ and $\text{var}(Y|Z)$, from the Bayesian HM are not available in closed form. Then sampling from the predictive distribution, $[Y|Z]$, is a way to solve this problem (e.g., using MCMC).

There are many challenges associated with building HMs and then carrying out valid inferences. A broad perspective is that there is subjectivity involved with the specification of *all* model components, specifically here the data model and the process model. However, it is not always clear what “subjective” means in this context. For example, it might be “subjective” to use deterministic relationships to motivate a stochastic model, such as for tropical winds (e.g., Wikle et al. 2001), yet the science upon which such a model is based comes from Newton’s laws of motion. Thus, we believe that it is not helpful to try to classify probability distributions that determine the statistical model, as subjective or objective. It would be better to ask about the sensitivity of inferences to model choices and whether such choices make sense scientifically.

Given that a modeler brings so much information to the table when developing models, the conditional probability framework presented earlier can be used to recognize that this information, say I , is part of what is involved in the conditioning. For the HM, we have

$$[Y|Z, \theta, I] \propto [Z|Y, \theta, I][Y|\theta, I].$$

A major challenge in this paradigm is, to the extent possible, acknowledgement of the importance of this information, I . It is often the case that a team of researchers at the table has a collective “ I ” that is better quantified and more appropriate than any individual’s “ I .”

In the HM approach, there are certainly cases where models have to be simplified due to practical concerns. Perhaps the computational issues in a given formulation are limiting, which usually leads to a modification of the model. Such practical concerns apply to all statistical inferences in complicated modeling scenarios. This tension between the model you want and the model with which you can compute is healthy, and in modern statistical computing it has led to algorithms that only *approximate* valid inferences. However, user beware! Approximations to approximations can lead to a serious propagation of errors.

Data hold so much potential, but unless they can be organized into a database they are an entropic collection of digits or bits. With the ability in a database to structure, search, filter, query, visualize, and summarize, the data begin to contain *information*. Some of this information comes from judicious use of statistics (i.e., summaries). Then, in going from information to *knowledge*, Science (and, with it, Statistical Science) takes over. Statistical Science makes contributions at all levels

of the data-information-knowledge pyramid, but it has often stopped short of the summit where knowledge is used to determine policy. At the interface between Science, Statistics, and Policy, there is an enormous need for decision-making in the presence of uncertainty.

Finally, it is the responsibility of the research team to temper the tendency to fit ever-more-complicated models, and to use model-selection criteria (e.g., AIC, BIC, DIC, etc.) that concentrate on the twin pillars of predictability and parsimony (e.g., Spiegelhalter et al. 2002; Wikle et al. 2019). But these criteria do not address the third pillar, namely scientific interpretability (i.e., knowledge). Our approach to spatio-temporal-econometric modeling is to use the hierarchical-modeling paradigm and, where possible, choose statistical models based on this third pillar, while not ignoring the other two.

Acknowledgements Material from Chaps. 1, 2, and 6 of Cressie and Wikle (2011) is used in Sect. 2 almost entirely, Sect. 3 almost entirely, and the last half of Sect. 6, with permission from the publisher: Copyright ©2011 by John Wiley & Sons, Inc. All rights reserved. Cressie’s research was supported by Australian Research Council Discovery Project DP190100180. Wikle’s research was supported by U.S. National Science Foundation grants SES-1853096 and DMS-1811745.

Appendix

Throughout the chapter, we have referred to the predictive distribution $[Y|Z]$ that arises from a joint Gaussian distribution, $[Y, Z]$. Specifically, we have claimed that $[Y|Z]$ is Gaussian and the first two moments can be obtained analytically without resort to iteration, simulation, or approximation. This claim is due to the following lemma from multivariate analysis (e.g., Rencher and Christensen 2012, p. 97).

Lemma *Consider the Gaussian random vector, $U \equiv (U_1', U_2')'$, and its first two moments:*

$$E(U) \equiv \mu \equiv (\mu_1', \mu_2')', \text{ and } var(U) \equiv \Sigma \equiv \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Then the conditional distribution, $[U_1|U_2]$ is also Gaussian with mean vector,

$$E(U_1|U_2) = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(U_2 - \mu_2)$$

and variance-covariance matrix,

$$var(U_1|U_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

References

- Ali, M. M. (1979). Analysis of stationary spatial-temporal processes: Estimation and prediction. *Biometrika*, 66, 513–518.
- Allcroft, D. J., & Glasbey, C. A. (2003). A latent Gaussian Markov random-field model for spatio-temporal rainfall disaggregation. *Journal of the Royal Statistical Society, Series C*, 52, 487–498.
- Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht: Kluwer.
- Arbia, G. (2006). *Spatial econometrics: Statistical foundations and applications to regional convergence*. Berlin: Springer.
- Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis, forecasting, and control*. Oakland: Holden-Day.
- Bradley, J. R., Holan, S. H., & Wikle, C. K. (2015). Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *Annals of Applied Statistics*, 9, 1761–1791.
- Bradley, J. R., Wikle, C. K., & Holan, S. H. (2020). Hierarchical models for spatial data with errors that are correlated with the latent process. *Statistica Sinica*, 30, 81–109.
- Burden, S., Cressie, N., & Steel, D. G. (2015). The SAR model for very large datasets: A reduced-rank approach. *Econometrics*, 3, 317–338.
- Chatfield, C. (1989). *The analysis of time series: An introduction* (4th ed.). London: Chapman and Hall.
- Cressie, N. (1993). *Statistics for spatial data* (rev. edn). New York: Wiley.
- Cressie, N., & Davidson, J. L. (1998). Image analysis with partially ordered Markov models. *Computational Statistics and Data Analysis*, 29, 1–26.
- Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Hoboken: Wiley.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Gibbons, S., & Overman, H. G. (2012). Mostly pointless spatial econometrics. *Journal of Regional Science*, 52, 172–191.
- Goulard, M., Laurent, T., & Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost-optimal strategies. *Spatial Economic Analysis*, 12, 304–325.
- Griffith, D. A. (2000). A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems*, 2(2), 141–156.
- Griffith, D. A. (2003). *Spatial autocorrelation and spatial filtering: Gaining understanding through theory and scientific visualization*. Berlin: Springer.
- Hanks, E. M., Schliep, E. M., Hooten, M. B., & Hoeting, J. A. (2015). Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26, 243–254.
- Hodges, J. S., & Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64, 325–334.
- Hotelling, H. (1927). Differential equations subject to error and population estimates. *Journal of the American Statistical Association*, 22, 283–314.
- Hughes, J., & Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B*, 75, 139–159.
- LeSage, J., & Pace, R. L. (2009). *Introduction to spatial econometrics*. Boca Raton: Chapman Hall/CRC Press.
- Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25, 107–125.
- Paelinck, J. H. P., & Klaasen, L. H. (1979). *Spatial econometrics*. Farnborough: Saxon House.
- Pfeifer, P. E., & Deutch, S. J. (1980). A three-stage iterative procedure for space-time modeling. *Technometrics*, 22, 35–47.
- Press, S. J. (1989). *Bayesian statistics: Principles, models, and applications*. New York: Wiley.
- Reich, B. J., Hodges, J. S., & Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62, 1197–1206.

- Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (3rd ed.). Hoboken: Wiley.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, *64*, 583–639.
- Tjostheim, D. (1978). Statistical spatial series modelling. *Advances in Applied Probability*, *10*, 130–154.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*, 234–240.
- Ver Hoef, J., Hanks, E., & Hooten, M. (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spatial Statistics*, *25*, 68–85.
- Wikle, C. K., Milliff, R. F., Nychka, D., & Berliner, L. M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association*, *96*, 382–397.
- Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019). *Spatio-temporal statistics with R*. Boca Raton: Chapman & Hall/CRC Press.

Guidelines on Areal Interpolation Methods



Van Huyen Do, Thibault Laurent, and Anne Vanhems

Abstract The objective of this article is to delve deeper into the understanding and practical implementation of classical areal interpolation methods using **R** software. Based on a survey paper from Do et al. (Spat Stat 14:412–438, 2015), we focus on four classical methods used in the area-to-area interpolation problem: point-in-polygon, areal weighting interpolation, dasymetric method with auxiliary variable and dasymetric method with control zones. Using the departmental election database for Toulouse in 2015, we find that the point-in-polygon method can be applied if the sources are much smaller than the targets; the areal interpolation method provides good results if the variable of interest is related to the area, but otherwise, a good alternative is to use the dasymetric method with another auxiliary variable; and finally, the dasymetric method with control zones allows us to benefit from both areal interpolation and dasymetric method and, from that perspective, seems to be the best method.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-73249-3_20) contains supplementary material, which is available to authorized users.

V. H. Do

Independent researcher, Toulouse School of Economics, CNRS, University of Toulouse, 1 Esplanade de l'université, 31080 cedex 06 Toulouse, France
e-mail: huyendvmath@gmail.com

T. Laurent

Toulouse School of Economics, CNRS, University of Toulouse, 1 Esplanade de l'université, 31068 cedex 06 Toulouse, France
e-mail: thibault.laurent@tse-fr.eu

A. Vanhems (✉)

TBS Business School, 1 place Alphonse Jourdain, 31068 Toulouse, France
e-mail: a.vanhems@tbs-education.fr

1 Introduction

1.1 Motivation

When working with spatial data, one often faces a situation where several datasets are independently collected by various organizations with different objectives.

To simultaneously use those spatially incompatible datasets, one needs to merge them on one spatial support. This type of problem is called the areal interpolation problem. Areal interpolation methods are variously applicable in socioeconomics (Goodchild et al. 1993), satellite imagery (Fisher and Langford 1996), GIS (Flowerdew et al. 1991; Flowerdew and Green 1993), political science (Grasland et al. 2000), population dynamics (Gregory 2002), epidemiology (Kelsall and Wakefield 2002) and many other fields. Readers are referred to Van Huyen et al. (2015) for an overview of areal interpolation methods and Van Huyen et al. (2015) for a theoretical comparison of the accuracy of these methods.

In this chapter, we consider four classic methods widely used in the areal interpolation problem: point-in-polygon, areal weighting interpolation, dasymetric method with auxiliary variable and dasymetric method with control zones, and we provide suggestions and advice concerning practical questions such as spatial scales, types of target variable and border incompatibility. The implementation is performed with **R** software (R Core Team 2020). Most of the functions used are included in the **sf** package (Pebesma 2018). All **R** codes, graphs and tables are gathered in a supplementary material resource which is available online (see <http://www.thibault.laurent.free.fr/code/areal>).

1.2 Context

The database we consider is the 2015 departmental elections in France, and our objective is to use socio-demographic covariates to explain the extreme right party score after the first round of the election.

The election takes place based upon geographic division. We consider 2054 zones, and voters vote at polling places. The election results are released by the Ministry of Interior in open access at the polling scales. However, the socio-demographic covariates we use are provided by the French statistical institute INSEE at various scales: cells with different sizes (since 2019), iris (a subdivision of communes), communes, etc. These scales are different from the polling places.

Therefore, INSEE data needs to be transferred into the form of polling places. To do this, one needs to use statistical methods to estimate the variables of interest at the polling places given information for the cells (or iris) or given auxiliary information for some additional zones, called control zones. In addition to the socio-demographic covariates provided by INSEE, we also use the network structure of roads as auxiliary information to improve our estimation. This auxiliary information is available at very

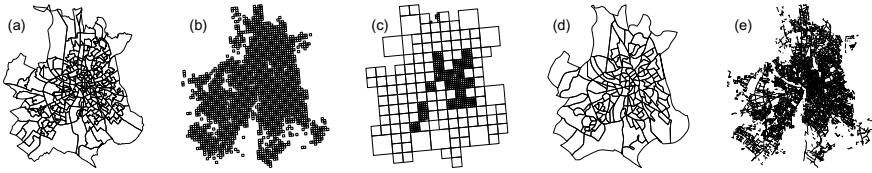


Fig. 1 From the left to the right: the polling places (a), the cells of small size (b), the cells of big size (c), the iris (d) and the network structure of the roads (e)

fine scale, which is very beneficial in our case. Figure 1 provides an overview of different spatial scales (from the left to the right): the polling places, the small cells, the large cells, the iris and the network structure of the roads (auxiliary information).

Nguyen and Laurent (2019) explain the 2015 departmental election’s extreme right votes through some socio-economic covariates with Compositional Data Regression Analysis. They use data from different open sources like the French Statistical Institute (INSEE) or Open Street Map (OSM) on various spatial scales. To merge those spatially incompatible data on the polling scale, the authors primarily utilize the areal weighting interpolation method.

In this article, however, we aim to present a broader set of area-to-area interpolation methods and apply them to the same dataset. Our article will be presented in the following order: Sect. 2 sets the classical notations of area-to-area change of the support problem; Sect. 3 presents the data; each of the four methods is then presented in Sects. 4, 5, 6 and 7 with a focus on practical issues. Finally, in Sect. 8, we present the results of a regression analysis in order to explain extreme right votes with respect to some socio-demographic characteristics (all variables being interpolated at the same geographical scale).

2 Notations

In this section, we briefly recall some classical definitions and notations used in the area-to-area interpolation literature.

The variable of interest that needs to be interpolated is called the *target variable* Y . In our setting, we consider only quantitative target variables. The value of the target variable Y for a given subzone A is denoted by Y_A . The objective is then to transfer the data available for a set of *source zones* to an independent set of *target zones*. In what follows, we denote by $S_s, s = 1, \dots, S$ the set of source zones and $T_t, t = 1, \dots, T$ the set of target zones. To simplify the notations, Y_s will denote the value of Y for the source S_s and Y_t denotes the value of Y for the target T_t .

The source zones and target zones are sometimes *nested* (see for instance, Do et al. 2014), but in general they are not and we will encounter boundary issues. We will offer some practical recommendations to address the issues through our case study.

We denote by $A_{s,t}$ the intersection zone between S_s and T_t , and $Y_{s,t}$ denotes the value of Y on $A_{s,t}$. We also denote by $|A|$ the area of a subregion A .

Let us now recall the difference between *extensive* and *intensive* variables. Consider a region Ω and Ω_k , $k = 1, \dots, p$, a partition of Ω . An extensive variable Y is such that its value for a region Ω is the sum of its values for each subzone Ω_k : $Y_\Omega = \sum_k Y_{\Omega_k}$. Any count variable is an extensive variable, such as population count or number of households. A variable is called *intensive* if $Y_\Omega = \sum_k w_k Y_{\Omega_k}$ with a set of weights w_k that are not all equal to 1. Proportions and rates are intensive variables. Note that it is possible to associate an intensive variable to a given extensive variable and the reverse (see, Do et al. 2015, for a detailed analysis of the transformation).

3 Data

As mentioned above, we study the extreme right vote of the 2015 French departmental election in Toulouse, which is available for polling places.

However, some covariates are not available at the polling scale. In rural cities where population density is large enough, a polling place might coincide with a commune, and the covariates provided by INSEE are available at the polling scale. However, in urban cities, polling places are often larger in order to contain enough voters. These polling place boundaries are defined by street layout. This geographical scale is never used by any other public administration. In this case, we need to interpolate data from INSEE scale to polling scale.

3.1 Target Zones

Our targets are polling places of the 2015 French departmental elections. Those boundaries are obtained from maps from the CarTElec project (Beauguitte et al. 2012). We only focus on Toulouse, but our codes could be used for any other regions.

There are $T = 256$ targets (polling places), and they vary greatly in size (see Table 1 in the supplementary material). Targets in the centre with denser population are smaller than those in the suburbs (see Fig. 1a). The smallest target area is 23 846 m², whereas the largest is 8 164 842 m². The latter is included in a low-density population industrial zone. We will see that one method may be more appropriate than another depending on the target area.

We then associate the election results to targets (the codes and data are available in our supplementary material). Two variables are available with respect to targets (polling places): percentage of extreme right vote (dependent variable in regression model at the end of the article) and percentage of turnout (covariate in our regression model). Summary statistics (Table 2), maps and scatter are in the supplementary material.

In the following subsections, we present the different geographical scales used as source zones.

3.2 *First Source Scale: The Cells*

The cell scale was introduced by INSEE in 2016. This technique consists in partitioning the territory into tiles to disseminate statistical information at a weakly aggregated scale. The INSEE data are obtained thanks to the income tax files (see the supplementary material to obtain the data).

There are two different cell scales:

1. The finest scale is the cell of dimension $200\text{ m} \times 200\text{ m}$ (Fig. 1b). There are $S = 2\,027$ small cells. All cells are approximately equal to $40\,000\text{ m}^2$ (see Table 1 in the supplementary material). Noise is added to data because of confidential issues.
2. The second scale is more aggregated (Fig. 1c) so that those cells have enough inhabitants to solve the confidential issues. There are $S = 591$ cells of 3 different sizes (473 of $200\text{ m} \times 200\text{ m}$, 109 of $1000\text{ m} \times 1000\text{ m}$ and 9 of $2000\text{ m} \times 2000\text{ m}$).

Note that the two datasets are not nested. Indeed, the total area covered by the small cells is $81\,132\,570\text{ m}^2$, approximately half of that covered by the large cells: $164\,026\,135\text{ m}^2$. The total number of inhabitants in the small cells is 404 497, whereas it is 457 031 for the large cells.

? Which scale of sources should we choose?

Theoretically, the finer the sources are, the better areal interpolation is. However, data for small cells can be noisy. Moreover, finding data at a very detailed scale is sometimes difficult. In our case, we will compare the usage of both kinds of sources (small cells and large cells) through four considered methods.

? What if sources and targets are not nested?

Sources and targets should cover the same region, i.e. the total area defined by the sources should coincide with that defined by the targets. In our case, the zones defined by the targets (polling places) are included in the city of Toulouse (voters have to register in a commune), whereas the INSEE data at cell scales are independent of the administrative boundary. A cell can cover one or several communes (cities), i.e. the sources and the targets are not nested. This situation requires some modifications to achieve a better interpolation.

Variables of interest (19 *target variables*) provided by INSEE are presented in Table 3 in the supplementary material. Note that all of those variables are *extensive*. Nguyen and Laurent (2019) also study the extreme right vote, but primarily use *intensive* variables (such as the unemployment rate and the proportion of people who own assets). Most of those intensive variables can be considered as a ratio of two extensive variables. In this work, we will first interpolate the extensive variables into targets (polling places), then calculate the intensive variables based on those extensive estimates.

However, many intensive variables are sometimes introduced without any information on their underlying extensive variables. To compare two scenarios, we consider two intensive variables (the proportion of inhabitants under 18 years old and the population density) and work directly with them without intermediate calculation. We are then able to compare both scenarios: one using directly intensive variables, and the other calculating the intensive variables through their underlying extensive variables. Summary statistics on these two intensive variables are presented in Table 4 in the supplementary material.

3.3 *Second Source Scale: The Iris*

This dataset contains only a few variables, and we would like to use additional covariates to better explain the extreme right vote: for example, the unemployment rate, or the proportion of foreign or immigrated people.

The finest scale that we can find in Open Access is the iris scale (the source of the data can be found in the supplementary material), which is a subdivision of a commune. The variables of interest are presented in Table 5 in the supplementary material. As for the previous dataset, these variables are also all *extensive*.

In this case, the number of sources is $S = 153$. Those sources are in general larger than the targets. This case is called the *disaggregation* problem. The variables of interest should be disaggregated first on the intersection zones and finally aggregated on the targets. In the case where sources are smaller than targets, the areal interpolation problem is called the *aggregation* problem.

3.4 *Variables to Estimate*

For modelling the extreme right vote, in line with the work of Nguyen and Laurent (2019) and with the INSEE dataset, we use the covariates presented in Table 6 in the supplementary material. All of these intensive covariates are ratios of extensive variables. Our approach consists of two steps: first interpolate the corresponding extensive variables and then calculate the (intensive) variables of interest.

4 Point-in-Polygon Method

In the Point-In-Polygon (*PIP*) method, the sources are points, such as postal addresses. If one source is a polygon, it is represented by a point. One could choose, for example, the centroid of the polygon. If the source is a cell, as is the case in our application, we could replace the centroid by one of the vertices of the cell. All sources points located in a target will be aggregated to the target (a graphical illustration is presented in the supplementary material).

This method is not costly from a computational point of view. Indeed, computing the intersection between a point and a polygon is much less demanding than building new geometries (the intersection zones) between two polygons.

4.1 Extensive Variables

In the case of extensive variables, the aggregation simply consists in summing up the values of all points located in the same target.

4.1.1 Border Effects

It may be the case that a point is not located in any target due to border effects (see the supplementary material for an illustration), especially when targets are not nested in sources. In our large cell case, this situation occurs frequently. There are two options:

1. **exclude:** The user decides that these points should not be included in any target. In that case, the sum of the extensive variables for the targets should be lower than the sum for the sources.
2. **include:** The user decides that all points should be associated with some target. It is possible to use the nearest neighbour algorithm to associate a point with its closest target. In this second case, we use the following two steps:
 - a. Identify the points which are not located in any target.
 - b. Use an appropriate **R** function to detect the closest neighbour and add the values to the corresponding targets.

In the supplementary material, we calculate the total number of inhabitants (variable **Ind**) at source scale and at target scale using the two methods proposed above. When sources are small cells, the total number of individuals for the sources is 404 497. In the **include** case, all inhabitants from the sources have been affected to the targets. In the **exclude** case, the estimated number of inhabitants for targets is 403 729. The difference between the two methods is not significant. The results are different in the large cell setting. Those numbers are 457 031 and 409 717, respectively. When a large cell is excluded, the magnitude is more important.

> Recommendation

The **include** method allows us to maintain the total value observed for the sources. It could be desirable when sources and targets cover the same population of interest. In our case, the sources are independent of the communes. For example, a source shared between Toulouse and another city contains the information of people in this area, no matter the city in which they live. Hence, we choose the **exclude** method to estimate the covariates used in our regression model.

4.1.2 Comparison Between Small and Big Cells Sources

For PIP, the smaller the sources, the better the estimation. Indeed, if a source is much smaller than a target, the probability that a source is completely included in a target is high. The variable of interest's value will be fully assigned to the targets.

In our application, the estimates with the small cells do seem to be better than those with the large cells: 28% of the small cells are fully included in the targets versus only 6% for the large cells. The scatter plot in Fig. 2 shows that the difference between using small and large cells can be huge. For example, there are three targets with estimated number of inhabitants larger than 7 500 when using large cells, whereas those values are smaller than 4 000 when using the small cells.

Moreover, the larger the sources, the larger the number of non-estimated targets. There are 57 unestimated targets in the case of large cells (grey in Fig. 2) versus only 2 for the small cells. This is caused by the fact that when a source overlaps several targets, its value is assigned to only one target, and some targets are left without any value assigned.

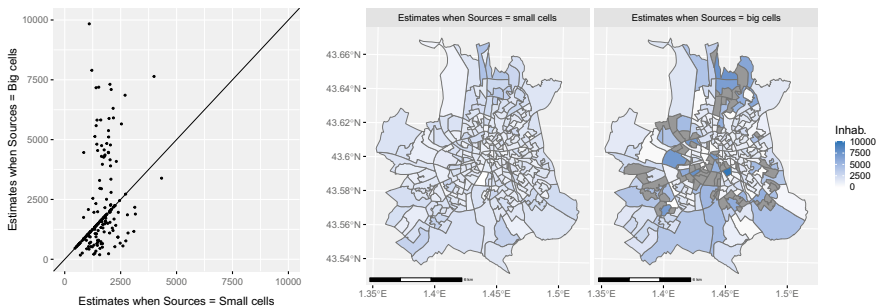


Fig. 2 From the left to the right: scatter plot and mapping of the *PIP* estimation of the number of individuals when the sources are the small size cells versus the big size cells

> **Important**

Unfortunately, we are in an unsupervised situation, and we are not able to compare our estimations with the true values.

4.2 Intensive Variables

If an intensive variable is defined by two known extensive variables, we strongly recommend following two steps: estimating the extensive variables first, then computing the intensive variables afterward. If those extensive variables are unknown, one possible solution is averaging all values of points located in the target. This solution is based on the assumption that all points have the same weight, which is rarely the case. For example, we consider a target containing a 1 000-inhabitant point with 10% of inhabitants under 18 and a 10-inhabitant point with 50% of inhabitants under 18: the estimate of the under 18 proportion with the same weights is 30% instead of the real value 10.4%.

Figure 3 shows a comparison of the under 18 proportion’s estimation in two scenarios: known and unknown underlying extensive variables. Even though those estimates are highly correlated in both scenarios, the small cell case exhibits some more substantial differences between two estimates. Indeed, there are often more points located in a target in the case of small cells, which makes the weight issue more serious. If only one source point is located in a target, both methods are equivalent.

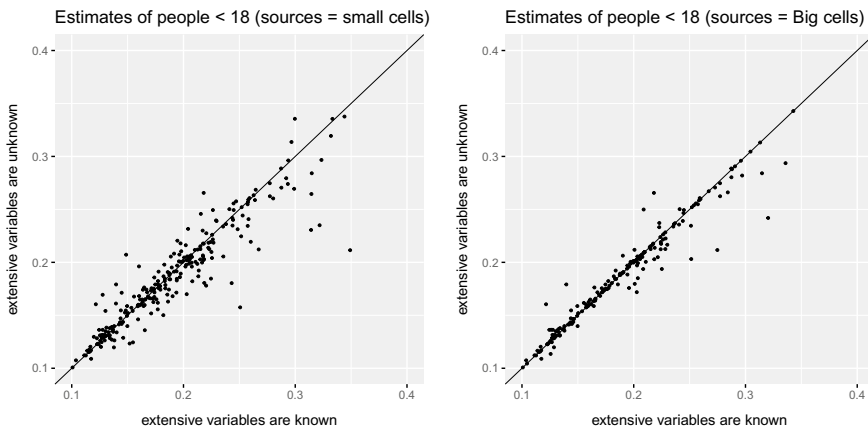


Fig. 3 Comparison of the *PIP* estimates depending on whether the intensive variable is defined as the ratio of two extensive variables or not; on the left, the sources are small cells and on the right, the sources are big cells

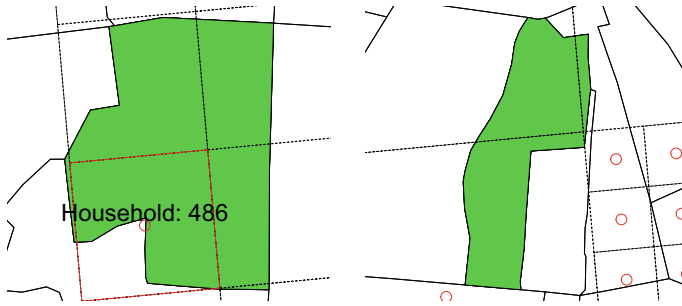


Fig. 4 On the left, the source (in red) allocates nothing to the target (in green) although it is mainly included inside it. On the right, the target (in green) does not receive anything from the sources as no centroid is located into it

4.3 Limitation of the Point-in Polygon Method

PIP is computationally inexpensive but has some limitations in our area-to-area case, i.e. the sources are polygons and not points. An example is illustrated in Fig. 4. The figure on the left shows a source (in red) which mostly overlaps with the target (in green), but its centroid is not located in the target: its 486 households are assigned to another target. In the case of comparable size between targets and sources, other methods, for example, areal weighting or dasymetric methods (see Sects. 5 and 6), should be more appropriate. In the figure on the right, no red point is located in the target (in green). Indeed, with those large cells, sources are sometimes larger than the target. This leads to an unestimated target here. In our application, the number of unestimated targets is 57 (resp. 2) when using large (resp. small) cells as sources.

> Recommendation

The *PIP* method seems to be efficient if and only if sources are much smaller than targets, like in the small cell case. Besides, the border issue should be taken into account when sources and targets do not coincide. At last, the method can be applied to both extensive and intensive variables. If the intensive variables are defined using observable extensive ones, it is better to estimate the extensive variables first and then calculate the intensive variables.

5 Areal Weighting Interpolation Method

When sources are polygons, PIP, where a source's value is presented by only one representative point, often exhibits many disadvantages. We should look for more suitable methods for our case of area-to-area interpolation. If we can reasonably

assume that our variable of interest is homogeneous for sources, then we can use the areal weighting interpolation (*DAW*) method (Goodchild and Lam 1980). This method can be applied to both extensive and intensive variables.

5.1 Extensive Variable

For an extensive variable, the *DAW* method’s homogeneity assumption is that $|Y_A|$ is proportional to the area $|A|$. The interpolation formula is defined by

$$\hat{Y}_t = \sum_s \hat{Y}_{s,t} = \sum_s \frac{|A_{s,t}|}{|S_s|} Y_s \tag{1}$$

From a computational point of view, one needs to compute the area of the intersection zones $\{A_{st}\}$ between the sources $\{S_s\}$ and the targets $\{T_t\}$.

Considering again the case in which one source overlaps at least two targets, *DAW* will disaggregate the variable’s value between the two targets proportionally to the area of the intersected zones. For example, if the source has 486 households, 75% of the source overlaps with Target 1 and the remaining part 25% overlaps with Target 2, and then 486×0.75 households will be assigned to Target 1 and 486×0.25 to Target 2.

5.1.1 Border Effect

As in PIP, the issue of border effect still occurs. There are two scenarios: the first case is when a source covers a not-target zone, and the second case is when a target is not completely covered by sources.

For the first case, we consider an example: if 30% of a source intersects with Target 1, 40% with Target 2, and 30% with an empty zone, how should we disaggregate the variable value $x = 100$ inhabitants? There are two possibilities:

- **exclude:** 30 inhabitants are allocated to Target 1, 40 inhabitants are allocated to Target 2 and 30 inhabitants are not allocated. In that case, the sum of values x for the targets will be lower than for the sources.
- **include:** $\frac{30}{30+40} \times 100 = 42.86$ inhabitants are allocated to Target 1 and $\frac{40}{30+40} \times 100 = 57.14$ inhabitants are allocated to Target 2.

In our application, we still use the **exclude** option because this is more coherent with our data. Indeed, our targets correspond only to the city of Toulouse, whereas the sources can overlap with several cities.

For the second case, we again consider an example: assume that a target is partially distant from all sources, i.e. $\sum_s |A_{s,t}| < |T_t|$. The target's estimate is therefore likely smaller than its true value. In this case, we also have two options:

- **First option:** If one believes that the off-zone (i.e. the part of the target area uncovered by sources) is unpopulated, the estimate might remain unchanged.
- **Second option:** If one believes that the off-zone is populated, the estimate should be modified. One possible modification is

$$\hat{Y}_t = \frac{|T_t|}{\sum_s |A_{s,t}|} \sum_s \frac{|A_{s,t}|}{|S_s|} Y_s \quad (2)$$

In our setting, if sources are small cells, then more targets are likely not to be fully covered by sources, and the second case might occur more frequently. The second option is then a possible solution. If sources are large cells, we might likely meet the first case, and as we discussed above, we can choose the 'exclude' method. However, because the large cells contain only 10% more population than the small cells, it is possible that the off-zones are sparsely inhabited.

➤ Recommendation

The performances of *DAW* using large cells and small cells are quite different (see the supplementary material for more details). Because *DAW* is based on a homogeneous assumption, the chance that this assumption is satisfied is higher for small zones than large zones. We therefore recommend the use of source zones with the most detailed geographical scale.

5.1.2 Comparison Between *PIP* and *DAW*

To compare *PIP* and *DAW*, we choose the variable number of households. Their estimates are presented in Fig. 5.

When using large cells as sources, the two methods perform very differently. Since *PIP* represents a source's scattered households into only one point, then locates the point to a single target, it makes estimation biased. In this case, *DAW* effectively corrects *PIP*'s weakness.

With small cells, the difference fades. Those small cells are likely to be nested into targets, so all households finally belong to only one target. Therefore, *PIP* and *DAW* perform similarly. In other words, the smaller sources are, the more similarly those two methods perform.

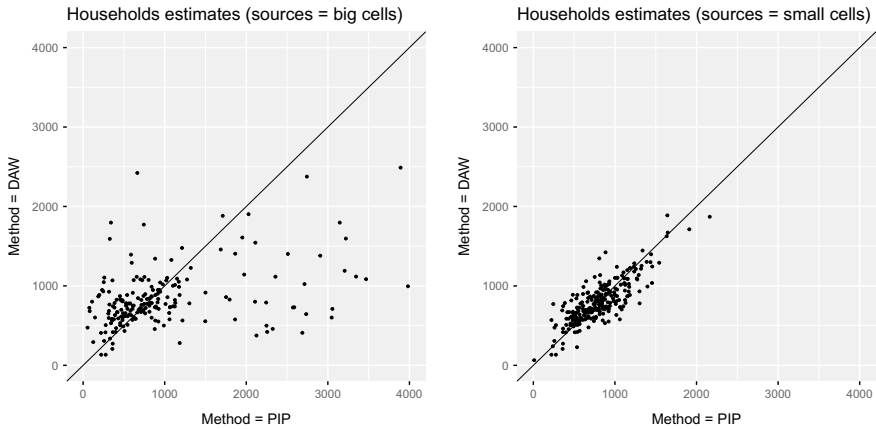


Fig. 5 Comparison between *DAW* and *PIP* estimates with the extensive variable number of households. On the left, the sources are big cells, on the right, the sources are small cells

5.2 Intensive Variable

5.2.1 Known Versus Unknown Underlying Extensive Variables

As we discussed above, if the underlying extensive variables defining an intensive variable are known, we recommend a two-step process: *DAW* is applied first on the extensive variables, then the intensive variable is computed based on those extensive estimates for target zones.

If those underlying extensive variables are unknown, we can still use *DAW*. *DAW*'s homogeneous assumption in this case is that Y is uniform for sources. In other words, we assume that $\hat{Y}_{s,t} = Y_s$. Then,

$$\hat{Y}_t = \sum \frac{|A_{s,t}|}{|T_t|} \hat{Y}_{s,t}. \tag{3}$$

5.2.2 Comparison of Both Cases

We compare hereafter the estimates in two scenarios: ‘extensive variables are known’ and ‘extensive variables are unknown’. The comparison is performed for two intensive variables, where one is directly related to area and the other is not. We choose population density (population/area) and proportion of inhabitants less than 18 years old, respectively. Figure 6a, b presents the results for the two variables under consideration.

We notice that the population density is clearly underestimated when applying the method without using underlying extensive variables. In fact, the estimate in this case is

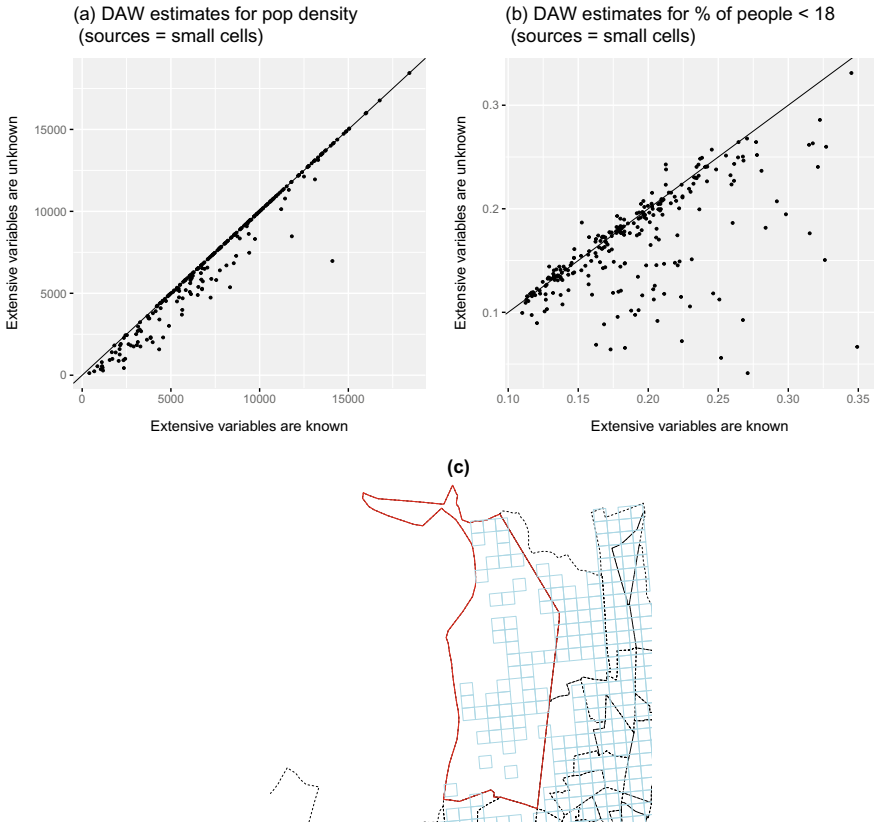


Fig. 6 From the left to the right: comparison of the *DAW* estimates for the intensive variables population density (a) and percentage of people with less than 18 y.o. (b). Example of underestimation (c): the non-intersected area between the sources in blue and the target in red is large

$$\hat{Y}_t = \frac{1}{|T_t|} \sum_s \frac{|A_{s,t}|}{|S_s|} X_s, \tag{4}$$

where X_s is the underlying extensive number of population variable. If we use two-step estimation, the result will be

$$\hat{Y}_t = \frac{1}{\sum_s |A_{s,t}|} \sum_s \frac{|A_{s,t}|}{|S_s|} X_s. \tag{5}$$

We encounter here again the second case of the border effect mentioned in Sect. 5.1.1. If a target is completely covered by sources, the two estimates are the same (points in the diagonal line); otherwise, points are under the line, which indicates underestimation.

The effect is a bit different for the variable percentage of inhabitants under 18. On one hand, there are points above the diagonal line, but some are far under it. We take a look into one of the most underestimated points, where the estimate is 34.9% in the ‘extensive variables are known’ scenario and 6.7% in the ‘extensive variables are unknown’ one. We represent in red this target in Fig. 6c.

It appears that the non-intersected zone area between sources and the target is large. In that case, \hat{Y}_t will be underestimated due to the total weights $\sum_s \frac{|A_{s,t}|}{|T_t|} < 1$. We therefore recommend replacing the term $|T_t|$ by $\sum_s |A_{s,t}|$, i.e.

$$\hat{Y}_t = \sum_s \frac{|A_{s,t}|}{\sum_s |A_{s,t}|} \hat{Y}_{s,t} \tag{6}$$

for intensive estimation.

To better understand why some points are above the diagonal line, we consider a simplified case: assume that one target T is built by two equal sources S_1, S_2 , i.e. $T = S_1 \cup S_2$. X, Z denote the number of inhabitants under 18 and population, respectively; X_1, X_2 are the numbers of inhabitants under 18 for zones S_1, S_2 . We use similar notation for population Z . The intensive variable in this case is $Y = \frac{X}{Z}$. It is easy to calculate that the difference of estimates by the two scenarios is $\frac{1}{2} \frac{(Z_1 - Z_2)}{Z_1 + Z_2} (\frac{X_1}{Z_1} - \frac{X_2}{Z_2})$. We can see that, unlike the population density variable, the proportion of inhabitants under 18 can be underestimated or overestimated depending on the sign of $(Z_1 - Z_2)(\frac{X_1}{Z_1} - \frac{X_2}{Z_2})$.

> Recommendation

DAW seems to correct many weaknesses of *PIP*. When the sizes of the sources are much smaller than the sizes of targets, *PIP* is still valid. However, when the sizes of sources and targets are comparable, *DAW* is much more precise. Note that this conclusion corresponds to the setting where the sources and targets are not nested. If the variable is extensive, we have decided not to take into account the part of the source which does not intersect with the target. This choice could be different if the user requires the sum of observed values for the sources to be equal to the sum of the estimates obtained for the targets. If one part of a target is not covered by any source, we recommend modifying the initial estimate by multiplying a correction ratio $\frac{|T_t|}{\sum_s |A_{s,t}|}$. When the variable is intensive, the user should rather first estimate the extensive variables defining the intensive variable. If this is not possible, we again advise users to use the correction ratio above. Obviously, the main drawback of the *DAW* method is that the target variable values should be related to the area. The purpose of the method presented in the next section is to relax this requirement and instead use auxiliary information.

6 Dasymetric Method with Auxiliary Variable X

The class of dasymetric methods (DAX) comprises generalizations of areal weighting methods. In order to improve upon areal weighting, the idea is to remove the assumption of the count density being uniform throughout the source zones because this assumption is almost never accurate.

It is instead assumed that the target variable is proportional to some auxiliary information for any subregion. In order to apply this method, the user needs to have information regarding some auxiliary variable X at the source scale and at the intersection scale.

However, it might be difficult to find such an auxiliary variable. For example, in our application, it seems very difficult to obtain socio-economic variables at the intersection scale. The national institute INSEE has individual data concerning housing (see the supplementary material) that could be used to obtain information at the intersection scale, but it is not open access.

One potential source of data which is easily accessible at the intersection scale is road data or Open Street Map data. These data are not areal: the information is given for points or lines, but it is easily interpolated into any zone scale. If the data are points, the user can obtain the information at the intersection scale by using the PIP method. For the road data, we simply compute the length of roads belonging to the intersection zones.

Note that by using this source of data, we make the assumption that the INSEE variables are correlated with the density of the roads. In other words, the higher the concentration of roads is, the greater population the zone has.

6.1 Extensive Variables

Because of the correlation assumption between our target variable and the auxiliary information, the estimator is

$$\hat{Y}_t = \sum_s \hat{Y}_{s,t} = \sum_s \frac{X_{s,t}}{X_s} Y_s \quad (7)$$

From a programming point of view, the DAX method is very close to the DAW method. The areas in DAW formulae are replaced by the auxiliary variable information. The computational time required to obtain the intersections between the roads and the intersected zones A_{st} is demanding, but it can be performed within a reasonable time (few seconds).

Here, we do not take into account the portion of the sources which does not intersect with any target (in other words, we consider the case **exclude** presented above).

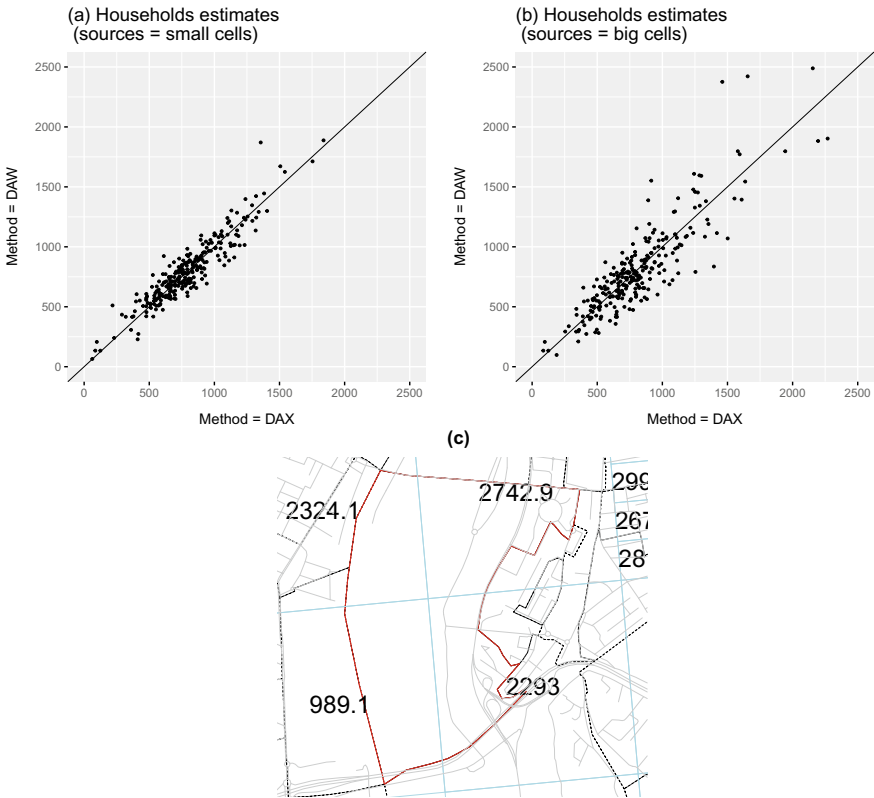


Fig. 7 Comparison of the *DAX* estimates for the extensive variable number of household in the case of small cells sources (a) and big cells sources (b). Example of target underestimated with the *DAX* method compared to *DAW* (c)

We compare the results obtained with *DAX* and *DAW* methods for number of households. In the case where small cells have been taken as sources (Fig. 7a), *DAW* and *DAX* seem to provide similar results.

In the case where large cells are taken as sources (Fig. 7b), it appears that the variability between the two methods is larger. We attempt to better understand the difference between the two methods. We consider the target with the largest difference between *DAX* and *DAW* estimates (Fig. 7c). The value estimated for this target (represented in red) is 1 460 with *DAX* and 2 376 with *DAW*. We observe that the density of the roads is very low in the target, whereas the area covered by the target is significant. This explains why the value is underestimated with *DAX* as compared to the *DAW* method.

6.2 Intensive Variables

If the extensive variables which define the intensive variables are known, the *DAX* method first interpolates those extensive variables into targets, and then intensive variables are computed from the extensive estimates.

If those extensive variables are otherwise unknown, it works similarly as the *DAW* method, except that we replace the area by the auxiliary information. The formula is

$$\hat{Y}_t = \sum \frac{X_{s,t}}{\sum_s X_{st}} Y_s. \quad (8)$$

Note that we choose $\sum_s X_{st}$ instead of X_t for a similar reason as in the *DAW* section. In the supplementary material, we compare the estimates using both cases: ‘extensive variables are known’ and ‘extensive variables are unknown’. When the sources are the small cells, we note that the two cases are quite similar. When the sources are the large cells, both cases also fit quite well.

➤ Recommendation

In our application, the *DAX* and *DAW* methods seem to produce very similar results. The choice of one method depends on the information available. If the variables are related to the area, the *DAW* method should be precise enough. If this is not the case, and if the user has additional auxiliary information related to the target variables, then the *DAX* method should be more efficient. Moreover, in the case of extensive variables, the smaller the sources are, the more similarly *DAW* and *DAX* exhibit. One reason for this is that the auxiliary information is more homogeneous for small zones than large zones, which makes replacing area by the auxiliary information less significant.

7 Dasymetric Method with Control Zones

7.1 Presentation of the Method

DAX needs the auxiliary information available on intersection scale, which is sometimes challenging. The dasymetric method with control zones (*DAC* hereafter) is used to relieve this constraint. It allows us to work on any auxiliary information for some new set of zones called control zones.

In this section, we consider the iris as the sources, whereas the targets are still the polling places and the control zones are the cell data at the smallest scale. The auxiliary information in this case is the ‘number of inhabitants’ which is available for small cells. We use the two-step *DAC* algorithm which consists of

1. Step 1 aims to prepare for *DAX* in step 2. To use *DAX*, one needs auxiliary information available on the intersection scale. One solution is using *DAW*. The variable of interest in the step is the ‘number of inhabitants’. Practically, we first define the geometries of the intersections between iris and polling places, then apply *DAW* by considering the small cells as sources and the intersections as targets.
2. Thanks to the first step, we now have the auxiliary information ‘the number of inhabitants’ at intersection scale. An aggregation step calculates the number of inhabitants on iris scale (sources). We hence can apply *DAX* for our target variable from sources (iris) to targets (the polling places).

7.2 Comparison Between DAC and DAX

In Fig. 8a, we present results of *DAC* and *DAX* for the variable ‘number of unemployed people’. We observe some targets with very different estimates. We examine what is happening for one of those targets. This target is represented in red in Fig. 8b. The estimated value with *DAX* is 674, and 197 with *DAC*. In *DAX*, the main contribution to this value comes from the source represented in purple, which shares 82% of the roads (in grey) with the target.

What happens with the *DAC* 2-step method? We remark in Fig. 8c that the intersected zone between the target in red and the source in purple has a sparse population, since it contains few control zones (in light blue). In this case, the source will attribute a large portion of its value to the non-intersected zone with the target, which is a denser zone.

➤ Recommendation

When some control zones can be used, allowing us to obtain auxiliary information (supposed to be related to the target variables) at a more detailed geographical scale than the sources, the 2-step *DAC* likely improves the estimation with regard to the *DAX* or *DAW* methods. This is why we have selected this method to estimate the covariates used in our regression model.

8 Regression Modelling

8.1 Covariates and Exploratory Analysis

To obtain the covariates presented in Table 6 in the supplementary material, we first estimate the extensive variables and then compute the ratios to obtain the intensive estimates.

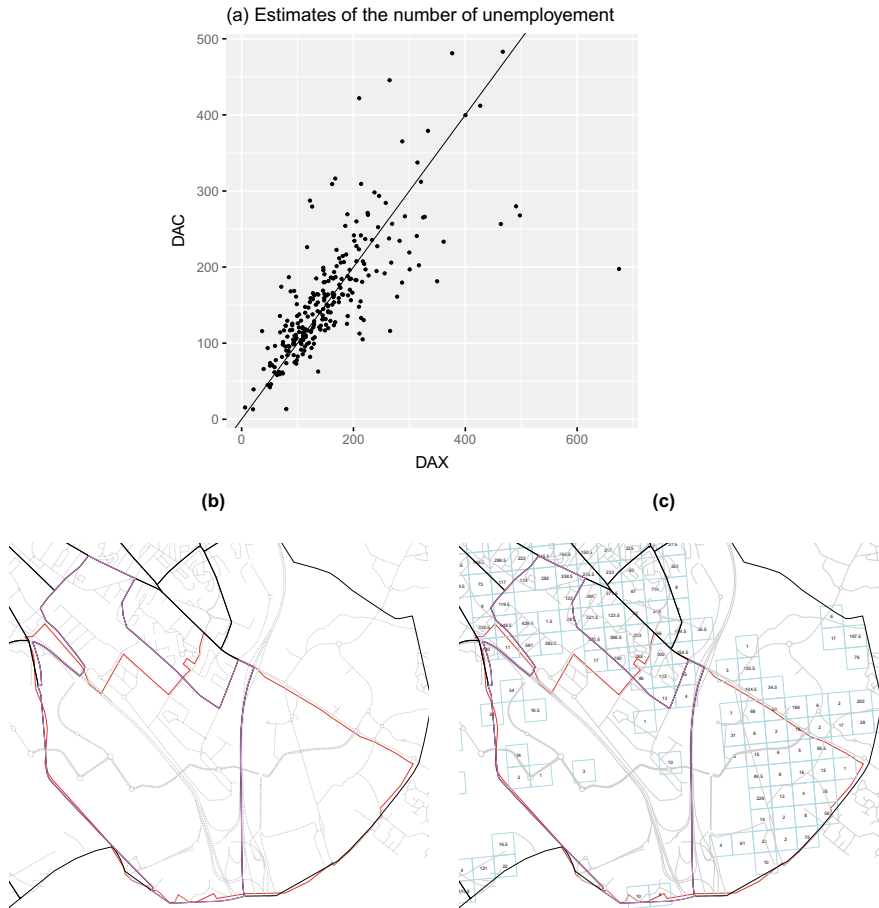


Fig. 8 Comparisons between the *DAX* and *DAC* 2 steps methods for the variable ‘number of unemployed people’ when the sources are the small cells

We use the *DAX* on the variables provided by INSEE at the cell scale and the *DAC* 2-step method on the variables provided by INSEE at the iris scale.

We have represented scatter plots and the correlation plot in the supplementary material. The percentage of workers seems to be highly positively correlated with the extreme right score. On the contrary, the percentage of highly qualified people and population density seem to be highly negatively correlated with the extreme right score.

Two additional remarks can be made: first, the covariates are strongly correlated, which might induce a collinearity issue. Second, the links between the extreme right vote and the covariates might not be linear (for example, when we examine the scatter plot of the extreme right voting with respect to the unemployment rate). Linear

modelling is based on the hypothesis that the covariates are not strongly linearly correlated with each other, which is not the case in our dataset. For that reason, we propose two regression modelling approaches: Linear modelling and Regression tree.

8.2 *Linear Modelling*

We apply several methods (*PIP*, *DAW*, *DAX*, *DAC*) with different sources (small cells, large cells, iris). The best model has an adjusted R^2 value of 56.58%, and the Mean Square Error (MSE) equals 11.475. We apply *DAX* with small cells as sources to estimate the covariates which are available on small cells. For the ones which are only available for iris zones, they are estimated by 2-step *DAC* with iris as sources and small cells as controls. The results are presented in Table 6 in the supplementary material.

The comparison of the adjusted R^2 and MSE allows us to confirm the main results obtained in our study:

1. The smaller the sources, the better
2. The *DAX* method performs better than the *DAW* method, which also performs better than the *PIP* method.

We notice one unexpected result: that the unemployment rate regression coefficient is negative. This is probably due to the collinearity among covariates, and also the non-linear link between the dependent variable and some covariates.

8.3 *Regression Tree*

Because of the collinearity issue between the covariates and possible non-linear link between the dependent variable and the covariates, it could be interesting to also use a regression tree (Breiman et al. 1984) in order to explain the extreme right vote. The regression tree is presented in the supplementary material. The MSE equals 8.456, which confirms that a regression tree is probably preferable to the linear modelling.

In each node, the first value is the average mean of the extreme right voting in that node, and the second and third values correspond to the number and percentage of observations included in the node, respectively.

It is shown that the first variable which splits the root node into two groups is proportional to workers. The left side (71% of polling places) corresponds to polling places with less than 28% workers. The average mean of extreme right votes for this group is 14%, while the right side (polling places with more than 28% workers) vote averages 21% for the extreme right party. The splitting continues until nodes contain a minimum number of polling places.

Let us now focus on the leaves of the tree, beginning with the last leaf at the right side. It corresponds to the group of polling places with the largest predicted value (equal here to 26%). This node contains 9 polling places represented in the last Figure on the left in the supplementary material. We note that they are located in the suburbs of the city in every direction. If we follow the tree from the top to this terminal node, we can see that it corresponds to observations with a proportion of workers larger than 28%, with a proportion of immigrants lower than 35% and with a proportion of highly qualified jobs lower than 14%.

On the other hand, let us now focus on the first leaf on the left side. It corresponds to the polling places with the lowest predicted value (equal here to 11%). The 52 polling places are represented in red in the last Figure on the right in the supplementary material. We note that they are mainly located in the city centre. If we follow the tree from the top to the bottom, we can see that it corresponds to polling places with a proportion of workers lower than 20% (at the first node, it is 28%, then it appears again at 20% at the third node), population density larger than 2 887 inhabitants per square kilometre, and a large proportion of people aged between 18 and 40 years old.

Acknowledgements The authors are grateful to Christine Thomas-Agnan who introduced them to this topic of spatial interpolation methods. They also thank two anonymous referees and the editors for their helpful comments. Thibault Laurent acknowledges funding from ANR under grant ANR-17-EURE-0010 (Investissements d’Avenir program).

References

- Beauguitte, L., Ruso, L., & Rivière, J. (2012). L’ANR Cartelec: analyse des choix électoraux et des inégalités sociales à l’échelle des bureaux de vote. In *Quatrième Congrès de Cartographie Statistique exploratoire, Tours, 27 janvier 2012*.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey: Wadsworth and Brooks.
- Do, V. H., Thomas-Agnan, C., & Vanhems, A. (2014). Testing areal interpolation methods with US census 2010 data. *Region et Développement, 40*, 83–96.
- Do, V. H., Thomas-Agnan, C., & Vanhems, A. (2015). Accuracy of areal interpolation methods for count data. *Spatial Statistics, 14*, 412–438.
- Fisher, P. F., & Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: A study of areal interpolation by dasymetric mapping. *The Professional Geographer, 48*(3), 299–309.
- Flowerdew, R., & Green, M. (1993). Developments in areal interpolation methods and GIS. In *Geographic information systems, spatial modelling and policy evaluation* (pp. 73–84). Berlin: Springer.
- Flowerdew, R., Green, M., & Kehris, E. (1991). Using areal interpolation methods in geographic information systems. *Papers in Regional Science, 70*(3), 303–315.
- Goodchild, M. F., Anselin, L., & Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning A, 25*(3), 383–397.
- Goodchild, M. F., & Lam, N. S.-N. (1980). *Areal interpolation: A variant of the traditional spatial problem*. Department of Geography, University of Western Ontario London, ON, Canada.

- Grasland, C., Vincent, J.-M., et al. (2000). Multiscalar analysis and map generalisation of discrete social phenomena: Statistical problems and political consequences. *Statistical Journal of the United Nations Economic Commission for Europe*, 17(2), 157–188.
- Gregory, I. N. (2002). The accuracy of areal interpolation techniques: Standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems*, 26(4), 293–314.
- Kelsall, J., & Wakefield, J. (2002). Modeling spatial variation in disease risk: A geostatistical approach. *Journal of the American Statistical Association*, 97(459), 692–701.
- Nguyen, T., & Laurent, T. (2019). Coda methods and the multivariate Student distribution with an application to political economy, preprint.
- Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data. *The R Journal*, 10(1), 439–446.
- R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Van Huyen, D., Thomas-Agnan, C., & Vanhems, A. (2015). Spatial reallocation of areal data-another look at basic methods. *Revue d'économie régionale et urbaine*, 1, 27–58.

Predictions in Spatial Econometric Models: Application to Unemployment Data



Thibault Laurent and Paula Margaretic

Abstract In the context of localized unemployment rates in France, we study the issue of prediction of spatial econometric models for areal data, by applying the prediction formulas gathered and derived in Goulard et al. (Spatial Economic Analysis, 12(2–3), 304–325, 2017), (2017). To model regional unemployment taking into account local interactions, we estimate several spatial econometric model specifications, namely, the spatial autoregressive SAR and SDM models, as well as the SLX model. We consider both types of predictions, namely, in-sample and out-of-sample prediction. We show that the prediction can be a complementary method to testing procedures for model comparison.

1 Introduction

Prediction is at the heart of spatial econometrics literature. Not only as a prediction problem by itself, but also in the context of new or missing information.

In conventional econometrics, a sample of n individuals is observed. If values are missing on some individuals, they are generally excluded from the analysis.¹

¹If there are no selection issues due to non-response, this reduces the size of the sample but does not prevent the econometric methods from being implemented.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-73249-3_21) contains supplementary material, which is available to authorized users.

T. Laurent

Toulouse School of Economics, CNRS, University of Toulouse, 1 Esplanade de l'université, 31080 Toulouse cedex 06, France
e-mail: thibault.laurent@tse-fr.eu

P. Margaretic (✉)

University of San Andrés, Victoria, Argentina
e-mail: pmargaretic@udesa.edu.ar

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_21

In contrast, in spatial econometrics, if the observation of the spatial distribution is incomplete (there are missing values), it might be impossible to estimate the model.

In this chapter, we apply the prediction formulas gathered and derived in Goulard et al. (2017) to model localized unemployment rates, by French employment zones, taking into account local interactions.² Regional economists have long been interested in understanding local unemployment differentials, and in predicting the likely impact of local shocks due, for example, to regional policy measures (Molho 1995).

A specific feature of regional labor markets is their correlation over space (Lottmann 2012, summarizes the evidence documenting this correlation). The presence of spatial correlation implies that the level of regional unemployment in one particular area is correlated with that of neighboring regions. As some manifestations of this phenomenon, firms do not restrict their recruiting activities to their resident location and job searchers might accept a job in a different area.

To model regional unemployment accounting for local interactions, we estimate several spatial econometric model specifications, namely, the spatial autoregressive model or SAR, the spatial autoregressive Durbin model or SDM, and the spatial lag of X model or SLX. As structural determinants, we follow the literature and include the characteristics of the labor market, captured by the proportion of low-educated working-age adults, by the proportion of working-age adults between 15 and 30, and by the labor force participation rate; the socio-economic structure, as measured by the proportion of industrial employment and public employment, as well as the logged population density; the housing market, as captured by the annual growth of unoccupied houses between 2006–2011 or 2011–2016, which is a proxy for the costs of migration related to housing.

We consider two types of predictions. First, we use in-sample prediction as a measure of model fit. Second, we make out-of-sample prediction. Specifically, to implement out-of-sample prediction, we use a k -fold cross-validation approach. We randomly split the full sample of employment zones into 10 subsamples. For each subsample i , we declare it as out-of-sample and predict the unemployment rate on these out-of-sample locations based on the model estimates obtained using the information of the remaining sites, the “in-sample zones”, assuming that we observe the structural characteristics of both the in-sample and the out-of-sample sites.

Note that while in this application we do not have a missing information problem (but instead, sequentially choose at random some zones and declare them as out-of-sample), the out-of-sample prediction formulas we apply here could be implemented in other contexts, where information is missing or new information is needed. As illustrations of the latter possibility, consider a geomarketing application, where analysts are evaluating the possible impact in sales of opening a new store in a certain region. Alternatively, consider the problem of an airline, which needs to assess

²Goulard et al. (2017) address the problem of prediction in the spatial autoregressive (SAR) model for areal data, they study what a best linear prediction or BLUP is in this context and they introduce new variants of out-of-sample prediction formulas.

whether it opens a new route.³ Being able to predict local sales or air passenger counts, based on the information of neighboring sites, appears to be particularly attractive.

The chapter is organized as follows. Section 2 summarizes the notation, the spatial autoregressive model specification, and the prediction formulas we apply following Goulard et al. (2017). In turn, Sect. 3 first discusses the theoretical explanations for local unemployment differentials; it then presents the data and the way we construct the neighborhood matrix describing local relations between employment zones. Section 4 presents the estimation results for the various spatial econometric model specifications, whereas Sect. 5 relies on the model estimates of the previous section to make in-sample and out-of-sample predictions. Finally, Sect. 6 concludes. An online appendix (http://www.thibault.laurent.free.fr/code/CT_honor/) provides the **R** codes to reproduce the results included in this chapter.

1.1 Related Literature

From a methodological point of view, the empirical literature can be divided into two strands of literature.

On the one hand, models for regional unemployment using (non-spatial) panel data techniques. Examples are Partridge and Rickman (1997); Taylor and Bradley (1997). On the other hand, studies applying spatial econometric models in cross-sectional settings. The first article in this direction is Molho (1995), which provides evidence of significant spillovers in the adjustment to local shocks using data on local labor market areas in Great Britain. Further examples for this second strand of literature are Aragon et al. (2003), relying on district-level data for the Midi-Pyrénées region of France, and Cracolici et al. (2007), for Italy. Finally, Elhorst (2003) provides a survey on theoretical models and explanatory variables for regional unemployment differences.

We contribute to this second strand of literature by focusing on the prediction of unemployment rates. To the best of our knowledge, we are the first to investigate prediction at the local unemployment level.

2 Notation, Models, and Prediction Formula

Section 2 presents the notation and summarizes the prediction formulas detailed in Goulard et al. (2017).

Note that in contrast to Goulard et al. (2017) who focus on the SAR model, the formulas in this chapter are written relative to the spatial autoregressive Durbin model.

³Margaretic et al. (2017) show evidence of spatial dependence in air passenger traffic.

2.1 Notation and the Spatial Autoregressive Durbin Model

Consider the classical homoscedastic spatial autoregressive Durbin model. Given a row-normalized spatial weight matrix \mathbf{W} and exogenous variables \mathbf{X} , the model writes as

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \alpha \mathbf{t}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (1)$$

where \mathbf{t}_n represents an $n \times 1$ vector of ones, with n being the sample units, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

The conditional mean of \mathbf{Y} is given by

$$\boldsymbol{\mu} = (\mathbf{I} - \rho \mathbf{W})^{-1} (\alpha \mathbf{t}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\gamma}) \quad (2)$$

and its covariance structure by

$$\boldsymbol{\Sigma} = \text{Var}(\mathbf{Y} \mid \mathbf{X}) = [(\mathbf{I} - \rho \mathbf{W}')(\mathbf{I} - \rho \mathbf{W})]^{-1} \sigma^2. \quad (3)$$

The SDM can be written as a spatial autoregressive SAR model by defining $\mathbf{Z} = [\mathbf{t}_n \ \mathbf{X} \ \mathbf{W} \mathbf{X}]$ and $\boldsymbol{\delta} = [\alpha \ \boldsymbol{\beta} \ \boldsymbol{\gamma}]'$, which leads to

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{Z} \boldsymbol{\delta} + \boldsymbol{\varepsilon}. \quad (4)$$

When ρ is known, the best linear unbiased estimator (BLUE) of $\boldsymbol{\mu} = (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{Z} \boldsymbol{\delta}$ is $\hat{\boldsymbol{\mu}} = (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{Z} \hat{\boldsymbol{\delta}}$, where $\hat{\boldsymbol{\delta}}$ is the best linear unbiased estimator of $\boldsymbol{\delta}$, as well as its maximum likelihood estimator in the gaussian case.

We distinguish between two types of prediction situations, that is, the in-sample and out-of-sample cases. In the in-sample prediction problem, we have n spatial units for which we observe the dependent variable \mathbf{Y} , as well as the independent variables \mathbf{X} ; we then want to predict the value of \mathbf{Y} at the observed sites after fitting the model.

In the out-of-sample case, there are two types of spatial units: The in-sample units, for which we observe the dependent variable \mathbf{Y}_S , as well as the independent variables \mathbf{X}_S , and the out-of-sample units, for which we only observe the independent variables \mathbf{X}_O . In the latter situation, we want to predict the variable \mathbf{Y}_O from the knowledge of \mathbf{Y}_S , \mathbf{X}_S and \mathbf{X}_O . Also, in the out-of-sample case, we distinguish according to the number of spatial units to be predicted simultaneously. If there is only one such unit, we refer to the single out-of-sample prediction case; otherwise, to the multiple out-of-sample prediction case.

2.2 In-Sample and Out-of-Sample Units

Let n_o and n_s denote the number of out-of sample and in-sample units, respectively, with $n = n_o + n_s$.

We can partition \mathbf{Z} and \mathbf{Y} in $\mathbf{Z} = (\mathbf{Z}_S' | \mathbf{Z}_O')'$ and $\mathbf{Y} = (\mathbf{Y}_S' | \mathbf{Y}_O')'$, where \mathbf{Z}_S (respectively, \mathbf{Y}_S) of dimension $n_s \times p$ ($n_s \times 1$) denotes the matrix of components of \mathbf{Z} (the vector of components of \mathbf{Y}) corresponding to the in-sample spatial units; \mathbf{Z}_O (\mathbf{Y}_O) of dimension $n_o \times p$ ($n_o \times 1$) denotes the matrix of components of \mathbf{Z} (the vector of components of \mathbf{Y}) corresponding to the out-of-sample spatial units. Similarly $\boldsymbol{\mu} = (\boldsymbol{\mu}_S' | \boldsymbol{\mu}_O')'$.

Let \mathbf{J} be a set of indices. The vector \mathbf{V}_J will correspond to the vector of components of \mathbf{V} relative to the indices in \mathbf{J} . For the case of the spatial weight matrix, variance, and precision matrices, we need a double index for initialization. Precisely, for two sets of indices \mathbf{I} and \mathbf{J} , and a matrix \mathbf{A} , the matrix $\mathbf{A}_{\mathbf{I}\mathbf{J}}$ will denote the block extracted from \mathbf{A} by selecting the rows corresponding to row indices in \mathbf{I} and column indices in \mathbf{J} .

Denote \mathbf{W}_{SS} the $n_s \times n_s$ submatrix corresponding to the neighborhood structure of the n_s in-sample sites; \mathbf{W}_{OO} , the $n_o \times n_o$ submatrix corresponding to the neighborhood structure of the n_o out-of-sample sites; \mathbf{W}_{OS} , the $n_o \times n_s$ submatrix indicating the neighbors of the out-of-sample units among the in-sample units; and finally, \mathbf{W}_{SO} the $n_s \times n_o$ submatrix indicating the neighbors of the in-sample units among the out-of-sample units. The partition of the spatial weight matrix \mathbf{W} follows:

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{SS} & \mathbf{W}_{SO} \\ \mathbf{W}_{OS} & \mathbf{W}_{OO} \end{pmatrix}. \tag{5}$$

For out-of-sample prediction, Goulard et al. (2017) assume that there is an overall model driving the in-sample and out-of-sample units. Given that partition, together with Eq. (1), the overall model M for the n observations of (\mathbf{Z}, \mathbf{Y}) becomes

$$\begin{pmatrix} \mathbf{Y}_S \\ \mathbf{Y}_O \end{pmatrix} = \rho \begin{pmatrix} \mathbf{W}_{SS} & \mathbf{W}_{SO} \\ \mathbf{W}_{OS} & \mathbf{W}_{OO} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_S \\ \mathbf{Y}_O \end{pmatrix} + \alpha \begin{pmatrix} t_{ns} \\ t_{no} \end{pmatrix} + \begin{pmatrix} \mathbf{X}_S \\ \mathbf{X}_O \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{W}_{SS} & \mathbf{W}_{SO} \\ \mathbf{W}_{OS} & \mathbf{W}_{OO} \end{pmatrix} \begin{pmatrix} \mathbf{X}_S \\ \mathbf{X}_O \end{pmatrix} \boldsymbol{\gamma} + \begin{pmatrix} \boldsymbol{\epsilon}_S \\ \boldsymbol{\epsilon}_O \end{pmatrix}.$$

The sub-model M_S driving the data $(\mathbf{Z}_S, \mathbf{Y}_S)$ is

$$\mathbf{Y}_S = [(\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{Z} \boldsymbol{\delta}]_S + [(\mathbf{I} - \rho \mathbf{W})^{-1} \boldsymbol{\epsilon}]_S, \tag{6}$$

where the error term has a variance equal to $(Var(\mathbf{Y}))_{SS}$.

Since only in-sample \mathbf{Y}_S observations are available, a feasible approximation to (6) (only based on in-sample units) after row-normalization of \mathbf{W}_{SS} is

$$\mathbf{Y}_S = [(\mathbf{I} - \rho \mathbf{W}_{SS})^{-1} \mathbf{Z}_S \boldsymbol{\delta}] + [(\mathbf{I} - \rho \mathbf{W}_{SS})^{-1} \boldsymbol{\epsilon}_S], \tag{7}$$

Table 1 In-sample prediction formulas

Predictor	Formula
<i>TC</i>	$\check{\mathbf{Y}}_S^{TC} = (\mathbf{I} - \hat{\rho}\mathbf{W})_{SS}^{-1}\mathbf{Z}_S\hat{\delta}$
<i>TS</i>	$\check{\mathbf{Y}}_S^{TS} = \mathbf{Z}_S\hat{\delta} + \hat{\rho}\mathbf{W}_{SS}\mathbf{Y}_S$
<i>BP</i>	$\check{\mathbf{Y}}_S^{BP} = (\mathbf{I} - \hat{\rho}\mathbf{W}_{SS})^{-1}\mathbf{Z}_S\hat{\delta} - \text{Diag}(\hat{\mathbf{Q}}_{SS})^{-1}\tilde{\mathbf{Q}}_{SS}(\mathbf{Y} - (\mathbf{I} - \hat{\rho}\mathbf{W}_{SS})^{-1}\mathbf{Z}_S\hat{\delta})$

where $\mathbf{Z}_S\hat{\delta} = \alpha\iota_{n_s} + \mathbf{X}_S\boldsymbol{\beta} + \mathbf{W}_{SS}\mathbf{X}_S\boldsymbol{\gamma}$.

Exact compatibility of the two models thus requires the following two constraints: $(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{Z}_S = (\mathbf{I} - \hat{\rho}\mathbf{W}_{SS})^{-1}\mathbf{Z}_S$ for the mean and $(\text{Var}(\mathbf{Y}))_{SS} = \text{Var}(\mathbf{Y}_S)$ for the variance.

2.3 In-Sample Prediction Formulas

For in-sample prediction, we assume that the sample units are driven by Eq. (7). Let $\check{\mathbf{Y}}_S$ denote the in-sample predictions. As in Goulard et al. (2017), we consider three in-sample predictors, namely, the “trend-corrected predictor”, indicated with upper index TC in $\check{\mathbf{Y}}_S$; the “trend-signal-noise” predictor, with upper index TS in $\check{\mathbf{Y}}_S$, and finally, the Goulard et al. (2017) alternative in-sample predictor, with upper index BP.⁴ Table 1 summarizes the alternative in-sample prediction formulas we consider.

In Table 1, $\text{Diag}(\hat{\mathbf{Q}}_{SS})$ denotes the diagonal matrix containing the diagonal of the precision matrix $\hat{\mathbf{Q}}_{SS}$ of the SDM model given by $\hat{\mathbf{Q}}_{SS} = \frac{1}{\hat{\sigma}^2}(\mathbf{I} - \hat{\rho}\mathbf{W}_{SS}')(\mathbf{I} - \hat{\rho}\mathbf{W}_{SS})$, $\hat{\sigma}^2$ is the gaussian maximum likelihood estimate of the variance, and $\tilde{\mathbf{Q}}_{SS} = \hat{\mathbf{Q}}_{SS} - \text{Diag}(\hat{\mathbf{Q}}_{SS})$.

2.4 Out-of-Sample Prediction Formulas

Let $\hat{\mathbf{Y}}_S$ refer to the out-of-sample predictions.

We consider five out-of-sample predictors: The classical Goldberger formula, indicated with upper index BP in $\hat{\mathbf{Y}}_S$; the trend-corrected predictor; the trend-signal-noise predictor, in the case of a single prediction; Goulard et al. (2017) extension of the Kelejian and Prucha (2007) predictor, indicated with upper index BP_W ; and finally, Goulard et al. (2017) predictor, with upper index BP_N . Table 2 summarizes the alternative out-of-sample predictors. For details on the formulas, refer to Goulard et al. (2017).

⁴Goulard et al. (2017) use the upper index BP to emphasize that the predictor is based on some kind of best prediction practice.

Table 2 Out-of-sample prediction formulas

Predictor	Formula
BP	$\hat{Y}_O^{\hat{B}P} = \hat{Y}_O^{\hat{T}C} - \hat{Q}_{OO}^{-1} \hat{Q}_{OS} \times (\mathbf{Y}_S - \hat{Y}_S^{\hat{T}C})$
TC	$\hat{Y}_O^{\hat{T}C} = [(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1} \hat{\mathbf{Z}}\hat{\boldsymbol{\delta}}]_O$
TS^1	$\hat{Y}_O^{TS^1} = \mathbf{Z}_O \hat{\boldsymbol{\delta}} + \hat{\rho} \mathbf{W}_{OS} \mathbf{Y}_S$
BP_W	$\hat{Y}_O^{\hat{B}P_W} = \hat{Y}_O^{\hat{T}C} + \hat{\boldsymbol{\Sigma}}_{OS} \mathbf{W}'_{OS} (\mathbf{W}_{OS} \hat{\boldsymbol{\Sigma}}_{SS} \mathbf{W}'_{OS})^{-1} (\mathbf{W}_{OS} \mathbf{Y}_S - \mathbf{W}_{OS} \hat{Y}_S^{\hat{T}C})$
BP_N	$\hat{Y}_O^{\hat{B}P_N} = \hat{Y}_O^{\hat{T}C} - \hat{Q}_{OO}^{-1} \hat{Q}_{OJ} (\mathbf{Y}_J - \mathbf{Q}_J^{\hat{T}C})_J$
TC^1	$\hat{Y}_O^{TC^1} = \text{row } o \text{ of } \{\mathbf{I}_{n_S+1} - \hat{\rho} \mathbf{W}^1\}^{-1} \begin{pmatrix} \mathbf{Z}_S \\ \mathbf{Z}_O \end{pmatrix} \hat{\boldsymbol{\delta}}$
BP^1	$\hat{Y}_O^{BP^1} = \hat{Y}_O^{TC^1} - \hat{Q}_{oo}^{-1} \hat{Q}_{oS} (\mathbf{Y}_S - \hat{Y}_S^{TC^1})$
BP^1_W	$\hat{Y}_O^{BP^1_W} = \hat{Y}_O^{TC^1} + \hat{\boldsymbol{\Sigma}}_{oS} \mathbf{W}'_{oS} (\mathbf{W}_{oS} \hat{\boldsymbol{\Sigma}}_{SS} \mathbf{W}'_{oS})^{-1} (\mathbf{W}_{oS} \mathbf{Y}_S - \mathbf{W}_{oS} \hat{Y}_S^{TC^1})$
BP^1_N	$\hat{Y}_O^{BP^1_N} = \hat{Y}_O^{TC^1} - \hat{Q}_{oo}^{-1} \hat{Q}_{oJ} (\mathbf{Y}_J - \hat{Y}_J^{TC^1})$

There are four elements to mention regarding the out-of-sample predictors in Table 2. To begin with, the Goldberger best prediction formula in Table 2 is written in terms of the precision matrix \mathbf{Q} , with

$$\mathbf{Q} = \frac{1}{\sigma^2} (\mathbf{I} - \rho(\mathbf{W}' + \mathbf{W}) + \rho^2 \mathbf{W}'\mathbf{W}) = \begin{pmatrix} \mathbf{Q}_{SS} & \mathbf{Q}_{SO} \\ \mathbf{Q}_{OS} & \mathbf{Q}_{OO} \end{pmatrix}$$

Second, the trend-corrected predictor can be extended for out-of-sample prediction because it only involves the values of \mathbf{Z} (and not \mathbf{Y}) for the out-of-sample units.

Third, Goulard et al. (2017) predictor consists of using the precision version of the Goldberg formula and replaces the set \mathbf{S} by \mathbf{N} , with \mathbf{N} being the set of all sites in \mathbf{S} which are neighbors in the sense of \mathbf{W} of at least one site in \mathbf{O} . The intuition is to only use among the sample locations, the neighbors of the out-of-sample sites in

order to predict. Let \mathbf{J} be the set of indices of such neighbors and n_J its size.⁵ Also, $\hat{\mathbf{Y}}_J^{TC}$ is obtained by extracting the rows corresponding to units in J from $\hat{\mathbf{Y}}^{TC}$.

Finally, because the single prediction formulas are simpler, when p out-of-sample units have to be predicted, Goulard et al. (2017) propose to apply the “single out-of-sample” formula to each of the out-of-sample units separately, ignoring at each stage the remaining $p - 1$ units. Hence, Table 2 also exhibits the predictors for location o , which are denoted by $\hat{Y}_o^{TC^1}$, $\hat{Y}_o^{TS^1}$, $\hat{Y}_o^{BP^1}$, $\hat{Y}_o^{BP_W^1}$ and $\hat{Y}_o^{BP_N^1}$.

3 Application

Section 3 starts with a summary of the theoretical explanations for regional unemployment differentials, which helps us identify the explanatory variables to use in the regression analysis for local unemployment rates.

It then presents the data and the way we define the spatial neighbors.

3.1 Theoretical Explanations for Regional Unemployment Differentials

In the literature, there are two different views explaining the regional unemployment differentials.

The equilibrium view assumes the existence of a stable labor market equilibrium in which regions have different unemployment rates. According to Molho (1995), this equilibrium is characterized by uniform utility across areas for homogeneous labor groups. Under this view, households (and firms) need to be compensated for high (low) unemployment by other positive factors, so-called amenities. These amenities are, for example, better climate, reasonable housing prices or higher quality of life.

⁵For clarification on the expression for $\mathbf{Y}_O^{BP_N}$, denote $\mathbf{W}_{\mathbf{J}\cup\mathbf{O}}$ as the neighborhood matrix for sites which are in \mathbf{J} or \mathbf{O} and \mathbf{W}^* its row-normalized version. The partition of \mathbf{W}^* is

$$\mathbf{W}^* = \left(\begin{array}{c|c} \overset{n_J}{\longleftrightarrow} & \overset{n_O}{\longleftrightarrow} \\ \mathbf{W}_{\mathbf{J}\mathbf{J}}^* & \mathbf{W}_{\mathbf{J}\mathbf{O}}^* \\ \hline \mathbf{W}_{\mathbf{O}\mathbf{J}}^* & \mathbf{W}_{\mathbf{O}\mathbf{O}}^* \end{array} \right) \cdot \begin{array}{l} \uparrow n_J \\ \downarrow n_O \end{array}$$

The partition of the precision matrix corresponding to sites in $\mathbf{J} \cup \mathbf{O}$ becomes

$$\hat{\mathbf{Q}}_{\mathbf{J}\cup\mathbf{O}} = \frac{1}{\hat{\sigma}^2} (\mathbf{I}_{\mathbf{J}\cup\mathbf{O}} - \hat{\rho}(\mathbf{W}^* + \mathbf{W}^{*'}) + \hat{\rho}^2(\mathbf{W}^{*'}\mathbf{W}^*)) = \begin{pmatrix} \hat{\mathbf{Q}}_{\mathbf{J}\mathbf{J}} & \hat{\mathbf{Q}}_{\mathbf{J}\mathbf{O}} \\ \hat{\mathbf{Q}}_{\mathbf{O}\mathbf{J}} & \hat{\mathbf{Q}}_{\mathbf{O}\mathbf{O}} \end{pmatrix}$$

Hence, the equilibrium rate of unemployment in region i is a function of the amenity endowment in this region (Marston 1985).

Contrary to the equilibrium view, the disequilibrium view assumes that regional unemployment will equalize in the long run. However, the adjustment process might be slow. The speed of adjustment depends on different factors that are connected to both labor supply and labor demand side.

Among these factors, younger people and better-educated workers are more likely to migrate in response to local economic opportunities (Aragon et al. 2003). Younger people are more likely to migrate as they have a lower opportunity cost of migrating, they can look forward to a longer period of payoff from migrating, and they may have less risk aversion (Gabriel et al. 1993). In the case of better-educated workers, it may be because the labor market for skilled workers tends to be geographically larger; hence, the payoff from moving is likely to be larger also for these workers (McCormick and Sheppard 1992).

In addition, other things being equal, the unemployment rate should be lower in urban areas (because it is easier for a person after a shock, to find another job locally that makes use of the same skills or for a firm to hire a new employee with similar skills to the one that previously left the company) and it may converge more slowly if the system of unemployment compensation is generous (Aragon et al. 2003). Finally, labor mobility should decrease with the costs of migration, such as transport or housing costs (Marston 1985).

Wrapping up, these two different views help us identify the main variables that we expect to influence unemployment rates. In what follows, we present the data we use to proxy some of these structural factors.

3.2 Data and Definition of Neighborhood Structure

To model local unemployment rates in France, we run the analysis at the employment zone level, excluding Corsica.

An employment zone is a geographic area within which most workers live and work, and in which establishments can find the bulk of the labor force required to fill the jobs offered. It corresponds to the aggregation of several communes (Floch and Le Saout 2018).

As structural determinants of local employment rates, we follow the literature and consider, the characteristics of the labor market, captured by the proportion of low-educated working-age adults, by the proportion of working-age adults between 15 and 30, and by the labor force participation rate; the socio-economic structure, as measured by the proportion of industrial employment and public employment, as well as the logged population density; the housing market, as captured by the average annual growth of unoccupied houses between 2006–2011 or 2011–2016, which is a proxy for the costs of migration related to housing.

We consider two estimation periods. The first sample period uses the unemployment rate in 2013 and the explanatory variables as of 2011. The objective is to repro-

Table 3 Descriptive statistics of the 2018 unemployment rate and 2016 structural factors

	Obs	Mean	SD	Min	P25	P50	P75	Max
Unemployment	297	8.77	2.23	4.50	7.20	8.40	9.80	16.50
Participation rate	297	73.88	2.50	67.00	72.40	73.80	75.20	83.50
Prop low-educated workers	297	31.93	4.44	20.40	29.30	32.20	35.10	43.70
Prop workers 15 – 30	297	15.51	2.29	10.50	14.00	15.10	16.70	23.90
Prop industrial employment	297	17.22	7.80	3.02	11.33	15.98	21.91	42.56
Prop public employment	297	34.32	6.63	15.65	30.18	34.30	38.41	54.23
Growth vacant houses	297	3.13	1.52	-1.21	2.20	3.06	4.16	7.48
Pop density	297	184.41	612.56	12.50	49.70	79.70	147.20	9179.00

Notes Obs and SD stand for number of observations and standard deviation, respectively. P25, P50, and P75 correspond to the 25, 50, and 75 percentiles, respectively, of the empirical distribution of each variable

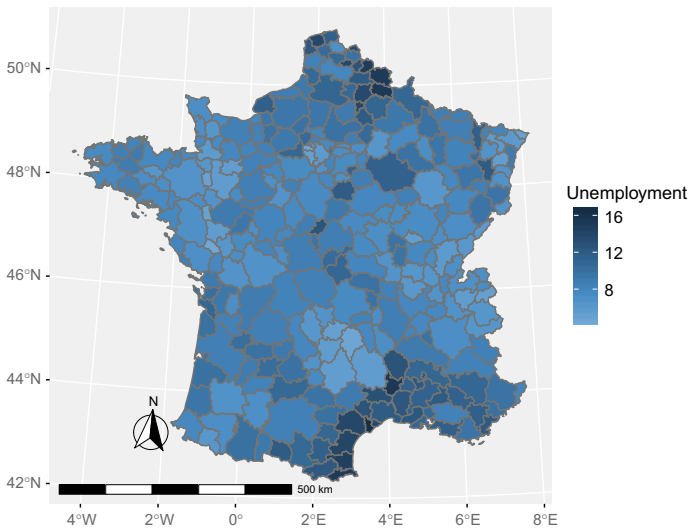
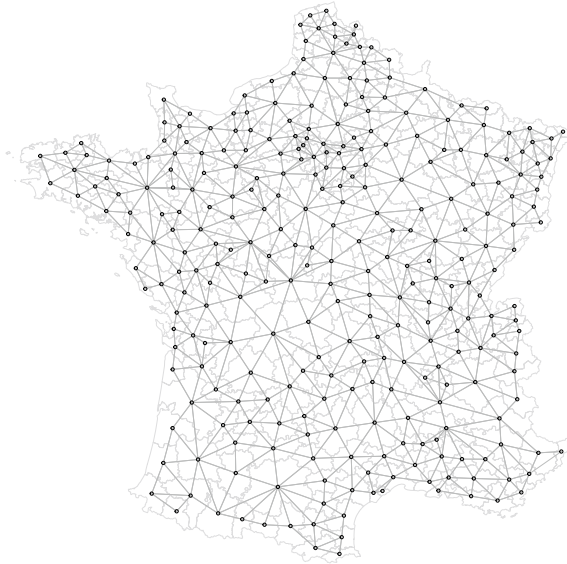


Fig. 1 2018 Unemployment rate, by employment zones

duce some of the results in Floch and Le Saout (2018). The second data set consists of more recent observations, with the unemployment rate, as of 2018, and the structural determinants, for 2016. The reason for lagging the explanatory variables is to limit the possibility of reverse causality. A causal interpretation nonetheless remains impossible.

Table 3 presents the descriptive statistics of the 2018 unemployment rates and the structural factors observed in 2016. In turn, Fig. 1 plots unemployment, by employment zones.

Fig. 2 Neighborhood structure, by employment zones



To begin with, Table 3 shows that population density, followed by the annual growth of unoccupied houses and the proportion of industrial employment are the factors with the strongest variation across employment zones. On top of that, Fig. 1 provides evidence of spatial heterogeneity for the unemployment rates: The North of France (former Nord Pas-de-Calais) and Languedoc-Roussillon exhibit higher unemployment rates, whereas the regions neighboring Germany, Alsace, and Auvergne have lower rates. Also, the employment zones close to these sites show similar unemployment patterns, thus indicating spatial correlation. The online appendix displays the Moran unemployment rate graph and associated map.

Finally, to define the neighborhood matrix describing local relations between employment zones, we combine two methods, namely, contiguity and two nearest neighbors. The reason for that is to allow that each employment zone has at least two neighbors. We then row-normalize the resulting spatial weight matrix. Figure 2 displays the neighborhood structure.

4 Estimation Results

Section 4 presents the full sample estimation results.

To begin with, Table 4 exhibits OLS, SAR, SLX, and SDM model estimates for the 2013 unemployment rate, relying on the same structural factors than in Floch and Le Saout (2018), that is, the labor force participation rate, the proportion of low-educated working-age adults, the proportion of working-age adults aged between 15 and 30, and the share of industrial and public employment. Second, Table 5

Table 4 Model estimates for 2013 unemployment rates, following specifications in Floch and Le Saout (2018)

	OLS	SAR	SLX	SDM
	(1)	(2)	(3)	(4)
Participation rate	-0.617*** (0.040)	-0.450*** (0.038)	-0.514*** (0.052)	-0.507*** (0.043)
Prop low-educated workers	0.212*** (0.030)	0.150*** (0.025)	0.169*** (0.039)	0.167*** (0.032)
Prop workers 15 – 30	0.115*** (0.038)	0.051 (0.032)	0.101** (0.044)	0.073** (0.037)
Prop industrial employment	-0.062*** (0.014)	-0.040*** (0.012)	-0.028 (0.017)	-0.021 (0.014)
Prop public employment	-0.064*** (0.018)	-0.064*** (0.015)	-0.047** (0.019)	-0.047*** (0.015)
$W \times$ Participation rate			-0.208*** (0.077)	0.190** (0.074)
$W \times$ Prop low-educated workers			0.032 (0.055)	-0.098** (0.047)
$W \times$ Prop workers 15 – 30			0.121* (0.068)	0.034 (0.057)
$W \times$ Prop industrial employment			-0.094*** (0.028)	-0.036 (0.023)
$W \times$ Prop public employment			-0.038 (0.036)	0.003 (0.030)
Constant	52.043***	36.791***	59.784***	26.380***
ρ		0.497***		0.601***
R^2	0.623		0.664	
Adjusted R^2	0.617		0.652	
σ^2	1.467	1.479	1.398	1.340
AIC	1078.24	991.157	1054.744	980.847
F Statistic	96.329***		56.403***	
	df = 5; 291		df = 10; 286	
Wald Test (df = 1)		110.300***		108.815***
LR Test (df = 1)		89.083***		75.898***

Notes *p<0.1; **p<0.05; ***p<0.01. Data for 2013 unemployment rates and 2011 structural factors

introduces two variations to model estimates in Table 4, that is, it considers more recent observations, with the unemployment rate, as of 2018, and the covariates for 2016; and it augments the model specifications in Table 4 with the logarithm of population density and the annual growth of unoccupied houses.

There are several findings to highlight from Table 4. To begin with, the model estimates in Table 4 are strongly consistent with the findings in Floch and Le Saout (2018), in spite of the differences in the way they and we compute the neighborhood structure (to build the neighborhood matrix, Floch and Le Saout 2018, use the inverse distance). It thus indicates that in this unemployment application, results are not sensitive to the choice of the spatial weight matrix.

Second, Table 4 shows that the characteristics of the labour force are statistically significant. However, while the sign of the estimated coefficients for the proportion of low-educated workers and for the labor force participation rate is in line with the theory, the sign for the proportion of working-age adults between 15 and 30 contrasts with its expected sign, according to the disequilibrium view explaining local unemployment: Zones with a high proportion of young people appear to be associated with higher unemployment rates. The latter thus suggests that young people might not be that flexible to migrate to take advantage of job opportunities in other regions. This result is also in line with Aragon et al. (2003) findings.

Third, Table 4 exhibits a negative relationship between unemployment and industrial and public employment. In the case of industrial employment, its coefficient is not statistically significant when we add spatially lagged structural factors, thus suggesting that the industry's capacity to absorb labor and thus reduce local unemployment rates, is influenced by the importance of industry in the neighboring regions. Regarding public employment, the significantly negative association with unemployment, which is stable across estimations, may be the indication of public jobs being more stable and less dependent on the business cycle.

The first conclusion to extract from Table 5 is that, regardless of the spatial econometric model considered, unemployment appears to be higher in areas more densely populated, which contrasts the intuition that job searching and matching should be easier in urban areas. Paraphrasing Aragon et al. (2003), the positive coefficient is consistent with the amenities view of equilibrium unemployment, according to which if urban areas are considered to be more interesting places in which to live, people may remain there longer as they search for work, instead of migrating to areas with more job opportunities.

On top of that, the inclusion of population density results in the proportion of working-age adults between 15 and 30 and public employment being no longer statistically significant, showing that the former factor dominates. Finally, in line with Aragon et al. (2003) findings, the annual growth of unoccupied houses capturing the housing market does not appear to have a significant effect on local unemployment rates.

The estimated parameter ρ reflects the spatial dependence inherent in our sample data, measuring the average influence on observations of their neighboring data points. It has a positive effect and it is highly significant in both the SAR and SDM models. As a result, we conclude that the general model fit has improved (relative to the OLS estimates), as indicated by the lower values of AIC.⁶

⁶In the supplementary material, we display a moran plot of residuals from the SAR and SDM model which show that the spatial models have indeed taken the correlation into account.

Table 5 Model estimates for 2018 unemployment rates

	OLS	SAR	SLX	SDM
	(1)	(2)	(3)	(4)
Participation rate	-0.546*** (0.048)	-0.366*** (0.041)	-0.424*** (0.060)	-0.396*** (0.046)
Prop low-educated workers	0.101*** (0.028)	0.075*** (0.022)	0.109*** (0.036)	0.113*** (0.028)
Prop workers 15 – 30	-0.062 (0.055)	-0.066 (0.043)	0.008 (0.068)	-0.038 (0.053)
Prop industrial employment	-0.045*** (0.016)	-0.024* (0.013)	-0.023 (0.017)	-0.016 (0.013)
Prop public employment	-0.020 (0.020)	-0.026 (0.016)	-0.016 (0.020)	-0.014 (0.016)
Growth vacant houses	-0.090 (0.060)	-0.060 (0.047)	-0.039 (0.060)	0.003 (0.047)
log(Pop density)	0.599*** (0.149)	0.426*** (0.118)	0.569*** (0.213)	0.689*** (0.163)
$W \times$ Participation rate			-0.310*** (0.094)	0.144* (0.079)
$W \times$ Prop low-educated workers			-0.043 (0.054)	-0.095** (0.042)
$W \times$ Prop workers 15 – 30			0.177 (0.117)	0.131 (0.090)
$W \times$ Prop industrial employment			-0.079** (0.034)	-0.026 (0.026)
$W \times$ Prop public employment			-0.032 (0.042)	-0.007 (0.032)
$W \times$ Growth vacant houses			-0.287** (0.114)	-0.096 (0.088)
$W \times$ log(Pop density)			-0.398 (0.364)	-0.744*** (0.280)
Constant	45.845*** (4.692)	28.950*** (3.889)	61.602*** (8.010)	21.420*** (6.736)
ρ		0.580***		0.680***
R^2	0.555		0.610	
Adjusted R^2	0.544		0.591	
σ^2	1.504	1.362	1.425	1.198
AIC	1095.055	977.777	1069.903	964.472
F Statistic	51.498***		31.504***	
	df = 7; 289		df = 14; 282	
Wald Test (df = 1)		163.313***		181.516***
LR Test (df = 1)		119.279***		107.431***

Notes * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Data for 2018 unemployment rates and 2016 structural factors

One question that remains is which spatial econometric model is the most likely to describe the data. As the first piece of evidence, both the LR and Wald test statistics reported in Table 5 reject the null of absence of spatial dependence. In addition, we follow Margaretic et al. (2017)’s testing procedure to examine whether we can reduce the SDM to the SAR or the SEM (which includes a spatially lagged error term, instead of spatially lagged independent variables or a spatially lagged dependent variable). When testing whether we can simplify the SDM to the SAR, the null hypothesis of the likelihood ratio (LR) test is $\gamma = 0$. Instead, if we test whether we can reduce the SDM to SEM (known as the common factor restriction), the null hypothesis is $\gamma + \rho\beta = 0$. Both tests follow a χ^2 distribution.

Both LR tests reject the spatial LAG (Log-likelihood= -437.03) or the SEM (Log-likelihood= -471.815), in favor of the SDM (Log-likelihood= -424.347). We thus conclude that the SDM seems to be the one that best describes the data. In what follows, we compare the spatial econometric model estimates, in terms of their predictive efficiency.

5 Predictions

To compare the model estimates of Table 5 in terms of their predictive efficiency, we compute the average, across employment zones, of the total mean square error (MSE) of in-sample or out-of-sample predictions, as

$$MSE_k = \frac{1}{n}(\mathbf{Y} - \hat{\mathbf{Y}}^k)'(\mathbf{Y} - \hat{\mathbf{Y}}^k), \tag{8}$$

for each in-sample prediction formula $k = BP, TS, TC$, or out-of-sample prediction formula $k = BP, BP_N, BP_N^1, BP_W, BP_W^1, TC, TC^1, TS, TS^1$. Note that in the case of the SLX model, there is no correction for spatial autocorrelation in the dependent variable in the prediction formulas.

Regarding out-of-sample prediction, the procedure we implement follows: First, we split the full sample into 10 subsamples and randomly assign the employment zones to each of the 10 subsamples. For each subsample i , we declare it as out-of-sample and estimate the model specifications in Table 5 using the information of the remaining sites, the “in-sample zones”. Assuming that we observe the structural characteristics of both the in-sample and the out-of-sample sites, we finally predict the unemployment rate of the out-of-sample locations based on the model estimates obtained with the in-sample sites.

Table 6 reports the average MSE for each of the model specifications, distinguishing between in-sample prediction formulas. Table 7, in turn, reports the average MSE of the out-of-sample predictions, exhibited in decreasing order of efficiency.

There are three conclusions to highlight from Tables 6 and 7. To begin with, they confirm the preference of the SDM over the SAR model or the SLX model, a conclusion we obtained when applying the testing procedure described in Sect. 4. Second,

Table 6 shows that in-sample predictors including a correction for spatial correlation (BP and TS) tend to perform better. Third, regarding out-of-sample prediction, Table 7 indicates that in the case of the SDM estimates, the performances of BP_N , BP_W , BP_W^1 , BP^1 , BP_N^1 are close to that of the best prediction BP and much better than that of TC, TC^1 , and TS^1 .

6 Conclusion

In this chapter, we study the issue of prediction of spatial econometric models for areal data, in the context of localized unemployment rates. Specifically, we apply the prediction formulas gathered and derived in Goulard et al. (2017) to model localized unemployment rates, by French employment zones. To the best of our knowledge, we are the first to investigate prediction at local unemployment level.

Researchers have long been interested in understanding differences in local unemployment rates and in predicting the likely impact of local shocks and/or regional policy measures on unemployment. This is important, because unemployment is, among others, a widely used indicator for the economic well-being of a country. A specific feature of regional labor markets is their correlation over space (Lottmann 2012), which implies that the level of unemployment in one particular region is correlated with that of neighboring regions.

To model regional unemployment accounting for local interactions, we estimate several spatial econometric model specifications, namely, the spatial autoregressive model or SAR, the spatial autoregressive Durbin model or SDM, and the spatial lag of X model or SLX. As structural determinants, we consider the characteristics of

Table 6 MSE of in-sample predictions

	OLS	SAR TS	SAR TC	SAR BP	SLX	SDM TS	SDM TC	SDM BP
MSE	2.200	1.362	2.141	1.230	1.929	1.198	1.792	1.139

Table 7 MSE of out-of-sample predictions

SDM BP	SDM BP_N	SDM BP_W	SDM BP_W^1	SDM BP^1	SDM BP_N^1	SAR BP
1.245	1.258	1.260	1.276	1.282	1.288	1.296
SAR BP_N	SAR BP_N^1	SAR BP^1	SAR BP_W	SAR BP_W^1	SDM TC	SDM TC^1
1.299	1.301	1.309	1.309	1.313	1.948	2.015
SLX	SAR TC	SAR TC^1	OLS	SAR TS^1	SDM TS^1	
2.095	2.204	2.235	2.355	2.370	2.472	

the labor market and the socio-economic structure, among other factors. Note that while we lag the previously described structural characteristics two years (relative to the unemployment data), a causal interpretation of our results remains impossible. We then use our model estimates to produce both single and multiple in-sample and out-of-sample predictions.

We can extract two main conclusions from our results. First, our results indicate that the SDM is the most likely to describe the data; hence, we should take this into account when predicting the reaction of regional unemployment to shocks. The second main conclusion we can derive from our findings is that prediction can be a complementary method to testing procedures for model comparison and model choice.

From a policy and applied standpoint, while in this application we do not have a missing information problem in itself (as we have full information for all the zones), the out-of-sample prediction formulas we use here could be implemented in other applications, where information was indeed missing or new information was needed. Having a methodology that is able to predict local sales, counts or whatever variable of interest, based on the information of neighboring sites, appears to be particularly useful and attractive.

One venue of future work could be to assess whether the performance of the various predictors changes (if any) with the spatial neighborhood structure, for instance, the sparseness of the spatial weight matrix. Another venue could be to apply the same methodology to other contexts, for example, a geomarketing application, with clients in one district going to shop in other zones.

Acknowledgements The authors are grateful to Christine Thomas-Agnan who introduced them to spatial econometrics. They also thank two anonymous referees and the editors for their helpful comments. Thibault Laurent acknowledges funding from ANR under grant ANR-17-EURE-0010 (Investissements d' Avenir program).

References

- Aragon, Y., Haughton, D., Haughton, J., Leconte, E., Malin, E., Ruiz-Gazen, A., et al. (2003). Explaining the pattern of regional unemployment: The case of the Midi-Pyrénées region. *Papers in Regional Science*, 82(2), 155–174.
- Cracolici, F. M., Cuffaro, M., & Nijkamp, P. (2007). Geographical distribution of unemployment: An analysis of provincial differences in Italy. *Growth and Change*, 38(4), 649–670.
- Elhorst, J. P. (2003). The mystery of regional unemployment differentials: Theoretical and empirical explanations. *Journal of Economic Surveys*, 17(5), 709–748.
- Floch, J. M., & Le Saout, R. (2018). Spatial econometrics - common models. In V. Loonis (Ed.), *Handbook of spatial analysis*. Insee Méthodes.
- Gabriel, S. A., Shack-Marquez, J., & Wascher, W. L. (1993). Does migration arbitrage regional labor market differentials? *Regional Science and Urban Economics*, 23(2), 211–233.
- Goulard, M., Laurent, T., & Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost optimal strategies. *Spatial Economic Analysis*, 12(2–3), 304–325.

- Kelejian, H. H., & Prucha, I. R. (2007). The relative efficiencies of various predictors in spatial econometric models containing spatial lags. *Regional Science and Urban Economics*, 37(3), 363–374.
- Lottmann, F. (2012). Explaining regional unemployment differences in Germany: A spatial panel data analysis. *SFB 649 discussion paper 2012-026*.
- Margaretic, P., Thomas-Agnan, C., & Doucet, R. (2017). Spatial dependence in (origin-destination) air passenger flows. *Papers in Regional Science*, 96(2), 357–380.
- Marston, S. T. (1985). Two views of the geographic distribution of unemployment. *The Quarterly Journal of Economics*, 100(1), 57–79.
- McCormick, B., & Sheppard, S. (1992). A model of regional contraction and unemployment. *The Economic Journal*, 102(411), 366–377.
- Molho, I. (1995). Spatial autocorrelation in British unemployment. *Journal of Regional Science*, 35(4), 641–658.
- Partridge, M. D., & Rickman, D. S. (1997). The dispersion of US state unemployment rates: The role of market and non-market equilibrium factors. *Regional Studies*, 31(6), 593–606.
- Taylor, J., & Bradley, S. (1997). Unemployment in Europe: A comparative analysis of regional disparities in Germany. *Italy and the UK. Kyklos*, 50(2), 221–245.

Lagrangian Spatio-Temporal Nonstationary Covariance Functions



Mary Lai O. Salvaña and Marc G. Genton

Abstract The Lagrangian reference frame has been used to model spatio-temporal dependence of purely spatial second-order stationary random fields that are being transported. This modeling paradigm involves transforming a purely spatial process to spatio-temporal by introducing a transformation in the spatial coordinates. Recently, it has been used to capture dependence in space and time of transported purely spatial random fields with second-order nonstationarity. However, under this modeling framework, the presence of mechanisms enforcing second-order nonstationary behavior introduces considerable challenges in parameter estimation. To address these, we propose a new estimation methodology which includes modeling the second-order nonstationarity parameters by means of thin plate splines and estimating all the parameters via two-step maximum likelihood estimation. In addition, through numerical experiments, we tackle the consequences of model misspecification. That is, we discuss the implications, both in the stationary and nonstationary cases, of fitting Lagrangian spatio-temporal covariance functions to data generated from non-Lagrangian models, and vice versa. Lastly, we apply the Lagrangian models and the new estimation technique to analyze particulate matter concentrations over Saudi Arabia.

1 Introduction

The need for models that explain spatio-temporal dependencies of environmental processes has been answered with a growing number of studies on spatio-temporal covariance functions. A number of the established spatio-temporal covariance func-

M. L. O. Salvaña · M. G. Genton (✉)
King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900,
Saudi Arabia
e-mail: marc.genton@kaust.edu.sa

M. L. O. Salvaña
e-mail: marylai.salvana@kaust.edu.sa

tions can only model spatio-temporal random fields that are second-order stationary in space and time. The list includes the spatio-temporal separable stationary covariance functions, spatio-temporal stationary mixture models (Ma 2003a), and the Gneiting class of spatio-temporal stationary covariance functions (Gneiting 2002). However, environmental processes are notorious for exhibiting second-order nonstationarity in space and/or time. The number of available spatio-temporal nonstationary covariance functions catering to this challenging second-order nonstationary behavior is slowly increasing but still lags behind its stationary counterpart. The construction approaches that define the current state-of-the-art for spatio-temporal nonstationary covariance functions modeling include the spatio-temporal dimension expansion (Shand and Li 2017), the spatio-temporal convolution (Garg et al. 2012), and the nonstationary Archimedean spectral densities (Porcu et al. 2009). Some spatio-temporal nonstationary models built from spatio-temporal stationary covariances and intrinsically stationary variograms were also proposed in Ma (2003b). Several other works on incorporating spatial nonstationarity focused on allowing the parameters in the covariance function to vary in space (Higdon et al. 1999; Neto et al. 2014; Paciorek and Schervish 2006; Stein 2005). These types of nonstationary covariance functions belong to a wider class of kernel convolution methods. Risser (2016), Sampson et al. (2001) feature comprehensive overviews of this wider class. Another flexible class of spatio-temporal nonstationary models termed the spatio-temporal random effects (STRE) models was put forward in Cressie et al. (2010). STRE combines the utilities of basis function approximations and Kalman filtering to achieve dimension reduction in space and fast and dynamic predictions in time. This class is highly useful in modeling large space-time nonstationary data.

A distinct class of spatio-temporal covariance functions has been championed for capturing a special behavior of a subset of spatio-temporal random fields. The class of Lagrangian spatio-temporal covariance functions was developed to model spatio-temporal dependence of transported purely spatial random fields through the use of the Lagrangian reference frame. Models springing from this technique obtain higher covariances along the direction of transport than the covariances lying in the other directions. However, much of the progress in this area was done in stationary variants such as Cox and Isham (1988), where this modeling technique was first proposed, and Salvaña et al. (2020), where the multivariate extension was explored. A recent treatment of this modeling scheme in the multivariate nonstationary setup was provided in Salvaña and Genton (2020). In this work, we formally establish the univariate nonstationary variant of the Lagrangian approach to spatio-temporal covariance construction. Moreover, we propose an efficient estimation methodology such that the novelty of the Lagrangian spatio-temporal nonstationary models translates to usability.

The rest of this paper is organized as follows. Section 2 reviews the developments in the Lagrangian spatio-temporal modeling and formulates the univariate nonstationary extension. Section 3 proposes a practical estimation procedure for nonstationary covariance models of the Lagrangian type. Section 4 presents some simulation studies designed to illustrate the advantages of Lagrangian spatio-temporal models over

other established spatio-temporal models. Section 5 details the application of the new models to a spatio-temporal particulate matter dataset. Section 6 draws a conclusion.

2 Lagrangian Spatio-Temporal Covariances

Under second-order stationarity of the purely spatial random field, Cox and Isham (1988) established that a new class of spatio-temporal stationary covariance functions can be constructed from purely spatial stationary covariance functions by utilizing the principles of Lagrangian reference frame. That is, define a spatio-temporal second-order stationary random field

$$Z(\mathbf{s}, t) = \tilde{Z}(\mathbf{s} - \mathbf{V}t), \quad (\mathbf{s}, t) \in \mathbb{R}^d \times \mathbb{R}, \mathbf{V} \in \mathbb{R}^d, d \geq 1,$$

such that $\tilde{Z}(\mathbf{s})$ is a purely spatial second-order stationary random field. Here \mathbf{V} is a random vector, independent from the purely spatial random field, that describes the velocity of the transport of $\tilde{Z}(\mathbf{s})$ and is often called the advection velocity vector. The resulting spatio-temporal stationary covariance function of $Z(\mathbf{s}, t)$ is

$$\text{cov}\{Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2)\} = \text{cov}\{\tilde{Z}(\mathbf{s}_1 - \mathbf{V}t_1), \tilde{Z}(\mathbf{s}_2 - \mathbf{V}t_2)\} = E_{\mathbf{V}}\{C^S(\mathbf{h} - \mathbf{V}u)\}, \quad (1)$$

where $\mathbf{h} = \mathbf{s}_1 - \mathbf{s}_2$, $u = t_1 - t_2$, and $C^S(\cdot)$ is the purely spatial stationary covariance function of $\tilde{Z}(\mathbf{s})$ on \mathbb{R}^d . By introducing a transformation on the spatial arguments of $C^S(\cdot)$, the number of available spatio-temporal stationary covariance functions would greatly expand by as much as the number of valid purely spatial stationary covariance functions.

The model in (1) can be extended to accommodate multiple variables of interest as shown in Salvaña et al. (2020). That is, suppose at each spatio-temporal location (\mathbf{s}, t) there are $p > 1$ observations corresponding to p different features. This means that the purely spatial second-order stationary random field is now vector valued, i.e., $\tilde{\mathbf{Z}}(\mathbf{s}) = \{\tilde{Z}_1(\mathbf{s}), \dots, \tilde{Z}_p(\mathbf{s})\}^\top$. A multivariate spatio-temporal random field can be similarly defined as above, i.e., $\mathbf{Z}(\mathbf{s}, t) = \tilde{\mathbf{Z}}(\mathbf{s} - \mathbf{V}t) = \{\tilde{Z}_1(\mathbf{s} - \mathbf{V}t), \dots, \tilde{Z}_p(\mathbf{s} - \mathbf{V}t)\}^\top$, with matrix-valued spatio-temporal stationary cross-covariance function

$$\text{cov}\{\mathbf{Z}(\mathbf{s}_1, t_1), \mathbf{Z}(\mathbf{s}_2, t_2)\} = \text{cov}\{\tilde{\mathbf{Z}}(\mathbf{s}_1 - \mathbf{V}t_1), \tilde{\mathbf{Z}}(\mathbf{s}_2 - \mathbf{V}t_2)\} = E_{\mathbf{V}}\{\mathbf{C}^S(\mathbf{h} - \mathbf{V}u)\}, \quad (2)$$

where $\mathbf{C}^S(\cdot)$ is the $p \times p$ matrix-valued purely spatial stationary cross-covariance function of $\tilde{\mathbf{Z}}(\mathbf{s})$ on \mathbb{R}^d . This newly defined multivariate spatio-temporal random field is second-order stationary in space and time.

Using these two previous developments of spatio-temporal covariance functions, a recent review paper further developed the Lagrangian approach in the multivariate nonstationary arena. Salvaña and Genton (2020) established that the model in (2) can be tailored to accommodate an underlying cross-covariance function \mathbf{C}^S that is

nonstationary. This is particularly useful when the multivariate purely spatial random field being transported has nonnegligible second-order nonstationarity. Models arising from their proposal have the form

$$\text{cov}\{\mathbf{Z}(\mathbf{s}_1, t_1), \mathbf{Z}(\mathbf{s}_2, t_2)\} = \text{cov}\{\tilde{\mathbf{Z}}(\mathbf{s}_1 - \mathbf{V}t_1), \tilde{\mathbf{Z}}(\mathbf{s}_2 - \mathbf{V}t_2)\} = \text{Ev} \left\{ \mathbf{C}^S(\mathbf{s}_1 - \mathbf{V}t_1, \mathbf{s}_2 - \mathbf{V}t_2) \right\}, \quad (3)$$

where $\mathbf{C}^S(\cdot, \cdot)$ is a matrix-valued purely spatial nonstationary cross-covariance function of $\tilde{\mathbf{Z}}(\mathbf{s})$ on \mathbb{R}^d .

The models in Eqs. (1)–(3) suggest how the Lagrangian framework can be used to create spatio-temporal covariance functions when one has at one's disposal purely spatial covariance functions that are either univariate stationary, multivariate stationary, or multivariate nonstationary. The univariate nonstationary formulation of the Lagrangian construction can be readily established from (3) when $p = 1$. For completeness, we state this as a theorem below.

Theorem 1 *Let \mathbf{V} be a random vector on \mathbb{R}^d . If $C^S(\mathbf{s}_1, \mathbf{s}_2)$ is a valid purely spatial nonstationary covariance function on \mathbb{R}^d , then,*

$$C(\mathbf{s}_1, \mathbf{s}_2; t_1, t_2) = \text{Ev} \left\{ C^S(\mathbf{s}_1 - \mathbf{V}t_1, \mathbf{s}_2 - \mathbf{V}t_2) \right\}, \quad \mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^d, t_1, t_2 \in \mathbb{R}, \quad (4)$$

is a valid spatio-temporal nonstationary covariance function on $\mathbb{R}^d \times \mathbb{R}$ provided that the expectation exists.

The validity of this theorem follows because it is a special case ($p = 1$) of a theorem proved for general p in Salvaña and Genton (2020). The construction approach in Theorem 1 requires a purely spatial nonstationary covariance function, $C^S(\cdot, \cdot)$, and returns a spatio-temporal covariance function that is nonstationary in both space and time. Theorem 1 implies a purely spatial random field with second-order nonstationarity that is transported to new locations at a velocity \mathbf{V} . The transport behavior, dictated by the velocity \mathbf{V} , influences the covariance through shifting the original spatial arguments of $C^S(\cdot, \cdot)$ by $\mathbf{V}t$. The derived Lagrangian spatio-temporal nonstationary covariance function $C(\mathbf{s}_1, \mathbf{s}_2; t_1, t_2)$ is nonstationary in space, as its fundamental building block is a purely spatial nonstationary covariance function, and is also nonstationary in time, as the transformation from purely spatial to spatio-temporal depends on time t .

There is a rich literature on valid purely spatial nonstationary covariance functions from which we can choose $C^S(\cdot, \cdot)$ including the dimension expansion (Bornn et al. 2012), deformation approach (Sampson and Guttorp 1992), kernel-based methods (Higdon et al. 1999), convolution-based methods (Heaton et al. 2014; Higdon 1998, 2002), spectral methods (Fuentes 2002), orthogonal expansions (Nychka and Saltzman 1998), spatially varying parameters (Neto et al. 2014; Paciorek and Schervish 2006; Gelfand et al. 2004), piece-wise Gaussian process (Kim et al. 2005), covariate-driven approaches (Schmidt et al. 2011), and basis function models (Nychka et al. 2002; Wikle 2010; Chang et al. 2010). Other purely spatial nonstationary models to which Theorem 1 can be applied are discussed in Sampson et al. (2001), Risser (2015), and Stephenson et al. (2004).

Lagrangian spatio-temporal random fields can be classified into two general categories, namely, frozen and non-frozen random fields. The former characterizes Lagrangian spatio-temporal random fields with a constant advection velocity, that is, $\mathbf{V} = \mathbf{v}$. Meanwhile, Lagrangian spatio-temporal random fields that are termed non-frozen are those transported with a random advection velocity \mathbf{V} . Salvaña and Genton (2020) showed realizations of frozen Lagrangian spatio-temporal random fields simulated from (3) when $p = 2$ using prominent classes of purely spatial nonstationary cross-covariance functions, such as the multivariate spatially varying parameters and the multivariate deformation models. Realizations of frozen Lagrangian spatio-temporal nonstationary random fields from the model in (4) can be obtained similarly by assuming that Z_1 and Z_2 in Fig. 2 of Salvaña and Genton (2020) are independent. In the following figures, we show non-frozen Lagrangian spatio-temporal random fields for two models when $\mathbf{V} \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Figure 1a plots the simulated $Z(\mathbf{s}, t)$ from the model

$$C(\mathbf{s}_1, \mathbf{s}_2; t_1, t_2) = \text{E}_{\mathbf{V}} \left(\sigma(\mathbf{s}_1 - \mathbf{V}t_1, \mathbf{s}_2 - \mathbf{V}t_2) \mathcal{M}_{\nu} \left[\{\mathbf{s}_1 - \mathbf{s}_2 - \mathbf{V}(t_1 - t_2)\}^{\top} \times \mathbf{D}(\mathbf{s}_1 - \mathbf{V}t_1, \mathbf{s}_2 - \mathbf{V}t_2)^{-1} \{\mathbf{s}_1 - \mathbf{s}_2 - \mathbf{V}(t_1 - t_2)\} \right]^{1/2} \right), \quad (5)$$

where $\sigma(\mathbf{s}_1 - \mathbf{V}t_1, \mathbf{s}_2 - \mathbf{V}t_2)$ is the spatially varying variance parameter and the matrix $\mathbf{D}(\mathbf{s}_1 - \mathbf{V}t_1, \mathbf{s}_2 - \mathbf{V}t_2)$ serves as the spatially varying scale parameter (Kleiber and Nychka 2012). Here $\mathcal{M}_{\nu}(\cdot)$ is the univariate Matérn correlation with smoothness parameter $\nu > 0$, $\mathbf{D}(\mathbf{s}_1, \mathbf{s}_2) = \frac{1}{2} \{\mathbf{D}(\mathbf{s}_1) + \mathbf{D}(\mathbf{s}_2)\}$, and $\sigma(\mathbf{s}_1, \mathbf{s}_2) = |\mathbf{D}(\mathbf{s}_1)|^{1/4} |\mathbf{D}(\mathbf{s}_2)|^{1/4} |\mathbf{D}(\mathbf{s}_1, \mathbf{s}_2)|^{-1/2}$. The matrix $\mathbf{D}(\mathbf{s})$ is parameterized through its spectral decomposition, i.e.

$$\mathbf{D}(\mathbf{s}) = \begin{bmatrix} \cos \{\phi(\mathbf{s})\} & -\sin \{\phi(\mathbf{s})\} \\ \sin \{\phi(\mathbf{s})\} & \cos \{\phi(\mathbf{s})\} \end{bmatrix} \begin{bmatrix} \lambda_1(\mathbf{s}) & 0 \\ 0 & \lambda_2(\mathbf{s}) \end{bmatrix} \begin{bmatrix} \cos \{\phi(\mathbf{s})\} & \sin \{\phi(\mathbf{s})\} \\ -\sin \{\phi(\mathbf{s})\} & \cos \{\phi(\mathbf{s})\} \end{bmatrix}.$$

Figure 1b illustrates the random field generated from the non-frozen Lagrangian deformation

$$C(\mathbf{s}_1, \mathbf{s}_2; t_1, t_2) = \text{E}_{\mathbf{V}} \left[\sigma^2 \mathcal{M}_{\nu} \{a \|\mathbf{f}(\mathbf{s}_1 - \mathbf{V}t_1) - \mathbf{f}(\mathbf{s}_2 - \mathbf{V}t_2)\|\} \right], \quad (6)$$

where $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a deterministic nonlinear smooth bijective deformation function and σ^2 and a are the variance and scale parameters, respectively. In the example in Fig. 1b, $\sigma^2 = a = \nu = 1$.

To illustrate the effect of the advection velocity $\mathbf{V} \sim \mathcal{N}_2 \{(0.1, 0.1)^{\top}, 0.01 \times \mathbf{I}_2\}$ on the space-time dependence of the random fields in Fig. 1, we examine two locations, marked with ‘×’, which we call “reference locations”. We plot as heatmaps the covariance between the observations at each reference location and the observations at all locations, including the reference locations themselves. For example, in Fig. 2a, the first image in the first row gives the covariance between $Z(\mathbf{s}_{\text{Ref Loc } 1}, 1)$ and $Z(\mathbf{s}_l, 1)$, at every pixel location $\mathbf{s}_l, l = 1, \dots, 2500$. The second image in the first

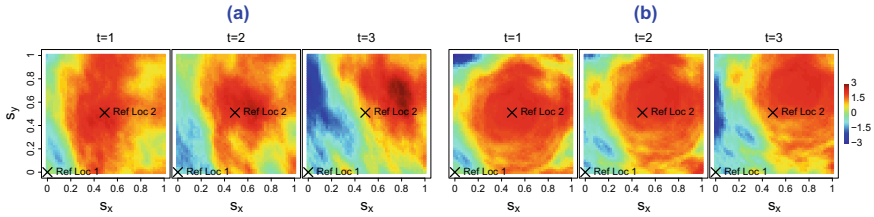


Fig. 1 Simulated realizations in the unit square on a 50×50 grid from the non-frozen Lagrangian nonstationary covariance models in (5) and (6) with $\mathbf{V} \sim \mathcal{N}_2 \{ (0.1, 0.1)^\top, 0.01 \times \mathbf{I}_2 \}$, \mathbf{I}_2 is the 2×2 identity matrix. (a) The spatially varying parameters have the following representations: for $\mathbf{s} = (s_x, s_y)^\top$, $\phi(\mathbf{s}) = (s_x - 0.5) + 2(s_y - 0.5) + (s_y - 0.5)^2$, $\lambda_1(\mathbf{s}) = -3 - 6(s_x - 0.5)^2 - 7(s_y - 0.5)^2$, and $\lambda_2(\mathbf{s}) = -5 + (s_x - 0.5)^2 - 4(s_y - 0.5)^2$. (b) The deformation function assumed is the point-source deformation, i.e., $\mathbf{f}(\mathbf{s}) = \mathbf{b} + (\mathbf{s} - \mathbf{b})\{1 + \exp(-0.5\|\mathbf{s} - \mathbf{b}\|^2)\}$, $\mathbf{b} = (0.15, 0.15)^\top$. Reference locations 1 and 2 are marked with ‘x’

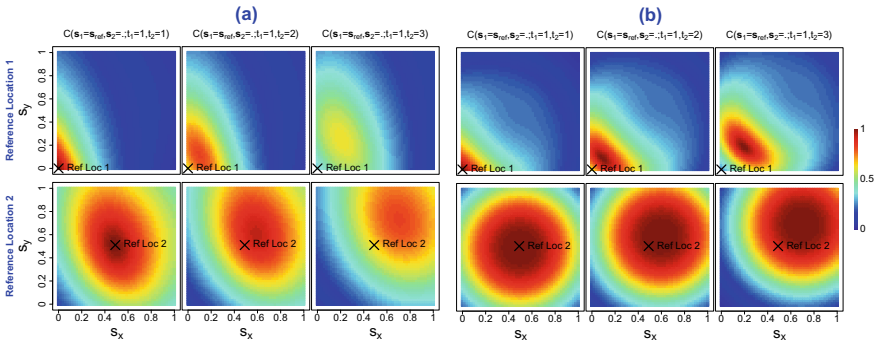


Fig. 2 Heatmaps of the non-frozen Lagrangian nonstationary covariance models in (5) and (6) observed at two reference locations marked with ‘x’. (a) shows the strengths of dependence between any two locations in space and time under the spatially varying parameters model and (b) under the deformation model. See Fig. 1 for the corresponding random field realizations

row plots the covariance between $Z(\mathbf{s}_{\text{Ref Loc } 1}, 1)$ and $Z(\mathbf{s}_l, 2)$, at every pixel location \mathbf{s}_l , $l = 1, \dots, 2500$. Lastly, the third image in the first row plots the covariance between $Z(\mathbf{s}_{\text{Ref Loc } 1}, 1)$ and $Z(\mathbf{s}_l, 3)$, at every pixel location \mathbf{s}_l , $l = 1, \dots, 2500$. All the other plots are organized in the same manner. Notice that among the covariances taken at the same temporal locations, i.e., $t_1 = t_2$, the maximum covariance occurs at the reference location. However, among the covariances taken between any two space-time locations that are one time step apart, the maximum covariance no longer occurs at the reference location. Instead, it can be observed at a spatial location $(0.1, 0.1)^\top$ away from the reference location. A similar observation can be made when taking covariances between any two space-time locations that are two time steps apart.

3 Estimation

The parameters for any spatio-temporal nonstationary covariance functions spawned by the Lagrangian approach include both purely spatial and advection velocity parameters. The estimation methods to recover the former depend on the form of C^S and are already fully developed in their respective references; see Sect. 2. Here we propose a way to extend those estimation methods to space-time in order to recover both the purely spatial and the additional advection velocity parameters. We focus on an estimation strategy that operates on the spatio-temporal nonstationary covariance matrix built using all the spatio-temporal locations. This allows inferences regarding the second-order nonstationarity structure of the transported purely spatial random field possible. However, alternative estimation strategies which involve fitting local spatio-temporal stationary models can also be considered (Kuusela and Stein 2018).

3.1 Thin Plate Splines

Throughout the remainder of this work, we narrow our attention to Lagrangian spatio-temporal nonstationary models whose C^S are the deformation and spatially varying parameters models. We focus on these two classes because their second-order nonstationarity parameters can be considered a surface and we aim to leverage a technique used to model surfaces, namely, thin plate splines (TPS). The TPS is a basis function and is used to interpolate surfaces using a predetermined set of landmarks or the locations where the basis functions are centered (Bookstein 1989; Wahba 1990; Donato and Belongie 2002; Chen and Geman 2014). TPS is a central topic in morphometrics and has found a wide range of applications including biomedical, computer vision, data mining, and engineering (Whitbeck and Guo 2006; Hegland et al. 1997; Tenakoon et al. 2013; Chen et al. 2017; Bazen and Gerez 2003). This section describes how TPS can be appropriately applied to model the second-order nonstationarity parameters of the Lagrangian spatio-temporal nonstationary models.

Suppose $\psi(\mathbf{s})$ is an unknown second-order nonstationarity parameter of interest at spatial location \mathbf{s} . This parameter might be the x – or y –coordinate in the new spatial domain for the deformation model or the spatially varying parameters $\lambda_1(\mathbf{s})$, $\lambda_2(\mathbf{s})$, or $\phi(\mathbf{s})$. The TPS model for $\psi(\mathbf{s})$ is

$$\psi(\mathbf{s}) = A_1 + A_2 s_x + A_3 s_y + \sum_{i=1}^L w_i U(\|\mathbf{s}_i^* - \mathbf{s}\|^2), \quad (7)$$

where $U(h) = h^2 \log h$, for $h > 0$, and zero otherwise, is a basis function, $\mathbf{A} = (A_1, A_2, A_3)^\top \in \mathbb{R}^3$ and $\mathbf{w} \in \mathbb{R}^L$ are the parameters responsible for the affine and nonlinear components of the transformation, respectively, and L is the number of landmarks. Sampson (2015) pointed out several problems springing from the formulation in (7), including multiple local maxima in the log-likelihood function and

highly correlated parameters. Hence, following their recommendation, we adopt the form in (7) with $w_i = \sum_{j=1}^{L-3} \beta_j g_{i,j}$, such that $\mathbf{g}_j = (g_{1,j}, \dots, g_{L,j})^\top \in \mathbb{R}^L$, $j = 1, \dots, L-3$, also called the principal warps, are the last $L-3$ eigenvectors of the bending energy matrix \mathbf{B} corresponding to its $L-3$ nonzero eigenvalues. The bending energy matrix \mathbf{B} is the upper left $L \times L$ sub-matrix of $\mathbf{B} = [\mathbf{D} \mathbf{P}; \mathbf{P}^\top \mathbf{O}]^{-1} \in \mathbb{R}^{(L+d+1) \times (L+d+1)}$ with elements:

- $\mathbf{D} \in \mathbb{R}^{L \times L}$ such that for $l, r = 1, \dots, L$, $D_{lr} = d_{lr}^2 \log(d_{lr})$, if $l \neq r$, and $D_{lr} = 0$, otherwise, where $d_{lr} = \|\mathbf{s}_l^* - \mathbf{s}_r^*\|$,
- $\mathbf{P} \in \mathbb{R}^{L \times (d+1)}$, where the l -th row of \mathbf{P} is $(1, \mathbf{s}_l^\top)$, $\mathbf{s} \in \mathbb{R}^d$, and $l = 1, \dots, L$, and
- \mathbf{O} is a zero matrix in $\mathbb{R}^{(d+1) \times (d+1)}$.

Together, the linear combinations of the coefficients, $\beta_{i,j}$, and the principal warps, \mathbf{g}_j , are termed partial warps.

A key ingredient in the TPS model is the set of landmarks, $\{\mathbf{s}_1^*, \mathbf{s}_2^*, \dots, \mathbf{s}_L^*\}$. The TPS model interpolates at these landmark points while preserving maximal smoothness (Bazen and Gerez 2003). The placement of these landmarks dictates the quality of the parameter estimates (Lewis et al. 2004). The landmarks and the number of landmarks are fixed prior to modeling and the choice is left to the discretion of the modeler. In the morphometrics literature, the landmarks are often positioned where important features can be observed (Gunz and Mitteroecker 2013). In the spatial statistics literature, the observation locations are commonly designated as landmarks (Kleiber et al. 2014).

In studying Lagrangian spatio-temporal random fields, there is a need to distinguish between the observation locations and the domain of the transported random field. The former refers to the predefined locations where measurements are obtained, e.g., regular latitude/longitude grid, wireless sensor networks, wind turbine sites, meteorological towers, and many others. The latter has its own coordinate system. The measurements contained in the transported random field get picked up by the data collection tools at the observation locations as the random field travels past them. In frozen Lagrangian spatio-temporal random fields, the measurement $Z(\mathbf{s}, t)$ collected at observation location \mathbf{s} at time t corresponds to the measurement $Z(\mathbf{s} - \mathbf{v}t)$ at spatial location $\mathbf{s} - \mathbf{v}t$ in the domain of the transported random field. Figure 3 shows a frozen Lagrangian spatio-temporal deformed random field traveling at a constant velocity of $\mathbf{v} = (0.5, 0.5)^\top$. While the observation locations are fixed at any time, the corresponding locations in the Lagrangian random field are not. Choosing the observation locations as landmarks, therefore, will not suffice in capturing the non-stationarity of the entire Lagrangian spatio-temporal random field as every region in the domain should be represented by these landmarks. Assuming that the domain of the Lagrangian spatio-temporal random field is larger than the domain of observation locations, we advocate to situate the landmarks on a regular grid that covers the entire Lagrangian spatio-temporal random field. In practice, unfortunately, the appropriate size and resolution of this regular grid of landmarks cannot be identified prior to modeling. However, cross-validation studies can be performed to determine the suitable positioning and number of landmarks.

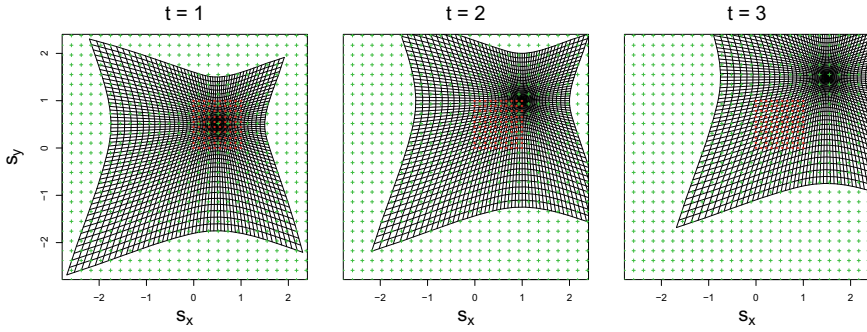


Fig. 3 Marked in red are the observation locations on a regular 10×10 grid. Superimposed in black are the spatial locations on the domain of the frozen Lagrangian spatio-temporal deformed random field which travels past the observation locations with an advection velocity $\mathbf{v} = (0.5, 0.5)^\top$, and in green are the landmarks. The landmarks (green) may or may not coincide with the observation locations (red)

3.2 Maximum Likelihood Estimation and Likelihood Approximations in the Temporal Domain

Having established the representation of the unknown nonstationarity parameters, we introduce the estimation procedure carried out in this work. Suppose $\mathbf{Z} = \{Z(\mathbf{s}_1, t_1), Z(\mathbf{s}_2, t_2), \dots, Z(\mathbf{s}_n, t_n)\}^\top$ is a zero mean measurement vector where $n \in \mathbb{Z}^+$ is the total number of space-time locations. Inference is performed through maximizing the log-likelihood

$$l(\boldsymbol{\Theta}; \mathbf{Z}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\Theta})| - \frac{1}{2} \mathbf{Z}^\top \boldsymbol{\Sigma}(\boldsymbol{\Theta})^{-1} \mathbf{Z} \tag{8}$$

with respect to all the parameters collected in $\boldsymbol{\Theta} \in \mathbb{R}^q$. Here $\boldsymbol{\Theta}$ includes all the purely spatial, advection velocity, and the TPS parameters, and q is the total number of parameters. The $n \times n$ covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\Theta})$ is formed by a parametric spatio-temporal nonstationary covariance function. Penalties can be introduced to Equation (8) such as the L_1 penalty for the deformation models in order to avoid folding of the surface (Sampson 2015).

For spatio-temporal measurements that are regularly spaced in time, \mathbf{Z} can be rewritten as $\mathbf{Z} = (\mathbf{Z}_1^\top, \dots, \mathbf{Z}_T^\top)^\top \in \mathbb{R}^{N \cdot T}$ such that $\mathbf{Z}_t = \{Z(\mathbf{s}_1, t), \dots, Z(\mathbf{s}_N, t)\}^\top \in \mathbb{R}^N$, for $t = 1, \dots, T$. Here N and T specify the number of spatial and temporal locations, respectively, and $n = N \cdot T$. Furthermore, the log-likelihood function above can be approximated as follows:

$$l(\boldsymbol{\Theta}; \mathbf{Z}_1, \dots, \mathbf{Z}_T) \approx l(\boldsymbol{\Theta}; \mathbf{Z}_{1,t^*}) + \sum_{j=t^*+1}^T l(\boldsymbol{\Theta}; \mathbf{Z}_j | \mathbf{Z}_{j-t^*, j-1}), \tag{9}$$

where $\mathbf{Z}_{a,b} = (\mathbf{Z}_a^\top, \dots, \mathbf{Z}_b^\top)^\top \in \mathbb{R}^{Mt^*}$, for $a < b$, and t^* specifies the number of consecutive temporal locations included in the conditional distribution. Here $l(\boldsymbol{\Theta}; \mathbf{Z}_j | \mathbf{Z}_{j-t^*,j-1})$ is the log-likelihood function based only on the vector of space-time measurements $\mathbf{Z}_{j-t^*,j-1} = (\mathbf{Z}_{j-t^*}^\top, \dots, \mathbf{Z}_{j-1}^\top)^\top$. This kind of approximation is usually preferred when T is large and the dependence in time relies heavily only on the more recent measurements (Stein 2005c).

3.3 Two-Step Maximum Likelihood Estimation

The inclusion of the nonstationarity parameters in the model increases the dimension of the estimation problem. This kind of setup is known to run into numerical difficulties and complications (Kathuria et al. 2019; Zhu and Wu 2010; Li and Sun 2018). Therefore, as a practical alternative to joint estimation of all the parameters, in this work, the estimation problem is split into two parts. First, a Lagrangian spatio-temporal stationary model is assumed and all the associated purely spatial and advection parameters are estimated by maximizing the approximated log-likelihood in (9). Second, fixing the estimates found in the first step, the nonstationary version of the model is assumed and the parameters involved in the TPS are estimated also by maximizing (9). After the second step, it is likely that the optimization routine may still not reach the global maximum of (9). Hence, assuming the nonstationary model, iterating between the two steps several times is pursued until a stopping criterion is satisfied.

4 Simulation Study: Lagrangian Versus Non-Lagrangian Spatio-Temporal Models

The Lagrangian spatio-temporal covariance functions are primarily used to model transported space-time data. There are other classes of spatio-temporal covariance functions that model space-time data that are not necessarily transported. In this section, we investigate the outcome of fitting a non-Lagrangian model to transported space-time data and the outcome of fitting a Lagrangian model to space-time data that are not transported. We conduct the study under both second-order stationarity and nonstationarity assumptions.

4.1 Second-Order Stationarity

For the Lagrangian spatio-temporal model, we hinge our simulation studies on a particular class of non-frozen models whose explicit forms were derived in Schlather

(2010). When $\mathbf{V} \sim \mathcal{N}_d(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V)$ and C^S is the stationary squared exponential covariance function, the model in (1) takes the form

$$C(\mathbf{h}, u) = \frac{1}{\sqrt{|\mathbf{I}_d + \boldsymbol{\Sigma}_V u^2|}} \exp \left\{ -a (\mathbf{h} - \boldsymbol{\mu}_V u)^\top (\mathbf{I}_d + \boldsymbol{\Sigma}_V u^2)^{-1} (\mathbf{h} - \boldsymbol{\mu}_V u) \right\}, \quad (10)$$

where $a > 0$ is a scale parameter in space inherited from C^S , and $\boldsymbol{\mu}_V$ and $\boldsymbol{\Sigma}_V$ are the Lagrangian parameters. When $\boldsymbol{\mu}_V = \mathbf{0}$ and $\boldsymbol{\Sigma}_V = \sigma_V^2 \mathbf{I}_d$, the Lagrangian model above reduces to

$$C(\mathbf{h}, u) = \frac{1}{(1 + \sigma_V^2 u^2)^{d/2}} \exp \left(-\frac{a \|\mathbf{h}\|^2}{1 + \sigma_V^2 u^2} \right), \quad (11)$$

which is a spatio-temporal isotropic covariance function under the Gneiting class (Gneiting 2002). The Gneiting model in (11), therefore, corresponds to a particular Lagrangian model wherein the advection velocity vector has mean zero and has independent components with common variance. While σ_V^2 is interpreted as the marginal variance of each component of \mathbf{V} in Lagrangian models, in non-Lagrangian models such as that in (11), σ_V^2 serves as a scale parameter in time, whose inverse controls the range of dependence in time.

A question of scientific interest is how the two models differ when the components of the advection velocity are no longer uncorrelated or when they do not share a common variance or when the advection velocity vector has a nonzero mean. To answer the first inquiry, we can scrutinize the form in (10) and compare it with (11). Suppose $d = 2$, $\boldsymbol{\mu}_V = \mathbf{0}$, and $\boldsymbol{\Sigma}_V = \sigma_V^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ then (10) reduces to

$$C(\mathbf{h}, u) = \frac{1}{\sqrt{(1 + \sigma_V^2 u^2)^2 - (\rho \sigma_V^2 u^2)^2}} \exp \left[-a \left\{ \frac{(h_x^2 + h_y^2)(1 + \sigma_V^2 u^2) - 2h_x h_y \rho \sigma_V^2 u^2}{(1 + \sigma_V^2 u^2)^2 - (\rho \sigma_V^2 u^2)^2} \right\} \right]. \quad (12)$$

Direct comparisons between (11) and (12) for different values of ρ are not straightforward since the terms bearing ρ involve the temporal lag u and the components of the spatial lag $\mathbf{h} = (h_x, h_y)^\top$. However, we can plot the values of (12) for different ρ , u , and \mathbf{h} , in order to visualize how the non-frozen Lagrangian spatio-temporal model deviates from the non-Lagrangian spatio-temporal model when the components of \mathbf{V} are correlated. Figure 4 provides such illustrations. It juxtaposes the covariance function values of the non-frozen Lagrangian spatio-temporal model, C^{LGR} for notational convenience, at different combinations of spatial lags with Euclidean norm equal to 1, at $u = 1, 2$, and 3, and at different strengths of dependence between the components of the advection velocity. In the plots, the values of the covariance function are plotted as the distance from the origin $(0, 0)$ to (h_x, h_y) . Note that the case $\rho = 0$ corresponds to the spatio-temporal Gneiting model in (11), denoted as C^G . The isotropy of C^G , at any u , manifests by the constant value of C^G when evaluated at any (h_x, h_y) . Another standout observation is that the value of C^{LGR} depends on

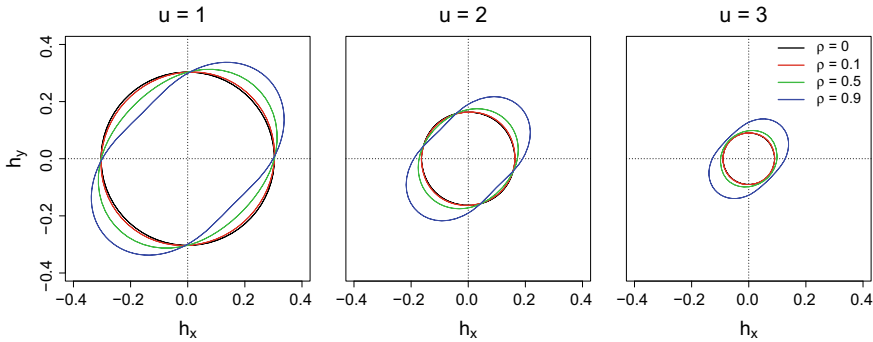


Fig. 4 Values of the non-frozen Lagrangian spatio-temporal covariance model in (12) for $\rho = 0, 0.1, 0.5,$ and $0.9,$ at temporal lags $u = 1, 2,$ and $3,$ at every $\mathbf{h} = (h_x, h_y)^\top$ such that $\|\mathbf{h}\|_2 = 1.$ Note that the case $\rho = 0$ corresponds to the non-Lagrangian Gneiting model in (11)

the signs of the components of the spatial lag and the magnitude of the correlation parameter $\rho.$

It can also be seen in the example in Fig. 4 that at $u = 1,$ when h_x and h_y have the same signs, C^G is less than $C^{LGR}.$ However, when h_x and h_y have different signs, C^G is greater than $C^{LGR}.$ This relationship between C^G and C^{LGR} at $u = 1$ does not persist as the temporal lag increases as other scenarios are observed. At $u = 3,$ for example, C^G and C^{LGR} are almost identical when ρ is near zero. However, when $\rho = 0.5$ or $\rho = 0.9,$ C^G is less than C^{LGR} in any direction. The difference, therefore, between C^G and C^{LGR} under the presence of a nonzero dependence parameter between the components of \mathbf{V} is not clear-cut but can be explored under some scenarios. Nevertheless, the deviation of C^{LGR} from C^G gets more pronounced as ρ increases.

We turn to some numerical experiments to answer the other unexplored questions, including what happens when C^G is fitted to data simulated from $C^{LGR},$ denoted $D^{LGR},$ such that the components of \mathbf{V} have different marginal variances or \mathbf{V} has a nonzero mean. Suppose $T = 10, N = 100, d = 2.$ The values $D^{LGR} = \mathbf{Z} = (\mathbf{Z}_1^\top, \dots, \mathbf{Z}_{10}^\top)^\top,$ such that $\mathbf{Z}_t = \{Z(\mathbf{s}_1, t), \dots, Z(\mathbf{s}_{100}, t)\}^\top, (\mathbf{s}, t) \in \mathbb{R}^2 \times \mathbb{R},$ are simulated from (10), with $a = 5,$ on a 10×10 grid in the unit square, under the following distributions of $\mathbf{V}:$

- (a) $\mathbf{V} \sim \mathcal{N}_2 \left\{ \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$ at different values of $\rho;$
- (b) $\mathbf{V} \sim \mathcal{N}_2 \left\{ \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \right\}$ at different values of $\sigma_y^2;$
- (c) $\mathbf{V} \sim \mathcal{N}_2 \left\{ \boldsymbol{\mu} = (\mu, \mu)^\top, \boldsymbol{\Sigma} = \mathbf{I}_2 \right\}$ at different values of $\mu.$

We reserve the values of \mathbf{Z}_{10} for prediction purposes and use the remaining 900 spatio-temporal realizations for estimation. Given the small problem size, full maximum likelihood estimation is performed; see (8). At this point, we question the effect of different values of $\rho, \sigma_y^2,$ and μ on the estimates of σ_y^2 in the non-Lagrangian model

in (11). Figure 5 gives the boxplots of parameter estimates $\hat{\sigma}_{\mathbf{V}}^2$ for 100 rounds of fitting C^G on D^{LGR} . The values of $\hat{\sigma}_{\mathbf{V}}^2$ reflect the changing degree of dependence in space-time as we change the values of the different parameters associated to the distribution of \mathbf{V} . In the first panel in Fig. 5, for example, when $\rho = 0.9$, the median of the estimates is 0.887 which translates to a stronger dependence in time, a fact also established in Fig. 4. In the middle set of boxplots, interestingly, the median of $\hat{\sigma}_{\mathbf{V}}^2$ is approximately equal to $(1 + \sigma_y^2)/2$. This result cannot be easily explained mathematically. Numerically, however, this is expected as the optimization routine finds the isotropic model parameters that maximize the log-likelihood given data simulated from a model with elliptical contours that are stretched in the x-axis. Lastly, as the mean of \mathbf{V} gets farther from $\mathbf{0}$, the estimate for $\sigma_{\mathbf{V}}^2$ has to compensate for a faster decorrelation in time which explains the increasing median of $\hat{\sigma}_{\mathbf{V}}^2$ as μ increases. In the initial experiments concerning Fig. 5c, a number of experimental replicates obtained $\hat{\sigma}_{\mathbf{V}}^2$ with values greater than 100 as μ increases. To obtain more compact boxplots, we re-ran the experiments and bounded the values that $\hat{\sigma}_{\mathbf{V}}^2$ can take to 10. This does not alter the insights provided by the unconstrained version of the experiments for Fig. 5c and the results presented in Fig. 5a and b. That is, as the non-frozen Lagrangian spatio-temporal model deviates from the non-Lagrangian scenario, i.e., $\mathbf{V} \sim \mathcal{N}_2(\boldsymbol{\mu}_{\mathbf{V}}, \boldsymbol{\Sigma}_{\mathbf{V}})$, where $\boldsymbol{\mu}_{\mathbf{V}} = \mathbf{0}$, and $\boldsymbol{\Sigma}_{\mathbf{V}} = \sigma_{\mathbf{V}}^2 \mathbf{I}_2$, the more disparate the models (10) and (11) become.

Next, we study the effect of ρ on the predictions and the quality of those predictions. Often, the assessment of the quality of the predictions is done by computing the Mean Square Error (MSE)

$$\text{MSE} = \frac{1}{100} \sum_{l=1}^{100} \left\{ \hat{Z}(\mathbf{s}_l, 10) - Z(\mathbf{s}_l, 10) \right\}^2,$$

where $\hat{Z}(\mathbf{s}_l, 10)$ is the prediction for $Z(\mathbf{s}_l, 10)$ at spatial location $\mathbf{s}_l, l = 1, \dots, 100$, and temporal location $t = 10$. Assuming the mean of the measurement vector that was used to estimate the parameters is $\mathbf{0}$, i.e., $E(\mathbf{Z}_{1,9}) = \mathbf{0}$, where $\mathbf{Z}_{1,9} = (\mathbf{Z}_1^\top, \dots, \mathbf{Z}_9^\top)^\top$, predictions are computed using the simple kriging predictor

$$\hat{Z}(\mathbf{s}_l, 10) = \mathbf{c}_l^\top \boldsymbol{\Sigma}(\boldsymbol{\Theta})^{-1} \mathbf{Z}_{1,9}.$$

Here \mathbf{c}_l is the vector of $N \times (T - 1)$ covariance function values between $Z(\mathbf{s}_l, 10)$ and $\mathbf{Z}(\mathbf{s}_r, t), r = 1, \dots, N$ and $t = 1, \dots, 9$, i.e.

$$\mathbf{c}_l = \{C(\mathbf{s}_l, \mathbf{s}_1; 10, 1), \dots, C(\mathbf{s}_l, \mathbf{s}_N; 10, 1), C(\mathbf{s}_l, \mathbf{s}_1; 10, 2), \dots, C(\mathbf{s}_l, \mathbf{s}_N; 10, 9)\}^\top. \tag{13}$$

Nevertheless, the MSE is unable to give an appropriate measure of the loss of statistical efficiency in cases when a different model is used instead of the true model. In this regard, we turn to the proposed criteria of Stein (1999), namely, the Loss of Efficiency (LOE) and the Misspecification of the Mean Square Error (MOM). The

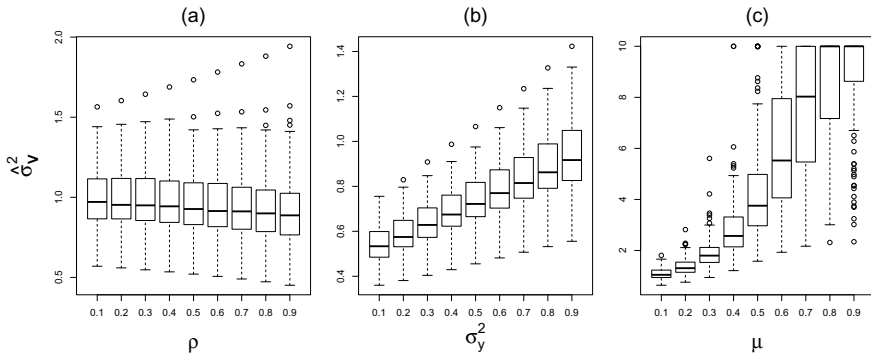


Fig. 5 Estimates of σ_y^2 in (11) when fitted to D^{LGR} generated using (10) with (a) $\mathbf{V} \sim \mathcal{N}_2 \left\{ \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$ at different values of ρ , (b) $\mathbf{V} \sim \mathcal{N}_2 \left\{ \boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \right\}$ at different values of σ_y^2 , and (c) $\mathbf{V} \sim \mathcal{N}_2 \left\{ \boldsymbol{\mu} = (\mu, \mu)^\top, \boldsymbol{\Sigma} = \mathbf{I}_2 \right\}$ at different values of μ

LOE and MOM at space-time location (\mathbf{s}_l, t) are computed as follows:

$$\text{LOE}(\mathbf{s}_l, t) = \frac{E_{tr,m}(\mathbf{s}_l, t)}{E_{tr}(\mathbf{s}_l, t)} - 1 \quad \text{and} \quad \text{MOM}(\mathbf{s}_l, t) = \frac{E_m(\mathbf{s}_l, t)}{E_{tr,m}(\mathbf{s}_l, t)} - 1, \quad (14)$$

where $E_{tr}(\mathbf{s}_l, t)$ and $E_m(\mathbf{s}_l, t)$ are the mean square errors of the predictors under the true, tr , and misspecified, m , models, respectively, and are calculated as follows:

$$E_j(\mathbf{s}_l, t) = C(\mathbf{s}_l, \mathbf{s}_l; t, t) - \mathbf{c}_l^{j\top} \boldsymbol{\Sigma}(\boldsymbol{\Theta}^*)^{-1} \mathbf{c}_l^j, \quad j = \{tr, m\}, \quad (15)$$

where \mathbf{c}_l^j and $\boldsymbol{\Sigma}(\boldsymbol{\Theta}^*)$ are computed using $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}$, for model tr , and $\boldsymbol{\Theta}^* = \hat{\boldsymbol{\Theta}}^m$ for model m . Here $\boldsymbol{\Theta}$ is the true parameter vector while $\hat{\boldsymbol{\Theta}}^m$ is the estimated parameter vector under the model m . On the other hand, $E_{tr,m}(\mathbf{s}_l, t)$ is the mean square error, with respect to the true model, of the predictor that is derived from the misspecified model, and is given as

$$E_{tr,m}(\mathbf{s}_l, t) = C(\mathbf{s}_l, \mathbf{s}_l; t, t) - 2\mathbf{c}_l^{tr\top} \boldsymbol{\Sigma}(\hat{\boldsymbol{\Theta}}^m)^{-1} \mathbf{c}_l^m + \mathbf{c}_l^{m\top} \boldsymbol{\Sigma}(\hat{\boldsymbol{\Theta}}^m)^{-1} \boldsymbol{\Sigma}(\boldsymbol{\Theta}) \boldsymbol{\Sigma}(\hat{\boldsymbol{\Theta}}^m)^{-1} \mathbf{c}_l^m. \quad (16)$$

Figure 6 plots the LOE and MOM values at every prediction location at $t = 10$. The LOE is closer to zero when ρ is near zero compared to the LOE when $\rho = 0.9$. An LOE near zero indicates that the misspecified model is similar to the true model. Furthermore, the change in the LOE at each prediction location as we increase ρ is different and is somehow dictated by the contours of the distribution of \mathbf{V} . This means that the quality of predictions is not equal everywhere and the worst misspecification occurs in the direction where the highest correlation under C^{LGR} occurs. The plots for the MOM convey the same story.

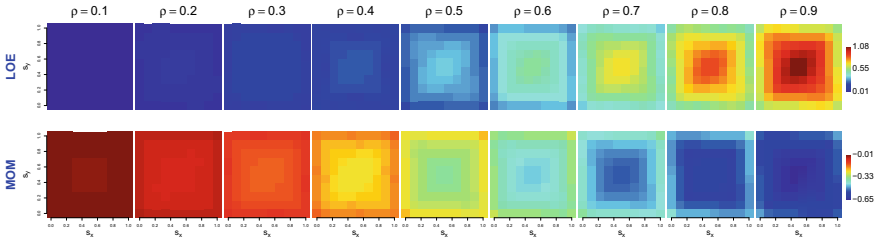


Fig. 6 Values of the LOE and MOM at every spatial location when C^G is fitted to D^{LGR} simulated with $\rho = 0.1, \dots, 0.9$. The closer the LOE values are to zero or the bluer the plots are, the better. Similarly, the closer the values of the MOM are to zero or the redder the plots are, the better

4.2 Second-Order Nonstationarity

Similar analyses cannot be easily adapted to the nonstationary counterparts of the models in the previous section since the covariances may depend on arbitrary nonstationarity parameters at each spatio-temporal location. However, we can draw insights on the consequences of fitting C^G to data generated from C^{LGR} and vice versa, under second-order nonstationarity, by again looking at the quality of predictions.

The non-Lagrangian nonstationary covariance model used in the succeeding numerical experiments, C_{NS}^G , is the nonstationary version of (11) proposed in Garg et al. (2012). It has the form

$$C(\mathbf{s}_1, \mathbf{s}_2; u) = \frac{\sigma(\mathbf{s}_1, \mathbf{s}_2)}{(1 + a_t u^2)^{d/2}} \mathcal{M}_v \left[\frac{\{(\mathbf{s}_1 - \mathbf{s}_2)^\top \mathbf{D}(\mathbf{s}_1, \mathbf{s}_2)^{-1} (\mathbf{s}_1 - \mathbf{s}_2)\}^{1/2}}{(1 + a_t u^2)^{1/2}} \right], \quad (17)$$

where $\sigma(\mathbf{s}_1, \mathbf{s}_2)$ and $\mathbf{D}(\mathbf{s}_1, \mathbf{s}_2)$ are defined in Sect. 2 and the parameter $a_t > 0$ is the scale parameter in time. Data generated from (17) are labeled D_{NS}^G . On the other hand, C_{NS}^{LGR} is the non-frozen Lagrangian spatio-temporal nonstationary model in (5) and data from this model are tagged as D_{NS}^{LGR} . We assess the quality of the predictions by comparing the mean LOEs (MLOE) and mean MOMs (MMOM) when C_{NS}^G is fitted to D_{NS}^{LGR} and when C_{NS}^{LGR} is fitted to D_{NS}^G (Hong et al. 2021). Figure 7 plots the medians of the computed MLOE and MMOM for both scenarios after 100 rounds of parameter estimation via maximization of the full log-likelihood at different values of ρ . It can be seen that at every ρ , the median MLOE is greater when C_{NS}^G is fitted to D_{NS}^{LGR} compared to the median MLOE when C_{NS}^{LGR} is fitted to D_{NS}^G . Moreover, both scenarios of model misspecification yield median MMOMs that are far from zero. However, the median MMOMs are more favorable in cases when C_{NS}^{LGR} is fitted to D_{NS}^G at larger values of ρ . This should advocate the use of Lagrangian models even when the random field does not appear to be transported.

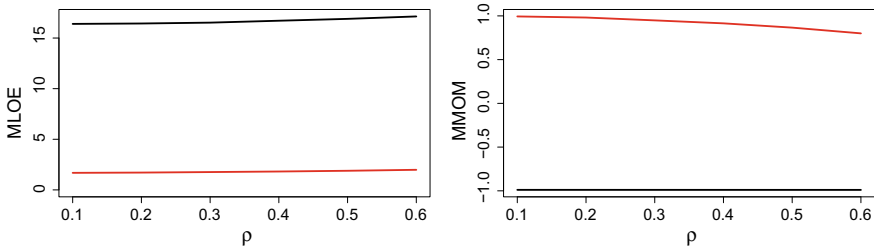


Fig. 7 Medians of the MLOE and MMOM when C_{NS}^G is fitted to D_{NS}^{LGR} (black) and when C_{NS}^{LGR} is fitted to D_{NS}^G (red) at different values of ρ

5 Application to Particulate Matter Data

A spatio-temporal process that is known to be heavily influenced by the presence of a transport medium is pollutant measurements. Pollutants in the atmosphere are transported by the wind to neighboring sites over time (National Research Council 2010). This behavior causes the pollutant measurements at one site to be strongly correlated to the pollutant measurements at a site along the path of transport. Thus, a model incorporating this transport behavior to its spatio-temporal dependence structure is physically reasonable.

5.1 PM 2.5 Data

We study the spatio-temporal dependence of log particulate matter (log PM 2.5) residuals. We retrieve the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) reanalyses dataset of hourly PM 2.5 measurements from NASA Earthdata. Preliminary processing of the raw PM 2.5 data was done to ensure that the resulting spatio-temporal residuals fulfill the modeling assumptions of Gaussianity and zero mean. We consider the first 744 hourly measurements for each year from 1980–2019, at 550 spatial locations, as spatio-temporally dependent, while measurements across years are regarded as spatio-temporally independent. Since the measurements between any two years are at least 11 months apart, this independence assumption is reasonable. Figure 8 maps the log PM 2.5 residuals at 550 locations in Saudi Arabia, at 4 h intervals, starting from 0:00 of January 1, 2017. The transport behavior is evident and can be identified when following the red, blue, and yellow blobs. The direction of transport at every spatial and temporal location appears to be different as the displacements of the red blob indicate transport to the South or South East direction while a North or North West movement can be detected from the yellow blob.

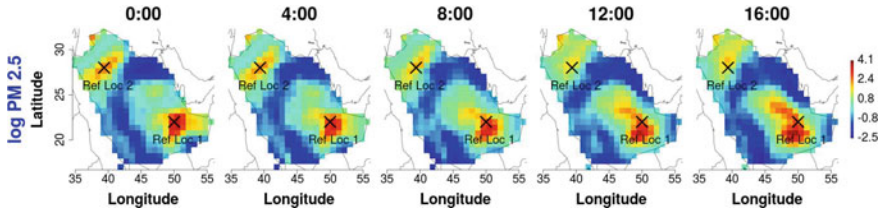


Fig. 8 Snapshots of the log PM 2.5 residuals on January 1, 2017. The spatial images are 4 h apart. Two reference locations are marked for ease of transport movement detection

5.2 Models

We fit six different spatio-temporal covariance functions with Matérn spatial margins. The models under consideration are the following:

- M1: Non-frozen Lagrangian spatio-temporal stationary covariance:

$$C(\mathbf{h}; u) = \sigma^2 \mathbb{E}_{\mathbf{V}} \{ \mathcal{M}_v (a \| \mathbf{s}_1 - \mathbf{s}_2 - \mathbf{V}u \|) \};$$

- M2: Non-frozen Lagrangian spatio-temporal spatially varying parameters model in (5);
- M3: Non-frozen Lagrangian spatio-temporal deformation model in (6);
- M4: Non-Lagrangian spatio-temporal stationary covariance:

$$C(\mathbf{h}; u) = \frac{\sigma^2}{(a_t |u|^{2\alpha} + 1)^{\beta d/2}} \mathcal{M}_v \left\{ \frac{a \| \mathbf{h} \|}{(a_t |u|^{2\alpha} + 1)^{\beta/2}} \right\},$$

where $\alpha \in (0, 1]$ is the smoothness parameter in time and $\beta \in [0, 1]$ is the space-time interaction parameter;

- M5: Non-Lagrangian spatio-temporal nonstationary model:

$$C(\mathbf{s}_1, \mathbf{s}_2; u) = \frac{\sigma(\mathbf{s}_1, \mathbf{s}_2)}{(a_t |u|^{2\alpha} + 1)^{\beta d/2}} \mathcal{M}_v \left[\frac{\{(\mathbf{s}_1 - \mathbf{s}_2)^\top \mathbf{D}(\mathbf{s}_1, \mathbf{s}_2)^{-1} (\mathbf{s}_1 - \mathbf{s}_2)\}^{1/2}}{(a_t |u|^{2\alpha} + 1)^{\beta/2}} \right],$$

a more flexible version of the model in (17); and

- M6: Non-Lagrangian spatio-temporal nonstationary covariance II:

$$C(\mathbf{s}_1, \mathbf{s}_2; t_1, t_2) = \frac{\sigma(\mathbf{s}_1, \mathbf{s}_2)}{\{|(t_1 - t_2)D(t_1, t_2)|^{2\alpha} + 1\}^\beta} \mathcal{M}_v \left[\frac{\{(\mathbf{s}_1 - \mathbf{s}_2)^\top \mathbf{D}(\mathbf{s}_1, \mathbf{s}_2)^{-1} (\mathbf{s}_1 - \mathbf{s}_2)\}^{1/2}}{\{|(t_1 - t_2)D(t_1, t_2)|^{2\alpha} + 1\}^{\beta/2}} \right],$$

where $D(t_1, t_2) = \frac{1}{2} \{D(t_1) + D(t_2)\}$ and $D(t)$ controls the temporally varying parameters. This is a more general nonstationary version of model M5; see Garg et al. (2012).

Table 1 A summary of the models fitted to the log PM 2.5 residuals and their corresponding AIC*, BIC*, and MSE. The lower the values, the better. The best scores are in bold. The number of parameters, q , are also reported

Model	q	AIC*	BIC*	MSE
M1 (S)	8	-13, 823, 238	-13, 823, 121	0.0050
M2 (NS)	37	-15, 051, 228	-15, 050, 688	0.0018
M3 (NS)	28	-14, 859, 980	-14, 859, 571	0.0023
M4 (S)	6	-13, 408, 544	-13, 408, 456	0.0171
M5 (NS)	35	-13, 808, 486	-13, 807, 975	0.0081
M6 (NS)	44	-14, 315, 594	-14, 314, 951	0.0035

The expectations in models M1, M2, and M3 are evaluated numerically with respect to $\mathbf{V} \sim \mathcal{N}_2(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V)$. Furthermore, the covariance matrix $\boldsymbol{\Sigma}_V$ is parametrized using its Cholesky decomposition to guarantee positive definiteness.

Each pixel in Fig. 8 is $0.5^\circ \times 0.625^\circ$. The spatial coordinates are transformed to their appropriate projections in kilometers (km). This means that the unit of the advection velocity is in km/hr. The minimum distance between any two stations is 16.9 km. Following the techniques presented in Sect. 3, we order the measurements based on their locations in time and group them into blocks of 6 consecutive purely spatial random fields and maximize the approximated log-likelihood in (9). Moreover, we perform a two-step estimation where we retrieve first the estimates of the space and time parameters of the stationary versions and plug in those estimates to the nonstationary models in the next round of maximizing the approximated log-likelihood with respect to the nonstationarity parameters.

To validate the models, we leave out the spatio-temporal observations in the last 5 h of January 31 and predict the measurements at all spatial locations. Table 1 reports the performance of the six models as measured by the MSE, Akaike (AIC*), and Bayesian information criteria (BIC*), where $\text{AIC}^* = -2l(\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2) + 2q$ and $\text{BIC}^* = -2l(\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2) + q \log(Mn)$. Here $l(\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2)$ is the value of the approximated log-likelihood function at the second estimation step with parameter estimates $\hat{\boldsymbol{\Theta}}_2$ while fixing the parameters $\hat{\boldsymbol{\Theta}}_1$ obtained at the first estimation step and M is the number of independent replicates of the spatio-temporal random field. The nonstationary models show more favorable AIC* and BIC* values compared to their stationary counterparts. The additional nonstationarity parameters provided the nonstationary models more flexibility to model the space-time data. In terms of prediction, the Lagrangian models report lower MSEs than the non-Lagrangian models. Overall, the non-frozen Lagrangian spatially varying parameters model M2 is the best performing model across all criteria. The estimated mean and covariance matrix of \mathbf{V} under M2 are $\hat{\boldsymbol{\mu}} = (-0.0003, 0.0017)^\top$ km/hr and $\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.719 & 2.257 \\ 2.257 & 3.301 \end{pmatrix} \times 10^{-5}$ km²/hr². This indicates that the estimated value of the correlation between the components of \mathbf{V} is $\hat{\rho} = 0.948$.

6 Conclusion

The theme of this work focused on the practicalities of using Lagrangian spatio-temporal covariance functions to model space-time data, especially under second-order nonstationarity assumptions. The work undertaken in this article aims to illustrate the usability and utility of Lagrangian spatio-temporal models.

We demonstrated the use of thin plate splines in modeling second-order nonstationarity parameters. We also advocated the maximization of the approximated log-likelihood function when data are available at regular time intervals. We showed through several numerical studies the effect of fitting Lagrangian models to data generated from non-Lagrangian models, and vice versa. We found that the predictions of non-Lagrangian models on Lagrangian data are of inferior quality compared to the quality of predictions of Lagrangian models on non-Lagrangian data. This should provide support to using Lagrangian models even when the spatio-temporal random field is not transported.

Further work would be to validate the estimated distribution of the advection velocity vector against the real wind data used as inputs to a partial differential equation which generated the PM 2.5 measurements under study. The equivalence between Lagrangian spatio-temporal models and physical models such as the advection-dispersion equations in Physics is not straightforward and is worth exploring.

The models used in this work as underlying purely spatial nonstationary covariance functions were limited to only two classes. There are other classes in the literature whose Lagrangian formulations deserve attention in terms of model interpretation and parameter estimation, such as the dimension expansion and basis functions models. Future work will focus on these other classes.

References

- Bazen, A. M., & Gerez, S. H. (2003). Fingerprint matching by thin-plate spline modelling of elastic deformations. *Pattern Recognition*, *36*(8), 1859–1867.
- Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(6), 567–585.
- Bornn, L., Shaddick, G., & Zidek, J. (2012). Modeling nonstationary processes through dimension expansion. *Journal of the American Statistical Association*, *107*(497), 281–289.
- Chang, Y.-M., Hsu, N.-J., & Huang, H.-C. (2010). Semiparametric estimation and selection for nonstationary spatial covariance functions. *Journal of Computational and Graphical Statistics*, *19*(1), 117–139.
- Chen, Y., Zhao, J., Deng, Q., & Duan, F. (2017). 3D craniofacial registration using thin-plate spline transform and cylindrical surface projection. *PLoS One*, *12*(10), e0185567. Public Library of Science, San Francisco, CA, USA.
- Chen, T.-L., & Geman, S. (2014). Image warping using radial basis functions. *Journal of Applied Statistics*, *41*(2), 242–258.
- Cox, D. R., & Isham, V. (1988). A simple spatial-temporal model of rainfall. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, *415*, 317–328.

- Cressie, N., Shi, T., & Kang, E. (2010). Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19(3), 724–745.
- Donato, G., & Belongie, S. (2002). Approximate thin plate spline mappings. In *European conference on computer vision* (pp. 21–31). Springer.
- Fouedjio, F., Desassis, N., & Romary, T. (2015). Estimation of space deformation model for non-stationary random functions. *Spatial Statistics*, 13(1), 45–61.
- Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika*, 89(1), 197–210.
- Garg, S., Singh, A., & Ramos, F. (2012). Learning non-stationary space-time models for environmental monitoring. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.
- Gelfand, A. E., Schmidt, A. M., Banerjee, S., & Sirmans, C. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST*, 13(2), 263–312.
- Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458), 590–600.
- Gunz, P., & Mitteroecker, P. (2013). Semilandmarks: A method for quantifying curves and surfaces. *Hystrix, the Italian Journal of Mammalogy*, 24(1), 103–109.
- Heaton, M., Katzfuss, M., Berrett, C., & Nychka, D. (2014). Constructing valid spatial processes on the sphere using kernel convolutions. *Environmetrics*, 25(1), 2–15.
- Hegland, M., Roberts, S., & Altas, I. (1997). *Finite element thin plate splines for data mining applications* Citeseer.
- Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues* (pp. 37–56). Springer.
- Higdon, D. (1998). A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, 5(2), 173–190.
- Higdon, D., Swall, J., & Kern, J. (1999). Non-stationary spatial modeling. *Bayesian Statistics*, 6(1), 761–768.
- Hong, Y., Abdulah, S., Genton, M. G., & Sun, Y. (2021). Efficiency assessment of approximated spatial predictions for large datasets. *Spatial Statistics*, 43, 100517.
- Kathuria, D., Mohanty, B. P., & Katzfuss, M. (2019). A nonstationary geostatistical framework for soil moisture prediction in the presence of surface heterogeneity. *Medical Imaging 2004: Image Processing*, 55(1), 729–753.
- Kaufman, C., Schervish, M., & Nychka, D. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484), 1545–1555.
- Kim, H.-M., Mallick, B. K., & Holmes, Co. (2005). Analyzing nonstationary spatial data using piecewise Gaussian processes. *Journal of the American Statistical Association*, 100(470), 653–668.
- Kleiber, W., & Nychka, D. (2012). Nonstationary modeling for multivariate spatial processes. *Journal of Multivariate Analysis*, 112, 76–91.
- Kleiber, W., Sain, S., & Wiltberger, M. (2014). Model calibration via deformation. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1), 545–563.
- Kuusela, M., & Stein, M. L. (2018). Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proceedings of the Royal Society A*, 474(2220), 20180400.
- Lewis, J. P., Hwang, H.-J., Neumann, U., & Enciso, R. (2004). Smart point landmark distribution for thin-plate splines. *Medical Imaging 2004: Image Processing*, 5370, 1236–1243.
- Li, Y., & Sun, Y. (2018). Efficient estimation of non-stationary spatial covariance functions with application to high-resolution climate model emulation. *Statistica Sinica*, 29, 1209–1231.
- Ma, C. (2003a). Spatio-temporal stationary covariance models. *Journal of Multivariate Analysis*, 86(1), 97–107.
- Ma, C. (2003b). Nonstationary covariance functions that model space-time interactions. *Statistics and Probability Letters*, 61(4), 411–419.
- National Research Council. (2010). *Global sources of local pollution: An assessment of long-range transport of key air pollutants to and from the United States*, National Academies Press.

- Neto, J. H. V., Schmidt, A. M., & Guttorp, P. (2014). Accounting for spatially varying directional effects in spatial covariance structures. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(1), 103–122.
- Nychka, D., & Saltzman, N. (1998). Design of air-quality monitoring networks. In *Case Studies in Environmental Statistics* (pp. 51–76). Springer.
- Nychka, D., Wikle, C., & Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4), 315–331.
- Paciorek, C. J., & Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5), 483–506.
- Porcu, E., Gregori, P., & Mateu, J. (2009). Archimedean spectral densities for nonstationary space-time geostatistics. *Statistica Sinica*, 19(1), 273–286.
- Risser, M. D. (2015). *Spatially-varying covariance functions for nonstationary spatial process modeling*. PhD thesis, The Ohio State University.
- Risser, M. D. (2016). Nonstationary spatial modeling, with emphasis on process convolution and covariate-driven approaches. *arXiv preprint arXiv:1610.02447*.
- Salvaña, M. L., & Genton, M. G. (2020). Nonstationary cross-covariance functions for multivariate spatio-temporal random fields. *Spatial Statistics*, 37, 100411.
- Salvaña, M. L., Lenzi, A., & Genton, M. G. (2020). Spatio-temporal cross-covariance functions under the Lagrangian framework with multiple advections. Unpublished Manuscript.
- Sampson, P. (2015). A partial warp parameterization for the spatial deformation model for nonstationary covariance. *Joint Statistical Meeting*.
- Sampson, P., Damian, D., & Guttorp, P. (2001). Advances in modeling and inference for environmental processes with nonstationary spatial covariance. In *GeoENV III - Geostatistics for environmental applications* (pp. 17–32). Springer.
- Sampson, P. D., & Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417), 108–119.
- Schlather, M. (2010). Some covariance models based on normal scale mixtures. *Bernoulli*, 16(3), 780–797.
- Schmidt, A. M., Guttorp, P., & O'Hagan, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22(4), 487–500.
- Shand, L., & Li, B. (2017). Modeling nonstationarity in space and time. *Biometrics*, 73(3), 759–768.
- Stein, M. L. (2005a). Nonstationary spatial covariance functions. *Unpublished technical report*.
- Stein, M. L. (1999). *Interpolation of spatial data: Some theory for kriging*. New York: Springer.
- Stein, M. L. (2005c). Statistical methods for regular monitoring data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5), 667–687.
- Stephenson, J., Holmes, C., Gallagher, K., & Pintore, A. (2005). A statistical technique for modelling non-stationary spatial processes. In *Geostatistics Banff 2004* (pp. 125–134). Springer.
- Tennakoon, R. B., Bab-Hadiashar, A., Suter, D., & Cao, Z. (2013). Robust data modelling using thin plate splines. *2013 international conference on digital image computing: Techniques and applications (DICTA)* (pp. 1–8).
- Wahba, G. (1990). *Spline models for observational data* (Vol. 59). SIAM.
- Whitbeck, M., & Guo, H. (2006). Multiple landmark warping using thin-plate splines. *IPCV*, 6, 256–263.
- Wikle, C. K. (2010). Low-rank representations for spatial processes. In *Handbook of Spatial Statistics* (pp. 114–125). CRC Press.
- Zhu, Z., & Wu, Y. (2010). Estimation and prediction of a class of convolution-based spatial nonstationary models for large spatial data. *Journal of Computational and Graphical Statistics*, 19(1), 74–95.

Compositional Data Analysis

Logratio Approach to Distributional Modeling



Peter Filzmoser, Karel Hron, and Alessandra Menafoglio

Abstract Distributional data, such as age distributions of populations, can be treated as continuous or discrete data, but the main interest is in the relative information, e.g., in terms of ratios (or logratios) between the different age classes. Here we present a unifying framework for the discrete and the continuous case based on the theory of Bayes spaces. While the discrete case is more widely treated in the literature, the continuous case allows to make a link to functional data analysis. Moreover, the methodological framework of Bayes spaces can also be used to develop methods for analyzing several distributional variables simultaneously. It turns out that the centered logratio transformation is a convenient tool for practical computations. Two real data examples illustrate the usefulness of this framework.

1 Introduction

Multivariate functional data arise frequently from distributing a given whole into a finite or infinite number of components. This is reflected, also, by the domain of such observations (data objects) which can be either countable or uncountable. In the latter case, the domain is represented by a subset of the real line; for the mul-

We dedicate this contribution to Christine Thomas-Agnan, a friend and collaborator since many years. We hope that this approach to compositional data analysis will be inspiring.

P. Filzmoser (✉)

Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology,
Wiedner Hauptstraße 8-10, 1040 Vienna, Austria
e-mail: peter.filzmoser@tuwien.ac.at

K. Hron

Department of Mathematical Analysis and Applications of Mathematics, Palacký University, 17.
listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: hronk@seznam.cz

A. Menafoglio

MOX, Department of Mathematics, Politecnico di Milano, Via Bonardi 9, 20133 Milano, Italy
e-mail: alessandra.menafoglio@polimi.it

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_23

tivariate observations, this must not necessarily be the case as components might represent disjoint categories. Clearly, these data have a relative character from their nature, but in contrast to univariate relative data, now the additional feature is the mentioned distribution of the whole among the parts. Accordingly, such observations can be called *distributional data*. Formally, distributional variables can be defined in terms of symbolic data analysis (Billard and Diday 2006; Diday 2016). Within this methodological framework, they can be understood intuitively as variables formed by realizations of a random variable and their respective relative frequencies, being the result of aggregating large-scale data. However, for practical reasons, it may be preferable to consider distributional data more generally as *objects* which carry relative information and to develop any related methodology for their statistical processing in the frame of object-oriented data analysis (Marron and Alonso 2014). Distributional data thus represent variables (objects) whose values contain quantitatively expressed relative contributions on a whole— not necessarily a distribution in the probabilistic or symbolic data sense. According to their domain (either countable or uncountable), distributional variables are classified into discrete and continuous. An example of the former is the relative structure of the GDP (Gross Domestic Product), the structure of death causes, or the chemical composition of a rock; instances of the latter are represented by population pyramids and particle size distributions. Each of these distributional variables can be analyzed separately. However, they can also occur together in a data set raising the challenge to analyze their relation to other quantitative or qualitative variables, or the association between themselves.

The actual representation of contributions on the whole (probabilities, concentrations, ppm, etc.) can be chosen arbitrarily without any loss of information, a property which is typical for compositional data (Aitchison 1986; Filzmoser et al. 2018; Pawlowsky-Glahn et al. 2015). The relative character of distributional variables is easy to imagine for the discrete case. For the continuous setting, the domain is characterized by a subset of the real line, which is typically a bounded (or, sometimes an unbounded) interval. In this case, the probability-like function is replaced by a density, i.e., a non-negative Borel measurable function with unit integral constraint. An example here is the age/income distribution in a certain region—the finite domain being replaced by an infinite one. Note that, even if we used a representation of the density that would lead to another integral value, the main feature—i.e., that the density conveys relative contributions of Borel sets (subsets of the domain) to the overall probability (weight, frequency)—remains unaltered. In other words, both compositional data and density functions as distributional variables share the property of *scale invariance*. Additionally, also their *relative scale* should be taken into account. For example, for the case of densities, the relative increase of a probability over a Borel set from 0.05 to 0.1 (multiple of two) differs from the increase of 0.5 to 0.55 (multiple of 1.1), although the absolute differences are the same in both cases. Accordingly, not only scale invariance, but also relative scale of distributional variables should be reflected by their statistical processing. The features of both discrete and continuous distributional variables are captured by the general framework of the Bayes space (van den Boogaart et al. 2014) which results in the Aitchison geometry on the simplex (Egozcue et al. 2003) when considering the special case of discrete

distributional data (i.e., compositional data). Notably, the logratio methodology of compositional data is also useful for the development of methods for symbolic data analysis (Hron et al. 2017).

Even though compositional data analysis belongs to multivariate statistics and the statistical processing of densities to functional data analysis (FDA, Ramsay and Silverman 2005), they both represent just *univariate* cases when considering them as distributional variables. Therefore, the main challenge in this setting is to extend the existing compositional methodology to handle more than one distributional variable (discrete or continuous) simultaneously. The aim of this chapter is to go a step forward in this direction. Therefore, the next section is devoted to the description of Bayes spaces and, as a special case, the Aitchison geometry for compositional data, that form the milestones to introduce the statistical analysis of compositions and density functions under a common framework through the logratio approach. Concrete aspects of their modeling, with extension to multivariate distributional variables, are discussed in Sect. 3. Two real-world data sets, representing discrete and continuous distributions, are used in Sect. 4 to illustrate the methodological developments. Finally, Sect. 5 concludes the work.

2 The Bayes Space Embedding for Compositional Vectors

2.1 An Introduction to Bayes Spaces

The distribution of a random variable is characterized by a σ -finite positive measure μ on a measurable space (Ω, \mathcal{A}) . Although in practice exclusively probability measures \mathbf{P} are considered for this purpose, the condition of normalization by $\mathbf{P}(\Omega) = 1$ is rather a convention than an actual need. In fact, any probability measure forms just a representation of a family of proportional measures $\mathcal{M} = \{\mu \mid \exists c > 0 : \forall \mathbf{A} \in \mathcal{A}, \mu(\mathbf{A}) = c \mathbf{P}(\mathbf{A})\}$, which are equivalent from the viewpoint of the *relative* information they provide. Indeed, a rescaling of the measure leaves the ratios (or logratios) between its “parts” unchanged—i.e., between the measure of the measurable subsets of Ω —which, in turn, is the only relevant information embedded into the measure itself. As such, two measures μ, ν are equivalent if they are proportional, denoted hereafter by $\mu =_{B(\lambda)} \nu$, where λ is a reference measure on (Ω, \mathcal{A}) . Given a measure μ , if there exists its Radon–Nikodym derivative with respect to λ (i.e., the density $d\mu/d\lambda$), it is identified with the measure μ itself. As long as $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is concerned, the reference measure λ is often set to the Lebesgue measure. However, any probability measure \mathbf{P} could be considered as well. Given a reference measure \mathbf{P} , the equivalence classes of σ -finite measures can be equipped with the geometrical structure of a Bayes space as in van den Boogaart et al. (2010), whose origin is precisely the reference measure \mathbf{P} . More specifically, a Bayes space is a space of $(B(\mathbf{P})$ -equivalence classes of σ -finite) measures on (Ω, \mathcal{A}) endowed with

a vectorial structure induced by the perturbation and powering operations (\oplus, \odot) , which are defined as

$$(\mu \oplus \nu)(\mathbf{A}) = {}_{B(\mathbf{P})} \int_{\mathbf{A}} \frac{d\mu}{d\mathbf{P}}(x) \frac{d\nu}{d\mathbf{P}}(x) d\mathbf{P}(x); \tag{1}$$

$$(\alpha \odot \mu)(\mathbf{A}) = {}_{B(\mathbf{P})} \int_{\mathbf{A}} \left(\frac{d\mu}{d\mathbf{P}}(x) \right)^\alpha d\mathbf{P}(x), \tag{2}$$

with μ, ν elements of the space and α a real number. Both perturbation and powering can also be expressed in terms of densities; for $f = d\mu/d\mathbf{P}$ and $g = d\nu/d\mathbf{P}$ we get

$$(f \oplus g)(x) = {}_{B(\mathbf{P})} f(x)g(x), \quad (\alpha \odot f)(x) = {}_{B(\mathbf{P})} f(x)^\alpha. \tag{3}$$

The result of both operations are densities again, possibly rescaled to unit integral constraint using the closure operation $C(f) = \frac{f}{\int f d\mathbf{P}}$. Subtraction (or perturbation-subtraction) of densities is then defined as $f \ominus g = f \oplus (-1 \odot g) = {}_{B(\mathbf{P})} f/g$. This operation can be used, e.g., to change the reference measure to \mathbf{P}_1 by employing the well-known chain rule, $(d\mu/d\mathbf{P}_1)(d\mathbf{P}_1/d\mathbf{P}) = d\mu/d\mathbf{P}$.

Given a reference measure \mathbf{P} , we call $B^2(\mathbf{P})$ the Bayes space whose elements are $(B(\mathbf{P})$ -equivalence classes of σ -finite) measures μ such that

$$\int \left| \ln \frac{d\mu}{d\mathbf{P}} \right|^2 d\mathbf{P} < +\infty.$$

Here, measures are identified with the corresponding Radon–Nikodym densities. In $B^2(\mathbf{P})$ an inner product was defined originally in van den Boogaart et al. (2014), reformulated later in Talská et al. (2020) in order to keep the dominance under a change of the reference measure

$$\langle f, g \rangle_{B^2(\mathbf{P})} = \frac{1}{2\mathbf{P}(\Omega)} \int \int \ln \frac{f(x)}{g(x)} \ln \frac{f(y)}{g(y)} d\mathbf{P}(x) d\mathbf{P}(y), \tag{4}$$

for f, g densities in $B^2(\mathbf{P})$. The induced notions of norm and distance are then

$$\|f\|_{B^2(\mathbf{P})} = \frac{1}{2\mathbf{P}(\Omega)} \int \int \ln^2 \frac{f(x)}{f(y)} d\mathbf{P}(x) d\mathbf{P}(y)$$

and

$$d_{B^2(\mathbf{P})}(f, g) = \frac{1}{2\mathbf{P}(\Omega)} \int \int \left(\ln \frac{f(x)}{f(y)} - \ln \frac{g(x)}{g(y)} \right)^2 d\mathbf{P}(x) d\mathbf{P}(y),$$

respectively. The space $B^2(\mathbf{P})$ equipped with the operations of perturbation and powering (\oplus, \odot) , and the inner product $\langle \cdot, \cdot \rangle$ is a separable Hilbert space (van den Boogaart et al. 2014).

The reference measure \mathbf{P} may be chosen according to convenience, although one should be aware that the scale of the reference measure matters for the value of the

inner product (van den Boogaart et al. 2014; Talská et al. 2020). Although several options are discussed in van den Boogaart et al. (2014), two cases are thoroughly considered in the literature: (a) the continuous uniform distribution \mathbf{P}_c (van den Boogaart et al. 2014; Hron et al. 2016; Menafoglio et al. 2014; Talská et al. 2018), and (b) the discrete uniform distribution \mathbf{P}_d , which leads to the Aitchison geometry (Aitchison 1986; Egozcue et al. 2003; Pawłowsky-Glahn et al. 2015). The continuous uniform measure, defined on the interval $I = [a, b](\equiv \Omega)$, is commonly represented by the Lebesgue measure, $\lambda(I) = b - a = \eta$, with the respective density

$$\frac{d\mathbf{P}_c}{d\lambda}(x) = \frac{d\lambda}{d\lambda}(x) = 1;$$

it can be considered as a reference for functional distributional variables (i.e., continuous densities) with bounded domain. Nevertheless, \mathbf{P}_c has often been considered as reference even for variables with (theoretically) unbounded domain, e.g., by neglecting subdomains with very rare occurrence. In all these cases, the inner product simplifies to

$$\langle f, g \rangle_{B^2(\mathbf{P}_c)} = \frac{1}{2\eta} \int_a^b \int_a^b \ln \frac{f(x)}{g(x)} \ln \frac{f(y)}{g(y)} dx dy,$$

and by virtue of the Weierstrass theorem, continuous densities belong to $B^2(\mathbf{P}_c)$.

In case of multivariate compositional data, the discrete uniform distribution is usually employed as a reference measure on $\Omega = \{m_1, \dots, m_D\}$ (Egozcue and Pawłowsky-Glahn 2016), i.e.,

$$\frac{d\mathbf{P}_d}{d\lambda}(x) = 1, \quad x \in \Omega,$$

thus obtaining the Aitchison geometry. Here, compositions with D parts, $\mathbf{x} = (x_1, \dots, x_D)'$, are identified with discrete probability functions over Ω (thus referring to a discrete distributional variable). Having set the unit sum representation of compositions, the sample space of compositional data becomes the $(D - 1)$ -dimensional simplex

$$\mathcal{S}^D = \left\{ \mathbf{x} = (x_1, \dots, x_D)', x_i > 0, \sum_{i=1}^D x_i = 1 \right\}.$$

In this setting, the closure operation reads $C(\mathbf{x}) = \mathbf{x} / \sum_{i=1}^D x_i$; as before, both \mathbf{x} and $C(\mathbf{x})$ belong to the same equivalence class. For two compositions $\mathbf{x}, \mathbf{y} \in \mathcal{S}^D$ and a real α , the operations of perturbation and powering read

$$\mathbf{x} \oplus \mathbf{y} =_{B^2(\mathbf{P}_d)} C(x_1 y_1, \dots, x_D y_D)', \quad \alpha \odot \mathbf{x} =_{B^2(\mathbf{P}_d)} C(x_1^\alpha, \dots, x_D^\alpha),$$

respectively, and the Aitchison inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$

This geometry is the basis of compositional data analysis of multivariate compositional vectors, based on the logratio approach. In the next subsection, we illustrate a practical strategy to employ the Bayes space geometry, either continuous or discrete, for the statistical analysis of compositions.

2.2 Statistical Analysis in Bayes Spaces

A statistical analysis of continuous or discrete density functions needs to properly account for both the data dimensionality and the geometrical structure governing Bayes spaces. In fact, although continuous densities are functional data and discrete compositions are multivariate observations, they both are featured by the basic properties of compositions (as scale invariance and relative scale), that are captured neither by FDA nor by classical multivariate methods. For instance, most methods of FDA rely on the assumption that the data belong to the space $L^2(\lambda)$ of squared-integrable functions with respect to the Lebesgue measure (the general case of $L^2(\mathbf{P})$ needs to be mapped to $L^2(\lambda)$ using a nonlinear transformation Talská et al. 2020). However, the geometrical structure of the space $L^2(\lambda)$ is not appropriate for compositions (e.g., the point-wise sum of compositions does not result in a composition). Similarly, most multivariate statistical methods are built in the Euclidean setting, which is not appropriate to analyze discrete compositions. Nevertheless, as long as the data are embedded in a separable Hilbert space, one can map the observations into $L^2(\lambda)$ or into the Euclidean space \mathbb{R}^D , and accordingly perform the statistical analysis via FDA or multivariate statistics, while accounting for the Bayes space geometry.

Let us first focus on the continuous case, having set the reference measure to a probability measure \mathbf{P} . As separable Hilbert spaces, an isometric isomorphism exists between $B^2(\mathbf{P})$ and a subspace of $L^2(\mathbf{P})$. An instance of such an isometry is provided by the centered logratio (clr) transformation, defined for a density $f = d\mu/d\mathbf{P}$ as

$$\text{clr}(f) = \ln f - \frac{1}{\mathbf{P}(\Omega)} \int \ln f \, d\mathbf{P}. \quad (5)$$

Consequently, for $\alpha \in \mathbb{R}$, $f, g \in B^2(\mathbf{P})$ the following relations hold,

$$\text{clr}(f \oplus g) = \text{clr}(f) + \text{clr}(g), \quad \text{clr}(\alpha \odot f) = \alpha \cdot \text{clr}(f),$$

$$\langle f, g \rangle_{B^2(\mathbf{P})} = \langle \text{clr}(f), \text{clr}(g) \rangle_{L^2(\mathbf{P})}.$$

We note that these relations enable one to handle clr transformed densities in the L^2 setting. Due to its construction, clr transformations fulfill the integral constraint $\int \text{clr}(f) d\mathbf{P} = 0$ that should be taken into account in any statistical analysis based on clr transformed data. The subspace of functions with zero integral wrt. \mathbf{P} is denoted hereafter as $L_0^2(\mathbf{P})$. Moreover, in case of a uniform reference measure $\mathbf{P} \equiv \mathbf{P}_c$, which is of primary importance in practical applications, (5) reads

$$\text{clr}(f)(t) = \ln f(t) - \frac{1}{b-a} \int_a^b f(\tau) d\tau. \tag{6}$$

Note that it would also be possible to get rid of the zero integral constraint resulting from the clr transformation, e.g., by expressing the densities via the Fourier coefficients of a basis in $B^2(\mathbf{P}_c)$ (such as Legendre polynomials Tolosana-Delgado et al. 2008 or orthogonal splines Machalová et al. 2020); though, most recent literature works propose clr-based methods (Hron et al. 2016; Menafoglio et al. 2014; Talská et al. 2018).

The situation is a bit different for compositional data (the case of \mathbf{P}_d), where the clr transformation of a composition \mathbf{x} (in fact, coefficients with respect to a generating system on the simplex) results in

$$\text{clr}(\mathbf{x}) = (y_1, \dots, y_D)' = \left(\frac{x_1}{\sqrt[D]{\prod_{i=1}^D x_i}}, \dots, \frac{x_D}{\sqrt[D]{\prod_{i=1}^D x_i}} \right)'. \tag{7}$$

An orthonormal coordinate system can be built on the hyperplane $y_1 + \dots + y_D = 0$, induced by clr coordinates, which is commonly called isometric logratio (ilr) coordinates (Egozcue et al. 2003). Sequential binary partitioning (SBP) (Egozcue and Pawlowsky-Glahn 2005) provides a range of possibilities to build interpretable ilr coordinates. Indeed, SBPs enables one to construct $D - 1$ coordinates with respect to an orthonormal basis of the simplex, on the basis of balances between groups of compositional parts, represented through their geometric means. The use of SBP usually requires some prior knowledge about the problem at hand. However, “automated” versions of orthonormal coordinates can be considered as well (Fišerová and Hron 2011). For instance, for a composition \mathbf{x} one can obtain the $(D - 1)$ -dimensional real vector $\mathbf{z} = (z_1, \dots, z_{D-1})'$, as Hron et al. (2012)

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^D x_j}}, \quad i = 1, \dots, D-1, \tag{8}$$

a permuted version of coordinates from Egozcue et al. (2003), known as *pivot coordinates* in compositional data analysis (see Filzmoser et al. 2018, p. 49). Note that only the first pivot coordinate contains the part x_1 in terms of its logratio to the remaining parts at hand, thus it conveys information about the dominance of x_1 “on average”. The remaining pivot coordinates (z_2, \dots, z_{D-1}) then represent the

subcomposition including the parts x_2, \dots, x_D . We notice that if the l -th part is of interest, one can consider a permutation of the parts in the input composition such that x_l , $l = 1, \dots, D$ takes the first position, the others being placed arbitrarily (different orthonormal coordinate systems are just rotations of each other Egozcue et al. 2003). In this case, the first element of the corresponding pivot coordinates, denoted by $\mathbf{z}^{(l)} = (z_1^{(l)}, \dots, z_{D-1}^{(l)})'$ would have the above interpretation.

An explicit relationship exists between the clr transformation and the first element of $\mathbf{z}^{(l)}$, as $y_l = \sqrt{\frac{D-1}{D}} z_1^{(l)}$. This relation can be used to support the interpretation of clr variables.

Once compositional data are expressed either via clr or in orthonormal coordinates, all the standard methods of multivariate statistics that rely on the Euclidean geometry (Eaton 1983) can be employed. We discuss this in more detail in the next section from the perspective of modeling of distributional variables.

3 Implications for Distributional Modeling

In both the discrete and the continuous settings (i.e. compositional data and density functions), most methodologies available for logratio modeling in Bayes spaces address just a *univariate* perspective from the viewpoint of object-oriented data analysis. Indeed, object-oriented methods open the possibility to cope with cases where each statistical unit is formed by a set of more than one distributional variable (i.e., a vector of distributional variables), that appear with increasing frequency in the applications. This raises an urgent need to provide sets of coordinates for compositional data and densities, that would enable one to perform joint analyses of discrete and continuous distributional data, through multivariate object-oriented statistical analyses.

The problem of building coordinates for compositional data can be addressed either through the centered logratio coefficients (7), or by using orthonormal coordinates (8). Although for some methods (e.g. principal component analysis and the associated compositional biplots Aitchison and Greenacre 2002) the clr coefficients are preferable, in other cases both options are allowed (e.g. cluster analysis, or regression analysis with compositional covariates Bruno et al. 2015). On the other hand, the link between clr coefficients and first pivot coordinates can be used also for the mentioned case of principal component analysis (Kynčlová et al. 2016). It follows that in practice, whenever possible, the orthonormal coordinates are employed. The reason for this relies on the fact that they guarantee a regular covariance matrix of the observations, which is a must for most robust multivariate methods (Filzmoser and Hron 2011; Filzmoser et al. 2018). In the continuous case, a set of coordinates can be obtained by using the Fourier coefficients of a basis in $B^2(\mathcal{P}_c)$ (Egozcue et al. 2006), or a B-spline representation of clr transformed densities (Machalová et al. 2016; Machalová et al. 2020). Note that, in the latter case, one should take care of the fact that B-splines should be orthonormalized first (Machalová et al. 2020). More-

over, except for particular cases (van den Boogaart et al. 2014), density functions need infinitely many coefficients to be described. Thus, in general, an appropriate dimensionality reduction needs to be performed prior to their statistical analysis.

Among the compositional methods which are suitable to be extended to the multivariate object-oriented context, we focus here on two special cases that illustrate the potential of the logratio approach to analyze distributional data. In the next section, linear regression with a real response and several compositional covariates is presented, followed by multivariate principal component analysis for density functions.

3.1 Linear Regression with Discrete Distributions as Covariates

In Wang et al. (2013), a regression model is presented, where both the response and the explanatory variables are compositional data. Although the model was not originally intended to provide a link between compositional and object-oriented data analysis, it is particularly well-suited for our purposes. In the following, we employ a simplified version of this model, based on p compositions $\mathbf{x}_1, \dots, \mathbf{x}_p$, containing D_1, \dots, D_p parts ($D := D_1 + \dots + D_p$), that explain a real response variable Y . Note that this setting represents a generalization of the so-called *experiments with mixtures* (Scheffé 1958), that has been adapted to the logratio methodology in Hron et al. (2012).

Instead of analyzing the original compositional data, we express these in orthonormal coordinates, $\mathbf{z}_1, \dots, \mathbf{z}_p$, where $\mathbf{z}_j = (z_{j,1}, \dots, z_{j,D_j-1})'$, $j = 1, \dots, p$, and consider the regression model

$$E(Y|\mathbf{z}_1, \dots, \mathbf{z}_p) = \beta_0 + z_{1,1}\beta_{1,1} + \dots + z_{1,D_1-1}\beta_{1,D_1-1} + \dots + z_{p,D_p-1}\beta_{p,D_p-1}. \tag{9}$$

The linear model for the observations is

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{10}$$

where the $n \times (D - p + 1)$ design matrix \mathbf{Z} is defined as

$$\mathbf{Z} = \begin{pmatrix} 1 & \mathbf{z}'_{1,1} & \dots & \mathbf{z}'_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & \mathbf{z}'_{n,1} & \dots & \mathbf{z}'_{n,p} \end{pmatrix}.$$

The model thus contains $D - p + 1$ regression parameters. Under the usual assumptions, the parameters can be estimated by a least squares (LS) method, i.e., by minimizing the sum of squared residuals RSS . This yields the estimates $\widehat{\beta}_0, \widehat{\beta}_{1,1}, \dots, \widehat{\beta}_{p,D_p-1}$. The result can then be used for prediction purposes, or for further statistical inference.

Under the Gaussian assumption, a series of tests can be performed. For instance, one may want to evaluate whether the j -th composition, $j = 1, \dots, p$, has a significant influence on the explanatory variable Y . For this purpose, the following test statistic can be employed:

$$Q_j = \frac{1}{(D_j - 1)S^2} \widehat{\boldsymbol{\beta}}_j' \mathbf{W}_j^{-1} \widehat{\boldsymbol{\beta}}_j, \quad j = 1, \dots, p, \tag{11}$$

where $S^2 = RSS/(D - p + 1)$, $\widehat{\boldsymbol{\beta}}_j = (\widehat{\beta}_{j,1}, \dots, \widehat{\beta}_{j,D_j-1})'$ and the $(D_j - 1) \times (D_j - 1)$ matrix \mathbf{W}_j is formed by the block of $(\mathbf{Z}'\mathbf{Z})^{-1}$ that corresponds to $\boldsymbol{\beta}_j$ as part of $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}'_1, \dots, \boldsymbol{\beta}'_p)'$. Under the null hypothesis $\boldsymbol{\beta}_j = \mathbf{0}$, the statistic Q_j follows a Fisher distribution with $D_j - 1$ and $n - D + p - 1$ degrees of freedom.

If, for the j -th composition (distributional variable), the above hypothesis is rejected, one may want to investigate which of its part(s) does have significant influence on Y . A solution can be provided again in terms of orthonormal coordinates. Indeed, one may take advantage of the interpretation of (8), leading to pivot coordinates $\mathbf{z}_j^{(l_j)} = (z_{j,1}^{(l_j)}, \dots, z_{j,D_j-1}^{(l_j)})'$ and the corresponding parameters $\boldsymbol{\beta}_j^{(l_j)} = (\beta_{j,1}^{(l_j)}, \dots, \beta_{j,D_j-1}^{(l_j)})'$. Here, only the first pivot coordinate of $\mathbf{z}_j^{(l_j)}$ and the corresponding regression parameter are of primary interest. Concretely, if the significance of the regression parameter $\beta_{j,1}^{(l_j)}$ is confirmed by the rejection of the corresponding hypothesis on a significance level α , then the relative information concerning the l_j -th part of the composition \mathbf{x}_j (resulting from summarizing logratios to the other parts of \mathbf{x}_j) has an influence on the response. The decision can be taken based on the test statistic

$$T_{jl_j} = \frac{\widehat{\beta}_{j,1}^{(l_j)}}{\sqrt{S^2 \{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(l_j,l_j)}}}, \quad j = 1, \dots, p, \quad l_j = 1, \dots, D_j, \tag{12}$$

where $\{(\mathbf{Z}'\mathbf{Z})^{-1}\}_{(l_j,l_j)}$ denotes the diagonal element of the matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$ which corresponds to the coefficient $\widehat{\beta}_{j,1}^{(l_j)}$. Under the null hypothesis $\beta_{j,1}^{(l_j)} = 0$, T_{jl_j} follows a Student's t distribution with $n - D + p - 1$ degrees of freedom. Note that for an exhaustive search for the significance of the coordinates in single explanatory parts, pD regression models would need to be built. Nevertheless, the estimate of the intercept parameter, as well as the coefficient of determination for the regression model (9) are always the same (Hron et al. 2012), due to the orthonormality of the pivot coordinates.

3.2 *Multivariate Functional Principal Component Analysis When Data Are Density Functions*

The focus of this subsection is on multivariate functional composition (mFCs). These are K -dimensional vectors whose elements are functional compositions (FCs, i.e., elements of the Bayes space B^2). For instance, Figure 1 represents a dataset of population pyramids in 57 districts of Upper Austria. These are instances of mFCs of dimension $K = 2$: they are coupled density functions, describing the age density of males and females in these regions. The aim of this subsection is to introduce a methodology to explore the variability of a dataset of mFCs, and consistently perform dimensionality reduction. To attain these goals in either multivariate statistics and functional data analysis, (functional) principal component analysis is widely employed. Here, the focus is posed on the main modes of variability of the sample, whose interpretation is often insightful in terms of the observed phenomenon. A similar problem is considered, in the case of multivariate discrete compositions, in Wang et al. (2015). In the recent literature, Hron et al. (2016) introduces the simplicial functional principal component analysis (SFPCA) as an extension of functional principal component analysis to the Bayes space setting (for *univariate* density functions). Here, we consider an approach similar to that introduced in Hron et al. (2016) to derive an extension of simplicial principal component analysis to the multivariate, simplicial and functional setting, that relies on the Hilbert space structure introduced in Sect. 2.

We first note that mFCs are not multivariate density functions: only the marginal densities are available, up to a scale factor. For instance, when population pyramids are concerned, the available observations do not represent the joint age distribution of male and female populations, but just their marginals. Instead, a mFC can be considered as an element of the space $[B^2]^K = B^2 \times \dots \times B^2$, which is a separable

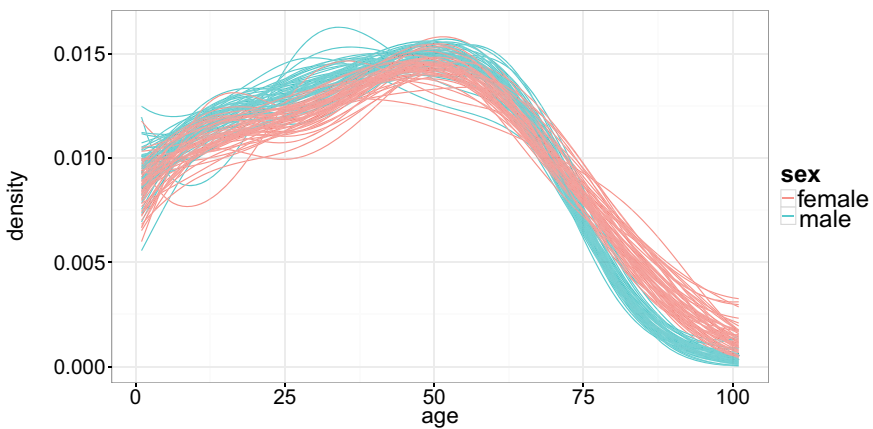


Fig. 1 Population pyramids in the 57 districts of Upper Austria

Hilbert space, if equipped with the component-wise B^2 operations:

$$(\mathbf{f} \oplus \mathbf{g})_i = f_i \oplus g_i, (\alpha \odot \mathbf{f})_i = \alpha \odot f_i, \mathbf{f} = (f_i), \mathbf{g} = (g_i) \in [B^2]^K, \alpha \in \mathbb{R},$$

and the inner product $\langle \mathbf{f}, \mathbf{g} \rangle_{[B^2]^K} = \sum_{i=1}^K \langle f_i, g_i \rangle_{B^2}$, for $\mathbf{f} = (f_i), \mathbf{g} = (g_i) \in [B^2]^K$.

Let $\mathbf{X}_1, \dots, \mathbf{X}_N$ be a dataset of mFCs, e.g. that displayed in Fig. 1. To simplify the notation and without loss of generality, hereafter we assume the dataset to be centered. Note that one can always consider the centered version of a given dataset, that is $\tilde{\mathbf{X}}_1, \dots, \tilde{\mathbf{X}}_N$, with $\tilde{\mathbf{X}}_i = \mathbf{X}_i \ominus \bar{\mathbf{X}}$ and $\bar{\mathbf{X}} = \frac{1}{N} \odot \bigoplus_{i=1}^N \mathbf{X}_i$. In this setting, multivariate SFPCA (mSFPCA) of $\mathbf{X}_1, \dots, \mathbf{X}_N$ aims to find the main modes of variability of the dataset. These are the orthogonal directions in $[B^2]^K$ —identified by a collection of orthogonal elements $\{\boldsymbol{\zeta}_j\}_{j \geq 1}$, $\boldsymbol{\zeta}_j \in [B^2]^K$, of unitary norm—that display the maximum variability of the dataset. That is, $\{\boldsymbol{\zeta}_j\}_{j \geq 1}$, $\boldsymbol{\zeta}_j \in [B^2]^K$, are found by maximizing the following objective functional:

$$\frac{1}{N} \sum_{i=1}^N \langle \mathbf{X}_i, \boldsymbol{\zeta} \rangle_{[B^2]^K}^2 \text{ subject to } \|\boldsymbol{\zeta}\|_{[B^2]^K} = 1; \langle \boldsymbol{\zeta}, \boldsymbol{\zeta}_k \rangle_{A^2} = 0, k < j, \quad (13)$$

where $\langle \mathbf{X}_i, \boldsymbol{\zeta} \rangle_{[B^2]^K}^2$ represents the projection of \mathbf{X}_i along the direction identified by $\boldsymbol{\zeta}$, and the orthogonality condition $\langle \boldsymbol{\zeta}, \boldsymbol{\zeta}_k \rangle_{[B^2]^K} = 0$, for $k < j$, is meaningful only for $j \geq 2$.

It is possible to show that, for each $j = 1, 2, \dots$, maximization of (13) leads to a unique solution in $[B^2]^K$, as this is a separable Hilbert space (Horváth and Kokoszka 2012, Theorem 3.2). Indeed, the principal components are uniquely found as the eigenfunctions of the sample covariance operator $V : [B^2]^K \rightarrow [B^2]^K$, that acts on $\mathbf{x} \in [B^2]^K$ as

$$V \mathbf{x} = \frac{1}{N} \odot \bigoplus_{i=1}^N \langle \mathbf{X}_i, \mathbf{x} \rangle_{[B^2]^K} \odot \mathbf{X}_i.$$

The operator V admits $N - 1$ non-zero eigenvalues, $\lambda_1 < \lambda_2 < \dots < \lambda_{N-1}$, that represent the variability of the dataset along its main modes of variability $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_{N-1}$. These can be displayed in a scree plot (as in multivariate PCA) to drive the dimensionality reduction, as done for population pyramids in Sect. 4.2 (see Fig. 3a).

For the actual computation of the eigenpairs $(\lambda_j, \boldsymbol{\zeta}_j)$, $j = 1, \dots, N - 1$, we propose to employ the clr transformation introduced in Sect. 2, in order to map the problem in L_0^2 and proceed as in the multivariate functional case. Specifically, we propose the following procedure:

1. **Transform:** For $i \in 1, \dots, N$, transform the i -th observed mFC as $\mathbf{clr}(\mathbf{X}_i)$, where the mapping \mathbf{clr} acts as a component-wise clr transformation: $\mathbf{clr}(\mathbf{f}) = (\mathbf{clr}(f_i)) \in [L_0^2]^K$, for $\mathbf{f} = (f_i) \in [B^2]^K$;
2. **Solve in $[L_0^2]^K$:** Compute the multivariate FPCs $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{N-1}$ in $[L_0^2]^K$ and the corresponding eigenvalues $\lambda_1, \dots, \lambda_{N-1}$;

3. **Back-transform:** Employ the inverse *clr*-transformation to ξ_1, \dots, ξ_{N-1} , i.e., apply component-wise the inverse of the *clr* transformation, and set $\zeta_j = \text{clr}^{-1}(\xi_j)$.

It is possible to prove that (i) the eigenvalues found at step (1) are the same as those of the operator V , and (ii) the introduced procedure leads to a correct characterization of the set of eigenpairs of V , since the *clr* transformation is an isometric isomorphism between B^2 and L_0^2 . The proof of these points can be obtained by generalizing the arguments presented in Hron et al. (2016) (not shown). The dimensionality reduction can then follow the same lines of the classical setting. For instance, one can employ the scree plot to determine the relevant mSFPCs in terms of the proportion of explained variability. The interpretation of the mSFPCs can be based on graphical displays, such as the plot of the eigenfunctions (possibly transformed via *clr*), or the perturbation of the mean via the eigenfunction perturbed by a coefficient. The former allows to single out contrasts between parts of the domains which are attributed different weights; the latter enables one to visualize the portion of variability around the mean which is captured by the corresponding principal component. These graphical displays shall be exemplified in Sect. 4.2, where the mSFPCA of population pyramids will be presented.

4 Case Studies

4.1 *Effect of GDP Components and Causes of Death on Life Expectancy*

Eurostat provides various data sets at <http://ec.europa.eu/eurostat/data> that refer to economy, population, health, education, etc., of the EU countries. For the purpose of illustrating the procedure outlined in Sect. 3.1, we consider the life expectancy as response variable, and two compositions as explanatory variables. The first composition includes the most important components of the GDP (Gross Domestic Product), namely the *private final consumption expenditure* (private), the *government final consumption expenditure* (government), the *gross fixed capital formation* (capital), the *exports*, and the *imports*. All these data are taken from the year 2011, for the EU countries, as well as for Norway and Switzerland, and we use the data reported in absolute values (million Euros). The second composition contains the most relevant causes of death. Again, we use data from 2011, for the same countries as before, and take the absolute numbers as the initial representation of the compositions. The following groups are considered (the abbreviations in brackets refer to the ICD codes, and to the abbreviations we are using later on): *Certain infectious and parasitic diseases* (A00-B99) (infect), *Malignant neoplasms* (C00-C97) (neoplasm), *Endocrine nutritional and metabolic diseases* (E00-E90) (nutrition), *Mental and behavioral disorders* (F00-F99) (mental), *Diseases of the nervous system and the sense organs*

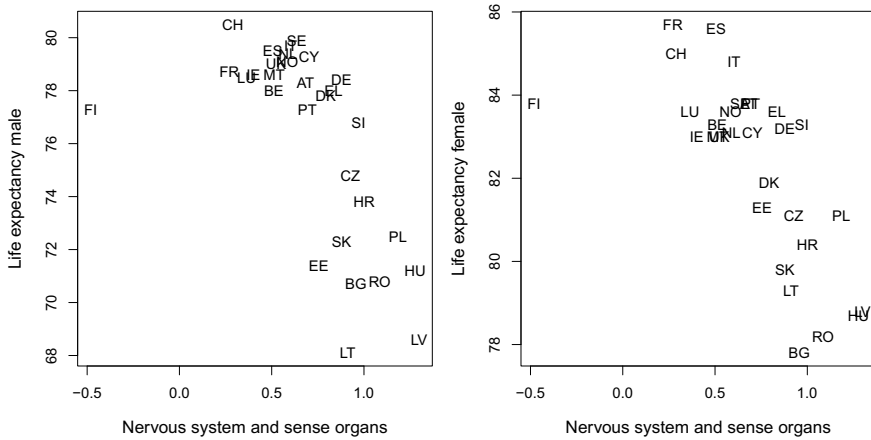


Fig. 2 Relation between all relative information to the diseases of the nervous system and the sense organs, expressed in terms of the respective first pivot coordinate, with the life expectancy; left for males, right for females

(G00-H95) (nervous), *Diseases of the circulatory system* (I00-I99) (circulatory), *Diseases of the respiratory system* (J00-J99) (respiratory), and *Diseases of the digestive system* (K00-K93) (digestive).

The life expectancy as well as the causes of death are available for the total population, and for males and females separately. Therefore, in the analyses below, we investigate models for these three cases separately. The GDP composition is of course unchanged.

A first impression about the data structure is provided in Fig. 2. We compare the relative dominance of death by diseases of the nervous system and the sense organs (nervous) with the life expectancy, separately for males (left) and females (right). Thus, in the second composition, the variable *nervous* is put to the first position, and the first coordinate after applying Equation (8) represents all relative information about *nervous*. According to the figure, high values on this coordinate correspond to the dominance of the disease *nervous*, which relates to low life expectancy, and vice versa. The most important diseases covered by *nervous* are Alzheimer and Parkinson.

The regression model (10) is now applied to the problem, and the idea is to identify economic and/or health information that relates to life expectancy. The regression models which are considered here (total, male, female) lead to multiple R^2 values of more than 0.9. Also the squared correlations between the responses and the leave-one-out cross-validation predictions are above 0.7, indicating meaningful models. We apply the test statistic (11) to the different settings, and the resulting p -values are reported in Table 1. Both compositions have significant influence for the models based on the total and on the female population, whereas for the males we do not obtain significance.

Table 1 Results of the test (11) for both compositions, for models based on the total population, the males, and the females, respectively. Shown are the resulting *p*-values of the test for the two compositions

	Total	Males	Females
GDP compositions	< 0.001	0.22	< 0.001
Causes of death	< 0.001	0.13	< 0.001

Table 2 Results of the test (12) for the first composition, for models based on the total population, the males and the females, respectively. Shown are the resulting *p*-values of the test, and the regression coefficients (coeff.) for the parts of the first composition

	Total		Males		Females	
	<i>p</i> -value	Coeff.	<i>p</i> -value	Coeff.	<i>p</i> -value	Coeff.
Private	0.50	-1.32	0.98	0.08	0.58	-0.85
Government	0.18	2.42	0.47	1.64	0.22	1.66
Capital	0.56	-1.19	0.43	-2.18	0.47	-1.19
Exports	0.68	1.57	0.92	0.52	0.49	2.00
Imports	0.72	-1.49	0.99	-0.06	0.61	-1.61

The second test according to (12) tests for significance of the single parts in the compositions via their corresponding coordinates. For this purpose, 5+8=13 regression models were built, where in each model the test concerns the first pivot coordinate representing the part of interest. The (statistical) interpretation of these tests wrt. rejecting and non-rejecting the hypothesis is then as usual, just reflecting the specific interpretation of the first pivot coordinates (see Sect. 3.1 for details).

The results are presented in Table 2 for the first composition and in Table 3 for the second composition. We realize that none of the parts in the first composition is significant on its own. In order to get the significance, we would need to go for other coordinates from (8) or even to consider a more complex coordinate system (Egozcue and Pawlowsky-Glahn 2005), to find such rotation of the orthonormal coordinates where the significance in one or more coordinates appears. In contrast, several parts from the second composition are contributing significantly. For example, the part *nervous* that was under consideration in Fig. 2 has significant contribution in all settings (total, male, female), and the regression coefficient is negative, as it was expected from the plot. So, the dominance of this disease (and subsequent death) refers to countries with lower life expectancy. The dominance of *neoplasm* for females also relates to low life expectancy, while the dominance of the other significant diseases *circulatory* and *digestive* are in relation to countries with higher life expectancy.

Note that the above analysis would lead to exactly the same results if for expressing in pivot coordinates the absolute values of the compositions (million Euro for GDP composition, numbers of death causes) would have been expressed in relative units, like proportions or percentages.

Table 3 Results of the test (12) for the second composition, for models based on the total population, the males, and the females, respectively. Shown are the resulting p -values of the test, and the regression coefficients (coeff.) for the parts of the second composition

	Total		Males		Females	
	p -value	Coeff.	p -value	Coeff.	p -value	Coeff.
Infect	0.60	0.40	0.10	1.60	0.66	0.24
Neoplasm	0.17	-3.71	0.52	-2.09	0.01	-5.06
Nutrition	0.21	-0.85	0.07	-1.67	0.49	-0.37
Mental	0.99	0.003	0.78	0.18	0.29	0.42
Nervous	0.004	-2.69	0.005	-3.70	< 0.001	-2.76
Circulatory	0.03	3.40	0.04	3.35	< 0.001	3.77
Respiratory	0.42	-1.056	0.17	-2.74	0.39	1.02
Digestive	0.005	4.49	0.004	5.06	0.01	2.74

4.2 Dimensionality Reduction of Population Pyramids via mSFPCA in Bayes Spaces

We demonstrate the results of the methodology proposed in Sect. 3.2 on the dataset of population pyramids displayed in Fig. 1, and presented in Hron et al. (2016). For performing the computations, we resort to numerical integration to deal with clr transforms and we solve numerically the eigenproblem in $[L_0^2]^K$ involved in step (2). Another strategy may be employed as well, e.g., to represent the data via a functional basis and express the solution via the associated coefficients (Ramsay and Silverman 2005; Hron et al. 2016). Figure 3 summarizes the obtained results. Panel (a) shows a rapid decrease of the variance explained by the principal components, which suggests a possible dimensionality reduction to two or three mSFPCs. However, the variability of the estimated scores along the third component (i.e., of $\langle X_i, \xi_j \rangle_{[B^2]^K}^2$, with $j = 3$, $i = 1, \dots, N$) appears affected by the presence of an outlier. Hence, we focus on the first two components for the scope of interpretation and dimensionality reduction. To ease the interpretation, Fig. 3c-d display the clr transformation of the elements of ξ_1 and ξ_2 , i.e., ξ_1 and ξ_2 obtained from step (2); colors are used to identify the gender. Indeed, the transformed eigenfunctions can be interpreted as in FPCA, e.g., looking for meaningful contrasts between portions of the domain. Notice that the clr transformed eigenfunctions are non-zero and fulfill the zero integral constraints which are the characteristics of clr transformed FCs. As such, contrasts are expected in all the ξ_i , $i \geq 1$. Considering the first mSFPCs, we notice that in both elements, a contrast exists between the oldest population (age > 80/75 years, for men and women, respectively) and the younger one. We note that this result is consistent with that of Hron et al. (2016), that analyzes separately men and women subpopulations. Hence, high scores along the first mSFPC are expected for the municipalities with a higher incidence of the elder population than the mean, and vice versa. This is evident also when observing the plot displayed in Fig. 3e. Here, the effect of the variability along

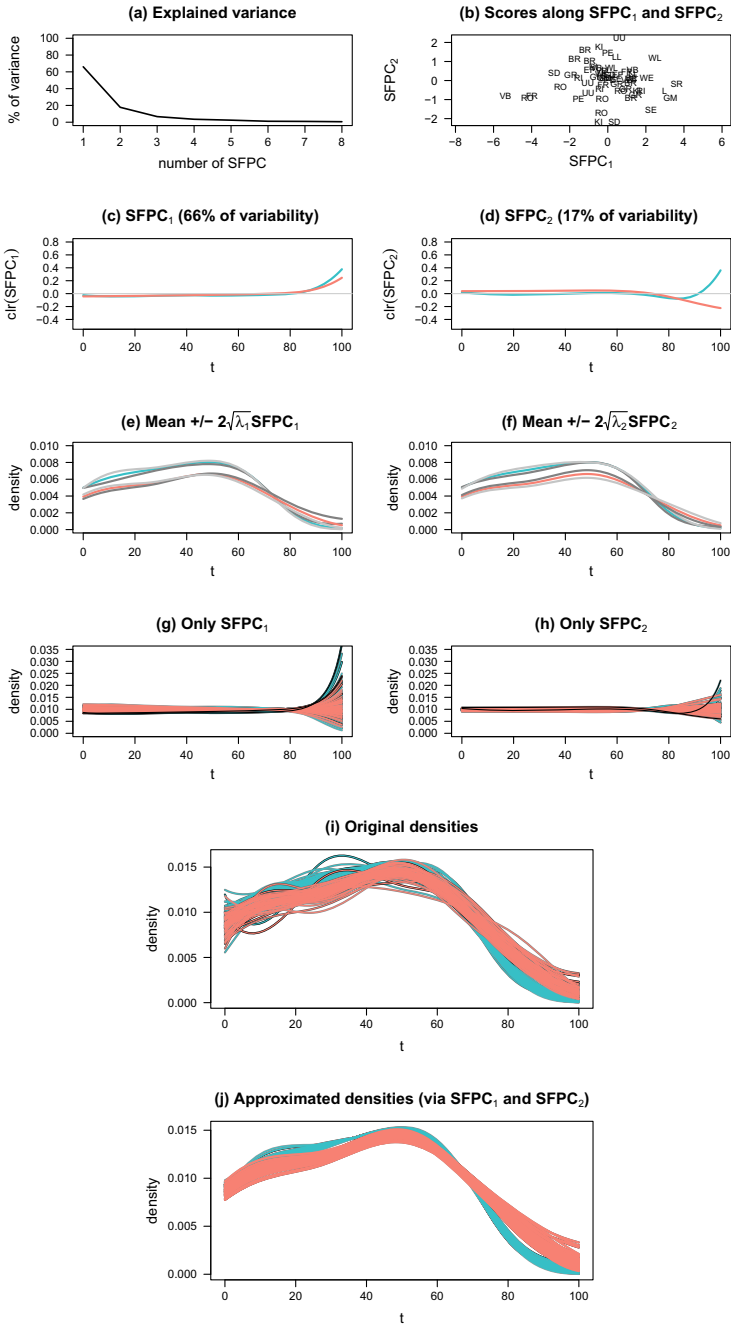


Fig. 3 Results of mSFPCA on the population pyramids. In panels (c) to (f): solid dark gray lines indicate the perturbation of the mean by the mSFPC ξ_j powered by $+2 \cdot \sqrt{\lambda_j}$, $j = 1, 2$; solid light gray lines indicate the perturbation of the mean by the mSFPC ξ_j powered by $-2 \cdot \sqrt{\lambda_j}$, $j = 1, 2$

the first principal component is visualized via the perturbation of the mean by the first mSFPC powered by $\pm 2 \cdot \sqrt{\lambda_1}$. Having fixed the sign of the eigenfunction ξ_1 as in panel (c), data with high corresponding scores (dark gray line) tend to have heavier tails than the mean and vice versa.

The interpretation of the second mSFPC in Fig. 3d is in terms of the contrast between men and women subpopulations, with a positive contribution in men for right tails (age > 93 years) higher than the mean, and a negative contribution in women's right tails (age > 75 years) higher than the mean. Overall, Fig. 3f shows that low scores along the second mSFPC associate with more pronounced peaks in the density functions and vice versa. Figure 3g and h display the contribution to the variability along the two mSFPCs: in each panel, the elements $\langle X_i, \zeta_k \rangle_{[B^2]^K} \odot \zeta_k, i = 1, \dots, N, k = 1, 2$, are represented. In agreement with the previous comments, these plots suggest that most variability is displayed within the right tails. In addition to this, further evidence of the previous interpretation is given by plotting the elements with maximum scores (black curves). Indeed, high scores along the first mSFPC in Fig. 3c correspond to a higher incidence of the old population than the mean in both men and women; instead, high scores along the second mSFPC correspond to a higher incidence of the old population than the mean in men and lower incidence in women. Similar interpretations—with opposite score signs—are obtained from the elements with minimum scores. In this sense, the second mSFPC provides a contrast between the behavior of men and women subpopulations for the elder ages. Finally, Fig. 3j displays the approximation of the densities which are attained via the first two SFPCs, that together explain more than 80% of the overall variability.

5 Conclusions

This contribution has been devoted to the logratio approach to the analysis of distributional data, objects carrying relative information. In our setting, the relevant information embedded in distributional data is being analyzed, based on the logratios between the values of the compositional parts (the discrete case) or densities (the continuous case). Here, we described a unifying framework for both the continuous and the discrete case, based on the theory of Bayes spaces. We illustrated the discrete case through a regression setting, for a real response modeled in terms of a number of compositions. We considered specific representations of the compositions in terms of coordinates, in order to use the classical tools for inference. To this end, we employed a particular type of the ilr coordinates, so called pivot coordinates. Orthonormal (ilr) coordinates are in general preferred in the discrete setting of distributional variables because they are convenient for the computations and avoid singularity issues, which occur with clr coefficients. However, when considering discrete and continuous distributional variables together, clr transformed densities are still represented in a functional space, and are thus easier to visualize, typically as a curve. Although one needs to be careful with the interpretation because of the zero integral, it is still more natural to link the resulting real function to the origi-

nal density, because they are both functional. In contrast, for an ilr representation, one would have to link the coordinates to the density through the functional basis, which is usually not straightforward. In line with this, in the continuous setting, we analyzed the multivariate distributional data in the form of densities by extending the multivariate functional principal component analysis. Here, the key to bringing theory to practice was to employ the clr transformation to simplify computations of eigenfunctions.

The examples on regression and functional principal component analysis served as illustrations of the great potential of this theory that enables one to deal with both compositional data and densities in the common framework of the Bayes space methodology, embedded in the broad context of (multivariate) object-oriented data analysis (and, in a narrower sense, of symbolic data analysis). This opens new views even to cope with mixed types of data (e.g., Euclidean, functional, compositional), that remains one of the greatest challenges for the future.

Acknowledgements The authors would like to thank the editors of this Festschrift for the initiative and all their work, as well as two anonymous reviewers for helpful comments. Karel Hron gratefully acknowledges the support by the Czech Science Foundation (GACR), GA 19-01768S.

References

- Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.
- Aitchison, J., & Greenacre, M. (2002). Biplots of compositional data. *Applied Statistics*, *51*, 375–392.
- Billard, L., & Diday, E. (2006). *Symbolic data analysis*. Chichester: Wiley.
- van den Boogaart, K. G., Egozcue, J. J., & Pawłowsky-Glahn, V. (2010). Bayes linear spaces. *SORT*, *34*, 201–222.
- van den Boogaart, K. G., Egozcue, J. J., & Pawłowsky-Glahn, V. (2014). Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, *56*, 171–194.
- Bruno, F., Greco, F., & Ventrucci, M. (2015). Spatio-temporal regression on compositional covariates: modeling vegetation in a gypsum outcrop. *Environmental and Ecological Statistics*, *22*, 445–463.
- Diday, E. (2016). Thinking by classes in data science: the symbolic data analysis paradigm. *Wiley Interdisciplinary Reviews: Computational Statistics*, *8*(5), 172–205.
- Eaton, M. L. (1983). *Multivariate Statistics. A Vector Space Approach*. New York: Wiley.
- Egozcue, J. J., Pawłowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, *35*, 279–300.
- Egozcue, J. J., & Pawłowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, *37*, 795–828.
- Egozcue, J. J., Díaz-Barrero, J. L., & Pawłowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica, English Series*, *22*, 1175–1182.
- Egozcue, J. J., & Pawłowsky-Glahn, V. (2016). Changing the reference measure in the simplex and its weighting effects. *Austrian Journal of Statistics*, *45*, 25–44.
- Filzmoser, P., & Hron, K. (2011). Robust statistical analysis. In V. Pawłowsky-Glahn & A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications* (pp. 59–72). Chichester: Wiley.
- Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis*. Cham: Springer.

- Fišerová, E., & Hron, K. (2011). On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, *43*, 455–468.
- Horváth, L., & Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Heidelberg: Springer.
- Hron, K., Filzmoser, P., & Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, *39*, 1115–1128.
- Hron, K., Menafoglio, A., Templ, M., Hružová, K., & Filzmoser, P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics and Data Analysis*, *94*, 330–350.
- Hron, K., Brito, P., & Filzmoser, P. (2017). Exploratory data analysis for interval compositional data. *Advances in Data Analysis and Classification*, *11*(2), 223–241.
- Kynčlová, P., Filzmoser, P., & Hron, K. (2016). Compositional biplots including external non-compositional variables. *Statistics*, *50*, 1132–1148.
- Machalová, J., Hron, K., & Monti, J. S. (2016). Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*, *43*, 1419–1435.
- Machalová, J., Talská, R., Hron, K., Gába, A. (2020) Compositional splines for representation of density functions. *Computational Statistics*, <https://doi.org/10.1007/s00180-020-01042-7>.
- Marron, J. S., & Alonso, A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal*, *56*, 732–753.
- Menafoglio, A., Guadagnini, A., & Secchi, P. (2014). A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment*, *28*, 1835–1851.
- Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. Chichester: Wiley.
- Ramsay, J., & Silverman, B. W. (2005). *Functional Data Analysis* (2nd ed.). New York: Springer.
- Scheffé, H. (1958). Experiments with mixtures. *Journal of the Royal Statistical Society - B*, *20*, 344–360.
- Talská, R., Menafoglio, A., Machalová, J., Hron, K., & Fišerová, E. (2018). Compositional regression with functional response. *Computational Statistics and Data Analysis*, *123*, 66–85.
- Talská, R., Menafoglio, A., Hron, K., Egozcue, J. J., Palarea-Albaladejo, J. (2020) Weighting the domain of probability densities in functional data analysis. *Stat* 9(1), e283.
- Tolosana-Delgado, R., van den Boogaart, K. G., Mikes, T., von Eynatten, H. (2008). Statistical treatment of grain-size curves and empirical distributions: Densities as compositions? In: Daunis-i-Estadella, J., Martín-Fernández, J. A. (Eds.), *Proceedings of CoDaWork 2008*. University of Girona, Girona.
- Wang, H., Shangquan, L., Wu, J., & Guan, R. (2013). Multiple linear regression modeling for compositional data. *Neurocomputing*, *122*, 490–500.
- Wang, H., Shangquan, L., Guan, R., & Billard, L. (2015). Principal component analysis for compositional data vectors. *Computational Statistics*, *30*(4), 1079–1096.

A Spatial Durbin Model for Compositional Data



Tingting Huang, Gilbert Saporta, and Huiwen Wang

Abstract A compositional linear model (regression of a scalar response on a set of compositions) for areal data is proposed, where observations are not independent and present spatial autocorrelation. Specifically, we borrow thoughts from the spatial Durbin model considering that it produces unbiased coefficient estimates compared to other spatial linear regression models (including the spatial error model, the spatial autoregressive model, the Kelejian-Prucha model, and the spatial Durbin error model). The orthonormal log-ratio (olr) transformation based on a sequential binary partition of compositions and maximum likelihood estimation method are employed to estimate the new model. We also check the proposed estimators on a simulated and a real dataset.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

T. Huang

School of Statistics, Capital University of Economics and Business, Beijing, China

e-mail: tingth@buaa.edu.cn

T. Huang · H. Wang

School of Economics and Management, Beihang University, Beijing, China

e-mail: wanghw@vip.sina.com

T. Huang

Beijing Key Laboratory of Emergence Support Simulation Technologies for City Operations, Beijing, China

G. Saporta (✉)

CNAM, Center for Studies and Research in Computer Science and Communication, Paris, France

e-mail: gilbert.saporta@cnam.fr

H. Wang

Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, China

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_24

1 Introduction

With rapid development of computer technology, a huge amount of data characterized with complex structures, such as functional data (Ramsay and Silverman 2002, 2005; Horváth and Kokoszka 2012; Wang et al. 2016), compositional data (Pawlowsky-Glahn and Buccianti 2011; Pawlowsky-Glahn et al. 2015; Martín-Fernández et al. 2019) and symbolic data (Billard and Diday 2012, 2020; Bock and Diday 2012; Ochs et al. 2016) have been stored. Among them, compositional data, which describes parts of some whole, receives much attention owing to its wide applications in biology, economics, survey analysis, and genomics (Pawlowsky-Glahn and Buccianti 2011; McKinley et al. 2020). The key problem of using statistical tools to analyze compositional data is that components of a composition are not independent. Aitchison (1986) has made a great contribution by introducing the concept of independence in the simplex, i.e., sample space of compositional data.

In compositional data analysis, linear regressions play an important role. Existing linear models developed for compositional data can be divided into two categories. For the first group, explanatory variables are compositional while dependent variable is scalar. Hron et al. (2012) used this type of model to study the relationship between life expectancy and GDP compositions for the European Union member states. As to the second type, covariates and responses are both compositional. Wang et al. (2013) employed model in this sub-category, multiple linear regression for compositional data, to learn how the employment and investment levels are related to the gross regional product. In this research, we are interested in Compositional Linear Model with Numerical Responses (CLMNR).

In the CLMNR, it is commonly presumed that observations are independent. However, this assumption can break down in practical issues, especially when disposing of data with spatial autocorrelation. For instance, to understand how three strata of industry are concerned in $PM_{2.5}$ concentration for 34 major cities of China, it is unreasonable to assume these cities are independent, as a city's $PM_{2.5}$ concentration can be influenced by nearby cities' because of air movements. There are mainly two kinds of spatially dependent data, point-referenced data and areal data (lattice data), which are separately considered (Anselin 2002). We focus on lattice data. Huang et al. (2019) proposed a spatial autoregressive model for compositional data (SARCD), for the purpose of modeling areal data whose observed covariates are compositional and scalar whereas responses are numerical. This new model is built upon the popular spatial autoregressive (SAR) model, which includes spatial dependence in the dependent variable. Nevertheless, as explained in Lesage and Pace (2009), SAR model is not the best model to be assumed when the true data generating process is uncertain. For example, if the true model is the spatial Durbin model (SDM), using SAR model will suffer from omitted variables bias. In contrast, SDM produces unbiased coefficient estimates when the true model is a spatial error model (SEM), an SAR model or a Kelejian–Prucha model. Therefore, constructing a SDM for compositional data (SDMCD) is more useful than building a model upon SAR. Based on this consideration, we put forward a new SDMCD.

The new model involves the effect of spatial average of neighboring responses and neighboring explanatory variables, thus more flexible than the existing SARCD.

To estimate our SDMCD, the orthonormal log-ratio (olr) transformation (Martín-Fernández et al. 2019) and maximum likelihood estimation (MLE) method are employed to handle compositional covariates and spatial terms, respectively. Numerical experiments show that our estimators are efficient. A real dataset of $PM_{2.5}$ concentration and three strata of industry is analyzed using our model.

The chapter is organized as follows. Section 2 reviews preliminaries of compositional data and introduces the SDMCD. Section 3 gives details of the estimation method. The simulation study and real data analysis is displayed in Sects. 4 and 5. The last section is a discussion.

2 Model Specification

About notations: normally, we use bold “ \mathbf{x} ” to denote a composition of D parts, i.e., $\mathbf{x} = (x_1, \dots, x_D)'$. Nevertheless, a d -dimensional vector of real numbers is often written as $\mathbf{x} = (x_1, \dots, x_d)'$ as well. To avoid confusion, we add a superscript “ D ”, i.e., $\mathbf{x}^D = (x_1^D, \dots, x_D^D)'$, to represent compositional data. Correspondingly, \mathbf{X}^D is a $n \times D$ matrix collecting compositions on n units.

2.1 Preliminaries of Compositional Data

In this subsection, we introduce the Aitchison geometry (Pawlowsky-Glahn et al. 2015, pp. 23–30), a geometric space made up of compositional data.

Firstly, the set formed by compositions \mathbf{x}^D with D parts is the simplex \mathcal{S}^D , where

$$\mathcal{S}^D = \{\mathbf{x}^D = (x_1^D, x_2^D, \dots, x_D^D)' \mid x_j^D > 0, \sum_{j=1}^D x_j^D = k, j = 1, 2, \dots, D\}.$$

Note that all the parts x_j^D of \mathbf{x}^D are non-negative and k is a rescaling parameter. Without loss of generality, we set $k = 1$, i.e., sum of all the proportions is 100%. For elements in the simplex \mathcal{S}^D , operations of adding and scalar multiplying are defined to produce a linear space for compositional data.

Denote any two compositions in \mathcal{S}^D by $\mathbf{x}^D = (x_1^D, x_2^D, \dots, x_D^D)'$ and $\mathbf{y}^D = (y_1^D, y_2^D, \dots, y_D^D)'$. The addition \oplus and scalar multiplication \odot are

$$\begin{aligned} \mathbf{x}^D \oplus \mathbf{y}^D &= C(x_1^D y_1^D, x_2^D y_2^D, \dots, x_D^D y_D^D), \\ \alpha \odot \mathbf{x}^D &= C((x_1^D)^\alpha, (x_2^D)^\alpha, \dots, (x_D^D)^\alpha), \end{aligned}$$

where α is a scalar value. We call \oplus and \odot perturbation and powering, respectively. Here, $C(\cdot)$ is the closure operation

$$C(x_1^D, x_2^D, \dots, x_D^D) = \left(\frac{x_1^D}{\sum_{j=1}^D x_j^D}, \frac{x_2^D}{\sum_{j=1}^D x_j^D}, \dots, \frac{x_D^D}{\sum_{j=1}^D x_j^D} \right)'.$$

It is known that the closure operation ensures that the resulting composition belongs to S^D . Based on perturbation and powering, operation of subtraction \ominus can be derived

$$\mathbf{x}^D \ominus \mathbf{y}^D = \mathbf{x}^D \oplus (-1 \odot \mathbf{y}^D) = C \left(\frac{x_1^D}{y_1^D}, \frac{x_2^D}{y_2^D}, \dots, \frac{x_D^D}{y_D^D} \right).$$

Notice the above operations are linear, hence inner product for compositional data should be also introduced, which is

$$\langle \mathbf{x}^D, \mathbf{y}^D \rangle_a = \sum_{j=1}^D \ln \frac{x_j^D}{g_m(\mathbf{x}^D)} \ln \frac{y_j^D}{g_m(\mathbf{y}^D)}$$

where $g_m(\cdot)$ is geometric average of all the parts of a composition, i.e., $g_m(\mathbf{x}^D) = (\prod_{j=1}^D x_j^D)^{\frac{1}{D}}$, $g_m(\mathbf{y}^D) = (\prod_{j=1}^D y_j^D)^{\frac{1}{D}}$. Here, the simplex together with operations of perturbation, powering, and inner product form a Hilbert space, called the Aitchison space. Subscript a in $\langle \cdot, \cdot \rangle_a$ implies the evaluation is within the Aitchison space.

To maintain consistency, in the following sections, $\langle \cdot, \cdot \rangle_a$ is also used to denote an inner product between a vector of compositions $\mathbf{X}^D = (\mathbf{x}_1^D, \dots, \mathbf{x}_n^D)'$ and a composition $\boldsymbol{\theta}^D$, i.e.,

$$\langle \mathbf{X}^D, \boldsymbol{\theta}^D \rangle_a = \begin{pmatrix} \langle \mathbf{x}_1^D, \boldsymbol{\theta}^D \rangle_a \\ \vdots \\ \langle \mathbf{x}_n^D, \boldsymbol{\theta}^D \rangle_a \end{pmatrix}.$$

Similarly, we use \odot to indicate a vector of compositions $\mathbf{X}^D = (\mathbf{x}_1^D, \dots, \mathbf{x}_n^D)'$ is multiplied by a scalar matrix \mathbf{W} ,

$$\mathbf{W} \odot \mathbf{X}^D = \begin{pmatrix} w_{11} & \dots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \dots & w_{nn} \end{pmatrix} \odot \begin{pmatrix} \mathbf{x}_1^D \\ \vdots \\ \mathbf{x}_n^D \end{pmatrix} = \begin{pmatrix} \bigoplus_{k=1}^n w_{1k} \odot \mathbf{x}_k^D \\ \vdots \\ \bigoplus_{k=1}^n w_{nk} \odot \mathbf{x}_k^D \end{pmatrix}.$$

2.2 The SAR Model and the SDM Model

Among spatial linear models, the SAR model is one of the most widely studied, which has the following expression:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n) \tag{1}$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ is response, $\mathbf{X} = (x_{ij})_{n \times p}$ is covariate that has p explanatory variables, $\mathbf{W} = (w_{ij})_{n \times n}$ is a spatial weight matrix and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)'$ is disturbance that follows multivariate normal distribution $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ where \mathbf{I}_n is the identity matrix. $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$, ρ, σ^2 are parameters to be estimated. Note that \mathbf{W} is predefined, which is constructed according to spatial scenarios. In general, we build \mathbf{W} based on adjacent relations (rook matrix, queen matrix), geographic distance or friend relationship (refer to Anselin 1998 for more details). Some unusual distance like economic distance has also been employed.

The SDM is another important model, which contains spatial dependence in both dependent variable and independent variables

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n), \tag{2}$$

where $\mathbf{y}, \mathbf{X}, \mathbf{W}, \rho, \boldsymbol{\beta}, \boldsymbol{\epsilon}$ are defined as those in (1) and $\mathbf{W}\mathbf{X}$ is the spatial lag of explanatory variables. Obviously, the SAR model is a special case of the SDM.

For the purpose of illustrating importance of the SDM, there is a need of introducing the Manski model, which additionally involves the spatial autocorrelation of residuals

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \quad \mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n). \tag{3}$$

At the first sight, (3) is a more general model that nests the SDM thus deserving more attention. However, Manski (1993) pointed out that the parameters in (3) are unidentified except that one of the spatial lag terms (i.e., $\rho \mathbf{W}\mathbf{y}, \mathbf{W}\mathbf{X}\boldsymbol{\theta}$ or $\lambda \mathbf{W}\mathbf{u}$) is excluded from the Manski model. It is found that excluding $\rho \mathbf{W}\mathbf{y}$ or $\mathbf{W}\mathbf{X}\boldsymbol{\theta}$ will lead to biased estimates if they are present in the true model. On the other side, ignoring $\lambda \mathbf{W}\mathbf{u}$ will only cause a loss of efficiency if it is involved in the true model (refer to Lesage and Pace (2009) (pp. 155–158) and Elhorst (2010) for more details). Therefore, the best strategy to solve the identification problem is omitting spatial dependence in the disturbances, which obtains an SDM. The SDM also nests the SEM, as it reduces to

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \rho \mathbf{W}\mathbf{u} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

when $\boldsymbol{\theta} = -\rho \boldsymbol{\beta}$. To sum up, the SDM is the best model to choose if we do not know the underlying data generating process.

2.3 The Spatial Durbin Model (SDM) for Compositional Data

Following Qu and Lee (2015), we suppose the spatial process takes place on an uneven lattice $L, L \subset \mathbf{R}^p, p \geq 1$. Besides, the distance between any two points on L is greater than 0. There are n units observed from L . For each unit i , the recorded data is $\{y_i, \mathbf{x}_i^D = (x_{i1}^D, x_{i2}^D, \dots, x_{iD}^D)\}$. Write $\mathbf{X}^D = (\mathbf{x}_1^D, \mathbf{x}_2^D, \dots, \mathbf{x}_n^D)'$, the new model is

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \langle \mathbf{W} \odot \mathbf{X}^D, \boldsymbol{\theta}^D \rangle_a + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n) \quad (4)$$

where ρ and \mathbf{W} are defined as those in (2), $\boldsymbol{\beta}^D$ and $\boldsymbol{\theta}^D$ are compositional parameters. We regard $\mathbf{W} \odot \mathbf{X}^D$, linear combination of neighborhood covariates, as a new independent variable in addition to \mathbf{X}^D . For the first component of the first element $\bigoplus_{k=1}^n w_{1k} \odot \mathbf{x}_k^D$ of $\mathbf{W} \odot \mathbf{X}^D$, it is the first part of spatial average of nearing \mathbf{x}_i^D s. $\boldsymbol{\beta}^D$ and $\boldsymbol{\theta}^D$ can be seen as the projection directions that most explain \mathbf{y} . Model (4) has much generality.

- When $\rho = 0$ and $\boldsymbol{\theta}^D = (\frac{1}{D}, \frac{1}{D}, \dots, \frac{1}{D})'$, the proposed model reduces to the common compositional linear model

$$\mathbf{y} = \langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \boldsymbol{\epsilon}.$$

- When $\boldsymbol{\theta}^D = (\frac{1}{D}, \frac{1}{D}, \dots, \frac{1}{D})'$ and numerical covariates present, our model is the SAR model for compositional data (Huang et al. 2019)

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X} = (x_{ij})_{n \times p}$ is a matrix of p numerical covariates.

3 Estimation Method

Although we have learned how to evaluate inner product of two compositions, it is not straightforward to estimate compositional parameters $\boldsymbol{\beta}^D, \boldsymbol{\theta}^D$ in (4). In this section, we first employ the *olr* transformation to deal with compositions, then use MLE method to handle the spatial terms.

3.1 Orthonormal Log-Ratio (*olr*) Transformation

With the sum-to-one constraint, components of a composition are not independent, which makes it hard to cope with in regressions. We adopt the *olr* transformation rather than additive log-ratio (*alr*) transformation here, as the *alr* is not distance preserving. The key idea of *olr* transformation is representing compositions by coordinates under an orthonormal basis.

Given a set of orthonormal compositions $\{\mathbf{e}_i^D\}_{i=1}^{D-1}$, any composition \mathbf{x}^D can then be expanded as

$$\mathbf{x}^D = \langle \mathbf{x}^D, \mathbf{e}_1^D \rangle_a \odot \mathbf{e}_1^D + \langle \mathbf{x}^D, \mathbf{e}_2^D \rangle_a \odot \mathbf{e}_2^D + \dots + \langle \mathbf{x}^D, \mathbf{e}_{D-1}^D \rangle_a \odot \mathbf{e}_{D-1}^D.$$

Thus, coordinates $\xi = (\xi_1, \xi_2, \dots, \xi_{D-1})'$ of the olr transformation is

$$\xi = olr(\mathbf{x}^D) = (\langle \mathbf{x}^D, \mathbf{e}_1^D \rangle_a, \langle \mathbf{x}^D, \mathbf{e}_2^D \rangle_a, \dots, \langle \mathbf{x}^D, \mathbf{e}_{D-1}^D \rangle_a)'$$

Different techniques have been raised to construct orthonormal basis in the Aitchison geometry. For the purpose of interpretability, we adopt the method introduced by Hron et al. (2012). Under this special basis, coefficients are

$$\begin{aligned} \xi_1 &= \sqrt{\frac{D-1}{D}} \ln \frac{x_1^D}{\sqrt{\prod_{k=2}^D x_k^D}} \\ \xi_j &= \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j^D}{\sqrt{\prod_{k=j+1}^D x_k^D}}, \quad j = 2, 3, \dots, D-1. \end{aligned} \tag{5}$$

Observing Eq. (5), ξ_1 is log-ratio of the first part x_1^D and geometric mean of the rest parts $\sqrt{\prod_{k=2}^D x_k^D}$ scaled by a coefficient $\sqrt{\frac{D-1}{D}}$. Thus, ξ_1 can be regarded as relative weight of the first part and average of the rest part, which on the other side represents relative information of x_1^D with respect to average of the rest parts of \mathbf{x}^D .

Suppose olr coordinates of β^D is $\beta = (\beta_1, \beta_2, \dots, \beta_{D-1})'$. And denote $olr(\mathbf{X}^D) = (olr(x_1^D), olr(x_2^D), \dots, olr(x_n^D))' = (\xi_1, \xi_2, \dots, \xi_n) = \Xi$ where $\xi_i = (\xi_{i1}, \xi_{i2}, \dots, \xi_{i(D-1)})'$. According to orthonormal property of the olr basis, we have

$$\langle \mathbf{X}^D, \beta^D \rangle_a = \Xi \beta,$$

where β is unknown and to be estimated. Similarly, we have

$$\langle \mathbf{W} \odot \mathbf{X}^D, \theta^D \rangle_a = \mathbf{W} \Xi \theta,$$

where $\theta = (\theta_1, \theta_2, \dots, \theta_{D-1})'$ is the coordinate of θ^D and to be estimated. Therefore, model (4) can be expressed by

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \Xi \beta + \mathbf{W} \Xi \theta + \epsilon. \tag{6}$$

Now the original compositional regression is rewritten by a scalar model. And from expression (5), we know ξ_1 represents the relative information of x_1^D as regard to average of the rest of parts. Therefore, by estimating β_1 , coefficient of $\xi_{.1} = (\xi_{11}, \xi_{21}, \dots, \xi_{n1})'$, we can understand how the first component x_1^D influences responses. Concretely, if β_1 is negative, increasing $\xi_{.1}$, i.e., log-ratio of x_1^D and $\sqrt{\prod_{k=2}^D x_k^D}$, will make \mathbf{y} decrease, which on the other hand means the first part is negatively related to \mathbf{y} . Similarly, to see impacts of the rest parts, Eq. (5) need to be changed accordingly. Taking x_2^D as an example, the first expression of (5) will be replaced by $\xi_1 = \sqrt{\frac{D-1}{D}} \ln \frac{x_2^D}{\sqrt{\prod_{k \neq 2}^D x_k^D}}$. This procedure needs to be done D times in

total to understand affects of all the components on y (details refer to Hron et al. (2012)). In Sect. 5, we will adopt the method to study how different industry sectors are related to $PM_{2.5}$ pollution.

3.2 Maximum Likelihood Estimation (MLE) Method

We briefly introduce the process of estimating model (6). Write $Z = (\Xi, W\Xi)$ and $\Delta = (\beta', \theta')'$, then (6) is

$$y = \rho Wy + Z\Delta + \epsilon. \tag{7}$$

Based on the fact that ϵ follows normal distribution, the log-likelihood function $L(\rho, \Delta, \sigma^2)$ of y can be obtained. The next step is to maximize $L(\rho, \Delta, \sigma^2)$ with respect to ρ, Δ and σ^2 . However, the optimization requires heavy computation and is not applicable. The log-likelihood function $L(\rho)$ concentrated with regards to Δ and σ^2 has many advantages over $L(\rho, \Delta, \sigma^2)$, and is generally used to get estimated parameters (more details refer to Lesage and Pace 2009).

Once $\hat{\Delta}$ is gained, $\hat{\beta}^D$ and $\hat{\theta}^D$ are evaluated through inverse of the olr transformation by $\hat{\beta}$ and $\hat{\theta}$. Take $\hat{\beta}^D$ as an example

$$\begin{aligned} \beta_1^D &= \exp\left(\frac{\sqrt{D-1}}{\sqrt{D}}\beta_1\right), \\ \beta_j^D &= \exp\left(-\sum_{k=1}^{j-1} \frac{1}{\sqrt{(D-k+1)(D-k)}}\beta_k + \frac{\sqrt{D-j}}{\sqrt{D-j+1}}\beta_j\right), \quad j = 2, \dots, D-1, \\ \beta_D^D &= \exp\left(-\sum_{k=1}^{D-1} \frac{1}{\sqrt{(D-k+1)(D-k)}}\beta_k\right). \end{aligned}$$

$\hat{\theta}^D$ can be similarly computed. $\hat{\rho}$ comes from optimization of $L(\rho)$.

4 Simulation Study

Several experiments are conducted to evaluate finite-sample performance of the estimators of ρ, β^D, θ^D . All of the computations were carried out in the R environment, and we used existing functions in the R packages ‘spdep’ (<https://r-forge.r-project.org/projects/spdep/>), ‘compositions’ (<http://www.stat.boogaart.de/compositions>), and ‘robCompositions’ (can be freely downloaded from CRAN).

The spatial scenario is designed as the rook case (Anselin 1998), where two units are neighbors if they share a common edge. We randomly apportioning n agents on a regular square grid of R rows and T columns; each agent occupies a cell on the grid. Thus, sample size $n = R \times T$. Agents in the inner field of the grid have four

neighbors while objects along the borders and in the corners have two and three. The spatial matrix \mathbf{W} is an adjacency matrix. We set $n = \{10 \times 30, 20 \times 25, 30 \times 30\}$ in the simulation. Besides, different levels of spatial dependency in regression are considered, i.e., $\rho = \{0, 0.5, 0.8\}$.

As for the data generating process, we produce $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ through the following three compositional spatial linear models:

- (1) spatial autoregressive model for compositional data (SARCD)

$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + 0.5\boldsymbol{\epsilon}),$$

- (2) spatial error model for compositional data (SEMCD)

$$\mathbf{y} = \langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \mathbf{u}, \quad \mathbf{u} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (0.5\boldsymbol{\epsilon}),$$

- (3) spatial Durbin model for compositional data (SDMCD)

$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\langle \mathbf{X}^D, \boldsymbol{\beta}^D \rangle_a + \langle \mathbf{W} \odot \mathbf{X}^D, \boldsymbol{\theta}^D \rangle_a + 0.5\boldsymbol{\epsilon}),$$

for the purpose of comparing properties of these models' estimators. Here

$$\boldsymbol{\beta}^D = \left(\frac{4}{9}, \frac{2}{9}, \frac{1}{3} \right)', \quad \boldsymbol{\theta}^D = \left(\frac{3}{8}, \frac{1}{8}, \frac{1}{2} \right)', \quad \epsilon_i \sim N(0, 1).$$

And the explanatory compositions $\mathbf{X}^D = (\mathbf{x}_1^D, \mathbf{x}_2^D, \dots, \mathbf{x}_n^D)'$ are simulated by the multivariate normal distribution $N_s(\boldsymbol{\mu}^D, \boldsymbol{\Sigma})$ (Pawlowsky-Glahn et al. 2015) (pp. 114–118) on the simplex, where

$$\boldsymbol{\mu}^D = \left(\frac{1}{6}, \frac{1}{3}, \frac{1}{2} \right)', \quad \boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1.5 \\ -1.5 & 2 \end{pmatrix}.$$

Estimation method of the SARCD is introduced in Huang et al. (2019). Parameters of the SEMCD can be obtained similarly by first representing compositional covariates using olr coordinates and then employing the MLE method for the SEM (Lesage and Pace 2009). For each simulated dataset, we use three models, the SARCD, the SEMCD, and the SDMCD to fit and estimate their parameters.

The experiment is repeated 500 times in each setting. We assess the performance of the estimator $\hat{\rho}$ by mean bias and standard deviation. As for $\boldsymbol{\beta}^D, \boldsymbol{\theta}^D$, their behavior is evaluated by the simplicial bias and the simplicial root mean square error (Pawlowsky-Glahn et al. 2015) (pp. 134–135). We mention that value of the simplicial bias is compositional while that of the simplicial root mean square error is scalar. Denote the empirical mean of $\hat{\boldsymbol{\beta}}^D$ by $\bar{\boldsymbol{\beta}}^D = \frac{1}{n} \odot \bigoplus_{k=1}^n \hat{\boldsymbol{\beta}}_k^D$, where $\hat{\boldsymbol{\beta}}_k^D$ is the estimated $\boldsymbol{\beta}^D$ for the kth simulated dataset. The simplicial bias of $\bar{\boldsymbol{\beta}}^D$ is

Table 1 The empirical average biases and standard deviations (in brackets) of $\hat{\rho}$

True model	Sample size	Fitting model	$\rho = 0$	$\rho = 0.5$	$\rho = 0.8$
SDMCD	$n = 300$	SDMCD	-0.0021 (0.0418)	-0.0019 (0.0290)	-0.0011 (0.0145)
		SARCD	0.4219 (0.0770)	0.3077 (0.0307)	0.1469 (0.0107)
		SEMCD	-0.0698 (0.4866)	0.3625 (0.0306)	0.1655 (0.0096)
	$n = 500$	SDMCD	0.0000 (0.0349)	-0.0019 (0.0230)	-0.0010 (0.0114)
		SARCD	0.4334 (0.0574)	0.3169 (0.0225)	0.1503 (0.0085)
		SEMCD	-0.0366 (0.4688)	0.3711 (0.0230)	0.1691 (0.0075)
	$n = 900$	SDMCD	-0.0013 (0.0243)	-0.0023 (0.0174)	-0.0002 (0.0083)
		SARCD	0.4376 (0.0396)	0.3202 (0.0168)	0.1534 (0.0060)
		SEMCD	-0.0181 (0.4281)	0.3744 (0.0176)	0.1721 (0.0053)
SARCD	$n = 300$	SDMCD	-0.0041 (0.0806)	-0.0126 (0.0590)	-0.0099 (0.0347)
		SARCD	-0.0014 (0.0514)	-0.0050 (0.0442)	-0.0053 (0.0275)
		SEMCD	-0.0038 (0.0811)	0.0595 (0.0629)	0.0767 (0.0284)
	$n = 500$	SDMCD	-0.0045 (0.0594)	-0.0063 (0.0469)	-0.0068 (0.0256)
		SARCD	-0.0010 (0.0418)	-0.0059 (0.0358)	-0.0044 (0.0201)
		SEMCD	-0.0046 (0.0598)	0.0646 (0.0508)	0.0792 (0.0214)
	$n = 900$	SDMCD	-0.0013 (0.0469)	-0.0091 (0.0354)	-0.0023 (0.0195)
		SARCD	-0.0034 (0.0314)	-0.0055 (0.0268)	-0.0020 (0.0156)
		SEMCD	-0.0012 (0.0471)	0.0629 (0.0388)	0.0831 (0.0154)
SEMCD	$n = 300$	SDMCD	-0.0041 (0.0806)	-0.0152 (0.0625)	-0.0147 (0.0405)
		SARCD	-0.0014 (0.0514)	-0.2265 (0.0598)	-0.2227 (0.0602)
		SEMCD	-0.0038 (0.0811)	-0.0117 (0.0623)	-0.0113 (0.0398)
	$n = 500$	SDMCD	-0.0045 (0.0594)	-0.0065 (0.0494)	-0.0097 (0.0299)
		SARCD	-0.0010 (0.0418)	-0.2287 (0.0462)	-0.2201 (0.0481)
		SEMCD	-0.0046 (0.0598)	-0.0040 (0.0493)	-0.0077 (0.0298)
	$n = 900$	SDMCD	-0.0013 (0.0469)	-0.0099 (0.0368)	-0.0035 (0.0231)
		SARCD	-0.0034 (0.0314)	-0.2299 (0.0350)	-0.2134 (0.0377)
		SEMCD	-0.0012 (0.0471)	-0.0086 (0.0369)	-0.0023 (0.0230)

Table 2 The simplicial biases and the simplicial root mean square error (in brackets) of $\hat{\beta}^D$ when fitting model is the SARCD

True model	n	$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$		
		$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$	$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$	$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$
SDMCD	300	0.3232 (0.0935)	0.3654	0.3114	0.3185 (0.0871)	0.3680	0.3135	0.3187 (0.0866)	0.3614	0.3200
	500	0.3214 (0.0695)	0.3665	0.3121	0.3175 (0.0634)	0.3682	0.3143	0.3219 (0.0673)	0.3597	0.3184
	900	0.3220 (0.0543)	0.3660	0.3120	0.3185 (0.0482)	0.3675	0.3140	0.3201 (0.0482)	0.3598	0.3201
SARCD	300	0.3332 (0.0424)	0.3336	0.3333	0.3326 (0.0430)	0.3334	0.3340	0.3330 (0.0453)	0.3334	0.3336
	500	0.3329 (0.0351)	0.3335	0.3336	0.3331 (0.0340)	0.3337	0.3333	0.3338 (0.0354)	0.3331	0.3331
	900	0.3331 (0.0261)	0.3335	0.3334	0.3334 (0.0246)	0.3334	0.3333	0.3332 (0.0254)	0.3335	0.3334
SEMCD	300	0.3332 (0.0424)	0.3336	0.3333	0.3326 (0.0462)	0.3333	0.3341	0.3334 (0.0534)	0.3330	0.3337
	500	0.3329 (0.0351)	0.3335	0.3336	0.3330 (0.0367)	0.3338	0.3332	0.3336 (0.0421)	0.3331	0.3333
	900	0.3331 (0.0261)	0.3335	0.3334	0.3333 (0.0263)	0.3334	0.3333	0.3333 (0.0307)	0.3335	0.3332

Table 3 The simplicial biases and the simplicial root mean square error (in brackets) of $\hat{\beta}^D$ when fitting model is the SEMCD

True model	n	$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$		
		$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$	$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$	$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$
SDMCD	300	0.3378 (0.2664)	0.3234	0.3387	0.2920 (0.0918)	0.4273	0.2806	0.2869 (0.0834)	0.4280	0.2850
	500	0.3338 (0.2490)	0.3293	0.3370	0.2913 (0.0681)	0.4268	0.2818	0.2897 (0.0674)	0.4259	0.2843
	900	0.3341 (0.2254)	0.3317	0.3342	0.2924 (0.0527)	0.4260	0.2816	0.2888 (0.0470)	0.4254	0.2858
SARCD	300	0.3333 (0.0427)	0.3334	0.3333	0.3243 (0.0453)	0.3430	0.3328	0.3143 (0.0451)	0.3547	0.3310
	500	0.3330 (0.0352)	0.3334	0.3336	0.3247 (0.0352)	0.3430	0.3322	0.3158 (0.0359)	0.3539	0.3304
	900	0.3332 (0.0262)	0.3335	0.3333	0.3252 (0.0255)	0.3426	0.3322	0.3151 (0.0253)	0.3539	0.3310
SEMCD	300	0.3333 (0.0427)	0.3334	0.3333	0.3326 (0.0417)	0.3335	0.3340	0.3329 (0.0410)	0.3335	0.3335
	500	0.3330 (0.0352)	0.3334	0.3336	0.3332 (0.0332)	0.3336	0.3332	0.3337 (0.0325)	0.3332	0.3330
	900	0.3332 (0.0262)	0.3335	0.3333	0.3333 (0.0237)	0.3334	0.3333	0.3332 (0.0230)	0.3334	0.3334

$$sBias(\bar{\beta}^D) = \bar{\beta}^D \ominus \beta^D.$$

Note that when $\bar{\beta}^D$ is an unbiased estimator of β^D , we have $sBias(\bar{\beta}) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$. The simplicial root mean square error $sRMSE$ is defined by the total variance $totvar$, i.e.,

$$sRMSE(\hat{\beta}^D) = \sqrt{totvar(\hat{\beta}^D)}$$

Table 4 The simplicial biases and the simplicial root mean square error (in brackets) of $\hat{\beta}^D$ when fitting model is the SDMCD

True model	n	$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$		
		$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$	$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$	$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$
SDMCD	300	0.3332 (0.0443)	0.3333	0.3334	0.3327 (0.0442)	0.3332	0.3341	0.3328 (0.0465)	0.3334	0.3337
	500	0.3329 (0.0365)	0.3334	0.3337	0.3331 (0.0349)	0.3335	0.3334	0.3337 (0.0359)	0.3331	0.3332
	900	0.3332 (0.0269)	0.3334	0.3334	0.3334 (0.0256)	0.3332	0.3334	0.3331 (0.0256)	0.3335	0.3334
SARCD	300	0.3332 (0.0428)	0.3335	0.3333	0.3328 (0.0433)	0.3332	0.3340	0.3331 (0.0461)	0.3332	0.3337
	500	0.3329 (0.0353)	0.3334	0.3336	0.3331 (0.0342)	0.3336	0.3332	0.3339 (0.0356)	0.3330	0.3331
	900	0.3332 (0.0261)	0.3335	0.3333	0.3334 (0.0247)	0.3333	0.3333	0.3332 (0.0255)	0.3334	0.3334
SEMCD	300	0.3332 (0.0428)	0.3335	0.3333	0.3326 (0.0431)	0.3334	0.3340	0.3328 (0.0455)	0.3336	0.3337
	500	0.3329 (0.0353)	0.3334	0.3336	0.3331 (0.0342)	0.3337	0.3332	0.3336 (0.0353)	0.3333	0.3331
	900	0.3332 (0.0261)	0.3335	0.3333	0.3333 (0.0246)	0.3334	0.3333	0.3331 (0.0253)	0.3335	0.3333

Table 5 The simplicial biases and the simplicial root mean square error (in brackets) of $\hat{\theta}^D$ when fitting model is the SDMCD

True model	n	$\rho = 0$			$\rho = 0.5$			$\rho = 0.8$		
		$\hat{\theta}_1^D$	$\hat{\theta}_2^D$	$\hat{\theta}_3^D$	$\hat{\theta}_1^D$	$\hat{\theta}_2^D$	$\hat{\theta}_3^D$	$\hat{\theta}_1^D$	$\hat{\theta}_2^D$	$\hat{\theta}_3^D$
SDMCD	300	0.3333 (0.0928)	0.3333	0.3334	0.3342 (0.0901)	0.3326	0.3332	0.3337 (0.0867)	0.3327	0.3336
	500	0.3344 (0.0678)	0.3331	0.3326	0.3329 (0.0700)	0.3338	0.3333	0.3328 (0.0700)	0.3333	0.3339
	900	0.3337 (0.0511)	0.3337	0.3326	0.3336 (0.0516)	0.3331	0.3333	0.3333 (0.0508)	0.3337	0.3330
SARCD	300	0.2108 (0.1006)	0.6311	0.1582	0.2123 (0.0912)	0.6293	0.1584	0.2123 (0.0886)	0.6292	0.1585
	500	0.2118 (0.0719)	0.6305	0.1577	0.2106 (0.0723)	0.6316	0.1579	0.2111 (0.0707)	0.6305	0.1584
	900	0.2107 (0.0557)	0.6319	0.1575	0.2114 (0.0536)	0.6305	0.1581	0.2107 (0.0517)	0.6315	0.1577
SEMCD	300	0.2108 (0.1006)	0.6311	0.1582	0.1653 (0.0901)	0.6923	0.1424	0.1407 (0.0867)	0.7269	0.1324
	500	0.2118 (0.0719)	0.6305	0.1577	0.1637 (0.0714)	0.6946	0.1417	0.1399 (0.0691)	0.7279	0.1322
	900	0.2107 (0.0557)	0.6319	0.1575	0.1644 (0.0527)	0.6936	0.1420	0.1397 (0.0505)	0.7287	0.1316

$$totvar(\hat{\beta}^D) = \sum_{j=1}^{D-1} var(\hat{\beta}_j), \quad olr(\hat{\beta}^D) = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{D-1})$$

The results are summarized in Table 1 ($\hat{\rho}$), Table 2 ($\hat{\beta}^D$ when fitting model is the SARCD), Table 3 ($\hat{\beta}^D$ when fitting model is the SEMCD), Table 4 ($\hat{\beta}^D$ when fitting model is the SDMCD) and Table 5 ($\hat{\theta}^D$ when fitting model is the SDMCD). Examination of Tables 1, 2, 3, 4, 5 leads to the following conclusions.

- (1) Comparing $\hat{\rho}$ of the SARCD, the SEMCD, and the SDMCD in Table 1, it can be seen $\hat{\rho}$ of the SDMCD always performs well whenever what the true model is. On the other side, the SARCD behaves poorly when true models are the SDMCD and the SEMCD. The SEMCD has similar problems as well.
- (2) See results of $\hat{\rho}$ when the fitting model is the SDMCD in Table 1, it can be found that the mean biases are small and the standard deviations decrease as sample size n increases. Thus, our proposed estimator for ρ is efficient.
- (3) Compare $\hat{\beta}^D$ when using the SARCD, the SEMCD, and the SDMCD as fitting models in Tables 2, 3, 4. Firstly, it can be observed that the *sBiases* of $\hat{\beta}$ of the SARCD are relatively large when the simulated data is generated by the SDMCD (see Table 2). And the *sBiases* of $\hat{\beta}^D$ of the SEMCD are far from neutral element $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$ when the true models are the SDMCD and the SARCD (see Table 3). $\hat{\beta}^D$ of the SDMCD, by contrast, has stable and rather small simplicial biases under all the cases (see Table 4).
- (4) Table 5 shows results of $\hat{\theta}^D$. We can find θ^D is accurately estimated if the true model is the SDMCD. However, under other settings, $\hat{\theta}^D$ is simplicial biased.
- (5) Checking Tables 4, 5 when the true model is the SDMCD, $\hat{\beta}^D$ and $\hat{\theta}^D$ behave similarly, i.e., having small simplicial biases and decreasing *sRMSE* with increasing n .

5 Real Data Analysis

In this section, the proposed model is employed to investigate how three strata of industry affect pollution of $PM_{2.5}$ (fine particulate matters with diameter smaller than 2.5 mm) in 34 major cities of China in 2016. The $PM_{2.5}$ data was collected from China National Environmental Monitoring Centre. The industry structure data (proportions of the primary sector, the secondary sector, and the tertiary sector in gross domestic product (GDP)) was from Statistical Communiqué of the People's Republic of China on the 2016 National Economic and Social Development (<http://www.tjcn.org/tjgb/00zg/>). We show locations of the 34 major cities on the map of China in Fig. 1. The Moran's I statistics is applied to test if there exists spatial autocorrelation among the $PM_{2.5}$ data. The resulting value of the Moran's I statistics is 0.63 with p-value equalling 0.00002 (selection of the spatial weight matrix W will be given in the second paragraph), which means spatial effects are significant and there is a need to utilize spatial models. Besides, it is obvious that the industry structure data is compositional. Therefore, compositional spatial linear models should be employed.

Specifically, we consider three models, the SARCD, the SEMCD, and the SDMCD to make a comparison, which are, respectively, formed as



Fig. 1 The 34 major cities on map of China. Reprinted from [26] by permission from Springer Nature Customer Service Centre GmbH, ©Springer-Verlag GmbH Germany, part of Springer Nature 2020

$$y_i = \rho \sum_{j=1}^n w_{ij} y_j + \langle \mathbf{x}_i^D, \boldsymbol{\beta}^D \rangle_a + \epsilon_i, \tag{8}$$

$$y_i = \langle \mathbf{x}_i^D, \boldsymbol{\beta}^D \rangle_a + u_i, \quad u_i = \rho \sum_{j=1}^n w_{ij} u_j + \epsilon_i, \tag{9}$$

$$y_i = \rho \sum_{j=1}^n w_{ij} y_j + \langle \mathbf{x}_i^D, \boldsymbol{\beta}^D \rangle_a + \langle \bigoplus_{j=1}^n w_{ij} \odot \mathbf{x}_j^D, \boldsymbol{\theta}^D \rangle_a + \epsilon_i, \tag{10}$$

where y_i is the annual mean concentration of $PM_{2.5}$ of the i th city, \mathbf{x}_i^D is a composition of the i th city, composed of proportions of the three sectors, and w_{ij} is the weight constructed according to the distance d_{ij} between centers of city i and j . Here, d_{ij} is computed using the haversine formula based on city centers' latitudes and longitudes. The inverse distance is employed, i.e., $w_{ij} = \frac{1}{d_{ij}}$. We have also considered two factors that may affect behavior of models (8)–(10). The first is distance threshold d_0 . If two cities are far from each other, spatial dependence between them will be small. In this case, we set $w_{ij} = 0$ when $d_{ij} > d_0$. The second is the number of nearest neighbors N_0 . If there are many cities close to city i , i.e., $d_{ij} < d_0$, it is important to decide value of N_0 so that neighbors that really count are included. We set $N_0 = \{1, 2, 3, 4, 5, 6, 7\}$ in this case. And for each N_0 , search for a d_0 which makes

Table 6 The values of the Moran’s I statistics under different values of N_0 and d_0

N_0	1	2	3	4	5	6	7
d_0 (km)	431.5	459.2	459.2	459.2	375.0	375.0	375.0
Moran’s I	0.51	0.66	0.63	0.61	0.62	0.62	0.62

Table 7 The R-square, value of the Moran’s I statistics of residuals, estimated parameters and their p-values (in brackets) for the SARCD

R-square	Moran’s I	$\hat{\rho}$	$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$
0.73	0.04	0.55	2.494×10^{-6}	0.9999975	6×10^{-9}
–	–	–	–2.80	12.99	–10.20
–	(0.33)	(0.0002)	(0.40)	(0.08)	(0.09)

Table 8 The R-square, value of the Moran’s I statistics of residuals, estimated parameters, and their p-values (in brackets) for the SEMCD

R-square	Moran’s I	$\hat{\rho}$	$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$
0.74	0.07	0.58	0.014872	0.985119	9×10^{-6}
–	–	–	1.32	6.46	–7.77
–	(0.26)	(0.0003)	(0.69)	(0.34)	(0.17)

Table 9 The R-square, value of the Moran’s I statistics of residuals, estimated parameters, and their p-values (in brackets) for the SDMCD

R-square	Moran’s I	$\hat{\rho}$	$\hat{\beta}_1^D$	$\hat{\beta}_2^D$	$\hat{\beta}_3^D$	$\hat{\theta}_1^D$	$\hat{\theta}_2^D$	$\hat{\theta}_3^D$
0.74	0.07	0.58	1.283×10^{-5}	0.9999871	7×10^{-8}	3×10^{-8}	0.9999979	2.07×10^{-6}
–	–	–	–2.00	11.80	–9.80	–10.11	13.07	–2.96
–	(0.26)	(0.0003)	(0.56)	(0.11)	(0.09)	(0.02)	(0.21)	(0.74)

the Moran’s I statistics take the greatest value. At last, we select the combination of N_0 and d_0 , with which the value of the Moran’s I statistics is the greatest. Table 6 summarizes results of the Moran’s I statistics when N_0 and d_0 take different values. It can be seen $N_0 = 2$ and $d_0 = 459.2$ is the best choice. As Urumchi and Lhasa are far from other cities, we exclude their records from the dataset. The estimation results of models (8)–(10) are summarized in Tables 7, 8, 9.

Comparing values of the R-square of the SARCD, the SEMCD, and the SDMCD in Tables 7, 8, 9, we can find the SEMCD and the SDMCD have better fitting results. Then observe the estimated compositional coefficient $\hat{\beta}^D = (\hat{\beta}_1^D, \hat{\beta}_2^D, \hat{\beta}_3^D)'$ of the SEMCD and the SDMCD in Tables 8, 9. It is obvious that the coefficient for the first coordinate ξ_1 of the SEMCD is 1.32, which means the first sector is positively related to the $PM_{2.5}$ pollution, while that of the SDMCD is –2.00, indicating a negative relationship (recall explanations of the coordinates in Sect. 3.1). As it is more convincing that the primary industry is environmentally friendly, we conclude

that the SDMCD provides a better interpretation for the first sector compared to the SEMCD.

At last, we examine other important parameters of the SDMCD in Table 9. The spatial autoregressive parameter $\hat{\rho}$ is significant with p-value being 0.0003, as we expected. And the value of the Moran's I statistics of residuals of the SDMCD is 0.07, which implies the spatial dependence in \mathbf{y} is eliminated. As for the coefficients for the second and the third coordinates of $\hat{\boldsymbol{\beta}}^D$, they are 11.80 and -9.80 , showing the second and the third parts of \mathbf{x}_i^D are positively and negatively connected with $PM_{2.5}$ concentration, respectively. The other coefficient $\hat{\theta}^D$ indicates how a city's air quality is related to near cities' industry structure. The coordinate coefficients are -10.11 , 13.07 , and -2.96 , can be similarly interpreted as those of $\hat{\boldsymbol{\beta}}^D$.

6 Conclusion and Discussion

The compositional linear model which associates a scalar response to a set of compositional covariates is an important part of compositional data analysis. However, literature related to this type of compositional regression and considering spatial autocorrelation is relatively scanty. To fill the gap, Huang et al. (2019) put forward a new model on the basis of an SAR model. Nevertheless, the SAR model produces biased estimators for coefficients when the spatial lag of explanatory variables presents in the true model. The spatial Durbin model is a better choice as it does not suffer from biased estimates when the true spatial linear model is unknown. Therefore, we propose a spatial Durbin model with compositional predictors. The *olr* transformation with interpretability of coordinates and the maximum likelihood estimation method has been used to obtain estimators. And the proposed method performs well on the simulated and the real dataset.

This new model can be easily generalized to a complex situation, where a mix of scalar, compositional, and functional predictors are involved, like in articles (Wang et al. 2019a, b). Besides, in this study, we focussed only on the quality of estimators, another important issue is about prediction: how well our model will generalize for new observations? The Goulard et al. methodology (2017) for finding optimal strategies is an interesting answer which needs to be adapted. Further developments in order to investigate the robustness are also necessary in the spirit of Huang et al. (2020).

Acknowledgements This research was financially supported by the National Natural Science Foundation of China under Grant No. 71420107025 and Capital University of Economics and Business under Grant No. XRZ2021040.

References

- Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.
- Anselin, L. (1998). *Spatial econometrics: Methods and models*. Berlin: Springer.
- Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27(3), 247–267.
- Billard, L., & Diday, E. (2020). *Clustering methodology for symbolic data*. New Jersey: Wiley Ltd.
- Billard, L., & Diday, E. (2012). *Symbolic data analysis: Conceptual statistics and data mining*. New Jersey: Wiley.
- Bock, H.-H., & Diday, E. (2012). *Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data*. Springer Science & Business Media.
- Elhorst, J. (2010). Applied spatial econometrics: Raising the bar. *Spatial Economic Analysis*, 5(1), 9–28.
- Goulard, M., Laurent, T., & Thomas-Agnan, C. (2017). About predictions in spatial autoregressive models: Optimal and almost optimal strategies. *Spatial Economic Analysis*, 12(2–3), 304–325.
- Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*. Springer Science & Business Media.
- Hron, K., Filzmoser, P., & Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, 39(5), 1115–1128.
- Huang, T., Saporta, G., Wang, H., & Wang, S. (2020). A robust spatial autoregressive scalar-on-function regression with t-distribution. *Advances in Data Analysis and Classification*,. <https://doi.org/10.1007/s11634-020-00384-w>.
- Huang, T., Wang, H., & Saporta, G. (2019). Spatial autoregressive model for compositional data. *Journal of Beijing University of Aeronautics and Astronautics*, 45(1), 93–98.
- Lesage, J., & Pace, R. K. (2009). *Introduction to spatial econometrics*. London: Chapman and Hall/CRC.
- Manski, C. F. (1993). Identification of endogenous social effects: the reflection problem. *Review of Economic Studies*, 60, 531–542.
- Martín-Fernández, J. A., Engle, M. A., Ruppert, L. F., & Olea, R. A. (2019). Advances in self-organizing maps for their application to compositional data. *Stochastic Environmental Research and Risk Assessment*, 33, 817–826.
- McKinley, J. M., Mueller, U., Atkinson, P. M., Ofterdinger, U., Jackson, C., & Cox, S. F., et al. (2020). Investigating the influence of environmental factors on the incidence of renal disease with compositional data analysis using balances. *Applied Computing and Geosciences*, 6.
- Ochs, M., Diday, E., & Afonso, F. (2016). From the Symbolic Analysis of Virtual Faces to a Smiles Machine. *IEEE Transactions on Cybernetics*, 46(2), 401–409.
- Pawlowsky-Glahn, V., & Buccianti, A. (2011). *Compositional data analysis: Theory and applications*. New Jersey: Wiley.
- Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. New Jersey: Wiley Ltd.
- Qu, X., Lee, L.-f.: Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics*, 184(2), 209–232 (2015).
- Ramsay, J. O., & Silverman, B. W. (2002). *Applied functional data analysis: Methods and case studies*. New York: Springer.
- Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. New York: Springer.
- Wang, J. L., Chiou, J. M., & Müller, H. G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3, 257–295.
- Wang, H., Huang, T., & Wang, S. (2019a). A flexible spatial autoregressive modelling framework for mixed covariates of multiple data types. *Communications in Statistics-Simulation and Computation*,. <https://doi.org/10.1080/03610918.2019.1626885>.
- Wang, H., Shangquan, L., Wu, J., & Guan, R. (2013). Multiple linear regression modeling for compositional data. *Neurocomputing*, 122, 490–500.

Wang, Z., Wang, H., Wang, S., Lu, S., & Saporta, G. (2019b). Linear mixed-effects model for longitudinal complex data with diversified characteristics. *Journal of Management Science and Engineering*, <https://doi.org/10.1016/j.jmse.2019.11.001>.

Compositional Analysis of Exchange Rates



Wilfredo L. Maldonado, Juan José Egozcue, and Vera Pawlowsky-Glahn

Abstract Triangular arbitrage in the foreign exchange market of a group of countries exists whenever it is possible to make profit by buying and selling their currencies using the spot exchange rates. Working in the framework of the Aitchison geometry, and using characterizations of the absence of triangular arbitrage, we present two applications to the currencies of Brazil (Real), the European Union (Euro), Great Britain (Pound Sterling), and the United States of America (US Dollar). The first application refers to the Special Drawing Rights, an asset created by the International Monetary Fund to provide liquidity to the member countries. The exchange rates matrix is projected onto the subspace of no-arbitrage exchange rate matrices, and its only eigenvector, associated with a non-null eigenvalue, is demonstrated to be compositional and close to the Special Drawing Rights. The second application studies the relative exchange rate bubbles among the countries. It uses the closest no-arbitrage matrix of an exchange rate matrix and the purchasing power parity values

Dedication. This contribution is dedicated to Christine Thomas-Agnan for her 65th birthday. Christine is one of the pioneers at the forefront of introducing compositional methods in market share studies and in economics in general. Here we acknowledge her initiative in the field.

W. L. Maldonado

Faculty of Economics, Management and Accounting, University of São Paulo. Av. Professor Luciano Gualberto, 908 - Butantã - São Paulo - SP. CEP 05508-010, São Paulo, Brazil
e-mail: wilfredo.maldonado@usp.br

J. J. Egozcue

Technical University of Catalonia, Department of Civil and Environmental Engineering, Campus Nord c/Jordi Girona 1-3, C-2, E-08034 Barcelona, Spain
e-mail: juan.jose.egozcue@upc.edu

V. Pawlowsky-Glahn (✉)

University of Girona, Department of Computer Science, Applied Mathematics and Statistics, Campus Montilivi, 17003 Girona, Spain
e-mail: vera.pawlowsky@udg.edu

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_25

489

for the fundamental exchange rates to analyze the dynamics of those bubbles. These applications show the potential the compositional approach has for the matrices of exchange rates.

1 Introduction

The Foreign Exchange (FX) market is one of the most liquid markets in the world. According to the Bank of International Settlements, the global average daily volume was USD 2.0 trillion in April 2013. The three major currencies negotiated on that market are the US Dollar (USD), the Euro (EU), and the Japanese Yen (JPY). Furthermore, the new information technologies make the dissemination of information almost instantaneous among the traders.

Despite such enormous liquidity and the speed of information propagation, an intriguing phenomenon has been recurrently present in that market: the existence of Triangular Arbitrage (TA). A TA is a financial operation with the spot prices of financial assets that allow for a strictly positive profit without any cost. For example, consider the following values of the exchange rates among the USD, EU, and JPY¹: 1 EU is worth 1.1148 USD, 1 USD is worth 105.30 JPY (both are ask prices), and 1 EU is worth 117.3674 JPY (bid price). Then, with one EU one could buy 1.1148 USD, and with those US Dollars one could buy 1.1148×105.30 JPY. Finally, one could sell those JPY and receive $1.1148 \times 105.30 / 117.3674 = 1.0002$ EU. Consequently, it would have been possible to obtain strictly positive gains with no cost.

In the literature, we can find several publications showing evidence of the existence of TA in the FX markets. Aiba et al. (2002) show the existence of TA among the USD, the EU, and the JPY. They propose a model including past effects of TA that satisfactorily fits the exchange rate movements. Regarding the duration and size of a TA, Fenn et al. (2009) and Ito et al. (2012) found that it occurs in less than one second and in very small magnitudes. Moreover, they showed that opportunities for TA have dramatically declined in the first decade of this century. More precisely, Gradojevic et al. (2019) found that on average 80–100 short duration (100–500 millisecond) arbitrage opportunities exist on a daily basis for the EUR/USD, USD/JPY, and EUR/JPY exchange rates.

An immediate consequence of TA is the non-verification of the Efficient Market Hypothesis, which proposes that asset prices fully reflect all available information. However, as the duration is short and the size is small, we can argue that the TAs are small adjustments toward the equilibrium prices of the exchange rates. Another consequence is the negative value for the auto-correlation function of each exchange rate when TA exists, as found in Aiba et al. (2003). TA may suggest the existence of a correlation between foreign exchange rates. However, Aiba and Hatano (2004) showed that such a correlation may exist even without actual TA transactions. In Choi (2011), empirical tests showed that small profitable arbitrage episodes occurred

¹These are approximated values on August 23, 2019, used just for illustration.

over very short intervals of uncertainty and turbulence. This fact was also verified in Maldonado et al. (2020), exhibiting the emergence of TA in Greece's most acute crisis period. The last consequence of the existence of TA we should mention is that pointed out by Cross and Kozyakin (2013), where it is found that in a 5-currency world, arbitrage sequences follow an exponential law and display periodicity; however, in higher order currency worlds, a double exponential law may emerge, increasing the instability of the efficiency market hypothesis.

As one can deduce, the TA is a consequence of a lack of or lagged information. For example, Lyons and Moore (2009) modeled the dynamics of exchange rates depending on the trade size that conveys information and showed that, in this situation, the TA is not significant. In a more descriptive work, Schaumburg (2014) observed the increasing incidence of algorithmic and high-frequency trading on the FX market. As a consequence, the rapid dissemination of information reduced TA opportunities. Parallel conclusions were found by Chaboud et al. (2014).

Finally, it is worth mentioning the methodologies with which to detect the existence of TA. In Bjønnes and Longarela (2014), a methodology was proposed to detect TA of any order based on a simple linear program. Using the Electronic Broking Services (EBS) platform, they found several short-lived arbitrage events involving up to five currencies. A theoretical and computational methodology based on the Perron-Frobenius theorem was provided by Cui et al. (2018) to detect and identify the presence of TA in the FX market. Using elements of compositional data analysis, Maldonado et al. (2020) defined a geometry of the space of the matrices of exchange rates. With such geometry, a distance to the no-arbitrage of exchange rate matrices subspace was defined and used here to detect the presence of TA.

This paper has four sections, including this Introduction. In Sect. 2, we define the primary concepts and resume the main results found in Maldonado et al. (2020) which we use in later sections. In Sect. 3, we present the main contribution of this work: two applications of the modeling of exchange rate matrices and the corresponding no-arbitrage matrices. The first one illustrates the closeness between the single eigenvector of the projection of a matrix of exchange rates onto the no-arbitrage subspace and the Special Drawing Rights exchange rates. As we will argue later, this is related to the negligible TA in a group of currencies. The second application is the analysis of the dynamics of the exchange rate bubbles in a group of countries. To this end, we firstly define the fundamental value of an exchange rate as that corresponding to the purchasing power parity between two countries; then, the bubble size is the deviation of the spot exchange rate from its fundamental. We develop a technique to calculate the levels of the fundamental values by minimizing the total size of the bubbles in the group. Finally, in Sect. 4 we discuss the main conclusions of this work, and in Appendix we prove the proposition given in Sect. 3.2.

2 Preliminaries

Since most of the analysis in the following sections uses the framework and results found in Maldonado et al. (2020), in this section we summarize the main findings and characterizations described in that work. The proofs of the propositions below can be found there. Consider a group of $N \geq 2$ countries indexed by $i = 1, 2, \dots, N$, and let $a_{ij} > 0$ be the exchange rate between the currencies in countries i and j , that is, to buy one unit of the currency of country j , an individual must pay a_{ij} units of the currency of the country i . It is clear that $a_{ii} = 1$ must be met for any $i = 1, 2, \dots, N$. Therefore, $A = [a_{ij}]$ is a *matrix of exchange rates* (MER) if all its entries are strictly positive and the diagonal is filled with 1s. We denote the set of matrices of exchange rates by \mathcal{E} .

Among the matrices of exchange rates, we have particular interest in those precluding TA. Thus, we have the following definition:

Definition 1 The matrix $A = [a_{ij}] \in \mathcal{E}$ is a *no-arbitrage matrix of exchange rates* (NAMER) if for any $k \geq 2$ and $i_1, i_2, \dots, i_k \in \{1, \dots, N\}$ the following condition is satisfied:

$$a_{i_1 i_2} a_{i_2 i_3} \cdots a_{i_{k-1} i_k} a_{i_k i_1} = 1. \tag{1}$$

If condition (1) is not satisfied (for example, if the left side is greater than 1) with one unit of the currency of i_1 buying consecutively the currencies in $i_k, i_{k-1}, \dots, i_2, i_1$, we will obtain more than one unit of the i_1 currency; this is known as *triangular arbitrage*. The set of no-arbitrage (in fact, no-triangular-arbitrage) matrices is denoted by \mathcal{E}' .

For any $\mathbf{u} \in \mathbb{R}_+^N$, the matrix $\mathbf{u}(\mathbf{u}^{-1})^\top$ satisfies (1) and therefore it is in \mathcal{E}' . In Maldonado et al. (2020), it is proven that any element in \mathcal{E}' has that form, as claimed below.

Proposition 1 *The set of no-arbitrage matrices of exchange rates is characterized by*

$$\mathcal{E}' = \{A \in \mathbb{R}_+^{N \times N} : A = \mathbf{u}(\mathbf{u}^{-1})^\top; \mathbf{u} \in \mathbb{R}_+^N\}. \tag{2}$$

Therefore, if $A = \mathbf{u}(\mathbf{u}^{-1})^\top \in \mathcal{E}'$, then it is easy to check that it has only two eigenvalues: N (associated with \mathbf{u}) with multiplicity one and 0 with multiplicity $N - 1$. Reciprocally, if $\mathbf{u} \in \mathbb{R}_+^N$ is an eigenvector of $A \in \mathcal{E}'$ associated with the eigenvalue N , then $A = \mathbf{u}(\mathbf{u}^{-1})^\top$. The vector $\mathbf{u} \in \mathbb{R}_+^N$ may be multiplied by any positive real number and the result $\mathbf{u}(\mathbf{u}^{-1})^\top$ remains unchanged. This means that it is a compositional vector (see Pawlowsky-Glahn et al. (2015), Egozcue and Pawlowsky-Glahn (2019)) and, therefore, the information it contains is relative. This fact is used to study the time evolution of NAMERs as a compositional time series.

Another remarkable finding in Maldonado et al. (2020) is the Euclidean space structure that can be defined in the set of exchange rate matrices: given $A = [a_{ij}]$, $B = [b_{ij}] \in \mathcal{E}$ and $r \in \mathbb{R}$, we can define addition, scalar multiplication and inner product by

$$A \oplus B = [a_{ij}b_{ij}], \quad r \odot A = [a_{ij}^r] \quad \text{and} \quad \langle A, B \rangle = \sum_{i,j} (\ln a_{ij})(\ln b_{ij}). \quad (3)$$

Thus, with the operations given in (3), \mathcal{E} is a Euclidean vector space, with the norm associated to the inner product, $\|A\| = \langle A, A \rangle^{1/2}$, and \mathcal{E}' is a vector subspace of the same. This Euclidean structure of $\mathbb{R}_+^{N \times N}$ is compatible with the Aitchison geometry of the simplex applied to MERs and NAMERS as proven in Pawlowsky-Glahn et al. (2015).

The importance of the Euclidean structure defined by (3) is that it allows us to define a distance in the space \mathcal{E} and, in particular, the projection of elements in \mathcal{E} onto the subspace \mathcal{E}' . The next proposition provides a formula to this end.

Proposition 2 *Let $A = [a_{ij}] \in \mathcal{E}$. The projection of A onto the vector subspace \mathcal{E}' is $A^* = [a_{ij}^*] \in \mathcal{E}'$ given by*

$$a_{ij}^* = \left(\frac{g_m(A_i)g_m(A^j)}{g_m(A^i)g_m(A_j)} \right)^{1/2}, \quad i, j = 1, \dots, N, \quad (4)$$

where A_i (A^j) represents the row i (column j) of the matrix A , and $g_m(v)$ is the geometric mean of the components of a vector $v \in \mathbb{R}_+^N$. As a consequence, if we define $u_i = (g_m(A_i)/g_m(A^i))^{1/2}$, we can write $A = \mathbf{u}(\mathbf{u}^{-1})^\top$.

Summing up, the set of matrices of exchange rates \mathcal{E} is a Euclidean vector space with the operations given in (3). The set of matrices of exchange rates that do not allow for TA \mathcal{E}' is a subspace of \mathcal{E} , and its elements are characterized by (2). Finally, the closest element in \mathcal{E}' to $A \in \mathcal{E}$ has its components given by (4). Here, closest refers to closest in the sense of the distance of the space \mathcal{E} .

3 Application to a Group of Countries

In this section, we present two applications of the framework and results presented in Sect. 2 for a group of countries. The first one is the empirical verification of the proximity of the eigenvector of the projection of an MER onto the NAMER subspace to the exchange rates of the currencies with the Special Drawing Rights (SDR). The SDR is an asset created by the International Monetary Fund to provide liquidity to the member countries. The second application is the analysis of the dynamics of the exchange rate bubbles in the group. This analysis allows us to describe the deviations from the fundamentals of the exchange rate of each pair of countries.

We consider the following currencies: the Brazilian Real (BRL), the Euro (EU), the Pound Sterling from the United Kingdom (GBP), and the US Dollar (USD). From now on, we refer to that group of countries as BEGU. All the currencies are negotiated

in the Central Banks of the respective country/region. The data was collected from the Central Bank of Brazil, the European Central Bank, the Bank of England, and the Federal Reserve System from August 1, 2011, to December 29, 2017, on a daily basis. We consider only the common days where all the currencies were actually traded and use the closing ask value of the exchange rates. As such, this and the time zone differences may produce some deviations from the no-arbitrage values.

In Maldonado et al. (2020), it was shown that the matrices of exchange rates of those countries in that period were close to their corresponding projections onto the no-arbitrage subspace of matrices of exchange rates. The relative errors or deviations with respect to the no-arbitrage composition is around 0.5%² and, as asserted before, may be a consequence of different time zones for the closure of trade and/or the fact of using only the ask price of the exchange rates. The exception was in a sub-period in the data, which corresponds to the most severe part of the Greek crisis, as explained in their work.

In the following subsections, we present two applications of the projections onto the subspace of the no-arbitrage matrices of exchange rates given in Sect. 2.

3.1 No-Arbitrage Matrix of Exchange Rates and SDR

As stated in Sect. 2, any NAMER can be written as $\mathbf{u}(\mathbf{u}^{-1})^\top$, where $\mathbf{u} \in \mathbb{R}_+^N$ is the eigenvector of that matrix associated with the eigenvalue N (see Proposition 1 and subsequent paragraph). Therefore, if $A^* = [a_{ij}^*]$ is the projection of an MER $A = [a_{ij}]$ onto the subspace \mathcal{E}' , then $a_{ij}^* = u_i u_j^{-1}$, for all i and j , where $\mathbf{u} \in \mathbb{R}_+^N$ is given in Proposition 2. Moreover, it is the eigenvector of A^* that is associated with the eigenvalue N .

In this way, the eigenvector \mathbf{u} (or its opposite composition \mathbf{u}^{-1}) of the MER projection onto the NAMERs subspace shows the relative exchange rates between the countries when TA is precluded. For example, the first row of A^* is $u_1(u_1^{-1}, u_2^{-1}, \dots, u_N^{-1}) = u_1^{-1}\mathbf{u}^{-1}$. If we normalize to $u_1 = 1$, it results that \mathbf{u}^{-1} provides the exchange rates without TA of country 1 with respect to all other countries. Reciprocally, \mathbf{u} provides the exchange rates without TA of all the countries with respect to that of country 1. Notice that the compositions \mathbf{u} and \mathbf{u}^{-1} can be normalized arbitrarily. A typical normalization consists of dividing by the sum of all components so that all normalized components add to one, as in Fig. 1. Accordingly, the time series of those eigenvectors describe the evolution of the NAMERs in the group. As shown in Fig. 1, the component u_1 has a sharp increase, whereas u_4 exhibits a slight decrease; therefore $a_{14}^* = u_1 u_4^{-1}$ has an intensive increment. Since a_{14}^* represents the exchange rate of the BRL (country 1) with respect to the USD (country 4) when TA is precluded, that intensive increment shows the strong devaluation of the BRL against the USD.

²They defined the relative error as $\|A \ominus A^*\|/\|A\|$, where $\|\cdot\|$ is the norm induced by the inner product defined in (3).

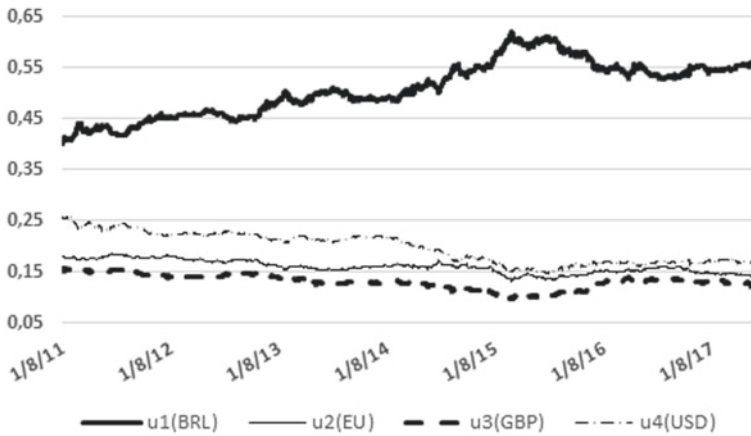


Fig. 1 Evolution of the eigenvector components (normalized for summing to one)

Likewise, we can observe the evolution of the exchange rates free of TA in the other countries.

In this subsection, we analyze the close relationship between the eigenvectors of the NAMERs projections and the Special Drawing Rights (SDR). According to the web page of the International Monetary Fund (IMF) (www.imf.org, 2019.08.29), the SDR is an international reserve asset created by the IMF in 1969 to supplement its member countries’ official reserves. So far SDR 204.2 billion (equivalent to about USD 291 billion) have been allocated to members, including SDR 182.6 billion allocated in 2009 in the wake of the global financial crisis. The value of the SDR was initially defined as equivalent to 0.8886 grams of fine gold (about one US Dollar). However, after the collapse of the Bretton Woods System in 1973, its value was defined through a basket of currencies. As of October 1, 2016, the SDR basket consists of the US Dollar, Euro, Chinese Renminbi, Japanese Yen, and Pound Sterling, quoted at noon each day on the London market. The weights used for each currency reflect their relative importance in the world’s trading and financial systems, and they are reviewed either every 5 years or if the circumstances warrant an earlier review. The current weights are 41.73% for USD, 30.93% for EUR, 8.33% for Chinese Renminbi, 8.09% for Japanese yen, and 10.92% for GBP.

To obtain the exchange rate of SDRs with respect to the currencies of other countries, first the IMF computes the exchange rate with respect to the US Dollar using the weights given above. Then, it uses the USD exchange rate with respect to the other countries to find the SDR exchange rate with respect to the local currency. As the exchange rates are close to those corresponding to no-arbitrage values (except for periods of liquidity shortage of a currency), the SDRs exchange rates must be close to the inverse eigenvector \mathbf{u}^{-1} of the projections onto the NAMER subspace. Figure 2 shows the time series of both compositional vectors, normalized to sum up to one.

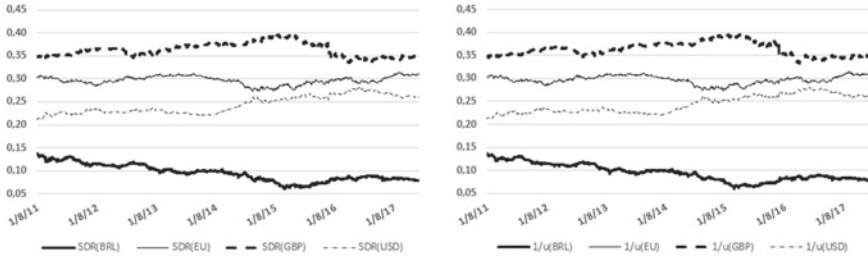


Fig. 2 Illustration of the similarity between the time series of normalized SDRs (left) and the normalized vectors \mathbf{u}^{-1} (right). Horizontal-axis: time. Vertical-axis: components of normalized SDRs and \mathbf{u}^{-1} proportions

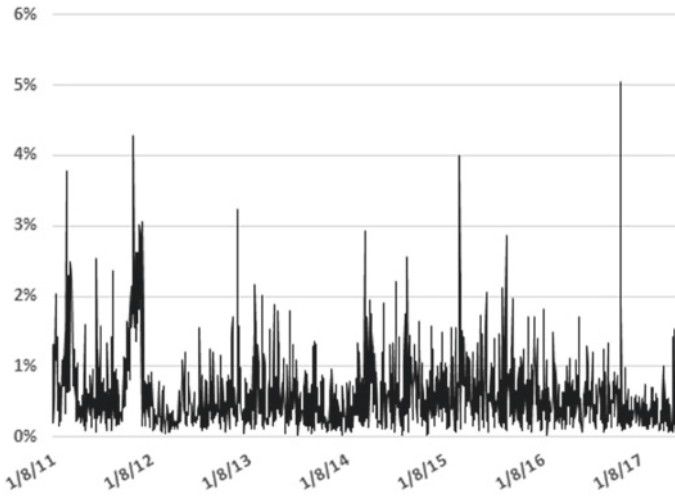


Fig. 3 Relative Aitchison distances Eq. (5) between SDRs and the vectors \mathbf{u}^{-1} expressed in percentages

We use the Aitchison distance (Aitchison 1983; Pawłowsky-Glahn et al. 2015) to measure the relative distance between the compositional vectors SDR and \mathbf{u}^{-1} . The relative distance is

$$\frac{d_a(\text{SDR}, \mathbf{u}^{-1})}{d_a(\text{SDR}, \mathbf{1})} = \frac{d_a(\text{SDR}, \mathbf{u}^{-1})}{\|\text{SDR}\|}, \tag{5}$$

where $d_a^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^N (\ln(x_i/g_m(\mathbf{x})) - \ln(y_i/g_m(\mathbf{y})))^2$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^N$, and $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}_+^N$. Figure 3 shows the time series of the relative distances between SDRs and \mathbf{u}^{-1} . They vary between 0 and 5%, with the quartiles being $Q_1 = 0.25\%$, $Q_2 = 0.46\%$, and $Q_3 = 0.77\%$, indicating that the vector \mathbf{u}^{-1} can be used as a proxy to the SDRs.

To justify the proximity between the SDR exchange rates and the vector \mathbf{u}^{-1} , recall the empirical result found in Maldonado et al. (2020). There it is shown that, with exception of the period May–June 2012, the exchange rates between the currencies in the BEGU group are close to those without TA, namely, $a_{ij} = a_{ij}^* = u_i u_j^{-1}$. On the other hand, by definition, one unit of the SDRs is computed as a geometric mean of the most important currencies in the world, expressed in USD. Let G be the value of 1 SDR in US Dollars. To obtain the value of 1 SDR in units of currency i , we have to multiply a_{i4} by G , so 1 SDR amounts to $G a_{i4}$ units of currency i . Therefore, using the result of Maldonado et al. (2020) mentioned above, we can approximately write $G a_{i4} = G u_i u_4^{-1}$ as the number of units of currency i that amounts to 1 SDR. Then, $(G u_i u_4^{-1})^{-1} = G^{-1} u_4 u_i^{-1}$ units of SDRs amount to one unit of currency i . Writing this in vector form, $G^{-1} u_4 \mathbf{u}^{-1}$ (or equivalently \mathbf{u}^{-1}) contains the exchange rates of the SDR with respect to all other currencies.

It is worth noting that, as in the case of the relative errors of the MERs, the relative distance oscillates around 0.5%, except in the same period of May–June 2012 as explained in the introduction of this section. That exception happened because during the Greek crisis, the exchange rates gradually differed more and more from their values without TA, until the crisis was overcome, as shown in Maldonado et al. (2020). Therefore, in normal periods, the SDRs can be defined using the inverse of the components of the NAMER eigenvector.

3.2 Exchange Rate Bubbles Using the NAMERs

Maldonado et al. (2020) verified that the matrices $A(t)$ of exchange rates in the Brazil, European Union, United Kingdom, and United States of America group are close to the no-arbitrage matrices of exchange rates. Their projections $A^*(t)$ onto the NAMER subspace are characterized by the eigenvectors time series in the 3–dimensional simplex (also called the 4–part simplex) $\mathbf{u}(t) \in \mathbb{S}^{(4-1)}$.³ Since the matrices $A^*(t)$ represent the exchange rates free of potential TA, we use the eigenvectors that define those matrices to estimate the level of the fundamental exchange rate in each country with respect to another.

The fundamental exchange rate between the currency of one country with respect to another can be defined in different ways, depending on the structural model that is considered for the exchange rate. Using the prices of tradable commodities, the interest rates, and the money supply defined by the central bank authority in each country, we can define different structural models and thus different fundamental exchange rates. In Van Norden (1996), a regime switching model was used to test the existence of bubbles in the exchange rates of Canada, Germany, and Japan with respect to the US Dollar. In that work, three fundamental values were used for the exchange rate: the Purchasing Power Parity (PPP) value, the uncovered interest parity value, and a value coming from a monetary model based on sticky prices.

³ $\mathbb{S}^{(4-1)} = \{(u_1, u_2, u_3, u_4) \in \mathbb{R}_+; \sum_{i=1}^4 u_i = 1\}$.

In Maldonado et al. (2012), three fundamental values (analogous to those of Van Norden) were defined for modeling the exchange rate bubble dynamics in Brazil. They proposed a methodology to find the level of the fundamental value, as well as to test the rational expectations hypothesis in the exchange rate future market. In a later study, Maldonado et al. (2016) analyzed the existence of cointegration among the exchange rate bubbles in the BRICS (Brazil, Russia, India, China, and South Africa) group countries. Finally, Hu and Oxley (2017) test the existence of bubbles in the exchange rates of BRICS and some Asian and G10 countries. All these studies analyzed the bubble dynamics of the exchange rates of local currencies with respect to the US Dollar. In this work, however, we propose analyzing such dynamics by considering all the exchange rates between the BEGU country group currencies.

Finding the level of the fundamental value for the exchange rate of country i with respect to country j is essential for measuring the exchange rate bubble size of i with respect to j . In doing that, we have to take into account that the bubble size of country i with respect to country j is the inverse of the bubble size of country j with respect to country i . Thus, the proposal here is the joint determination of the fundamental values in all countries, such that the bubble sizes (deviations of the exchange rates from their fundamentals) of all of them are minimized.

Recall that a_{ij} is the number of monetary units of country i needed to buy one monetary unit of the currency of country j , and consider the fundamental value for the exchange rate to be the one given by the PPP value. Thus, if p_i and p_j are the prices of the tradable commodity baskets in countries i and j , respectively, then the fundamental value of the exchange rate between countries i and j is $a_{ij}^f = p_i/p_j$. That fundamental value equalizes the price of a (representative) tradable good in countries i and j and the Law of One Price proposes that it should be the long-run value for the exchange rate between the currencies in both countries. These values can be arranged in a matrix A^f . In order to disregard the eventual presence of TA, and since the MERs are very close to the corresponding NAMER, we will use the corresponding projection a_{ij}^* rather than a_{ij} . In monetary economics, the bubble in the exchange rate is any deviation from its fundamental value. If the spot exchange rate is greater (lower) than its fundamental value, then the currency i is undervalued (overvalued) with respect to the currency in j . Thus, the bubble size in the exchange rate can be defined as the difference between its spot value and the corresponding fundamental value. Since we are working with compositional vectors and non-negative values, we choose to define the bubble size in the exchange rate of country i with respect to country j as $b_{ij} = a_{ij}^*/a_{ij}^f$. Therefore, this bubble size reports the gross rate of undervaluation (or overvaluation) of currency i with respect to currency j .

Definition 2 For each $t = 1, \dots, T$, let

$$A^*(t) = [a_{ij}^*(t)] = \mathbf{u}(t)(\mathbf{u}^{-1}(t))^\top \in \mathcal{E}' \subset \mathbb{R}_+^{N \times N}$$

be the NAMER projection of $A(t) \in \mathcal{E} \subset \mathbb{R}_+^{N \times N}$, and let $p_i(t)$, $p_j(t)$ be the prices of the tradable commodity baskets in countries i and j at time t , respectively. The

bubble matrix $B(t) \in \mathbb{R}_+^{N \times N}$ is

$$B(t) = [b_{ij}(t)] = A^*(t) \ominus A^f(t) = \begin{bmatrix} a_{ij}^* \\ a_{ij}^f \end{bmatrix} = \begin{bmatrix} u_i(t)p_j(t) \\ u_j(t)p_i(t) \end{bmatrix}, \tag{6}$$

which is also an element of \mathcal{E}' for any t . It is worth noting that the bubble matrix contains all the information regarding undervaluation and overvaluation among the currencies in the group of countries.

To measure the fundamental PPP value of the exchange rate, the use of the Wholesale Price Index (WPI) is recommended by Terra and Vahia (2008). For Brazil, the United Kingdom, and the United States of America, it is available in the corresponding Central Bank sites. However, for the European Union, the only time series data available is the Harmonized Index of Consumer Prices (HICP), so we will use those data. Since the price of the commodities basket is a multiple of the index, we will have that $p_i(t) = k_i \text{WPI}_i(t)$ where $\mathbf{k} = [k_1 \cdots k_N]^\top$ is a vector that we have to estimate. In terms of \mathbf{k} , the bubble matrix defined in (6) is given as

$$B(t; \mathbf{k}) = \begin{bmatrix} u_i(t)k_j \text{WPI}_j(t) \\ u_j(t)k_i \text{WPI}_i(t) \end{bmatrix} = \mathbf{k}^{-1} \mathbf{x}(t) ((\mathbf{k}^{-1} \mathbf{x}(t))^{-1})^\top, \tag{7}$$

where $\mathbf{x} = [x_i(t)] = [u_i(t)\text{WPI}_i(t)]$.

It is important to notice that the exchange rate bubble size of country i with respect to country j is

$$b_{ij}(t) = k_i^{-1} k_j x_i(t) (x_j(t))^{-1}.$$

As a consequence, decreasing the value of k_i reduces the bubble size of i with respect to j , but at the same time, the bubble size of all the other countries with respect to i increases. Therefore, we propose estimating the vector $\mathbf{k} \in \mathbb{R}_+^N$ as being the one that minimizes the joint bubble sizes in the whole group, namely, the minimum of the quadratic sum of the bubble matrix sizes.

Proposition 3 *The components of the vector $\mathbf{k}^* = [k_1^* \cdots k_N^*] \in \mathbb{R}_+^N$ which minimize*

$$\sum_{t=1}^N \|B(t; \mathbf{k})\|^2,$$

are

$$k_i^* = g_m(x_i(1), \dots, x_i(T)),$$

where $g_m(\mathbf{v})$ is the geometric mean of the components of \mathbf{v} .

To compute bubble sizes in the exchange rates of the BEGU group countries, we use the exchange rate values from the last business day of each month, because



Fig. 4 Exchange rate bubble dynamics with respect to the USD in %. Horizontal-axis: time. Vertical-axis: ratio to the USD

the price indexes are measured on the same day. Then, we compute the vector \mathbf{u} , which defines the NAMER. The levels of the fundamental values for each country of the group, the vector \mathbf{k}^* , are estimated using Proposition 3. As a result, we find $a_{ij}^f = p_i / p_j = k_i^* k_j^{*-1} \text{WPI}_i \text{WPI}_j^{-1}$, the fundamental exchange rate values of each country with respect to all other countries, as well as the exchange rate bubble dynamics. In Fig. 4, we show the evolution of the exchange rate bubbles of all the countries with respect to the US Dollar along the analyzed period. In that figure we are plotting the time series bubbles $b_{i,4}(t) = a_{i4}^*(t) / a_{i4}^f(t)$, for $i = 1, 2, 3$, namely, BRL, EU, and GBP, respectively. Level “1” (or close to it) of the bubble represents the no existence of it ($a_{i4}^* = a_{i4}^f$); it is the level at which the long-run behavior of the bubble sizes should converge, according to the PPP principle. Analogously, a value lower (greater) than 1 means that $a_{i4}^* < a_{i4}^f$ ($a_{i4}^* > a_{i4}^f$), then, the exchange rate of the currency i with respect to the USD is overvalued (undervalued) with respect to its fundamental value.

At this point, it is important to discuss the reasons for the appearance of exchange rate bubbles. The deviations of an exchange rate from its fundamental value are strongly related to the capital flight and/or to the excess of optimism regarding the financial perspective of a country. The pessimism triggers capital flight, thus exerting pressure over the demand for foreign currency, and this produces the undervaluation of the domestic currency. When the optimism of the financial market returns, investments in foreign currency arrive and the local currency recovers its value. That said, let us proceed to analyze the dynamics of bubbles in exchange rates in the group.

The Great Britain Pound keeps its spot value very close to its fundamental value until the middle of 2016, when the Brexit referendum triggers a continuous devaluation with respect to the US Dollar. For the Brazilian Real, we observe an overvalued currency up to the end of 2014. The perception of balanced growth and not too

deteriorated public accounts allowed for that overvaluation that then progressively diminishes. In fact, the overvaluation of BRL currency with respect to the others in the group experienced a slow decrease at the end of 2014. From the beginning of 2015 until the end of that year, the bubble size increased vigorously, leading the spot exchange rate to attain more than 140% of its fundamental value. Three events fed the rise of the bubble that year: (i) the deterioration of the public accounts (Holland 2019); (ii) the corruption cases deflagrated in different spheres of the current and former government⁴; and (iii) the uncertainty regarding the consolidation of the impeachment process of the president Dilma Rousseff (Nunes and Melo 2017). After August 2016 the bubble burst, following the market perception of the social security reforms approval. However, from the beginning of 2017, we observe a bubble growth following a global tendency of currency devaluations in emerging markets. Regarding the Euro (EU) and the Pound Sterling (GBP), the dynamics of the almost nonexistent bubble size with respect to the US Dollar are quite similar to that of the end of 2014. The devaluation of the Euro, and the consequent bubble rise from 2015 on, was correlated to three events: the likelihood of an increase in US interest rates from 0.2% at the beginning of 2016 to more than 2% in 2019⁵, the deepening of the crisis in Greece (Amini 2015), and the effect of the European Central Bank's quantitative easing programme, buying bonds from local banks, and creating in this way currency liquidity in the bank system (Claeys and Leandro 2016). However, prices were adjusted by inflation, so the bubble with respect to the US Dollar started to burst at the beginning of 2016. All those episodes meant the international investor preferred American investments, thus increasing the Euro exchange rate with respect to the US Dollar.

It is evident that the set of C-bubble (Compositional-bubble) sizes⁶ $B(t; \mathbf{k})$ are NAMERs and, therefore, they can be represented by the eigenvector $\mathbf{v}(t)$ such that $B(t; \mathbf{k}) = \mathbf{v}(\mathbf{v}^{-1})^\top$; note that, for notational simplicity, dependence on time has been dropped. The vector $\mathbf{v} \equiv \mathbf{v}(t)$ is a 4-part composition and the corresponding exploratory techniques can be used to show the evolution in time beyond the curves shown in Fig. 4. The compositional principal component analysis (Aitchison 1983) and the corresponding biplots (Aitchison and Greenacre 2002; Pawlowsky-Glahn et al. 2015, Chap. 5) provide a graphical tool to visualize the evolution of bubbles. Figure 5 shows the form biplots of the sample \mathbf{v} 's along time, both in the first and second (left panel) and first and third (right panel) principal components. The origin of rays is the center (compositional mean) of the C-bubbles along time; it can be considered as a null C-bubble computed using Proposition 3. The 3 axes follow the principal components and account for the whole variation of the sample which is 3-dimensional. Each point represents a value of $\mathbf{v}(t)$ in isometric log-ratio coordinates (Egozcue et al. 2003; Pawlowsky-Glahn et al. 2015) such that the distances to the origin are proportional to the size of the bubble; the distances from one point to

⁴<https://www.britannica.com/event/Petrobras-scandal>.

⁵<https://www.macrotrends.net/2015/fed-funds-rate-historical-chart>.

⁶C-bubble sizes is an alternative name for the bubble matrix defined in (6), where we highlight the compositional nature of that object.

another are their Aitchison distances, meaning that two nearby points stand for small deviations. The rays from the origin, and labeled with BRL, EU, GBP, and USD, are the projections of unitary vectors pointing to the directions of the increasing bubble of the currency, relative to all other countries. In Fig. 5, the unitary vectors representing each variable are projected on a plane. Then, a short ray, e.g. EU (Euro), in the first and second axes projection, means that the vector is quite orthogonal to that projection plane producing a visual reduction of the vector length. To better understand Fig. 5, the C-bubbles have been colored by year. In the right panel, we observe that from 2011 to 2015 there is a linear trend along the first axis meaning a continuous increase of the C-bubble mainly caused by overpricing BRL, relative to USD and GBP. In 2016 and 2017, this trend changes its direction dramatically with the role of the GBP, which suddenly increases its bubble with respect to the other currencies, especially with respect to the BRL which has a decreasing bubble in these years (see also Fig. 4). Additionally, Fig. 5 shows a noticeable continuity in the evolution of the C-bubble, well described by the low variability of the third axis (right panel) which is dominated by the contrast between EU and GBP.

4 Conclusions

In this work, we present two applications of the theory developed in Maldonado et al. (2020) to analyze the existence and the measurement of TA in the exchange rates of a group of countries. Using data on a daily basis from the foreign exchange (FX) markets of Brazil, the European Union, United Kingdom, and United States of America (here called the BEGU group) for the period August 2011 to December 2017, we perform the following empirical analyses.

First, we find the proximity between the eigenvector of the matrix of exchange rates projection onto the no-arbitrage subspace and the exchange rates of the Special Drawing Rights (SDR), an asset created by the International Monetary Fund to provide liquidity to the member countries with respect to the currencies of the countries in the BEGU group. The exchange rate between one SDR and the USD is determined through a basket of currencies where the weights are defined by the relative importance of the country in the world's trading and financial systems. Using the computed exchange rate of the SDR with respect to the USD, its exchange rates with respect to the other currencies in the BEGU group are computed multiplying their spot exchange rates with respect to the USD (the last column of the matrix of exchange rates of the BEGU's currencies). Since the TA between the BEGU's currencies is almost negligible, all the columns in the exchange rate matrix of the group are nearly proportional between them and close to the single eigenvector of the projection matrix onto the NAMER subspace. Thus, the SDR exchange rates with respect to the BEGU's currencies are also close to that eigenvector. Hence, we are able to empirically check the close proportionality between the SDR's exchange rates and the eigenvector of the exchange rates matrix of the group of countries.

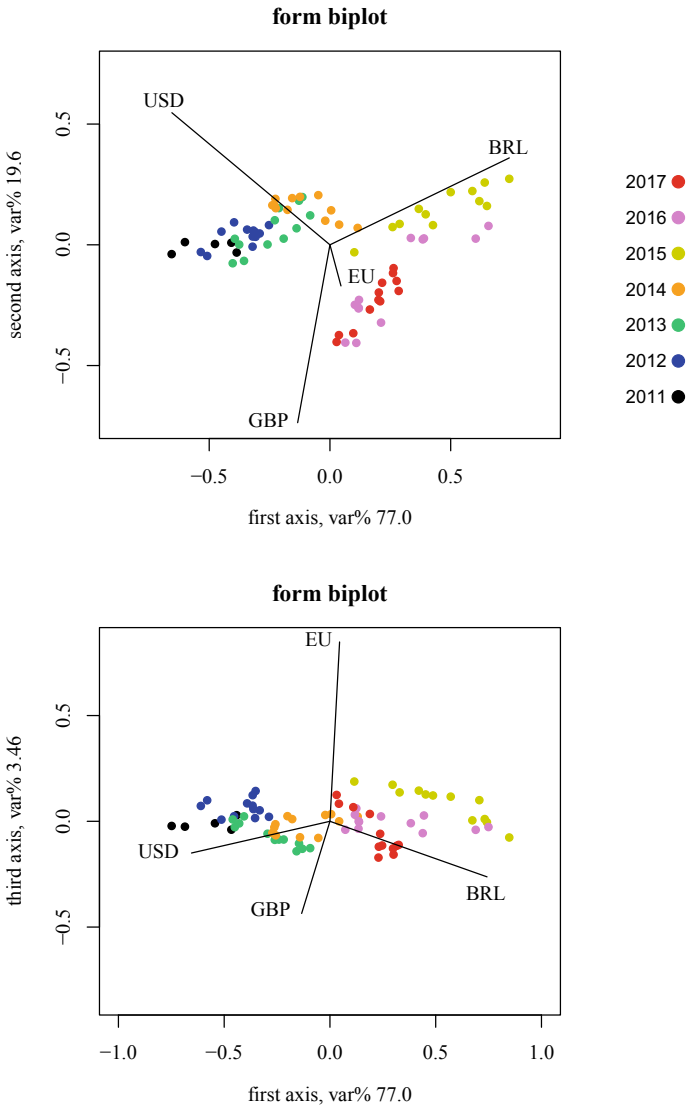


Fig. 5 Compositional form biplot of bubble eigenvectors (C-bubbles). Upper panel: first and second principal components; Lower panel: first and third principal components. Points are bidimensional projections of $\mathbf{v}(t)$, and their interdistances are optimal representations of the Aitchison distances. Colors indicate years. Rays (BRL, EU, GBP, and USD) are projections of unitary vectors in the directions representing the increasing bubbles of a particular currency with respect to other currencies

Second, using the exchange rates free of TA, we compute the level of the fundamental exchange rate values, considering as the fundamental value the one corresponding to the Purchasing Power Parity. Those levels are calculated in such a way that the distance between the spot exchange rate and its fundamental value is minimized for the whole group (according to the Law of One Price). With those levels in mind, we simultaneously analyze the exchange rate bubble dynamics in the whole group. This allows us to take into account that the bubble size in currency i with respect to currency j is the inverse of the bubble size in currency j with respect to currency i , because in minimizing the bubble size of one of them, the bubble size of the other increases. The movements of the bubble sizes in each country in the group correspond to economic and financial events (crises or liquidity problems) that occurred in the countries in the group during the analyzed period, as discussed in Sect. 3.

Some open issues arise from the analysis in this work. First, it was shown that, in general, the MER is close to the corresponding NAMER projection (except for periods of strong liquidity shortage). However, the NAMER and its characteristic eigenvector significantly vary over time. The inclusion of economic variables like prices, interest rates, Gross Domestic Product, imports/exports, and short-run debts in the explanation of the dynamics of the NAMER projection eigenvector could shed light on how the exchange rates in a group of countries may vary in the short run. This is a very important issue for financial economics. Also, the analysis performed here assumed that the bid-ask spread of prices in the FX market trading is zero. A more accurate treatment would consist of considering that there is a difference in buying and selling prices and, consequently, reformulating the definition and characterizations of no-arbitrage. Another interesting issue is the analysis of the possible influence of TA on the future values of the exchange rates. This has been proposed and tested by Gradojevic et al. (2019) and, if theoretically modeled, it could establish a connection between TA and inter-temporal arbitrage. Finally, despite the fact that the analysis proposed here is based on the FX markets, other markets (financial or non-financial) can be considered in order to capture the inefficiencies or the transaction costs that can produce TA.

Acknowledgements Wilfredo L. Maldonado thanks the CNPq of Brazil for financial support 306473/2018-6, as well as the FAPDF. Vera Pawlowsky-Glahn and Juan José Egozcue received financial support through the project METHods for COMpositional analysis of DATA (CODAMET), Ministerio de Ciencia, Innovación y Universidades (Ref: RTI2018-095518-B-C21, 2019-2021).

Appendix

Proposition 3. *The components of the vector $\mathbf{k}^* = [k_1^* \cdots k_N^*] \in \mathbb{R}_+^N$ which minimize*

$$\sum_{t=1}^N \|B(t; \mathbf{k})\|^2,$$

are

$$k_i^* = g_m(x_i(1), \dots, x_i(T)),$$

where $g_m(\mathbf{v})$ is the geometric mean of the components of \mathbf{v} .

Proof The (i, j) -component of the bubble matrix is

$$b_{ij}(t) = k_i^{-1} k_j x_i(t) x_j^{-1}(t)$$

and, taking logarithms,

$$\ln(b_{ij}(t)) = \ln(x_i(t)) - \ln(x_j(t)) - \ln(k_i) + \ln(k_j).$$

Consider $d_i(t) = \ln(x_i(t))$ and $z_i = \ln(k_i)$. The function to be minimized is

$$\sum_{t=1}^T \sum_{i,j} (d_i(t) - d_j(t) - z_i + z_j)^2.$$

Note that if (z_1, z_2, \dots, z_N) is a solution then, for any real number r , $(z_1 + r, z_2 + r, \dots, z_N + r)$ is a solution as well. Thus, we may suppose that $\sum_i z_i = 0$ and, after finding a solution, we may even add a constant if it leads to a simplification.

The first-order condition with respect to the component i_0 is

$$\sum_{t=1}^T \left[-2 \sum_j (d_{i_0}(t) - d_j(t) - z_{i_0} + z_j) + 2 \sum_i (d_i(t) - d_{i_0}(t) - z_i + z_{i_0}) \right] = \lambda,$$

where $\lambda \geq 0$ is a Lagrange multiplier. Rearranging the terms inside the brackets and using $\sum_i z_i = 0$ yields

$$4 \sum_{t=1}^T \left(-N d_{i_0}(t) + N z_{i_0} + \sum_i d_i(t) \right) = \lambda.$$

Summing up on i_0 and using $\sum_i z_i = 0$, we obtain $\lambda = 0$. Therefore,

$$z_{i_0} = \frac{\sum_{t=1}^T d_{i_0}(t)}{T} - \frac{1}{N} \sum_{t=1}^T \sum_{i=1}^N d_i(t).$$

If we add the amount $r = N^{-1} \sum_{t=1}^T \sum_{i=1}^N d_i(t)$ to each component of the vector (z_1, z_2, \dots, z_N) , we get

$$z_{i_0} = \frac{\sum_{t=1}^T d_{i_0}(t)}{T}.$$

Substituting the values of $d_i(t) = \ln(x_i(t))$ and $z_i = \ln(k_i)$, we complete the proof. \square

References

- Aiba, Y., & Hatano, N. (2004). Triangular arbitrage in the foreign exchange market. *Physica A*, 344, 174–177.
- Aiba, Y., Hatano, N., Takayasu, H., Marumo, K., & Shimizu, T. (2002). Triangular arbitrage as an interaction among foreign exchange rates. *Physica A*, 310, 467–479.
- Aiba, Y., Hatano, N., Takayasu, H., Marumo, K., & Shimizu, T. (2003). Triangular arbitrage and negative auto-correlation of foreign exchange rates. *Physica A*, 324, 253–257.
- Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70(1), 57–65.
- Aitchison, J., & Greenacre, M. (2002). Biplots for compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 51(4), 375–392.
- Amini, B. (2015). A chronology of the european sovereign debt crisis. *Socialism and Democracy*, 29(3), 166–178.
- Bjønnes, G., & Longarela, I. (2014). Arbitrage violations in currency markets. *Working Paper*.
- Chaboud, A., Chiquoine, B., Hjalmarsson, E., & Vega, C. (2014). Rise of the machines: Algorithmic trading in the foreign exchange market. *The Journal of Finance*, 69(5).
- Choi, M. S. (2011). Momentary exchange rate locked in a triangular mechanism of international currency. *Applied Economics*, 43(16), 2079–2087.
- Claeys, G., & Leandro, A. (2016). The European Central Bank's quantitative easing programme: Limits and risks. Technical report, Bruegel Policy Contribution, No. 2016/04.
- Cross, R., & Kozyakin, V. (2013). Double exponential instability of triangular arbitrage systems. *Discrete and Continuous Dynamical Systems Series B*, 18(2), 349–376.
- Cui, Z., Qian, W., Taylor, S., & Zhu, L. (2018). Detecting arbitrage in the foreign exchange market. *Stevens Institute of Technology School of Business Research Paper*.
- Egozcue, J. J., & Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure (with discussion). *TEST*, 28(3), 599–638. <https://doi.org/10.1007/s11749-019-00670-6>.
- Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figuera, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279–300.
- Fenn, D., Howison, S., MacDonald, M., Williams, S., & Johnson, N. (2009). The mirage of triangular arbitrage in the spot foreign exchange market. *International Journal of Theoretical and Applied Finance*, 12(8), 1105–1123.
- Gradojevic, N., Gençay, R., & Erdemlioglu, D. (2019). A new wavelet-based ultra-high-frequency analysis of triangular currency arbitrage. *Economic Modelling*, (forthcoming).
- Holland, M. (2019). Fiscal crisis in Brazil: Causes and remedy. *Brazilian Journal of Political Economy*, 39(1), 88–107.
- Hu, Y., & Oxley, L. (2017). Are there bubbles in exchange rates? some new evidence from G10 and emerging market economies. *Economic Modelling*, 64, 419–442.
- Ito, T., Yamada, K., Takayasu, M., Takayasu, H. (2012). Free lunch arbitrage opportunities in the foreign exchange markets. *NBER Working Paper 18541*.
- Lyons, R., & Moore, M. (2009). An information approach to international currencies. *Journal of International Economics*, 79(2), 211–221.

- Maldonado, W., Tourinho, O. A., & de Abreu, J. A. (2016). Cointegrated periodically collapsing bubbles in the exchange rate of BRICS. *Emerging Markets Finance and Trade*, 54, 54–70.
- Maldonado, W., Tourinho, O. A., & Valli, M. (2012). Exchange rate bubbles: Fundamental value estimation and rational expectations test. *Journal of International Money and Finance*, 31, 1033–1059.
- Maldonado, W. L., Egozcue, J. J., & Pawlowsky-Glahn, V. (2020). No-arbitrage matrices of exchange rates: Some characterizations. *International Journal of Economic Theory (forthcoming)*.
- Nunes, F., & Melo, C. R. (2017). Impeachment, political crisis and democracy in Brazil. *Revista de Ciência Política*, 37(2), 281–304.
- Pawlowsky-Glahn, V., Egozcue, J. J., & Lovell, D. (2015). Tools for compositional data with a total. *Statistical Modelling*, 15(2), 175–190.
- Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. *Statistics in practice* (p. 272). Chichester, UK: Wiley.
- Schaumburg, E. (2014). Has automated trading promoted efficiency in the fx spot market? *Federal Reserve Bank of New York*.
- Terra, M. C., & Vahia, A. L. (2008). A note on purchasing power parity and the choice of price index. *Revista Brasileira de Economia*, 62(1), 95–102.
- Van Norden, S. (1996). Regime switching as a test for exchange rate bubbles. *Journal of Applied Econometrics*, 11(3), 219–251.

Log-contrast and Orthonormal Log-ratio Coordinates for Compositional Data with a Total



Josep Antoni Martín-Fernández and Carles Barceló-Vidal

Abstract Compositional data require an appropriate statistical analysis because they provide the relative importance of the parts of a whole. Methods based on log-ratio coordinates give a consistent framework for analyzing this type of data. Any statistical model including variables created using the original parts should be formulated according to the geometry of the simplex. This geometry includes the log-contrast: a simple way to express a set of log-ratios in a linear form. Basic concepts and properties of log-ratios, log-contrasts, and orthonormal coordinates are revisited. In addition, we introduce an approach that includes both the log-ratio orthonormal coordinates and an auxiliary variable carrying absolute information. We illustrate the approach through the principal component analysis and discriminant analysis of real data sets.

1 Compositional Analysis and a Typical Linear Combination of Variables

When the components w_1, \dots, w_D of a real vector \mathbf{w} of \mathbb{R}_+^D (the strictly positive octant of \mathbb{R}^D) describe the disjoint *parts of a whole*, one says that \mathbf{w} is a *D-part composition* (Aitchison 1986). The generic term *compositional data* (CoDa) refers to a data set consisting of compositions. Sometimes, the original units of the components of \mathbf{w} represent absolute magnitudes (e.g., grams, euros, and liters); in other situations, they represent relative magnitudes (e.g., proportions, percentages, ppm, and mg/L). When the parts w_1, \dots, w_D constitute an exhaustive partition of a whole, one assumes that one is dealing with *full* compositions where the absolute total amount for the realizations (*samples*) can be a constant K , equal to $K = w_1 + \dots + w_D$, or can be different throughout the samples. In the case of a fixed total, the composition is only characterized by the *relative* information because the absolute values of the

J. A. Martín-Fernández (✉) · C. Barceló-Vidal
Department of IMAE, University of Girona, Campus Montilivi, Edif. P4, 17003 Girona, Spain
e-mail: josepantoni.martin@udg.edu

C. Barceló-Vidal
e-mail: carles.barcelo@udg.edu

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_26

components are not informative. Typical examples are the data from day-time-use surveys where the components of all D -part vectors of the data set refer to a constant absolute total equal to $K = 24$ h (or 1440 min). Another typical example are the data from experiments with mixtures in industry where $K = 1$ (or 100). On the other hand, in other very common situations the parts w_1, \dots, w_D do not constitute an exhaustive partition of a fixed whole because they form a *subcomposition*, that is, a composition not including all parts into which the whole has been divided. This case occurs frequently in geochemistry or air pollution data, among others, where the CoDa are concentrations of a few specific chemical elements. In other situations, the whole simply does not take a constant value across the samples. For instance, in economics or waste management, among others, these data are very common. In these two scenarios, where the total is not a fixed value, one can assume that the information provided by a composition can be split into two different parts: *relative* and *absolute* (Martín-Fernández et al. 2020). For example, the relative information provided by the vectors (0.03, 0.08, 0.10), (0.12, 0.32, 0.40) of a 3-part subcomposition of three chemical elements (in mg/L) is the same but the two vectors provide different absolute information because the concentrations in the second sample are four times the concentration in the former composition. Barceló-Vidal and Martín-Fernández (2016) coined the term *compositional analysis* (CoAn) to refer to the analysis of the relative information provided by compositional data, that is, the log-ratio analysis proposed in Aitchison (1986). In this chapter, we propose an approach to complete a CoAn by adding the analysis of the absolute information.

In a CoAn, the components of a composition represent relative magnitudes that stop being independent of each other, since any change in one of the parts necessarily causes the relative change in one or more of the other components. This effect is evident if \mathbf{w} is a full composition with a fixed total. However, this lack of independence between the components is also present even when one is doing a CoAn with subcompositions. This lack of independence between the components of a compositional data set prevents us from applying the standard procedures of statistical analysis. Aitchison (1986) showed that this serious inconvenience can be overcome if one analyzes the ratios w_i/w_j between the components of the compositions instead of their individual values. For strictly mathematical requirements, Aitchison proposed working with the logarithms of these ratios (*log-ratios*). Ratios and logarithms force us to avoid the case of zero values in CoAn. We consider the zero as a special value that deserves a particular analysis according to its nature. The reason why a zero value is present in CoDa is because it is informative and determines the approach to be applied (Palarea-Albaladejo and Martín-Fernández 2015). The requirement that a CoAn must be based on the ratios of the components of the compositions is equivalent to a *scale invariance* requirement (Barceló-Vidal and Martín-Fernández 2016). That is, the results of any analysis of CoDa must be the same if one replaces any D -part vector \mathbf{w} of the data set by the vector $k\mathbf{w}$, for any value $k > 0$. One considers that \mathbf{w} and $k\mathbf{w}$ are *compositionally equivalent* or, in mathematical terms, that they are members of the same equivalence class. All members of an equivalence class provide the same relative information, that is, the information given by the ratios between the components. For example, the 3-part

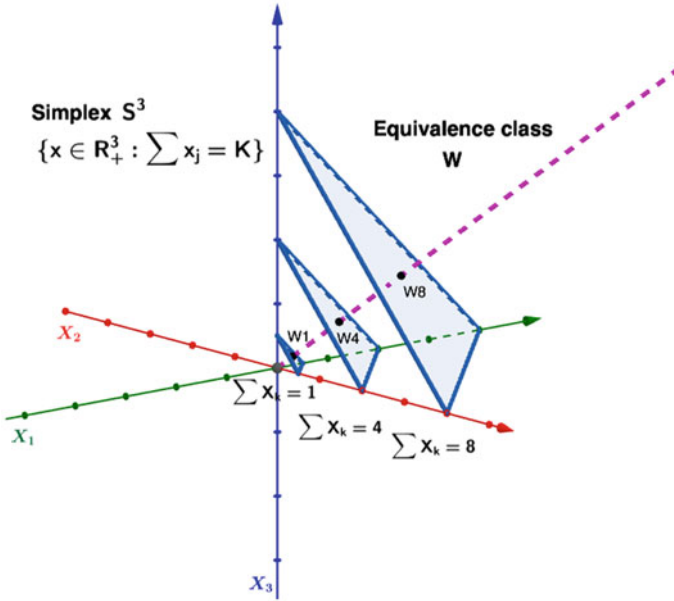


Fig. 1 A 3-part composition \mathbf{w} as an equivalence class (violet dashed line): three representatives \mathbf{w}_1 , \mathbf{w}_4 , and \mathbf{w}_8 in the simplex when the constant sum constraint, respectively, equals 1, 4 and 8

vectors $(0.3, 0.5, 0.2)$, $(1.2, 2.0, 0.8)$, $(2.4, 4.0, 1.6)$, $(7.2, 12, 4.8)$, and $(30, 50, 20)$ provide the same relative information although they, respectively, sum 1, 4, 8, 24, and 100. In other words, they are members of the same equivalence class. Consequently, in CoAn, the sample space is not \mathbb{R}_+^D , it is the quotient space \mathbb{R}_+^D / \sim , where \sim symbolizes the compositional equivalence relation. Observe that the different expressions—absolute, proportions, percentages, ppm...—of a D -part vector \mathbf{w} belong to the same equivalence class. Despite the D -part vector chosen to represent the equivalence class being irrelevant in CoAn, it is usual to choose the representative whose components' sum is equal to 1. The operation closure $C\mathbf{w} = \mathbf{w} / \sum_{j=1}^D w_j$ (Aitchison 1986) provides this representative. This criterion could be generalized to representatives with a sum equal to 100 or any other positive value. Figure 1 shows an equivalence class \mathbf{w} (violet dashed line) and three of its representatives \mathbf{w}_1 , \mathbf{w}_4 , and \mathbf{w}_8 with constant sum constraints, equal to 1, 4, and 8. The set of vectors of \mathbb{R}_+^D whose sum is equal to one is the *unit simplex* (Aitchison 1986):

$$S^D = \{\mathbf{x} = (x_1, \dots, x_D)^t : x_1 > 0, \dots, x_D > 0; x_1 + \dots + x_D = 1\}.$$

The fundamentals of CoAn were introduced in Aitchison (1986) and have been completed and justified mathematically in Pawłowsky-Glahn and Buccianti (2011), Barceló-Vidal and Martín-Fernández (2016), and Egozcue and Pawłowsky-Glahn (2019), among many other papers.

In statistics, the sample space of a random vector \mathbf{x} is of crucial importance for any linear model. These models, which are based on the linear combinations of the components of \mathbf{x} , are the fundamentals of, among others, regression models, principal component analysis (PCA), and discriminant analysis (DA). For example, consider the linear combinations for a typical PCA (Krzanowski 2000) defined as

$$\mathbf{z} = \mathbf{\Lambda} \cdot (\mathbf{x} - \boldsymbol{\mu}) , \quad (1)$$

where $\boldsymbol{\mu}$ is the arithmetic mean of \mathbf{x} , $\mathbf{\Lambda} = (\lambda_{ij})$ is a $(D \times D)$ matrix of constants (*loadings*), whose rows are orthogonal unit vectors of length D . $\mathbf{z} = (z_1, \dots, z_D)^t$ is a latent random vector with mean zero and a covariance matrix $\boldsymbol{\Psi} = \text{diag}(\psi_1^2, \dots, \psi_D^2)$. Note that the structure of this matrix indicates that the variables z_j are uncorrelated with each other.

When one wants to analyze the relative information of the random vector \mathbf{x} , the expression in Eq. (1) should provide the same information regardless of the representative selected for the composition. Consequently, when one uses the representative in the unit simplex it holds that $\sum_{j=1}^D x_j = 1$, $\sum_{j=1}^D \mu_j = 1$, and $\sum_{j=1}^D (x_j - \mu_j) = 0$. Therefore, from Eq. (1), for each z_i , $i = 1, 2, \dots, D$, it holds that

$$z_i = \lambda_{i1}(x_1 - \mu_1) + \lambda_{i2}(x_2 - \mu_2) + \dots + \lambda_{iD}(x_D - \mu_D) ,$$

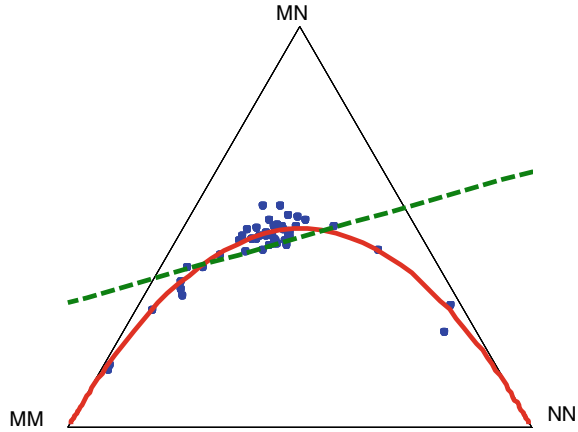
and, because $\sum_{j=1}^D (x_j - \mu_j) = 0$,

$$z_i = (\lambda_{i2} - \lambda_{i1})(x_2 - \mu_2) + (\lambda_{i3} - \lambda_{i1})(x_3 - \mu_3) + \dots + (\lambda_{iD} - \lambda_{i1})(x_D - \mu_D) ,$$

suggesting that at least one variable, z_i , should be equal to zero (i.e., $\lambda_{i1} = \lambda_{i2} = \dots = \lambda_{iD}$), because otherwise one would have D uncorrelated orthonormal variables in a $D - 1$ -dimensional sample space (simplex), which is absurd.

The *bloodMN* data set records the absolute frequencies of the genotypes MN, MM, and NN observed in 32,572 blood samples coming from 49 different ethnic groups around the world (Boyd 1950, pages 234–235). For example, in 300 Italian people, the genotype MN was observed 144 times, genotype MM 96 times, and genotype NN 60 times. Therefore, the (MN, MM, NN)-part vector of the Italian ethnic group is (144, 96, 60) or (0.48, 0.32, 0.20), when the unit-sum representative is considered. Figure 2 shows the ternary diagram associated with the compositional analysis of the *bloodMN* data set. Following Aitchison (1986), the ternary diagram is a popular representation of the simplex space for $D = 3$ (Fig. 1). The convex shape of the point cloud suggests that any typical latent variable model that consists of standard linear forms will not be able to explain this kind of association. In fact, the first component of the standard PCA based on the covariance matrix on the set of compositions expressed as proportions (green dashed line in Fig. 2) explains 80% of total variance. The equation of the standard PCA expressed in matrix form is

Fig. 2 Ternary diagram: MN, MM, and NN blood type composition in 49 ethnic populations. Green dashed line: typical linear PCA. Red solid line: log-ratio linear PCA



$$\mathbf{z} = \begin{pmatrix} 0.222 & -0.791 & 0.570 \\ 0.786 & -0.201 & -0.585 \\ 0.577 & 0.577 & 0.577 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} -0.071 \\ 0.164 \\ 0.577 \end{pmatrix}, \quad (2)$$

where $\mathbf{x} = (x_1, x_2, x_3)^t$ represents a (MN,MM,NN)-part composition expressed in proportions. From Eq. (2) one can derive that z_3 is equal to zero because $x_1 + x_2 + x_3 = 1$. The results shown in Eq. (2) as well as the other results in this chapter have been obtained using the programming language R (R Core Team 2019).

It is well-known that using the expression of a PCA (Eq.(1)) one can find a potential sample \mathbf{x} which has given pc scores. For example, taking extreme values for the pc scores one can characterize what the potential outliers are. Using Eq. (2) to detect the sample with pc scores (0.5, 0, 0), one obtains the composition (0.557, -0.039, 0.482). Despite being mathematically consistent because the unit-sum constraint is verified, the result cannot be a composition because a value $MM < 0$ is not in the sample space (Fig. 2).

This fact suggests that the typical PCA is not appropriate for CoAn because its assumptions are not fulfilled. Many authors (Mooijaart et al. 1999) suited this kind of typical linear techniques, like PCA, factor-analysis, and least squares, to the CoDa peculiarities. However, these typical linear techniques are not appropriate for modeling common associations between compositional parts (Fig. 2). At this point, two possible alternatives appear: either modify and adapt a more complex typical technique, or apply techniques that are appropriate for the particular geometry of the simplex. Although we admit that the first alternative could provide reasonable results in some scenarios, we use the second because we believe that the more coherent the method is, the more reasonable the results will be.

2 Log-Contrast and Log-Ratio Variables

Currently, there is a general agreement among CoAn researchers that the geometry of the simplex is based on log-ratio coordinates (Pawlowsky-Glahn and Buccianti 2011). For example, the red solid line in Fig. 2 suggests that a log-ratio linear PCA fits better than the typical (green dashed line) PCA does.

Let \mathbf{w} be a D -part vector and $\mathbf{a} = (a_1, \dots, a_D)^t$ be a vector in $\mathbb{R}_{\neq 0}^D$, being $\mathbf{0}$ a vector of zeros. We symbolize by $\text{ll}(\mathbf{a}; \mathbf{w})$ the log-linear (ll) combination $a_1 \ln w_1 + \dots + a_D \ln w_D$ of the components of $\ln \mathbf{w}$. The log-linear combination $\text{ll}(\mathbf{a}; \mathbf{w})$ can be written in matrix form as $\mathbf{a}^t \cdot \ln \mathbf{w}$. We say that $\text{ll}(\mathbf{a}; \mathbf{w})$ is *unitary* if $\mathbf{a}^t \cdot \mathbf{a} = 1$.

If $a_1 + \dots + a_D = 0$, then $\text{ll}(\mathbf{a}; \mathbf{w})$ is scale invariant and is called a *log-contrast* ($\text{lc}(\mathbf{a}; \mathbf{w})$) (Aitchison 1986). In fact, $\text{lc}(\mathbf{a}; k\mathbf{w}) = \text{lc}(\mathbf{a}; \mathbf{w})$, for any $k > 0$, in particular, $\text{lc}(\mathbf{a}; \mathbf{w}) = \text{lc}(\mathbf{a}; \mathbf{x})$, where \mathbf{x} is the unit-sum representative. Importantly, any linear combination of pairwise log-ratios can be expressed as a log-contrast. For example, for $D = 3$, the linear combination $\alpha \ln(w_1/w_2) + \beta \ln(w_1/w_3) + \lambda \ln(w_2/w_3)$ can be rewritten as $\text{lc}(\mathbf{a}; \mathbf{w})$, with $\mathbf{a} = (\alpha + \beta, \lambda - \alpha, -\beta - \lambda)^t$. Inversely, any log-contrast $\text{lc}(\mathbf{a}; \mathbf{w})$ can be rewritten in different ways as a linear combination of pairwise log-ratios. For example, if one considers $a_D = -a_1 - \dots - a_{D-1}$ then $\text{lc}(\mathbf{a}; \mathbf{w}) = \mathbf{a}^t \cdot \ln \mathbf{w} = a_1 \ln(w_1/w_D) + \dots + a_{D-1} \ln(w_{D-1}/w_D)$. Consequently, linear combinations of pairwise log-ratios and of log-contrasts are equivalent concepts.

The most common log-ratio scores used in CoAn to represent a composition are simply a set of log-contrasts. For example, the clr-score:

$$\text{clr } \mathbf{w} = (\ln(w_1/g(\mathbf{w})), \dots, \ln(w_D/g(\mathbf{w})))^t,$$

where $g(\mathbf{w})$ is the geometric mean of \mathbf{w} , can be formulated in terms of D log-contrasts:

$$\text{clr } \mathbf{w} = \left(\frac{D-1}{D} \ln w_1 - \sum_{j \neq 1} \frac{1}{D} \ln w_j, \dots, \frac{D-1}{D} \ln w_D - \sum_{j \neq D} \frac{1}{D} \ln w_j \right)^t. \quad (3)$$

Thus, the first component of $\text{clr } \mathbf{w}$ is the log-contrast $\text{lc}(\mathbf{a}_1; \mathbf{w})$, where $\mathbf{a}_1 = (1 - 1/D, -1/D, \dots, -1/D)^t$. Equation (3) can be written in matrix form as $\text{clr } \mathbf{w} = \mathbf{G} \cdot \ln \mathbf{w}$, where $\mathbf{G} = \mathbf{I} - \frac{1}{D}\mathbf{J}$, with \mathbf{I} being the $D \times D$ identity matrix and \mathbf{J} the $D \times D$ matrix of ones.

Observe that the sum of the components of vector $\text{clr } \mathbf{w}$, as well as the columns and rows of the symmetric $D \times D$ matrix \mathbf{G} , is equal to 0. In fact, if we interpret clr as a map from \mathbb{R}_+^D to \mathbb{R}^D , the image $\text{clr } \mathbb{R}_+^D$ (and also $\text{clr } \mathcal{S}^D$) is the subspace $\mathcal{V} = \{\mathbf{y} \in \mathbb{R}^D : \sum_{i=1}^D y_i = 0\}$ of dimension $D - 1$. Then, given an orthonormal basis $\{\mathbf{v}_1, \dots, \mathbf{v}_{D-1}\}$ of \mathcal{V} , we call the *orthonormal log-ratio coordinates* (olr) of a composition, \mathbf{w} , the coordinates of $\text{clr } \mathbf{w}$ relative to $\{\mathbf{v}_1, \dots, \mathbf{v}_{D-1}\}$ (Martín-Fernández 2019). The concept of orthonormal coordinates was firstly introduced by Egozcue et al. (2003) when defining the isometric log-ratio transformation. The

matrix expression of the corresponding coordinate vector $\text{olr } \mathbf{w}$ is equal to

$$\text{olr } \mathbf{w} = \mathbf{V}^t \cdot \text{clr } \mathbf{w} = \mathbf{V}^t \cdot \mathbf{G} \cdot \ln \mathbf{w} = \mathbf{V}^t \cdot \ln \mathbf{w} , \tag{4}$$

where the matrix \mathbf{V} is equal to $\mathbf{V} = (\mathbf{v}_1 : \dots : \mathbf{v}_{D-1})$, in that $\mathbf{V}^t \cdot \mathbf{V} = \mathbf{I}$ and $\mathbf{V} \cdot \mathbf{V}^t = \mathbf{G}$. Since the sum of the components of \mathbf{v}_i are equal to 0, the olr coordinates of \mathbf{w} are log-contrasts. The olr and clr scores of a composition \mathbf{w} are related since it holds that $\text{olr } \mathbf{w} = \mathbf{V}^t \cdot \text{clr } \mathbf{w}$ and $\text{clr } \mathbf{w} = \mathbf{V} \cdot \text{olr } \mathbf{w}$. This relation can be introduced in the expression of a log-contrast to obtain the expression $\text{lc}(\mathbf{a}; \mathbf{w}) = \mathbf{a}^t \cdot \ln \mathbf{w} = (\mathbf{G} \cdot \mathbf{a})^t \cdot \ln \mathbf{w} = \mathbf{a}^t \cdot \mathbf{G} \cdot \ln \mathbf{w} = \mathbf{a}^t \cdot \text{clr } \mathbf{w} = (\mathbf{V}^t \cdot \mathbf{a})^t \cdot \text{olr } \mathbf{w}$, where one states that any log-contrast is a linear combination of log-ratios.

The particular geometry of the simplex on which CoAn is based has three basic elements: the operations perturbation and powering, and an inner product. All three can be defined from the clr scores or the olr scores of compositions. For example, the inner product between two compositions \mathbf{w} and \mathbf{w}^* is defined as the usual inner product between the vectors $\text{clr } \mathbf{w}$ and $\text{clr } \mathbf{w}^*$ in \mathbb{R}^D or between the vectors $\text{olr } \mathbf{w}$ and $\text{olr } \mathbf{w}^*$ in \mathbb{R}^{D-1} . In this way, the simplex is structured as a Euclidean space of dimension $D - 1$ (Barceló-Vidal and Martín-Fernández 2016). This allows one to apply the well-known properties of Euclidean spaces to compositions. For example, one can make orthogonal projections, define angles, and calculate ellipses. In short, one can properly apply any multivariate method to analyze the relative information in CoDa (Mateu-Figueras et al. 2011).

The covariance structure of a random composition \mathbf{w} can be defined from the covariance matrix $\mathbf{\Gamma}$ of $\text{clr } \mathbf{w}$ as

$$\mathbf{\Gamma} = (\gamma_{ij}) = (\text{cov}\{\text{clr } \mathbf{w}_i, \text{clr } \mathbf{w}_j\}) .$$

It could also be defined from the covariance matrix $\mathbf{\Delta}$ of $\text{olr } \mathbf{w}$, that is,

$$\mathbf{\Delta} = (\delta_{ij}) = (\text{cov}\{\text{olr } \mathbf{w}_i, \text{olr } \mathbf{w}_j\}) .$$

The two covariance matrices are well related because it holds that

$$\begin{aligned} \mathbf{\Gamma} &= \mathbf{V} \cdot \mathbf{\Delta} \cdot \mathbf{V}^t, \\ \mathbf{\Delta} &= \mathbf{V}^t \cdot \mathbf{\Gamma} \cdot \mathbf{V}. \end{aligned}$$

Both covariance matrices have the same trace which is a measure of the *total variance* of \mathbf{w} .

As in standard PCA, the first principal component (PC1) associated with \mathbf{w} will be the unitary log-contrast $\text{lc}(\mathbf{a}_1; \mathbf{w})$ which maximizes $\text{var}\{\text{lc}(\mathbf{a}; \mathbf{w})\}$, that is, which maximizes $\mathbf{a}^t \cdot \mathbf{\Gamma} \cdot \mathbf{a}$ (Aitchison 1986). The procedure for calculating the other principal components $\text{lc}(\mathbf{a}_2; \mathbf{w}), \dots, \text{lc}(\mathbf{a}_D; \mathbf{w})$ is similar to the typical algorithm. Thus, Eq. (1) becomes

$$\mathbf{z} = (\mathbf{a}_1 : \cdots : \mathbf{a}_D)^t \cdot (\text{clr } \mathbf{w} - \boldsymbol{\mu}), \quad (5)$$

where $\boldsymbol{\mu}$ is the mean of the random variable $\text{clr } \mathbf{w}$.

For example, the log-ratio PCA of the *bloodMN* data set is equal to

$$\mathbf{z} = \begin{pmatrix} 0.008 & -0.711 & 0.703 \\ 0.816 & -0.401 & -0.415 \\ 0.577 & 0.577 & 0.577 \end{pmatrix} \cdot \begin{pmatrix} (\text{clr } \mathbf{w})_1 \\ (\text{clr } \mathbf{w})_2 \\ (\text{clr } \mathbf{w})_3 \end{pmatrix} - \begin{pmatrix} -0.487 \\ 0.546 \\ 0 \end{pmatrix}, \quad (6)$$

or in terms of log-contrasts

$$\mathbf{z} = \begin{pmatrix} 0.008 & -0.711 & 0.703 \\ 0.816 & -0.401 & -0.415 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \ln w_1 \\ \ln w_2 \\ \ln w_3 \end{pmatrix} - \begin{pmatrix} -0.487 \\ 0.546 \\ 0 \end{pmatrix}, \quad (7)$$

where $z_3 = 0$ and where PC1 captures 98.4% of the total relative variance of the CoDa. The log-ratio PCA (Eq. (6)) could also have been performed using the covariance matrix $\mathbf{\Delta}$ of $\text{olr } \mathbf{w}$. That is, the resulting PCA equation, once the olr coordinates are expressed in logarithms, is equal to Eq. (7). Note that the expression of PC1 can be approximated by $0.707 \ln(w_3/w_1) + 0.487$ suggesting that the variance is retained by the information provided by the ratio between the genotypes NN and MM. If we consider an olr basis where the first vector is formed using this ratio, then the second vector of the basis is formed using the ratio between the genotype MN and the geometric mean of genotypes MM and NN. Because these two olr vectors are orthogonal, the second vector approximates the direction of the PC2. This component can be considered approximately constant because it captures only 1.6% of the variance. Consequently, the ratio between the genotype MN and the geometric mean of the genotypes MM and NN can be considered approximately constant, consistent with the Hardy–Weinberg equilibrium.

3 Log-contrast and Multiplicative Total

Let \mathcal{V}^\perp be the subspace of \mathbb{R}^D orthogonal to the subspace \mathcal{V} . Since the dimension of \mathcal{V} is $D - 1$, the dimension of \mathcal{V}^\perp is equal to 1. Therefore, the unit-norm vector $\mathbf{t} = \frac{1}{\sqrt{D}} \mathbf{1}$, where $\mathbf{1}$ is the D -vector of ones, is a basis of \mathcal{V}^\perp . Then, any vector \mathbf{y} in \mathbb{R}^D can be uniquely decomposed as the sum $\mathbf{y}_\mathcal{V} + \mathbf{y}_{\mathcal{V}^\perp}$ of two orthogonal vectors, one belonging to \mathcal{V} and the other to \mathcal{V}^\perp :

$$\mathbf{y}_\mathcal{V} = \mathbf{G} \cdot \mathbf{y} = \left(\mathbf{I} - \frac{1}{D} \mathbf{J} \right) \cdot \mathbf{y} \quad \text{and} \quad \mathbf{y}_{\mathcal{V}^\perp} = \frac{1}{D} \mathbf{J} \cdot \mathbf{y} = \frac{\sum_{j=1}^D y_j}{D} \mathbf{1}.$$

When the orthogonal decomposition is applied to the vector \mathbf{a} of a log-linear combination $\text{ll}(\mathbf{a}; \mathbf{w})$, a decomposition in terms of two log-linear combinations is obtained:

$$\begin{aligned}\text{ll}(\mathbf{a}; \mathbf{w}) &= (\mathbf{a}_{\mathcal{V}} + \mathbf{a}_{\mathcal{V}^\perp})^t \cdot \ln \mathbf{w} = \\ &= \mathbf{a}_{\mathcal{V}}^t \cdot \ln \mathbf{w} + \mathbf{a}_{\mathcal{V}^\perp}^t \cdot \ln \mathbf{w} = \\ &= \text{ll}(\mathbf{a}_{\mathcal{V}}; \mathbf{w}) + \text{ll}(\mathbf{a}_{\mathcal{V}^\perp}; \mathbf{w}),\end{aligned}\quad (8)$$

which can be expressed in terms of a log-contrast as

$$\begin{aligned}\text{ll}(\mathbf{a}; \mathbf{w}) &= \text{ll}(\mathbf{a}_{\mathcal{V}}; \mathbf{w}) + \text{ll}(\mathbf{a}_{\mathcal{V}^\perp}; \mathbf{w}) = \\ &= \mathbf{a}_{\mathcal{V}}^t \cdot \ln \mathbf{w} + \mathbf{a}_{\mathcal{V}^\perp}^t \cdot \ln \mathbf{w} = \\ &= \mathbf{a}^t \cdot \mathbf{G} \cdot \ln \mathbf{w} + \frac{\sum_{j=1}^D a_j}{D} \left(\sum_{j=1}^D \ln w_j \right) = \\ &= \text{lc}(\mathbf{G} \cdot \mathbf{a}; \mathbf{w}) + \frac{\sum_{j=1}^D a_j}{\sqrt{D}} \frac{\sum_{j=1}^D \ln w_j}{\sqrt{D}}.\end{aligned}\quad (9)$$

Consequently, the log-linear combination $\text{ll}(\mathbf{a}; \mathbf{x})$ decomposes into two parts. The *relative* (scale invariant) part corresponds to the log-contrast $\text{lc}(\mathbf{G} \cdot \mathbf{a}; \mathbf{w})$. The *absolute* part is the last term in Eq. (9). This involves the logarithm of the *multiplicative total* $t(\mathbf{w}) = \prod_{j=1}^D w_j^{1/\sqrt{D}}$ introduced in Pawlowsky-Glahn et al. (2015). That is,

$$\begin{aligned}\text{ll}(\mathbf{a}; \mathbf{w}) &= \text{lc}(\mathbf{G} \cdot \mathbf{a}; \mathbf{w}) + \frac{\sum_{j=1}^D a_j}{\sqrt{D}} \frac{\sum_{j=1}^D \ln w_j}{\sqrt{D}} = \\ &= \mathbf{a}^t \cdot \text{clr } \mathbf{w} + \frac{\sum_{j=1}^D a_j}{\sqrt{D}} \ln t(\mathbf{w}) = \\ &= (\mathbf{V}^t \cdot \mathbf{a})^t \cdot \text{olr } \mathbf{w} + \frac{\sum_{j=1}^D a_j}{\sqrt{D}} \text{tlog } \mathbf{w},\end{aligned}\quad (10)$$

where $\text{tlog } \mathbf{w}$ represents the total of logarithms.

Following (Coenders et al. 2017), this decomposition can be generalized. Indeed, given an orthonormal basis $\{\mathbf{v}_1, \dots, \mathbf{v}_{D-1}\}$ of the subspace \mathcal{V} , one can complete the set of $D - 1$ vectors to an orthonormal basis of \mathbb{R}^D by adding the vector $\mathbf{t} = \frac{1}{\sqrt{D}} \mathbf{1}$ because \mathbf{t} is a unit-norm vector that is orthogonal to the subspace \mathcal{V} . Let $\mathbf{U} = [\mathbf{v}_1 : \dots : \mathbf{v}_{D-1} : \mathbf{t}]$ be the $D \times D$ matrix whose columns are the vectors of the basis of \mathbb{R}^D . Since it is an orthonormal basis, it holds that $\mathbf{U}^t \cdot \mathbf{U} = \mathbf{U} \cdot \mathbf{U}^t = \mathbf{I}$. In particular, for any D -part vector \mathbf{w} in \mathbb{R}_+^D , it holds that $(\text{olr } \mathbf{w} : \text{tlog } \mathbf{w}) = \mathbf{U}^t \cdot \ln \mathbf{w}$. Regarding the scale invariance property, it holds that

$$\mathbf{U}^t \cdot \ln(k\mathbf{w}) = (\text{olr}(k\mathbf{w}) : \text{tlog}(k\mathbf{w})) = \left(\text{olr } \mathbf{w} : (\text{tlog } \mathbf{w} + \sqrt{D} \ln k) \right), \text{ for any } k > 0.$$

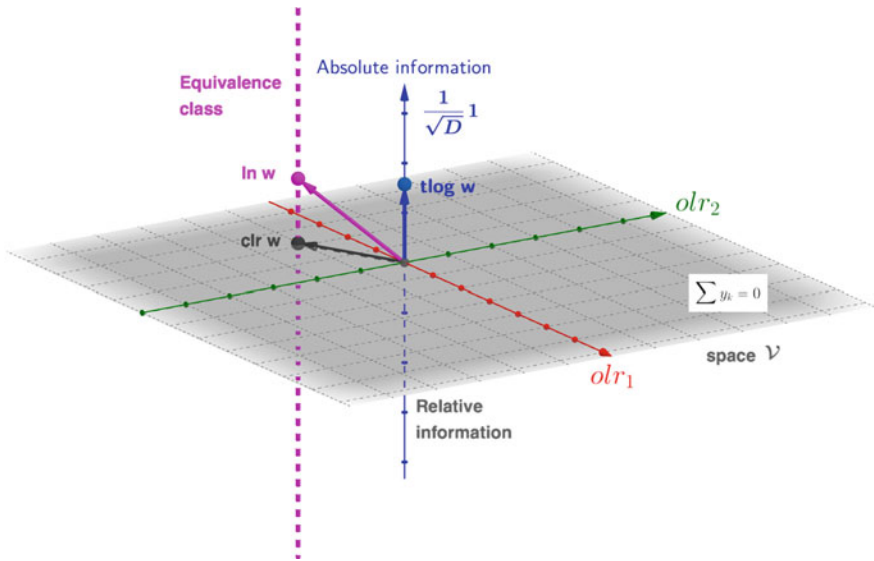


Fig. 3 Decomposition into relative and absolute information of a composition. All the compositions in the same equivalence class (*violet dashed line*) have the same relative information (*clr w*) but different absolute information (*tlog w*). The gray plane represents the space \mathcal{V} where an *olr* basis can be defined

That is, any composition \mathbf{w} can be decomposed into two terms providing, respectively, its *relative* and *absolute* information (Fig. 3).

Figure 3 shows how the information provided by a D -part composition \mathbf{w} is decomposed. When the log-transformed composition $\ln \mathbf{w}$ is projected to the vector $\text{clr } \mathbf{w}$, the relative information can be expressed in terms of an *olr* vector of coordinates. Projecting the $\ln \mathbf{w}$ vector into the direction of the vector \mathbf{t} gives the absolute information associated with the value $\text{tlog } \mathbf{w}$. All the points on the vertical violet dashed line share the same relative information (*clr* scores) because they belong to the same equivalence class. For example, let $\mathbf{v}_1 = (\frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}})$, $\mathbf{v}_2 = (0, \frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ be an *olr* basis of the space \mathcal{V} . The corresponding *olr* coordinates of the 3-part compositions (0.3, 0.5, 0.2), (1.2, 2.0, 0.8), (2.4, 4.0, 1.6), (7.2, 12, 4.8), and (30, 50, 20) are equal to $(-0.043, 0.648)$. However, their *tlog* scores are, respectively, $-2.025, 0.377, 1.577, 3.480,$ and 5.952 . In contrast, the compositions (30, 50, 20), (50, 20, 30), and (20, 30, 50), sharing the same *tlog* score (5.952), have different vectors of *olr* coordinates, respectively, $(-0.043, 0.648), (0.583, -0.286),$ and $(-0.540, -0.361)$.

Equation (10) can be also obtained using the expressions of the basis of \mathbb{R}^D :

$$\begin{aligned}
\|(\mathbf{a}; \mathbf{w})\| &= \mathbf{a}^t \cdot \ln \mathbf{w} = \\
&= \mathbf{a}^t (\mathbf{U} \cdot \mathbf{U}^t) \cdot \ln \mathbf{w} = \\
&= (\mathbf{U}^t \cdot \mathbf{a})^t \cdot (\mathbf{U}^t \cdot \ln \mathbf{w}) = \\
&= \left(\mathbf{V}^t \cdot \mathbf{a} : \frac{\sum_{j=1}^D a_j}{\sqrt{D}} \right)^t \cdot (\text{olr } \mathbf{w} : \text{tlog } \mathbf{w}) = \\
&= (\mathbf{V}^t \cdot \mathbf{a})^t \cdot \text{olr } \mathbf{w} + \frac{\sum_{j=1}^D a_j}{\sqrt{D}} \text{tlog } \mathbf{w} .
\end{aligned} \tag{11}$$

The equation-matrix of the PCA based on the covariance matrix of the logarithms of the (MN,MM,NN)-parts of the *bloodMN* data set expressed in its original units is equal to

$$\mathbf{z} = \begin{pmatrix} 0.572 & 0.458 & 0.680 \\ 0.083 & 0.793 & -0.604 \\ 0.816 & -0.402 & -0.415 \end{pmatrix} \cdot \begin{pmatrix} \ln w_1 \\ \ln w_2 \\ \ln w_3 \end{pmatrix} - \begin{pmatrix} 7.835 \\ 1.739 \\ 0.540 \end{pmatrix} \tag{12}$$

where $\mathbf{w} = (w_1, w_2, w_3)^t$ represents a generic (MN,MM,NN)-part vector expressed in absolute format. The PC retain, respectively, 80.4%, 19.2%, and 0.4% of the total variance of the data set. Note that PC3 can be considered as a constant pc score. In addition, the PC1 loadings (0.572, 0.458, 0.680) suggest a parallel direction to the vector \mathbf{t} , whereas the loadings of PC2 and PC3 are both approximately coefficients of a log-contrast.

When Eq. (9) is used for the PC, its log-linear combination decomposes into relative and absolute parts as

$$\mathbf{z} = \begin{pmatrix} 0.002 & -0.112 & 0.110 \\ -0.008 & 0.702 & -0.694 \\ 0.816 & -0.401 & -0.415 \end{pmatrix} \cdot \ln \mathbf{w} + \begin{pmatrix} 0.988 \\ 0.157 \\ -0.001 \end{pmatrix} \cdot \text{tlog } \mathbf{w} - \begin{pmatrix} 7.835 \\ 1.739 \\ 0.540 \end{pmatrix} \tag{13}$$

The directions provided by vectors (0.002, -0.112, 0.110) and (-0.008, 0.702, -0.694) of the two first log-contrasts are quite similar to the direction associated with the vector (0.008, -0.711, 0.703) (Eq. (7)), suggesting again that the relative information for PC1 and PC2 are associated with the ratio NN/MM. As expected, the largest coefficient of the absolute information (0.988) corresponds to PC1. However, because the decomposition is not based on orthonormal coordinates, one cannot evaluate the quality of the representation of the information in Eq. (13). To have an orthogonal decomposition of the PCA (Eq. (12)), one can express $\ln \mathbf{w}$ on the orthonormal basis of \mathbb{R}^D given by the matrix \mathbf{U} . Indeed, as suggested by the log-contrasts of PC1 and PC2 in Eq. (13), we select the basis

$$\mathbf{v}_1 = \left(0, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^t, \mathbf{v}_2 = \left(\frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}} \right)^t, \mathbf{t} = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right)^t, \tag{14}$$

with the coordinates of $\ln \mathbf{w}$ with respect to this basis being

$$\text{olr}_1 = \frac{1}{\sqrt{2}} \ln \frac{w_3}{w_2}, \quad \text{olr}_2 = \sqrt{\frac{2}{3}} \ln \frac{w_1}{(w_2 \cdot w_3)^{1/2}}, \quad \text{tlog} = \frac{1}{\sqrt{3}} \sum_{j=1}^3 \ln w_j. \quad (15)$$

The expression of the PCA (Eq. (12)) in this basis is

$$\mathbf{z} = \begin{pmatrix} 0.157 & 0.002 & 0.988 \\ -0.988 & -0.009 & 0.157 \\ -0.010 & 0.999 & -0.001 \end{pmatrix} \cdot \begin{pmatrix} \text{olr}_1 \\ \text{olr}_2 \\ \text{tlog} \end{pmatrix} - \begin{pmatrix} -0.848 \\ 5.467 \\ 5.839 \end{pmatrix}. \quad (16)$$

The vector (0.157, 0.002, 0.988) suggests that the tlog scores are very well represented by PC1 because its relative quality is 97.6% (0.988²). The best representation of the variable olr_1 is on PC2, whereas the constant pc scores provided by PC3 represent the variable olr_2 , consistent with the Hardy–Weinberg equilibrium.

Importantly, despite the PCA in Eq. (16) completing the expression obtained by CoAn (Eq. (7)), this analysis is recommended for the analysts who are interested only in the relative information. That is, an analyst, who is interested only in the proportions of the genotypes, should carry out a CoAn because in this analysis one assumes that CoDa provides only relative information.

4 Decomposition of a Linear Discriminant Analysis Model

In a clinical study, the amount of three metabolites—total cortisol, total corticosterone, and pregnanetriol + Δ -5-pregnenetriol—present in people's urine has been analyzed (Aitchison 1986, pages 363–364). The study involved 67 healthy people: 37 adults and 30 children. The 3-part composition $\mathbf{w} = (w_1, w_2, w_3)^t$ represents the amount (in mg) of the three metabolites present in the urine expelled for 24 h for each person. The aim is to analyze if the presence of these metabolites in the urine is different in adults and children. Possible differences may be in the absolute amounts of the metabolites, in their proportions or in both.

A linear discriminant analysis (LDA) may be the first approximation to the analysis. The linear discriminant function (LDF) of LDA on $\ln \mathbf{w}$, being \mathbf{w} in its original units, is

$$\text{LDF}_1 : \text{ll}(\mathbf{a}; \mathbf{w}) = (-0.303, -0.640, -1.036)^t \cdot \ln \mathbf{w}. \quad (17)$$

The leave-one-out cross-validated misclassification rate (MCR) of LDF_1 is equal to 4.5% (= 3/67). The structure matrix indicates that $\ln w_3$ is the coordinate most correlated with the LDF_1 scores. However, using the expression in Eq. (17), one is not able to know if this discriminating power of part w_3 is due to the different absolute values of this part in the two groups (adults and children) or if it is because of its

relative values or because of both causes. To assess the extent to which this misclassification rate is due to the absolute values of the metabolites, we can orthogonally decompose the log-linear combination (Eq. (17)):

$$\text{ll}(\mathbf{a}; \mathbf{w}) = (0.357, 0.020, -0.377)^t \cdot \ln \mathbf{w} - 1.143 \cdot \text{tlog } \mathbf{w} . \tag{18}$$

If we use as the linear discriminant function (LDF₂) the scores provided by the log-contrast $\text{lc}(\mathbf{a}_\Psi; \mathbf{w}) = (0.357, 0.020, -0.377)^t \cdot \ln \mathbf{w}$, MCR increases to 13.4% (= 9/67). In contrast, the MCR of the LDF₃ performed from the scores given by $-1.143 \cdot \text{tlog } \mathbf{w}$ —the absolute part of Eq. (18)—increases only up to 6.0% (= 4/67). These results suggest that the relative information discriminates more poorly than the absolute information does. Note that the log-contrast $\text{lc}(\mathbf{a}_\Psi; \mathbf{w}) = (0.357, 0.020, -0.377)^t \cdot \ln \mathbf{w} \approx 0.367 \ln(w_1/w_3)$, suggesting that the relative information is based on the ratio w_1/w_3 .

To simultaneously analyze the influence the relative and absolute information have in \mathbf{w} on the LDA, one can express $\ln \mathbf{w}$ on the orthonormal basis $\{\mathbf{v}_1, \dots, \mathbf{v}_{D-1}, \mathbf{t}\}$ of \mathbb{R}^D . To do this, we select the basis

$$\mathbf{v}_1 = \left(\frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}} \right)^t, \mathbf{v}_2 = \left(-\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, -\frac{1}{\sqrt{6}} \right)^t, \mathbf{t} = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}} \right)^t . \tag{19}$$

That is, the coordinates of $\ln \mathbf{w}$ with respect to this basis are

$$\text{olr}_1 = \frac{1}{\sqrt{2}} \ln \frac{w_1}{w_3}, \text{olr}_2 = \sqrt{\frac{2}{3}} \ln \frac{w_2}{(w_1 \cdot w_3)^{1/2}}, \text{tlog} = \frac{1}{\sqrt{3}} \sum_{j=1}^3 \ln w_j . \tag{20}$$

The coordinates olr_1 and olr_2 provide the relative information contained in \mathbf{w} , while the last coordinate (tlog score) gives the total of \mathbf{w} in a logarithmic scale. The LDF₁ on $\ln \mathbf{w}$ (Eq. (18)) expressed in this basis is $0.518 \text{olr}_1 + 0.024 \text{olr}_2 - 1.143 \text{tlog}$. Note that coherently with the log-contrast in Eq. (18), the coordinate olr_1 is the most relevant log-ratio. If one substitutes olr_1 , olr_2 , and tlog by their expressions (Eq. (19)), then the LDF₁ in Eq. (17) is obtained. However, using the new expression of LDF₁, one can calculate the structure matrix being the correlations of the discriminant scores with olr_1 , olr_2 , and tlog , respectively, -0.684 , -0.471 , and 0.974 . Consequently, one states that the absolute information (tlog coordinate) of \mathbf{w} is the most relevant part to discriminate between children and adults in relation to the metabolites contained in the urine.

Figure 4 shows the orthonormal coordinates of the metabolites data set. The olr coordinates represented in Fig. 4a suggest that, as regards the relative information, the two groups (adult: circles; children: triangles) mix in the center of the point cloud. On the other hand, the color and the size associated with the value of the tlog coordinate suggest that this score may better discriminate the two groups. Figure 4b shows the distribution of the tlog coordinate for each group. This figure corroborates

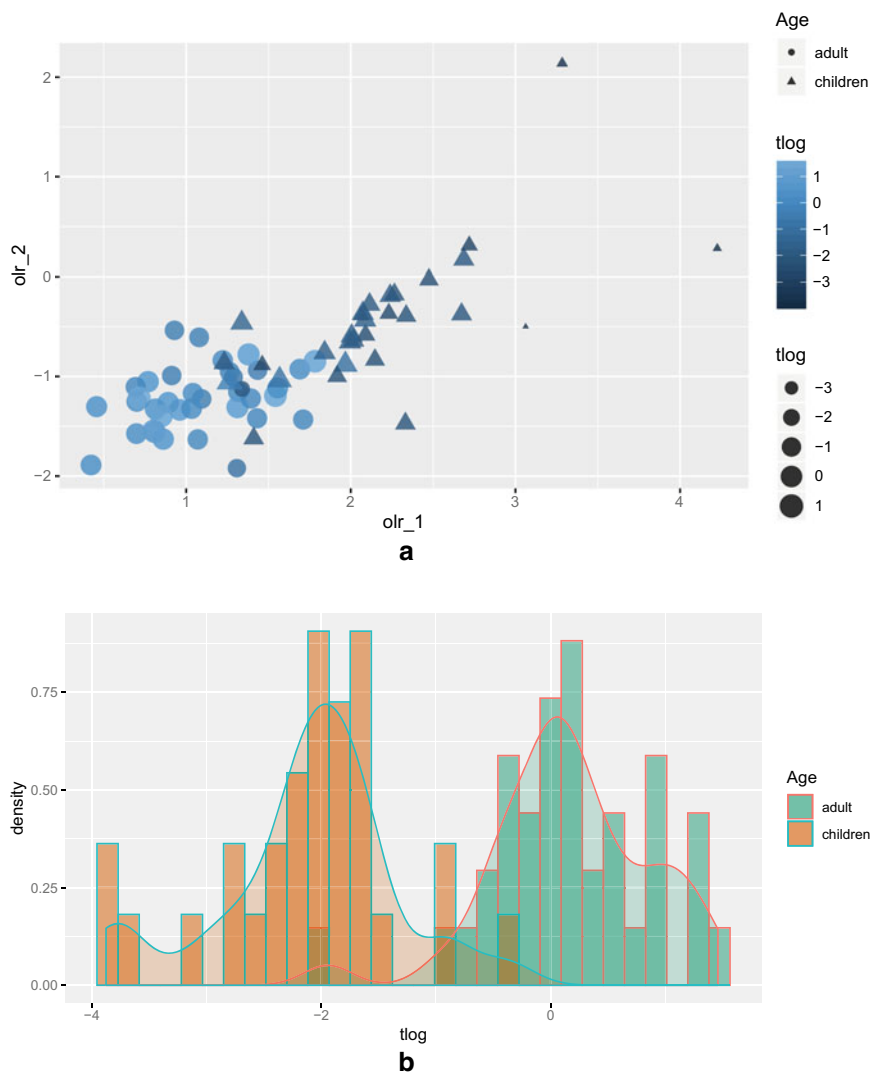


Fig. 4 Orthonormal coordinates of metabolites data set: **a** orl coordinates; **b** tlog coordinate

that only a few cases will be misclassified when using this coordinate to classify metabolite composition.

5 Final Concluding Remarks

When CoDa has a constant constraint sum or the analyst is interested only in analyzing the relative information, then CoAn is the analysis recommended. On the other hand, when CoDa does not have a constant constraint sum or the analyst is also interested in the absolute information, CoAn should be completed using the multiplicative total. Using the fundamentals of the Aitchison geometry, log-linear models (linear models applied to log-transformed data) can be decomposed into the relative and absolute information parts. This decomposition is based on the decomposition of log-linear combinations into log-contrasts and the log-score of the multiplicative total. As a result, a decomposition of a log-linear model into its relative and absolute parts allows the importance of each type of information to be evaluated. In this chapter, we have illustrated this new approach by performing a PCA and an LDA of real data sets.

Acknowledgements This work has been partially financed by the CODAMET project (Ministerio de Ciencia, Innovación y Universidades; Ref: RTI2018-095518-B-C21).

References

- Aitchison, J. (1986). *The statistical analysis of compositional data*. Monographs on statistics and applied probability. London: Chapman & Hall Ltd. (Reprinted in 2003 with additional material by The Blackburn Press) 416 p.
- Barceló-Vidal, C., & Martín-Fernández, J. A. (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics*, 45, 57–71.
- Boyd, W. C. (1950). *Genetics and the races of man: An introduction to modern physical anthropology* (p. 453). Toronto: McClelland & Stewart Ltd.
- Coenders, G., Martín-Fernández, J. A., & Ferrer-Rosell, B. (2017). When relative and absolute information matter: Compositional predictor with a total in generalized linear models. *Statistical Modelling*, 17(6), 494–512.
- Egozcue, J. J., & Pawłowsky-Glahn, V. (2019). Compositional data: The sample space and its structure. *Test*, 28(3), 599–638.
- Egozcue, J. J., Pawłowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300.
- Krzanowski, W. (2000). *Principles of multivariate analysis* (2nd ed., Vol. 23). Oxford statistical science series. Oxford, UK. 563 p.
- Martín-Fernández, J. A. (2019). Comments on: Compositional data: The sample space and its structure. *Test*, 28(3), 653–657.
- Martín-Fernández, J. A., Egozcue, J. J., Olea, R. A., & Pawłowsky-Glahn, V. (2020). Units recovery methods in compositional data analysis. *Natural Resources Research* (in revision)
- Mateu-Figueras, G., Pawłowsky-Glahn, V. & Egozcue, J. J. (2011). The principle of working on coordinates. In V. Pawłowsky-Glahn & A. Buccianti (Eds.) (pp. 31–42). 378 p.
- Mooijaart, A., Van der Heijden, P. G. M., & Van der Ark, L. A. (2015). A least squares algorithm for a mixture model for compositional data. *Computational Statistics and Data Analysis*, 30, 359–379.

- Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2015). zCompositions - R package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemometrics and Intelligent Laboratory Systems*, *143*, 85–96.
- Pawlowsky-Glahn, V. & Buccianti, A. (Eds.). (2011). *Compositional data analysis: Theory and applications*. Hoboken: Wiley. 378 p.
- Pawlowsky-Glahn, V., Egozcue, J. J., & Lovell, D. (2015). Tools for compositional data with a total. *Statistical Modelling*, *15*, 175–190.
- R Core Team. (2019). R: A language and environment for statistical computing, Vienna, Austria. <https://www.R-project.org/>.

Independent Component Analysis for Compositional Data



Christoph Muehlmann, Kamila Fačevicová, Alžběta Gardlo, Hana Janečková,
and Klaus Nordhausen

Abstract Compositional data represent a specific family of multivariate data, where the information of interest is contained in the ratios between parts rather than in absolute values of single parts. The analysis of such specific data is challenging as the application of standard multivariate analysis tools on the raw observations can lead to spurious results. Hence, it is appropriate to apply certain transformations prior to further analysis. One popular multivariate data analysis tool is independent component analysis. Independent component analysis aims to find statistically independent components in the data and as such might be seen as an extension to principal component analysis. In this paper, we examine an approach of how to apply independent component analysis on compositional data by respecting the nature of the latter and demonstrate the usefulness of this procedure on a metabolomics dataset.

C. Muehlmann · K. Nordhausen (✉)
Institute of Statistics & Mathematical Methods in Economics,
Vienna University of Technology, Wiedner Hauptstr. 7, 1040 Vienna, Austria
e-mail: klaus.nordhausen@tuwien.ac.at

C. Muehlmann
e-mail: christoph.muehlmann@tuwien.ac.at

K. Fačevicová
Department of Mathematical Analysis and Applications of Mathematics,
Palacký University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: kamila.facevicova@gmail.com

A. Gardlo
Department of Clinical Biochemistry, University Hospital Olomouc and Palacký University
Olomouc, I.P. Pavlova 185/6, 779 00 Olomouc, Czech Republic
e-mail: alzbetagardlo@gmail.com

H. Janečková
Laboratory for Inherited Metabolic Disorders, Department of Clinical Biochemistry,
University Hospital Olomouc and Palacký University Olomouc, I.P. Pavlova 185/6,
779 00 Olomouc, Czech Republic
e-mail: janeckovah@gmail.com

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics
and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_27

1 Introduction

Independent component analysis (ICA) is a well-established data analysis method in signal processing with the goal of recovering hidden signals that are usually meant to have a physical meaning. In recent years, ICA methods have attracted increasing interest in the statistics community as an extension of normality-based multivariate methods that only use second-order moments. In principle, ICA can be seen as a refinement of principal component analysis where, after removing second-order information, higher order moments are used to search for hidden structures which are not visible in the principal components. Classical ICA methods are mainly developed for independent and identically distributed observations in a Euclidean space. Nevertheless, these methods are also applied, for example, on time series, spatial data, etc. but to the best of our knowledge not on iid compositional data.

Compositional data is special in the way that the entries (parts) of a d -variate vector are positive and carry relative rather than absolute information about the respective observation of interest. Moreover, the parts of the compositional vector are by nature not independent and in some specific situations, e.g. when all parts are bounded by a constant sum constraint, a spurious correlation between them is present. Therefore, compositional data lies on a simplex and does not follow the real Euclidean geometry. Examples of compositional data are geochemical data where the chemical composition of soil samples is of interest, the composition of nutrients of food intake or the distribution of market shares. For further details and examples of compositional data, see, for example, Aitchison (1986), Egozcue and Pawlowsky-Glahn (2019), Fačevićová et al. (2016), Filzmoser et al. (2018), Morais et al. (2018), Pawlowsky-Glahn and Buccianti (2011), Trinh et al. (2019).

It is well established that standard multivariate methods should not be applied directly to compositional data. Either methods which take the geometry of compositional data into account or methods that transform compositional data in such a way that standard multivariate analysis tools can be applied are appropriate. In this paper, we take the latter approach.

We review some basic ICA methods in Sect. 2. In Sect. 3, we describe compositional data and methods to transform such data into the real space. Based on the former two sections, we present how ICA can be performed on compositional data in Sect. 4 and conclude the paper with the analysis of a metabolomics dataset from healthy newborns in Sect. 5 and a discussion in Sect. 6.

2 Independent Component Analysis

From a statistical perspective, independent component analysis is usually formulated as a latent variable model as follows.

Definition 1 An observable p -vector \mathbf{x} follows the independent component (IC) model if

$$\mathbf{x} = \mathbf{A}\mathbf{z} + \mathbf{b},$$

where \mathbf{A} is a $p \times p$ non-singular matrix, \mathbf{b} a p -vector, and the latent p -variate random vector \mathbf{z} satisfies

- (A1) $\mathbf{E}(\mathbf{z}) = \mathbf{0}$ and $\mathbf{COV}(\mathbf{z}) = \mathbf{I}_p$,
- (A2) the components of \mathbf{z} are independent, and
- (A3) at most one component of \mathbf{z} is Gaussian.

Thus $\mathbf{E}(\mathbf{x}) = \mathbf{b}$ and $\mathbf{COV}(\mathbf{x}) = \mathbf{A}\mathbf{A}^\top$. The goal of ICA is to find a $p \times p$ matrix \mathbf{W} such that $\mathbf{W}\mathbf{x}$ has independent components. Note however that in general it will not hold that $\mathbf{W}(\mathbf{x} - \mathbf{b}) = \mathbf{z}$ as the IC model assumptions only fix the location and scale of \mathbf{z} but not the signs or the order of the components. Therefore, for every solution \mathbf{W} , also $\mathbf{P}\mathbf{J}\mathbf{W}$ is a solution, where \mathbf{P} is a $p \times p$ permutation matrix (1 per row and column, 0 elsewhere) and \mathbf{J} is a $p \times p$ sign-change matrix (a diagonal matrix with ± 1 on its diagonal).

There are many suggestions in the literature on how to estimate \mathbf{W} based on a sample $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, and for recent reviews see, for example, Comon and Jutten (2010), Nordhausen and Oja (2018). Almost all ICA methods make, however, use of the following result:

Key result Let \mathbf{x} follow the IC model and denote $\mathbf{x}^{st} = \mathbf{COV}(\mathbf{x})^{-1/2}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$, then there exists an orthogonal $p \times p$ matrix \mathbf{U} such that

$$\mathbf{U}^\top \mathbf{x}^{st} = \mathbf{z}.$$

This result implies that after estimating $\mathbf{COV}(\mathbf{x})$ and $\mathbf{E}(\mathbf{x})$, the problem is reduced from finding a general $p \times p$ matrix to a $p \times p$ orthogonal matrix. Also note that this means that the performance of ICA methods does not depend on the values of \mathbf{A} and \mathbf{b} , as these are accounted for when standardizing the data. An unmixing matrix estimate is therefore obtained as $\mathbf{W} = \mathbf{U}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ and different ICA approaches differ in the way they estimate \mathbf{U} . In the following, we will show how some popular ICA methods estimate this rotation.

2.1 FOBI

Fourth-order blind identification (FOBI), presented in Cardoso (1989), was one of the first ICA methods but is still popular as it has a closed-form solution. For FOBI, we need to define the scatter matrix of fourth-order moments

$$\mathbf{COV}_4(\mathbf{x}) = \frac{1}{p+2} \mathbf{E} \left((\mathbf{x} - \mathbf{E}(\mathbf{x}))^\top \mathbf{COV}(\mathbf{x})^{-1} (\mathbf{x} - \mathbf{E}(\mathbf{x})) (\mathbf{x} - \mathbf{E}(\mathbf{x})) (\mathbf{x} - \mathbf{E}(\mathbf{x}))^\top \right).$$

Then we can define the following:

Definition 2 The FOBI unmixing matrix is $\mathbf{W}_{\text{FOBI}} = \mathbf{U}_{\text{FOBI}}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ where the columns of \mathbf{U}_{FOBI} are given by the eigenvectors of $\mathbf{COV}_4(\mathbf{x}^{st})$.

From denoting $\mathbf{U}_{\text{FOBI}} \mathbf{D} \mathbf{U}_{\text{FOBI}}^\top$ to be the eigendecomposition of $\mathbf{COV}_4(\mathbf{x}^{st})$ which is needed to compute \mathbf{W}_{FOBI} , it is obvious that FOBI is only unique when the eigenvalues contained in the diagonal matrix \mathbf{D} are distinct. One can actually show that these eigenvalues are linked to the kurtosis values of the independent components. For FOBI to be well-defined, Assumption (A3) from the IC model needs to be replaced by the stronger assumption:

(A4) The kurtosis values of the independent components must be distinct.

FOBI is often the first ICA method applied as it is quick to compute, gives a fast first impression, and its statistical properties are well known; see, for example, Miettinen et al. (2015), Nordhausen and Virta (2019) for more details. FOBI can also be of interest outside the IC model and can be seen as an invariant coordinate selection method (Tyler et al. 2009).

2.2 JADE

Assumption (A4) is considered highly restrictive. Joint approximate diagonalization of eigenmatrices (JADE) can be seen as an extension of FOBI which relaxes this strict assumption, Cardoso and Souloumiac (1993).

For JADE, we have to define the fourth-order cumulant matrices

$$\mathbf{C}_{ij}(\mathbf{x}) = \mathbf{E} \left((\mathbf{x}^{st \top} \mathbf{E}_{ij} \mathbf{x}^{st}) \mathbf{x}^{st} \mathbf{x}^{st \top} \right) - \mathbf{E}_{ij} - \mathbf{E}_{ij}^\top - \text{tr}(\mathbf{E}_{ij}) \mathbf{I}_p,$$

where $\mathbf{E}_{ij} = \mathbf{e}_i \mathbf{e}_j^\top$ with \mathbf{e}_i being a vector of dimension p with the i th element equals 1 and 0 otherwise. As i and j range from 1 to p , there are in total p^2 such cumulant matrices. In the IC model, $\mathbf{C}_{ij}(\mathbf{z}) = 0$ if $i \neq j$ and for the case where $i = j$ $\mathbf{C}_{ii}(\mathbf{z})$ corresponds to the kurtosis of the i th component. The matrix of fourth moments can actually be expressed as

$$\mathbf{COV}_4(\mathbf{x}) = \frac{1}{p+2} \sum_{i=1}^p \mathbf{C}_{ii}(\mathbf{x}) + (p+2) \mathbf{I}_p,$$

meaning that it uses not all possible cumulant information. The idea of JADE is to exploit the information contained in all cumulant matrices.

Definition 3 The JADE unmixing matrix is $\mathbf{W}_{\text{JADE}} = \mathbf{U}_{\text{JADE}}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ where \mathbf{U}_{JADE} is the maximizer of

$$\sum_{i=1}^p \sum_{j=1}^p \|\text{diag}(\mathbf{U}^\top \mathbf{C}_{ij}(\mathbf{x}^{st}) \mathbf{U})\|_F^2.$$

Thus, JADE tries to maximize the diagonal elements of $\mathbf{U}^\top \mathbf{C}_{ij}(\mathbf{x}^{st}) \mathbf{U}$ which is equivalent to minimize the off-diagonal elements by the orthogonal invariance of the Frobenius norm $\|\cdot\|_F$. As in the IC model, only $\mathbf{C}_{ii}(\mathbf{z})$ is non-zero and corresponds to the kurtosis of z_i . This means that JADE relaxes the FOBI assumption (A4) to the following:

(A5) At most one independent component can have zero kurtosis.

For a finite sample, the joint diagonalization of more than two matrices needs to be carried out approximately; many algorithms that jointly diagonalize two or more matrices are available; see, for example, Illner et al. (2015). For the purpose of this paper, we will use an algorithm based on Givens rotations, Clarkson (1988).

The statistical properties of JADE are, for example, given in Miettinen et al. (2015); from an asymptotic point of view, FOBI is never superior compared to JADE. JADE is however computationally more expensive, especially when the number of independent components grows, as p^2 matrices need to be computed and jointly diagonalized.

As a compromise, k-JADE was suggested in Miettinen et al. (2013). The idea is to use not all matrices \mathbf{C}_{ij} , but only those whose indices are not too far apart, i.e. $|i - j| < k$. This requires however that the first step, the whitening step, is not done using just the covariance matrix but using \mathbf{W}_{FOBI} .

Definition 4 Denote $\mathbf{x}^{st'} = \mathbf{W}_{\text{FOBI}}(\mathbf{x} - \mathbf{E}(\mathbf{x}))$ and choose an integer $1 \leq k \leq p$, then the k-JADE unmixing matrix is $\mathbf{W}_{\text{kJADE}} = \mathbf{U}_{\text{kJADE}}^\top \mathbf{W}_{\text{FOBI}}$ where $\mathbf{U}_{\text{kJADE}}$ is the maximizer of

$$\sum_{|i-j|<k}^p \|\text{diag}(\mathbf{U}^\top \mathbf{C}_{ij}(\mathbf{x}^{st'}) \mathbf{U})\|_F^2.$$

The value k is basically a tuning parameter. The intuition is that the multiplicities of the distinct non-zero kurtosis values of the independent components are at most k , and that there is at most one component having kurtosis zero. Usually, k is simply chosen by the user based on expert knowledge. In Virta et al. (2020), some guidelines for the selection are offered, which are however not very practical. The statistical properties of k-JADE are given in Miettinen et al. (2013), Virta et al. (2020). It can be shown that for a value of k which fulfills the multiplicity condition, k-JADE is asymptotically as efficient as JADE but has, if k is small, a much smaller computational complexity.

2.3 FastICA

FOBI, JADE, and k-JADE are often called algebraic ICA methods. Another large group of ICA methods is based on projection pursuit ideas, where the most prominent one is FastICA. It was originally suggested in Hyvärinen (1999a). Some of the many FastICA variants are discussed below.

The general idea of FastICA is to find the column vectors $\mathbf{u}_1, \dots, \mathbf{u}_p$ of \mathbf{U} which maximize the non-Gaussianity of the components of $\mathbf{U}^\top \mathbf{x}^{st}$. Non-Gaussianity of a univariate random variable x is measured by $|E(G(x))|$ with some twice continuously differentiable and non-quadratic function G that satisfies $E(G(y)) = 0$ for $y \sim N(0, 1)$. The most popular choices for G are

pow3: $G(x) = (x^4 - 3)/4$,
 tanh: $G(x) = \log(\cosh(x)) - c_t$, and
 gauss: $G(x) = -\exp(-x^2/2) - c_g$.

The constants $c_t = E(\log(\cosh(y))) \approx 0.375$ and $c_g = E(-\exp(-y^2/2)) \approx -0.707$ are normalizing constants. The derivatives of G , denoted as g , are called non-linearities and are the name givers as pow3: $g(x) = x^3$, tanh: $g(x) = \tanh(x)$ and gauss: $g(x) = x \exp(-x^2/2)$.

2.3.1 Deflation-Based FastICA

FastICA was first suggested in Hyvärinen and Oja (1997) using the non-linearity pow3 and finding the column vectors of \mathbf{U}_{DF} one after another which is now known as deflation-based FastICA.

Definition 5 The deflation-based FastICA unmixing matrix is defined as $\mathbf{W}_{DF} = \mathbf{U}_{DF}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$, where the k th column of \mathbf{U} , \mathbf{u}_k , maximizes

$$|E[G(\mathbf{u}_k^\top \mathbf{x}_{st})]|$$

under the constraints $\mathbf{u}_k^\top \mathbf{u}_k = 1$ and $\mathbf{u}_j^\top \mathbf{u}_k = 0$, $j = 1, \dots, k - 1$.

To obtain estimates, a modified Newton-Raphson algorithm is used which iterates the following steps until convergence:

$$\begin{aligned} \mathbf{u}_k &\leftarrow \mathbf{E}[g(\mathbf{u}_k^\top \mathbf{x}_{st}) \mathbf{x}_{st}] - \mathbf{E}[g'(\mathbf{u}_k^\top \mathbf{x}_{st})] \mathbf{u}_k \\ \mathbf{u}_k &\leftarrow \left(\mathbf{I}_p - \sum_{l=1}^{k-1} \mathbf{u}_l \mathbf{u}_l^\top \right) \mathbf{u}_k \\ \mathbf{u}_k &\leftarrow \|\mathbf{u}_k\|^{-1} \mathbf{u}_k. \end{aligned}$$

The last two steps perform the Gram-Schmidt orthonormalization.

The properties of deflation-based FastICA have been studied in detail in Ollila (2010), Nordhausen et al. (2011). One issue with deflation-based FastICA is that besides the global maximum it has many local maxima and the order in which the vectors \mathbf{u}_k are found depends heavily on the initial value of the algorithm, where in turn the estimation performance depends on the order in which the vectors \mathbf{u}_k are found. Using asymptotic arguments, Nordhausen et al. (2011) suggested reloaded

Table 1 Table of default candidate set of non-linearities of adaptive deflation-based FastICA, where $(x)_+ = x$ if $x > 0$ and 0 otherwise, and $(x)_- = x$ if $x < 0$ and 0 otherwise

$g_1(x) = x^3$	$g_6(x) = (x)_+^2 + (x)_-^2$	$g_{11}(x) = (x - 1.0)_+^2 + (x + 1.0)_-^2$
$g_2(x) = \tanh(x)$	$g_7(x) = (x - 0.2)_+^2 + (x + 0.2)_-^2$	$g_{12}(x) = (x - 1.2)_+^2 + (x + 1.2)_-^2$
$g_3(x) = x \exp(-x^2/2)$	$g_8(x) = (x - 0.4)_+^2 + (x + 0.4)_-^2$	$g_{13}(x) = (x - 1.4)_+^2 + (x + 1.4)_-^2$
$g_4(x) = (x + 0.6)_-^2$	$g_9(x) = (x - 0.6)_+^2 + (x + 0.6)_-^2$	$g_{14}(x) = (x - 1.6)_+^2 + (x + 1.6)_-^2$
$g_5(x) = (x - 0.6)_+^2$	$g_{10}(x) = (x - 0.8)_+^2 + (x + 0.8)_-^2$	

FastICA, which estimates first the independent components using FOBI or k-JADE and then derives an optimal order based on the estimated independent components.

The idea of reloaded FastICA to fix the extraction order based on asymptotic arguments was extended in Miettinen et al. (2014) to also select an optimal non-linearity for each component out of a candidate set of possible non-linearities. This is known as adaptive deflation-based FastICA. We will denote the adaptive deflation-based FastICA unmixing matrix as \mathbf{W}_{ADF} . The candidate set of non-linearities suggested in Miettinen et al. (2014) contains, for example, the non-linearities presented in Table 1.

2.3.2 Symmetric FastICA

A FastICA variant estimating all directions in parallel was suggested in Hyvärinen (1999b).

Definition 6 The symmetric FastICA estimator $\mathbf{W}_{\text{SF}} = \mathbf{U}_{\text{SF}}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ uses as a criterion for \mathbf{U}_{SF}

$$\sum_{j=1}^p |\mathbf{E}[G(\mathbf{u}_j^\top \mathbf{x}_{st})]|$$

which should be maximized under the orthogonality constraint $\mathbf{U}_{\text{SF}}^\top \mathbf{U}_{\text{SF}} = \mathbf{I}_p$.

The steps of the iterative algorithm to compute \mathbf{U}_{SF} are

$$\begin{aligned} \mathbf{u}_k &\leftarrow \mathbf{E}[g(\mathbf{u}_k^\top \mathbf{x}_{st}) \mathbf{x}_{st}] - \mathbf{E}[g'(\mathbf{u}_k^\top \mathbf{x}_{st})] \mathbf{u}_k, \quad k = 1, \dots, p \\ \mathbf{U}_{\text{SF}}^\top &\leftarrow (\mathbf{U}_{\text{SF}}^\top \mathbf{U}_{\text{SF}})^{-1/2} \mathbf{U}_{\text{SF}}^\top. \end{aligned}$$

The first update step of the algorithm is similar to that of the deflation-based FastICA estimator. The orthogonalization step can be interpreted as taking an average

over the vectors of the first step. This differs from the deflation-based approach where errors made in the k th direction carry on to the following directions and therefore the errors accumulate. This is often the reason why symmetric FastICA is usually considered superior to the deflation-based FastICA. However, there are also cases where the accumulation is preferable to the averaging. This occurs when some independent components are easier to find than the others. Statistical properties of symmetric FastICA are given in Miettinen et al. (2015), Wei (2015), Miettinen et al. (2017).

2.3.3 Squared Symmetric FastICA

One of the most recent variants of FastICA is the squared symmetric FastICA estimator (Miettinen et al. 2017). The idea of this estimator is to replace the absolute values in the objective function of the symmetric FastICA with squared values.

Definition 7 The squared symmetric FastICA estimator $\mathbf{W}_{S2F} = \mathbf{U}_{S2F}^\top \mathbf{COV}(\mathbf{x})^{-1/2}$ obtains \mathbf{U}_{S2F} as the maximizer of

$$\sum_{j=1}^p (\mathbf{E}[G(\mathbf{u}_j^\top \mathbf{x}_{st})])^2$$

under the orthogonality constraint $\mathbf{U}_{S2F}^\top \mathbf{U}_{S2F} = \mathbf{I}_p$.

The steps of the resulting algorithm are

$$\begin{aligned} \mathbf{u}_k &\leftarrow \mathbf{E}[G(\mathbf{u}_k^\top \mathbf{x}_{st})](\mathbf{E}[g(\mathbf{u}_k^\top \mathbf{x}_{st})\mathbf{x}_{st}] - \mathbf{E}[g'(\mathbf{u}_k^\top \mathbf{x}_{st})]\mathbf{u}_k), \quad k = 1, \dots, p, \\ \mathbf{U}_{S2F}^\top &\leftarrow (\mathbf{U}_{S2F}^\top \mathbf{U}_{S2F})^{-1/2} \mathbf{U}_{S2F}^\top. \end{aligned}$$

Thus, the first step of the algorithm equals the first step in the symmetric algorithm with an additional multiplication by $\mathbf{E}[G(\mathbf{u}_k^\top \mathbf{x}_{st})]$. Hence, the squared symmetric variant puts more weight on components that are “more” non-Gaussian, which most often, but not always, is advantageous. The properties of the squared symmetric FastICA estimator as well as comparisons to the deflation-based and symmetric FastICA methods are given in Miettinen et al. (2017). In Miettinen et al. (2017), it is also shown that if the non-linearity `pow3` is used, symmetric squared FastICA is asymptotically equivalent to JADE.

Besides assumptions (A1)–(A3), deflation-based, symmetric, and squared symmetric FastICA need further assumptions based on G to ensure consistency. Assuming the order of the components is fixed as $|\mathbf{E}[G(z_1)]| \geq \dots \geq |\mathbf{E}[G(z_p)]|$, then it is required that for any $\mathbf{z} = (z_1, \dots, z_p)^\top$ with independent and standardized components and for any orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_p)$, the following holds.

For deflation-based FastICA:

(A6) For all $k = 1, \dots, p$, $|\mathbf{E}[G(\mathbf{u}_k^\top \mathbf{z})]| \leq |\mathbf{E}[G(z_k)]|$, when $\mathbf{u}_k^\top \mathbf{e}_j = 0$ for all $j = 1, \dots, k - 1$, where \mathbf{e}_i is a p -vector with i th element one and others zero,

for symmetric FastICA

(A7) $|\mathbf{E}[G(\mathbf{u}_1^T \mathbf{z})]| + \dots + |\mathbf{E}[G(\mathbf{u}_p^T \mathbf{z})]| \leq |\mathbf{E}[G(z_1)]| + \dots + |\mathbf{E}[G(z_p)]|$,

and for squared symmetric FastICA

(A8) $(\mathbf{E}[G(\mathbf{u}_1^T \mathbf{z})])^2 + \dots + (\mathbf{E}[G(\mathbf{u}_p^T \mathbf{z})])^2 \leq (\mathbf{E}[G(z_1)])^2 + \dots + (\mathbf{E}[G(z_p)])^2$.

It was proven, for example, in Miettinen et al. (2015), that all three conditions are fulfilled with `pow3`. On the other hand, in the case of non-linearities like `tanh` and `gauss` some of these conditions might be violated for certain source distributions.

From a computational point of view, the advantage of both symmetric versions is that the initial value of \mathbf{U} is not important when the sample size is large, as the algorithms converge usually to the global maxima.

To conclude this section we can point out that FOBI, JADE, k-JADE, symmetric FastICA, and squared symmetric FastICA are affine equivariant ICA methods which means that their performance does not depend on the mixing matrix. So, from this point of view, only deflation-based FastICA differs, which can be overcome when the reloaded version or adaptive version is used. Affine equivariance will be of relevance later when applying the ICA methods to compositional data.

3 Compositional Data and Its Real Space Representation

A specific family of d -dimensional vectors is present when each entry (part) of a vector is positive and carries information about its contribution to the whole. In the following, such multivariate observations are called (vector) compositional data, whose specifics were already described, utilized, and analyzed in a wide range of applications (Pawlowsky-Glahn and Buccianti 2011). The main property of compositional data is its relative nature, when the relevant information is contained in the ratios between parts rather than in the absolute values of the parts. Consider, e.g. a vector describing a geochemical structure of soil, where each part represents the quantity of the given element in the sample. The quantity can be given either in absolute scale, like in mg of the component contained in the sample, or some of its relative alternatives, typically ppm. While the mg representation depends on the overall size of the sample, the ppm one does not, despite the ratios between parts remaining unchanged. Both representations are therefore from the compositional point of view equivalent.

Due to the relative nature of compositional data, the sample space of representations of a d -part compositional vector \mathbf{x} forms a d -part simplex

$$\mathcal{S}^d = \left\{ \mathbf{x} = (x_1, \dots, x_d)^\top, \sum_{i=1}^d x_i = \kappa, \kappa > 0 \right\},$$

where the Aitchison geometry holds. The whole sample space is formed by equivalence classes of proportional vectors (Pawlowsky-Glahn et al. 2015, Chaps. 2, 3). Since most of the standard statistical methods are designed for real-valued data following the usual Euclidean geometrical structure, it is favorable to express compositional data in real coordinates prior to their analysis. One of the possible representations is the centered log-ratio (clr) transformation from \mathcal{S}^d to \mathbb{R}^d given by

$$\text{clr}(\mathbf{x})_i = \ln \frac{x_i}{g_m(\mathbf{x})} = \frac{1}{d} \sum_{j=1}^d \ln \frac{x_i}{x_j}, \quad \text{for } i = 1, \dots, d,$$

where $g_m(\mathbf{x})$ denotes the geometrical mean of all parts. The parts of the resulting clr vector can be interpreted in terms of the dominance of the compositional part in the numerator within the whole composition or equivalently as its mean dominance over each part of the whole composition. The use of logarithm symmetrizes this relationship. Let us stress here that the clr values depend on the set of compositional parts used for its computation and therefore the above interpretation holds true only when the whole composition is considered. Within the whole manuscript, the clr transformation based on all compositional parts will be of interest. On the other hand, from its construction, the clr coefficients/variables are not linearly independent, as they sum up to zero and, therefore, the whole clr vector falls in a $(d - 1)$ -dimensional subspace of \mathbb{R}^d . This feature prevents direct use of the clr representation within methods that require full rank data, like robust PCA (Filzmoser et al. 2009) or the above stated ICA methods.

One possible workaround is the isometric log-ratio (ilr) transformation, which represents the compositional vector \mathbf{x} in a system of $d - 1$ orthonormal real coordinates. This system can be obtained directly from the clr vector as

$$\text{ilr}(\mathbf{x}) = \mathbf{V}^\top \text{clr}(\mathbf{x}),$$

where the columns of the $d \times d - 1$ log-contrast matrix \mathbf{V} are given as $\mathbf{v}_i = \text{clr}(\xi_i)$ and the vectors ξ_i , $i = 1, \dots, d - 1$ constitute an orthonormal basis in \mathcal{S}^d . See Pawlowsky-Glahn and Buccianti (2011), Ch. 11 for details.

The system of basis vectors $\{\xi_1, \dots, \xi_{d-1}\}$ is not uniquely given and can be chosen according to the purpose of further analysis. Since each system of ilr coordinates can be obtained as an orthogonal rotation of the others, its specific choice does not affect the results of their analysis, like predictions of the regression model with a compositional regressor or scores of the robust PCA model (Filzmoser et al. 2009; Hron et al. 2012). When it is required, a specific coordinate system can be selected by some data-driven method, like hierarchical clustering of the compositional parts, or using expert knowledge. In both cases, the main aim is to obtain such an interpretation of the coordinates at hand, which is favorable according to the given problem (Egozcue

and Pawlowsky-Glahn 2005). Since a specific interpretation of the ilr coordinates is not the main purpose here, the same system as in Nordhausen et al. (2015) is used. The basis vectors ξ_i have the value $\exp(\sqrt{1/i(i+1)})$ at the first i positions, $\exp(-\sqrt{1/i(i+1)})$ at the position $i+1$, and 1 at the remaining ones. Consequently, the columns of the log-contrast matrix are

$$\mathbf{v}_i = \sqrt{\frac{i}{i+1}} \left(\frac{1}{i}, \dots, \frac{1}{i}, -1, 0, \dots, 0 \right)^\top, \quad i = 1, \dots, d-1.$$

The ilr coordinates have the form of balances between the i th part of the composition and all parts with lower indices

$$\text{ilr}(\mathbf{x})_i = \sqrt{\frac{i}{i+1}} \ln \left(\frac{(x_1 \cdots x_i)^{1/i}}{x_{i+1}} \right), \quad \text{for } i = 1, \dots, d-1.$$

Finally, the clr and ilr representations are mutually transferable through the contrast matrix \mathbf{V}

$$\text{clr}(\mathbf{x}) = \mathbf{V} \text{ilr}(\mathbf{x})$$

and also the back-transformation to the simplex is possible by using

$$\mathbf{x} = \exp(\text{clr}(\mathbf{x})) = \exp(\mathbf{V} \text{ilr}(\mathbf{x})).$$

4 ICA for Compositional Data

As described above, ICA is not reasonable for data following the Aitchison geometry in its raw form. Therefore, it is natural to transform the data first into the Euclidean space. As ICA methods start with whitening and therefore require full rank data, the ilr space is the most natural representation. Due to the affine equivariance property of the discussed ICA methods, the particular used basis for the ilr transformation at most affects the order and signs of the estimated independent components. Hence, for compositional ICA we have the following model assumption:

$$\text{ilr}(\mathbf{x}) = \mathbf{A}_{\text{ilr}} \mathbf{z} + \mathbf{b},$$

where \mathbf{A}_{ilr} is a $(d-1) \times (d-1)$ full rank mixing matrix specific for a chosen ilr basis, \mathbf{b} a $d-1$ -dimensional location vector, and $\mathbf{z} = (z_1, \dots, z_{d-1})^\top$ a random vector with independent components, which are standardized so that $\mathbf{E}(\mathbf{z}) = \mathbf{0}$ and $\text{COV}(\mathbf{z}) = \mathbf{I}_{d-1}$. When the unmixing matrix \mathbf{W}_{ilr} is estimated using one of the ICA methods described in Sect. 2, the system of independent components is given by

$$\mathbf{z} = \mathbf{W}_{\text{ilr}} (\text{ilr}(\mathbf{x}) - \mathbf{b}) = \mathbf{W}_{\text{ilr}} (\mathbf{V}^\top \text{clr}(\mathbf{x}) - \mathbf{b}).$$

As ilr coordinates are not directly related to the dominance of the original parts within the considered composition, the relationship between ilr and clr spaces can be exploited yielding a $(d - 1) \times d$ “clr” loading matrix $\mathbf{W}_{\text{clr}} = \mathbf{W}_{\text{ilr}}\mathbf{V}^T$, allowing interpretation of the independent components in the clr space. In the context of principal component analysis performed in the clr space, principal components lead to a new system of ilr coordinates (Pawlowsky-Glahn et al. 2011). This is not the case for ICA, as the unmixing matrix \mathbf{W}_{ilr} (and consequently also \mathbf{W}_{clr}) is generally not restricted to be orthogonal. Even if the independent component model does not hold, ICA transformations remain affine equivariant which means that \mathbf{z} can be seen as an intrinsic data representation with a coordinate system, whose components are as independent as possible.

After performing ICA, one is usually interested in either using \mathbf{z} itself for further analysis, such as classification and outlier identification, with possible interpretation in ilr or clr space using the former defined loading matrices \mathbf{W}_{ilr} or \mathbf{W}_{clr} , or, using ICA for noise or artifact removal. For that purpose, the components of \mathbf{z} are divided into a signal part \mathbf{z}_s and a noise/artifact part \mathbf{z}_n . This defines also the partition of the unmixing matrix \mathbf{W}_{ilr} into $\mathbf{W}_{\text{ilr}}^s$ and $\mathbf{W}_{\text{ilr}}^n$ and the mixing matrix $\mathbf{A}_{\text{ilr}} = (\mathbf{W}_{\text{ilr}})^{-1}$ into $\mathbf{A}_{\text{ilr}}^s$ and $\mathbf{A}_{\text{ilr}}^n$. $\mathbf{A}_{\text{ilr}}^s$ is formed only by those columns of \mathbf{A}_{ilr} that correspond to the signal components \mathbf{z}_s . The pure signal can then be restored in the ilr, clr, and original space by using

$$\text{ilr}(\mathbf{x})_s = \mathbf{A}_{\text{ilr}}^s \mathbf{z}_s + \mathbf{b}, \quad \text{clr}(\mathbf{x})_s = \mathbf{V} (\mathbf{A}_{\text{ilr}}^s \mathbf{z}_s + \mathbf{b}), \quad \text{and} \quad \mathbf{x}_s = \exp [\mathbf{V} (\mathbf{A}_{\text{ilr}}^s \mathbf{z}_s + \mathbf{b})],$$

respectively.

5 A Case Study in Metabolomics

In order to demonstrate the above-described methods, the data from a neonatal screening program in the Czech Republic was analyzed. Anonymous data were obtained from a retrospective study approved by the Ethics Committee of the University Hospital Olomouc which was part of a larger international study described in Fleischman et al. (2013). Newborn screening is a preventive program that allows for early detection of a selected spectrum of inborn metabolic diseases. At an age of 48–72 hours after birth, several drops of blood from the heel of the child were sampled on a special paper and sent for analysis to the screening laboratory. The data at hand were constituted by the metabolite profile of over 10 000 healthy newborns. For each neonate, the values of 48 metabolites were measured. Moreover, information about sex and birth weight was available. More specifically, the birth weight ranged from 300 to 5 570 grams and for newborns with very low birth weight (less than 1500 grams) a different metabolite structure can be expected, due to their prematurity and the artificial nutrition they receive. One of the main goals of metabolomics is to investigate interactions between metabolites, their dynamic changes, and responses to stimuli. Biofluids, e.g. blood or urine, and also tissues are used for the analysis. On the one

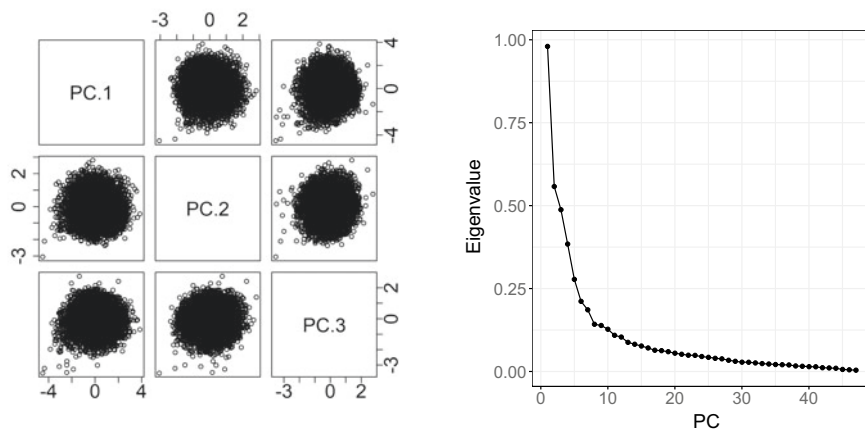


Fig. 1 Scatterplots of the first three principal components resulting from the compositional PCA (left) and scree plot of the respectively explained variability (right)

hand, the most frequently used approach for the data analysis is done through comparison of absolute values of biomarkers and reference ranges (data from the healthy population). On the other hand, the new trend of data evaluation is based on the use of ratios of metabolite data. Relative changes are more relevant/informative than absolute values in diagnostics based on profiling. Therefore, metabolomic data can be considered as observations carrying relative information, i.e. as compositional data (Kalivodová et al. 2018), and as such the above-discussed methods can be applied.

The following analysis was carried out in R 3.6.1 (R Core Team 2019) with the help of the packages JADE (Miettinen et al. 2017), fICA (Miettinen et al. 2018), compositions (van den Boogaart et al. 2019), and robCompositions (Templ et al. 2011). As the first step, standard principal component analysis (PCA) was performed on the clr transformed data. There were no significant patterns visible within the first three principal components; see Fig. 1, left. The whole dataset forms one quite compact cluster with no outliers. Moreover, the variance explained by the first components is low (around 20 % for the first PC) (Fig. 1, right), and therefore PCA does not seem to deal well with the issue of outlier detection, grouping, as well as dimension reduction in that case.

As PCA seems not to reveal any clear structure, we applied FOBI, k-JADE, with $k = 5$, and adaptive deflation-based FastICA to the ilr representation of the data (the dimension $p = 47$ was already too large for JADE). For easier comparison, the components from all three ICA methods were ordered according to their kurtosis values. As all three ICA methods showed similar results, we focus our presentation and discussion of the components on those from adaptive deflation-based FastICA.

Due to the kurtosis ordering, the first components show heavy-tailed distributions, and they are expected to find outliers or small groupings, while the last components show light-tailed distributions and hence might find more balanced groupings. Scores of the first and last three independent components are plotted in Fig. 2, and the chosen

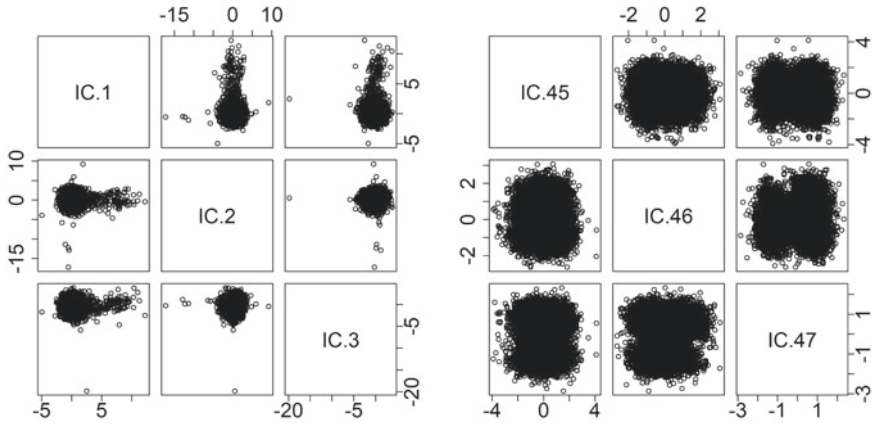


Fig. 2 Scatterplots of the first (left) and last (right) three independent components resulting from the compositional FastICA, using adaptive deflation-based FastICA

non-linearities are given in Table 2 for all independent components. According to the left plot of Fig. 2, one outlier is clearly detected due to its high negative value in the third component (IC.3). According to its loadings, which are collected in Table 3, IC.3 mostly reflects the relative dominance (with respect to concentrations of all 48 measured metabolites) of phenylalanine (Phe), hexadecanoylcarnitine (C16), octadecanoylcarnitine (C18:1), valine (Val), and hexadecenoyl- and octadecanoylcarnitines in the form of C16:1 and C18, respectively, when the higher dominance of the first three metabolites results in a decrease of IC.3 and vice versa for the last three stated metabolites. The high loadings of the clr coefficients of these six metabolites imply that IC.3 reflects mostly (but not solely) the balance between subcompositions formed by Phe, C16, C18:1, and Val, C16:1, C18. The value of this balance was for the outlier significantly lower than that within the rest of the sample. After a deeper investigation of the outlying sample, it turned out that it belongs to a newborn suffering from Phenylketonuria, a metabolic disease which is typically followed by distinctly high absolute blood concentrations of phenylalanine. The measured value was 1014.7 $\mu\text{mol/l}$, which significantly exceeds the upper norm value set on 120 $\mu\text{mol/l}$ (van Wegberg et al. 2017) and which is represented with the respective high clr value 6.76. The levels of the remaining metabolites were comparable with the other samples, but particularly the atypical high dominance of Phe over all measured metabolites, which for the rest of samples ranged from 5.72 to 3.58 for their clr values, resulted in the high negative value of the third component, and therefore clear identification of this non-standard observation.

The next interesting feature is presented by IC.1. According to Fig. 2, the values of this component are not very homogeneous across the whole dataset and therefore some specific groups of neonates might be identified. A deeper graphical analysis of the first component (presented in Fig. 3) shows that for newborns with a birth weight smaller than 1500 grams, higher values of IC.1 are typical. The independent

Table 2 Chosen non-linearities g_i for each independent component computed with the adaptive deflation-based FastICA algorithm. Non-linearities are ordered according to kurtosis values of the corresponding ICs. In the original ordering, IC.44 was the last component, thus no non-linearity is given. See Table 1 for the definitions of the functions g_i

IC	g_i	IC	g_i	IC	g_i	IC	g_i	IC	g_i
IC.1	g_2	IC.11	g_2	IC.21	g_5	IC.31	g_1	IC.41	g_5
IC.2	g_2	IC.12	g_9	IC.22	g_6	IC.32	g_5	IC.42	g_4
IC.3	g_6	IC.13	g_9	IC.23	g_9	IC.33	g_{14}	IC.43	g_4
IC.4	g_2	IC.14	g_8	IC.24	g_1	IC.34	g_5	IC.44	–
IC.5	g_6	IC.15	g_6	IC.25	g_1	IC.35	g_5	IC.45	g_5
IC.6	g_8	IC.16	g_{10}	IC.26	g_4	IC.36	g_5	IC.46	g_3
IC.7	g_8	IC.17	g_5	IC.27	g_{14}	IC.37	g_{12}	IC.47	g_6
IC.8	g_8	IC.18	g_6	IC.28	g_5	IC.38	g_{14}		
IC.9	g_9	IC.19	g_{10}	IC.29	g_{10}	IC.39	g_4		
IC.10	g_8	IC.20	g_6	IC.30	g_8	IC.40	g_{11}		

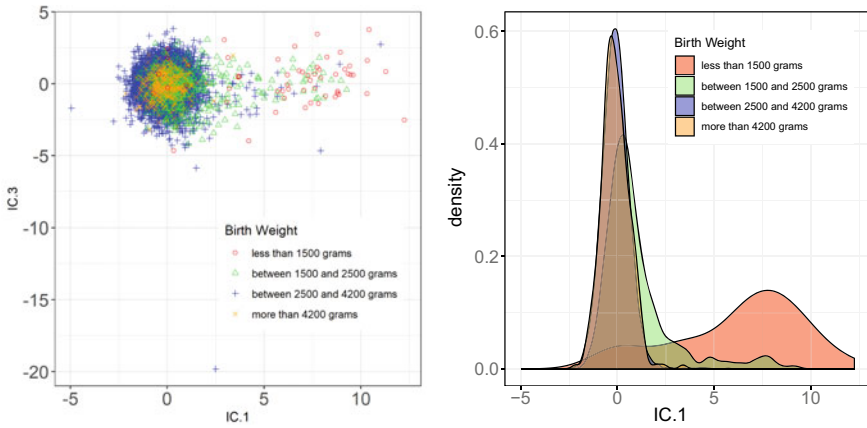


Fig. 3 Scatterplots of IC.1 and IC.3 (left) and the kernel density plot of IC.3 (right) with the groups defined according to the birth weight

component IC.1 is mostly formed by clr values of acylcarnitines dodecanoylcarnitine (C12), C16, and C18:1, whose high relative dominance over all measured metabolites results in low values of the component and, e.g. clr values of acylcarnitines isovalerylcarnitine/methylbutyrylcarnitine C5, and linoleoylcarnitine (C18:2) increase the IC.1 values. Even though there are also other metabolites contributing with a high weight to the values of IC.1 (all clr loadings are collected in Table 3), the clr values of the selected ones systematically differ for the group of the newborns with low birth weight, and therefore these acylcarnitines seem to be responsible for their separation from the remaining neonates. The differences in the selected metabolites are clearly visible in Fig. 4. Let us stress here that the immature neonates tend

to have different diet supplementation, therefore the metabolic profile can substantially differ within this group, but despite the proposed ICA method being able to find some similar patterns, detect the important metabolites, and separate the low birth weight newborns from the remaining ones. More specifically, artificial nutrition consists of amino acids, lipids, sugars, vitamins, etc. Essential unsaturated fatty acids including linoleic acid may be responsible for increased C18:2. The increased blood concentration of the long-chain acylcarnitines (C12, 16, C18:1) as well as of the short-chain C5 carnitine, which then results in high respective *clr* values, corresponds with previous studies. In Gucciardi et al. (2015), the significantly lower amounts of acylcarnitines except the branched-chain acylcarnitines (e.g. C5), which were significantly higher in preterm infants, were described. The latter mentioned are direct products of branched-chain amino acid (BCAA) catabolism, therefore its elevated levels may be related to BCAA overfeeding (Gucciardi et al. 2015; Wilson et al. 2014). The difference of several amino acids measured for the premature newborns compared to the others agrees with findings in Wilson et al. (2014), where increased levels of several amino acids (arginine, leucine, Orn, Phe, and Val) in the blood spots of premature infants were described. This observation may be related to the catabolic state of organisms in these children, amino acid supplementation, and immaturity of preterm infants (hepatic maturation, renal insufficiency, etc.) (Wilson et al. 2014; te Braake et al. 2005). The raw concentrations of valine (Val) and leucine/isoleucine (Xle) are known to be highly positively correlated, therefore the opposite signs of the respective loadings of IC.1 seem to be counter-intuitive at the first glance. However, the values of the loadings suggest that the resulting value of IC.1 is affected by the difference of *clr* values of the respective metabolites, or equivalently by the log-ratio of their measured concentrations, when the higher relative dominance of Val over Xle results in a higher value of IC.1. These findings agree with the data, since slightly higher values of the Val-Xle log-ratio are typical for newborns with a low birth weight (see Fig. 4). Finally, an even more complex interpretation can be based on the *ilr* loading matrix \mathbf{W}_{ilr} . According to the values of this matrix, IC.1 is mainly influenced by the balance between C18 and subcompositions C18:1, C18:OH, C18:2, and C18:2OH. This balance corresponds to the highest positive loading, and its values are systematically higher for the group of newborns with low birth weight than for the rest of the samples.

An even better visible pattern is formed by the last independent component IC.47, which clearly divides the whole dataset into two groups as seen in Fig. 5. According to the loadings (collected in Table 3), the most contributing are *clr* values of metabolites Xle, ornithine (Orn), and lysine (Lys) with a negative effect and methionine (Met), proline (Pro), and valine (Val) with a positive one. This suggests that the value of IC.47 is highly affected by the balance between subcompositions Met, Pro, Val and Xle, Orn, Lys. The dataset is roughly separated into two groups of observations with values of IC.47 higher and lower than -0.34 ; this value was chosen as the corresponding value of IC.47 at the local minimum in the middle of the density presented in Fig. 5 (this density was computed with Gaussian kernels and a bandwidth selection with Silverman's rule of thumb). The relative dominance of the six above-mentioned metabolites itself over all measured concentrations does not significantly differ in its

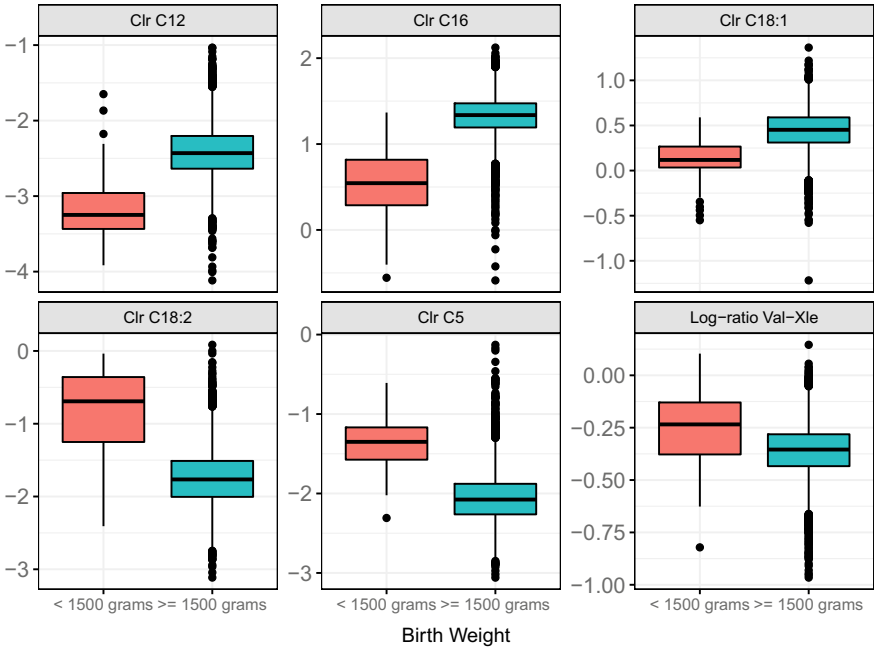
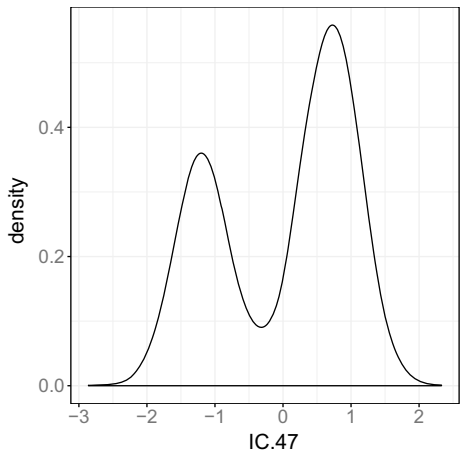


Fig. 4 Boxplots of clr as well as log-ratio values of the selected metabolites, which significantly differ for newborns with very low (< 1500 g) and normal (>= 1500 g) weight at birth

Fig. 5 Density plot of IC.47, the bimodal shape shows a clear grouping



values between the two groups. Therefore, the grouping effect of IC.47 is hidden in some of their more complex combinations, e.g. the suggested balance between Met, Pro, Val and Xle, Orn, Lys, which is distinctly higher by cases with IC.47 higher than -0.34 .

Table 3 The list of loadings for IC.1, IC.3, and IC.47 computed with the adaptive deflation-based FastICA algorithm regarding clr transformed data

	IC.1	IC.3	IC.47		IC.1	IC.3	IC.47
Ala	0.02	0.24	0.41	C5DC/C6OH	-0.08	0.14	0.04
Arg	0.13	-0.33	0.04	C5:1	-0.05	0.11	0.13
ArgSucc	0.00	-0.09	-0.01	C6	-0.22	0.22	0.07
Cit	-0.05	0.38	0.01	C8	0.84	-0.05	0.21
Glu	-0.16	0.70	0.20	C8:1	-0.05	-0.16	0.09
Gly	0.08	1.33	-0.16	C10	0.25	-0.21	-0.09
His	-0.56	0.44	0.55	C10:1	-1.05	-0.01	-0.21
Lys	0.20	0.38	-0.46	C10:2	0.28	0.20	0.06
Met	0.53	1.03	1.01	C12	-1.27	-0.01	-0.13
Orn	0.48	0.40	-0.78	C12:1	-0.09	0.10	-0.09
Phe	1.02	-7.30	-0.40	C14	-0.39	0.45	-0.38
Pro	-1.15	-0.66	0.65	C14:1	0.88	-0.46	0.04
Thr	-0.17	-1.71	-0.01	C14:2	0.15	-0.01	0.17
Trp	-1.33	-0.77	0.14	C14OH	0.03	0.27	0.08
Tyr	-0.28	1.25	-0.02	C16	-2.00	-1.49	-0.02
Val	3.09	3.86	0.59	C16:1	0.97	1.60	0.09
Xle	-1.73	-0.15	-1.03	C16OH	0.24	0.09	0.01
C0	0.51	0.60	0.17	C16:1OH	0.05	0.01	0.24
C2	0.08	-0.30	-0.25	C18	2.38	1.72	0.12
C3	-0.57	-0.34	-0.01	C18:1	-3.46	-2.64	0.00
C3DC/C4OH	0.29	-0.23	0.34	C18:2	1.80	0.41	0.16
C4	0.05	0.14	0.01	C18:1OH	-0.00	0.27	-0.15
C4DC/C5OH	0.07	0.49	-1.42	C18:2OH	0.19	0.23	-0.05
C5	0.35	-0.32	-0.08	C18OH	-0.33	0.20	0.09

6 Discussion

In this paper, we reviewed some classical independent component analysis methods and showed how these can be applied to compositional data. The key finding here is that when the ICA methods are affine equivariant it is most natural to use an ilr transformation, as the choice of the basis constituting the ilr coordinate system does not matter. For interpretability, the link between ilr coordinates and clr coefficients/variables can be easily exploited, which allows interpreting the results either in terms of the dominance of single compositional parts with respect to the whole composition, or, e.g. based on values of balances between subcompositions formed according to values of clr loadings. Finally, since the clr loadings are derived from the ilr ones, it is also possible to provide the interpretation directly in terms of the ilr coordinates. The proposed technique is demonstrated on a metabolomics dataset

where PCA, which is probably the most used multivariate transformation, reveals no specific feature on the first few components while ICA reveals several interesting features visible when exploiting the higher order moments information. Independent component analysis belongs to the larger class of blind source separation methods where for the separation of the latent components often also temporal or spatial information is used. In the context of compositional data such blind source separation methods are, for example, discussed in Nordhausen et al. (2015), Nordhausen et al. (2020). But these methods would not be applicable to the metabolomics dataset from Sect. 5 as there is no temporal or spatial information present. The current results, which were discussed mostly in terms of the relative dominance of a single compositional part respective to the highest loading of an IC, open new challenges for further research. An alternative interpretation can be reached, e.g. by adaptation the approach based on principal balances (Pawlowsky-Glahn et al. 2011). However, the loadings of ICs are in general not orthonormal and therefore the principal balances approach is not as straightforward as in the case of PCA. Finally, an extension of the dataset with a group of blood samples collected from neonates with a diagnosed disease can further prove the usefulness of the method.

Acknowledgements The work of CM and KN was supported by the Austrian Science Fund (FWF) Grant no. P31881-N32. The work of KF was supported by The Czech Science Foundation Grant no. 19-07155S. The work of HJ and AG was supported by The Czech Science Foundation Grant no. 18-12204S, the grant of Internal Grant Agency, Palacký University Olomouc no. IGA_LF_2019_006 and grant of the Ministry of Health no. NU20-08-00367. We also thank V. Pawlowsky-Glahn and an anonymous referee for their comments which helped to improve this article.

References

- Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman & Hall.
- Cardoso, J.-F. (1989). Source separation using higher order moments. *International Conference on Acoustics, Speech, and Signal Processing, 1989*, 2109–2112.
- Cardoso, J.-F., & Souloumiac, A. (1993). Blind beamforming for non-gaussian signals. *IEE Proceedings F (Radar and Signal Processing)*, 140, 362–370.
- Clarkson, D. B. (1988). A least squares version of algorithm AS 211: The FG diagonalization algorithm. *Journal of the Royal Statistical Society C*, 37, 317–321.
- Comon, P., & Jutten, C. (2010). *Handbook of blind source separation: Independent component analysis and applications*. Amsterdam: Academic.
- Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances. *Mathematical Geology*, 37, 795–828.
- Egozcue, J. J., & Pawlowsky-Glahn, V. (2019). Compositional data: The sample space and its structure. *Test*, 28, 599–638.
- Fačevicová, K., Bábek, O., Hron, K., & Kumpan, T. (2016). Element chemostratigraphy of the devonian/carboniferous boundary - a compositional approach. *Applied Geochemistry*, 75, 211–221.
- Filzmoser, P., Hron, K., & Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, 20, 621–632.
- Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis*. Cham: Springer.

- Fleischman, A., Thompson, J. D., & Glass, M. (2013). Systematic data collection to inform policy decisions: Integration of the region 4 stork (r4s) collaborative newborn screening database to improve ms/ms newborn screening in Washington State. In J. Zschocke, K. M. Gibson, G. Brown, E. Morava, & V. Peters (Eds.), *JIMD reports - case and research reports* (Vol. 13, pp. 15–21). Berlin: Springer.
- Gucciardi, A., Zaramella, P., Costa, I., Pirillo, P., Nardo, D., Naturale, M., et al. (2015). Analysis and interpretation of acylcarnitine profiles in dried blood spot and plasma of preterm and full-term newborns. *Pediatric Research*, *77*(1–1), 36–47.
- Hron, K., Filzmoser, P., & Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, *39*(5), 1115–1128.
- Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*, 626–634.
- Hyvärinen, A. (1999). Gaussian moments for noisy independent component analysis. *IEEE Signal Processing Letters*, *6*, 145–147.
- Hyvärinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*, 1483–1492.
- Illner, K., Miettinen, J., Fuchs, C., Taskinen, S., Nordhausen, K., Oja, H., et al. (2015). Model selection using limiting distributions of second-order blind source separation algorithms. *Signal Processing*, *113*, 95–103.
- Kalivodová, A., Hron, K., Filzmoser, P., Najdekr, L., Janečková, H., & Adam, T. (2018). PLS-DA for compositional data with application to metabolomics. *Journal of Chemometrics*, *29*, 21–28.
- Miettinen, J., Nordhausen, K., Oja, H., & Taskinen, S. (2013). Fast equivariant JADE. In *IEEE international conference on acoustics, speech and signal processing (ICASSP) 2013* (pp. 6153–6157).
- Miettinen, J., Nordhausen, K., Oja, H., & Taskinen, S. (2014). Deflation-based FastICA with adaptive choices of nonlinearities. *IEEE Transactions on Signal Processing*, *62*, 5716–5724.
- Miettinen, J., Nordhausen, K., Oja, H., Taskinen, S., & Virta, J. (2017). The squared symmetric FastICA estimator. *Signal Processing*, *131*, 402–411.
- Miettinen, J., Nordhausen, K., & Taskinen, S. (2017). Blind source separation based on joint diagonalization in R: The packages JADE and BSSasyp. *Journal of Statistical Software*, *76*(2), 1–31.
- Miettinen, J., Nordhausen, K., & Taskinen, S. (2018). fICA: FastICA algorithms and their improved variants. *The R Journal*, *10*, 148–158.
- Miettinen, J., Taskinen, S., Nordhausen, K., & Oja, H. (2015). Fourth moments and independent component analysis. *Statistical Science*, *30*, 372–390.
- Morais, J., Thomas-Agnan, C., & Simioni, M. (2018). Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics*, *47*, 1–25.
- Nordhausen, K., Fischer, G., & Filzmoser, P. (2020). Blind source separation for compositional time series. *To appear in Mathematical Geosciences* (pp. 1–21).
- Nordhausen, K., Ilmonen, P., Mandal, A., Oja, H., Ollila, E.: Deflation-based FastICA reloaded. In *Proceedings of 19th European signal processing conference* (pp. 1854–1858).
- Nordhausen, K., & Oja, H. (2018). Independent component analysis: A statistical perspective. *WIREs: Computational Statistics*, *10*, e1440.
- Nordhausen, K., Oja, H., Filzmoser, P., & Reimann, C. (2015). Blind source separation for spatial compositional data. *Mathematical Geosciences*, *47*, 753–770.
- Nordhausen, K., & Virta, J. (2019). An overview of properties and extensions of FOBI. *Knowledge-Based Systems*, *173*, 113–116.
- Ollila, E. (2010). The deflation-based FastICA estimator: Statistical analysis revisited. *IEEE Transactions on Signal Processing*, *58*, 1527–1541.
- Pawlowsky-Glahn, V., & Buccianti, A. (Eds.). (2011). *Compositional Data Analysis, Theory and Applications*. Chichester: Wiley.

- Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R. (2011). Principal balances. In J. J. Egozcue, R. Tolosana-Delgado, & M. I. Ortego (Eds.), *Proceedings of the 4th International Workshop on Compositional Data Analysis*.
- Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (Eds.). (2015). *Modelling and analysis of compositional data*. Chichester: Wiley.
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Te Braake, F. W., van den Akker, C. H., Wattimena, D. J., Huijmans, J. G., van Goudoever, J. B. (2005). Amino acid administration to premature infants directly after birth. *The Journal of Pediatrics*, *147*, 457–461.
- Templ, M., Hron, K., & Filzmoser, P. (2011). robCompositions: an R-package for robust statistical analysis of compositional data. In *Compositional data analysis: Theory and applications* (pp. 341–355). New York: Wiley.
- Trinh, H. T., Morais, J., Thomas-Agnan, C., & Simioni, M. (2019). Relations between socio-economic factors and nutritional diet in Vietnam from 2004 to 2014: New insights using compositional data analysis. *Statistical Methods in Medical Research*, *28*, 2305–2325.
- Tyler, D., Critchley, F., Dümbgen, L., & Oja, H. (2009). Invariant coordinate selection. *Journal of Royal Statistical Society B*, *71*, 549–592.
- van den Boogaart, K. G., Tolosana-Delgado, R., & Bren, M. (2019). *Compositions: Compositional data analysis*. R package version 1.40-3.
- van Wegberg, A., MacDonald, A., Ahring, K., Bélanger-Quintana, A., Blau, N., Bosch, A. M., et al. (2017). The complete European guidelines on phenylketonuria: Diagnosis and treatment. *Orphanet Journal of Rare Diseases*, *12*, 162.
- Virta, J., Lietzen, N., Ilmonen, P., Nordhausen, K. (2020). Fast tensorial JADE. *To appear in Scandinavian Journal of Statistics*.
- Wei, T. (2015). A convergence and asymptotic analysis of the generalized symmetric FastICA algorithm. *IEEE Transactions on Signal Processing*, *63*, 6445–6458.
- Wilson, K., Hawken, S., Ducharme, R., Potter, B. K., Little, J., Thébaud, B., et al. (2014). Metabolomics of prematurity: Analysis of patterns of amino acids, enzymes, and endocrine markers by categories of gestational age. *Pediatric Research*, *75*, 367–373.

Diet Quality and Food Sources in Vietnam: First Evidence Using Compositional Data Analysis



Michel Simioni, Huong Thi Trinh, Tuyen Thi Thanh Huynh, and Thao-Vy Vuong

Abstract Food environments have been evolving rapidly in lower-middle-income countries. Nevertheless, little is known about the impact of these changes on diet quality. Thanks to the availability of detailed data on Vietnamese household consumption, this chapter presents a set of first results on the association between food sources and diet quality. These results highlight the contrasts between three Vietnamese districts located on an urban to rural gradient. We used recent advances in compositional data analysis to take into account the compositional nature of the share data describing the different food sources: principal balances as a tool for summarizing information carried by share data and techniques to deal with observed zero-valued shares.

1 Introduction

In the face of economic development, urbanization, and tighter global connections, food systems in lower-middle-income countries have evolved (HLPE: Nutrition and food systems 2017). Food supply sources have changed and diversified. In more

Dedicated to Christine Thomas-Agnan with whom our collaboration has always been successful.

M. Simioni (✉)

MOISA, INRAE, University of Montpellier, Montpellier, France
e-mail: michel.simioni@inrae.fr

H. T. Trinh

Department of Mathematics and Statistics, Thuongmai University, Hanoi, Vietnam
e-mail: trinhthihuong@tmu.edu.vn

T. T. T. Huynh

International Center for Tropical Agriculture (CIAT)—Asia Office, Hanoi, Vietnam
e-mail: T.Huynh@cgiar.org

T.-V. Vuong

College of Agriculture and Life Sciences, Cornell University, Ithaca, USA
e-mail: vtv6@cornell.edu

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_28

547

populous urban areas, wholesalers and retailers have larger markets, and with a greater population density they can reach more consumers at a lower cost per unit. In many lower-middle-income countries, supermarkets now provide an additional food source to traditional grocery outlets, and the number of supermarkets continues to increase. This rapid change has been referred to as “food system revolution” (Reardon and Timmer 2012).

The impact of specific food sources on nutrition is not well-known. A brief overview of the literature looking at the impact of food supply on nutrition and public health indicators, such as obesity, has been provided in Qaim (2017). Existing work mainly focuses on the consequences of food purchases in supermarkets on nutritional diets in developed countries. Most of this work shows that shopping in supermarkets is associated with higher consumption of processed foods and lower consumption of unprocessed foods, including fruits and vegetables. High consumption of ultra-processed foods appears to be one of the major drivers of obesity epidemics (Monteiro et al. 2013). Research on the impact of supermarkets on consumer nutritional status in developing countries is rare (Demmler et al. 2018; Wertheim-Heck and Raneri 2019; Trinh et al. 2021).

This chapter aims to contribute to the existing literature on the impact of food sources on nutrition in Vietnam, a lower-middle-income country. Detailed data on dietary patterns of households in three Vietnamese districts makes it possible to investigate this impact according to a gradient ranging from urban to rural, in three very different food environments. The healthiness of the overall diet is assessed using the Diet Quality Index International (DQI-I) with modifications to adapt to the Vietnamese Dietary Guidelines (Kim et al. 2021). This index was chosen because it has been tested in a range of cultural contexts and validated for use in a range of countries with different dietary patterns.¹ Supply sources of the various foods eaten by households were identified during data collection. The relative importance of food sources is evaluated using their shares in total calorie intake, which are comparable across households.

The chapter first presents an exploratory analysis of food sources, highlighting differences between the three districts. Then, associations between food sources and quality of diet are investigated.

The methodological contribution of the chapter is to treat food sources as compositional data. By definition, compositional data describe parts of a whole and, consequently, convey only relative information (Pawlowsky-Glahn et al. 2015). They are usually recorded in closed forms such as proportions or percentages. Consequently, their particular numerical properties hamper the use of standard statistical methods designed for unconstrained variables. A methodology based on log-ratios has been proposed to deal with compositional data. The basic idea is to focus on the ratios between components, specifically on their log-ratios to leverage mathematical properties. Following the pioneering work of Aitchison (1986), a general framework has been proposed based on the characterization of the simplex—the sample space

¹See <https://index.nutrition.tufts.edu/data4diets/indicator/diet-quality-index-international-dqi-i?back=/data4diets/indicators>.

of compositions—and other subsets of the real space, as genuine Euclidean vector spaces equipped with their own geometry. This allows compositional data to be isometrically mapped into a space of real coordinates with respect to an orthonormal basis, where standard statistical methods such as principal component analysis and ordinary least squares can be used.

Compositional data analysis has only recently been applied to nutrition studies. In these studies, the associations between macronutrient balances and diseases were analyzed (Corrêa Leite 2016, 2019; Corrêa Leite and Prinelli 2017). They show how classical logistic regressions can be used when assessing the impact of macronutrient shares on metabolic syndromes. This approach can be extended by considering not only these shares but also total calorie intake as explanatory variables in regressions of obesity indicators (Beal et al. 2018), and by using recent results on the computation of elasticities in compositional regression models (Morais et al. 2018). Trinh et al. (2019) propose a measure of diet quality using a vector of macronutrient shares and their association to the socio-demographic characteristics of households. Moreover, balances—a central tool in compositional data analysis—has been shown to be a powerful exploratory tool when analyzing individual diets (Solans et al. 2019).

Recent advances of two important issues in compositional data analysis support the empirical analysis presented in this chapter. First, as mentioned above, compositions provide information about relative rather than absolute values. This observation has led to the development of particular methods based on the logarithms of ratios between parts (or groups of parts), which are a suitable means of transforming compositional data to allow for the use of standard statistical methods. There has been focused attention on the isometric log-ratio transformation, due to its properties documented in Egozcue et al. (2003). This transformation leads to the definition of new variables representing groups of parts and their relationships. The construction of isometric log-ratios is most often investigator-driven. Groups of components are defined based on external information, and comparisons are conducted using their geometric means. For example, it is usual to first compare fats with other macronutrients (carbohydrates and proteins), and then to compare carbohydrates and proteins, when dealing with the potential association of diet components with obesity indicators (Beal et al. 2018). In the absence of any external information, or in the presence of a large number of components, it is possible to extend the use of data-driven Principal Component Analysis (PCA) to compositional data (Aitchinson 1983). Principal components, referred to as principal component coordinates, with decreasing variances, are extracted from this method. Each principal component coordinate can be written as a log-ratio where absolute values of PCA loadings are the exponents of parts, with a loading sign defining the presence of the corresponding part, either in the numerator (if positive) or in the denominator (if negative). Principal component coordinates can be proven to fulfill all conditions for being isometric log-ratio coordinates. Nevertheless, principal component coordinates can be difficult to interpret as their numerators and denominators do not possess a clear interpretation in terms of geometric means. An intermediate approach which keeps the main properties of principal component coordinates and provides easy-to-interpret coordinates has been proposed: the principal balances approach (Pawlowsky-Glahn et al. 2011). This

data-driven approach will allow us to reveal contrasts in the sources of food supply chosen by households in the three districts. Principal balances identified during the exploratory analysis of food sources will then be used when measuring associations between food sources and diet quality.

Moreover, the log-ratio approach proposed for the statistical analysis of compositional data presents a serious limitation when certain components are zero. Zero values can be present for various reasons, as comprehensively described in Martín-Fernández et al. (2011). For instance, rounded zeros represent a prominent zero type in compositional data analysis. They occur frequently in environmental and chemical data, when either small values of components are rounded to zero, or a measurement device has a detection limit that sets values below the limit to zero. Rounded zeros cannot be considered as “true” or “essential” zeros due to the data generating process (like corner solutions in economic modeling), but rather as a result of precision issues. In other words, it is more meaningful to impute them with a reasonably small value and process the complete data set (Martín-Fernández et al. 2012). Here, dietary data has been collected using a 24-h recall period, which may not precisely represent the long-term dietary habits of the participants. Episodically consumed food is relatively likely to be misrepresented. For this reason, we propose to treat the zero values observed for these food sources as rounded zeros, and accordingly implement the imputation algorithm (Palarea-Albaladejo and Martín-Fernández 2015).

The chapter is organized as follows. Section 2 details algorithms used to compute principal balances and to deal with zero component values. Section 3 presents the framework of the study, the procedure followed to collect data, and the construction of diet quality indicator and food sources shares. Exploratory and regression analysis results are presented in Sect. 4. Concluding remarks are provided in Sect. 5.

2 Methodology

2.1 Principal Balances

Let a composition \mathbf{x} be a positive vector in D -dimensional real space, or

$$\mathbf{x} = (x_1, x_2, \dots, x_D), \text{ with } x_j > 0 \text{ for all } j = 1, 2, \dots, D, \quad (1)$$

where D is the number of components, in our case, food sources. In order to focus on the relative importance of the components, the *closure* of \mathbf{x} is commonly used:

$$\mathbf{y} = C(\mathbf{x}) \equiv \left(\frac{x_1}{\sum_{l=1}^D x_l}, \frac{x_2}{\sum_{l=1}^D x_l}, \dots, \frac{x_D}{\sum_{l=1}^D x_l} \right). \quad (2)$$

Let $y_j = x_j / \sum_{i=1}^D x_i$, $j = 1, 2, \dots, D$. By construction, \mathbf{y} is such that $\sum_{i=1}^D y_i = 1$. The vector \mathbf{y} resides in a subspace of \mathbb{R}_+^D which is constrained by positivity and a fixed sum, called the *simplex*, with operations, angles and distances different from those in the real space. For this reason, most statistical tools such as correlation or variance are meaningless when applied to \mathbf{y} .

As emphasized in the introduction of this chapter, a methodology based on log-ratios has been proposed to deal with compositional data. The general expression of a log-ratio is a log-contrast (Martín-Fernández et al. 2018):

$$\sum_{i=1}^D a_i \ln x_i = \ln \left(\prod_{i=1}^D x_i^{a_i} \right), \text{ with } \sum_{i=1}^D a_i = 0. \tag{3}$$

It is obvious that a log-contrast is a log-ratio of components because for $a_i > 0$, the corresponding component x_i appears in the numerator with the exponent a_i , but if $a_i < 0$ it appears in the denominator with the exponent $-a_i$, while for the components that do not contribute to the log-ratio $a_i = 0$ holds.

The overarching idea is to focus on the log-ratios of components and then defining mathematical properties that they must fulfill in order to define a one-to-one relationship between the simplex and the real space. It can be shown that log-ratios must satisfy the following requirements to capture all information in the compositional data set (Egozcue et al. 2003):

- They must define an orthonormal $(D - 1)$ -dimensional basis of the simplex,
- The sum of exponents in the numerator of the log-ratio must equal the sum of exponents in the denominator, and
- The sum of all squared exponents must be equal to one.

Such log-ratios are called isometric log-ratios. Results obtained by applying classical statistical techniques to isometric log-ratios can then be transferred onto compositions.

Principal balances are a particular class of isometric log-ratios. They are defined as follows (Martín-Fernández et al. 2018): principal balances are log-linear functions $z_k = \sum_{i=1}^D a_{ki} \ln x_i$, $k = 1, \dots, D - 1$, such that the vectors $\mathbf{a}_k = (a_{k1}, \dots, a_{kD})$ are constant and maximize the variance:

$$\text{var} \left(\sum_{i=1}^D a_{ki} \ln x_i \right), \tag{4}$$

under three conditions:

1. (*balance condition*) For $k = 1, \dots, D - 1$, the coefficients a_{ki} take one of the three values $(-c_1, 0, c_2)$, for some strictly positive c_1 and c_2 ,
2. (*zero sum and unit norm conditions*) for $k = 1, \dots, D - 1$, \mathbf{a}_k satisfies $\sum_{i=1}^D a_{ki} = 0$ and $\sum_{i=1}^D a_{ki}^2 = 1$, and

3. (*orthogonality condition*) for $k = 1, \dots, D - 1$, \mathbf{a}_k is orthogonal to previous $\mathbf{a}_{k-1}, \dots, \mathbf{a}_2, \mathbf{a}_1$, i.e.

$$\sum_{i=1}^D a_{ki} a_{(k-l)i} = 0, \quad l = 1, \dots, k - 1. \tag{5}$$

Each principal balance can be written as the following log-ratio:

$$z_k = \prod_{i=1}^{r_k} x_i^{c_{ik}^*} / \prod_{j=1}^{s_k} x_j^{c_{jk}^*} \tag{6}$$

where c_{1k}^* and c_{2k}^* are the solutions of the previous maximization program, and r_k (resp. s_k) is the number of components to which a positive (resp. negative) coefficient is associated. Thus principal balances compare groups of components using their geometric means. Moreover, given a sample of D -component random composition, it can be shown that, just like for principal component analysis, the total variance of the sample² can be decomposed into the sum of the variances associated with the principal balances. The first principal balance has maximum sample variance, and the k th principal balance has the maximum variance conditional to its balancing element, being orthogonal to the previous $(k - 1)$ balancing elements.

An algorithm to build principal balances has been proposed Martín-Fernández et al. (2018), and its implementation can be found in the BALANCE package of the R computing language (Quinn 2018).

Hereafter, principal balances calculated from data on food sources will be used as diet quality predictors. Associations between diet quality and food sources will be assessed using the OLS estimation technique.³

2.2 Zeros

The presence of zero components in compositional data precludes log-ratio calculations because the logarithm of zero is undefined. Hereafter, we consider the case where some components of a composition are believed to be present, but are not observed due to randomness or limitations of measurement.⁴ More precisely, we consider rounded zeros which are defined as follows: for each component x_j in a

²Total variance in a compositional data set is defined as the sum of variances of all centered log-ratios, i.e. the log-ratios of components to their geometric mean.

³See Pawlowsky-Glahn et al. (2015) for an introduction to linear regression models with compositional data.

⁴For instance, rounded zeros are often observed when data are collected using a retrospective food frequency questionnaire expressed in daily or weekly portions. This kind of questionnaire is known to fail to record food groups that are consumed infrequently.

composition $\mathbf{x} = (x_1, \dots, x_D)$, there is a threshold t_j such that observations with $x_{jn} < t_j$ are rounded to zero.⁵

Rounded zeros can be considered as missing values. Thus, imputation techniques can be used to replace these missing values with imputed values. The Expectation-Maximization (*EM*) algorithm proposed by Dempster et al. (2021) provides a reliable parametric method for missing data in real space. This algorithm was adapted to compositional data by Palarea-Albaladejo et al. (2007); Palarea-Albaladejo and Martín-Fernández (2008).

The log-ratio *EM* algorithm proceeds as follows when facing rounded zeros. Compositional data are first transformed into real data using an additive log-ratio (*alr*) transformation⁶:

$$q_j = \ln \left(\frac{x_j}{x_D} \right), j = 1, \dots, D - 1, \tag{7}$$

where, for ease of presentation, we assume that the last component does not exhibit any rounded zeros.⁷ Indeed, direct application of the *EM* algorithm to compositional data has been shown to generate serious distortions in imputed data (Martín-Fernández et al. 2003): zeros can be replaced by negative values or by values larger than specified thresholds, and the constant-sum constraint is not respected.

Observation of rounded zeros then corresponds to a censoring pattern with observed, \mathbf{q}_{obs} , and unobserved components, \mathbf{q}_{non} . The log-ratio *EM* algorithm is an iterative procedure where two steps are involved at each iteration t :

- Expectation-step or *E*-step: given the estimated parameters $\widehat{\theta}^{(t)}$, compute

$$\widehat{\mathbf{q}}_{\text{non}} = E \left[\mathbf{q}_{\text{non}} \mid \mathbf{q}_{\text{obs}}, \mathbf{q}_{\text{non}} < \psi; \widehat{\theta}^{(t)} \right], \text{ and}$$

- Maximization-step or *M*-step: find a new estimate $\widehat{\theta}^{(t+1)}$ based on the completed data set $[\widehat{\mathbf{q}}_{\text{non}}, \mathbf{q}_{\text{obs}}]$,

where ψ denotes the vector of censoring points, i.e. $\psi_j = \ln(t_j/x_D)$ when component j exhibits rounded zeros. More precisely, assuming multivariate normality of *alr*-transformed data, the expected value of q_{non} is computed at the t th iteration in the *E*-step as

$$\widehat{q}_{\text{non}}^{(t)} = \mathbf{q}_{\text{obs}} \widehat{\beta}^{(t)} - \widehat{\sigma}^{(t)} \frac{\phi \left((\psi - \mathbf{q}_{\text{obs}} \widehat{\beta}^{(t)}) / \widehat{\sigma}^{(t)} \right)}{\Phi \left((\psi - \mathbf{q}_{\text{obs}} \widehat{\beta}^{(t)}) / \widehat{\sigma}^{(t)} \right)}, \tag{8}$$

⁵In practice, threshold values are usually set below the smallest values that have been observed for the different components.

⁶An isometric log-ratio (*ilr*) transformation can also be used, both transformations producing same results. We focus here on the *alr* formulation to simplify the presentation.

⁷It can be shown that imputation results do not depend on the component used as an *alr* divisor (Palarea-Albaladejo and Martín-Fernández 2008). Algorithm implementation only needs one component with no rounded zeros to be able to compute *alr*-ratios.

where $\widehat{\beta}^{(t)}$ and $\widehat{\sigma}^{(t)}$ are maximum-likelihood estimates of the regression parameters at the $t - 1$ th iteration in the M -step. $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution functions of the standard normal distribution, respectively. The ratio on the right-hand side of Eq. (8)—the inverse Mills ratio (Amemiya 1985)—which is evaluated at $((\psi - \mathbf{q}_{\text{obs}}\widehat{\beta}^{(t)})/\widehat{\sigma}^{(t)})$, represents the censoring threshold required for Eq. (8) to consistently produce values below the ratio, as expected.

The log-ratio EM algorithm is iteratively repeated until convergence, i.e. the algorithm stops when the distance between $\widehat{\theta}^{(t)}$ and $\widehat{\theta}^{(t+1)}$ is lower than a fixed tolerance level. Once the convergence has been reached, the last completed data set obtained at the E -step is transformed back to the simplex using the inverse *alr* transformation, resulting in a completed compositional data set without zeros.

The log-ratio EM algorithm has been shown to produce only minimal distortion (Palarea-Albaladejo et al. 2007; Palarea-Albaladejo and Martín-Fernández 2008). This algorithm does not alter the log-ratios of non-zero components. The relative information conveyed by these components is fully preserved. Absolute values are modified to accommodate the closure, but the compositional information stays the same after imputation.

The log-ratio EM algorithm has been implemented in the `ZCOMPOSITIONS` package of the R computing language (Palarea-Albaladejo and Martín-Fernández 2015).

3 Data

3.1 Data Collection

Data was collected under the Consortium of International Agricultural Research Centers (CGIAR) Research Program on the Agriculture for Nutrition and Health (A4NH) project (Huynh et al. 2021). The aim of the project was to elucidate specific features of local Vietnamese food systems along a rural to urban gradient, focusing on (i) diets and nutrition, (ii) consumer behavior, and (iii) food flows (food sources).⁸ Three Vietnamese districts were selected to capture the rural to urban gradient. Figure 1 gives their geographical locations. Their main characteristics are as follows:

- Moc Chau District, a rural site, located in Son La Province—This district is characterized by its high diversity of ethnic groups, and a large volume of agricultural production for both home consumption and income generation.
- Dong Anh District, a peri-urban site, located in Hanoi Province—This district is characterized by rapid urbanization, intensive crop-livestock production, and food transformation next to the urban area, and a typical peri-urban population with a high percentage of migrants and a commuting labor force.

⁸The study protocol was approved by the Medical Research Ethics Committee of the National Institute of Nutrition of Vietnam (Number 223/VDD-QLKH).

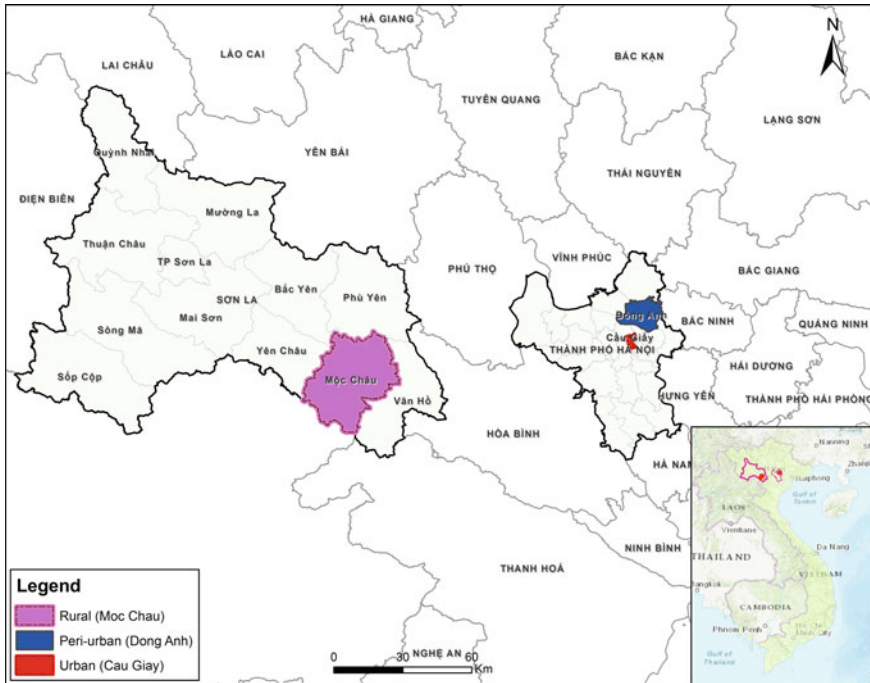


Fig. 1 Geographical location of the three districts in Vietnam. Credit: Phan Van Trong

- Cau Giay District, an urban site, located also in Hanoi Province—This district represents a typical urban space with mixed retail outlets ranging from street markets, formal wet markets to supermarkets.

Households were sampled in the three districts as follows:

- First, within the rural district (Moc Chau), the two main urban agglomerations were excluded to guarantee that the sampling target was, in fact, the rural population. In the two other districts (Dong Anh and Cau Giay), neighborhoods with high-income households were excluded to only sample primarily middle- and low-income households who were the targeted population of the project. In spite of these exclusions, the target population includes the majority of the population within the three districts.
- Thirty communes were randomly selected as Primary Sampling Units (PSUs) using a probability proportional to size (PPS) procedure where higher population villages had a greater probability of being selected, to meet the general purpose of the survey.⁹ For consumer behavior and dietary assessment, 10 PSUs were then randomly selected from the 30 previously selected ones. Once a PSU was

⁹For instance, the nutrition status (anthropometric) component was also collected but is not used in this chapter.

selected, a rapid enumeration of households was conducted and information about household composition (to identify parents with children up to the age of 5) was obtained from district health centers.

- Households were then selected at random from the lists collected at the previous step. Participation in the survey was completely voluntary, selected participants being asked to give their consent to participate in the survey. Substitutes for households that were originally chosen but those that did not want to participate in the survey were randomly drawn from the same lists.

For the consumer behavior component of the study, the interviewee was the person who was mainly responsible for household food purchases and/or preparation, hereafter defined as the household representative.

The wider project that this study is a part of is aimed at evaluating childhood malnutrition. This project focuses on children under the age of 5 because malnutrition is among the main causes of death for this group (nearly half of deaths), and is one of the most common factors threatening the lives and health of children in developing countries like Vietnam. To assess their dietary conditions, 24-h recall interviews¹⁰ were conducted with three individuals in each household: a child under 5 years old, their father, and their mother. Each survey participant was interviewed on 2 days. The first recall was conducted on a weekday and the second recall was done on either a weekday or a weekend. In addition, the source of each ingredient used when preparing meals was identified.

3.2 Sample Characteristics

Data on household characteristics: gender, age, and education level of household representative, with household size and income level, were also collected. Table 1 reports descriptive statistics for the three districts. As expected, the urban district is characterized by household representatives with higher education levels and by wealthier households relative to the other two districts.

3.3 Nutrition Knowledge

As emphasized by Wertheim-Heck and Raneri (2019), understanding food safety risk perceptions and trust in food safety as well as nutrition knowledge and attitudes are useful for gaining insight into peoples' personal determinants of their food shopping

¹⁰This method provides comprehensive, quantitative information on individual diets by querying respondents about the type and quantity of all food and beverages consumed during the previous 24-h period (Gibson et al. 2017).

and dietary habits. Food safety risk perceptions have been demonstrated to impact trust in food safety and ultimately shopping practices, what and when products are being purchased, and from whom or where.

The nutrition knowledge score was measured through a series of 30 questions about diet and nutrition. Questions about the following topics were addressed: micronutrient attitudes, diet diversity knowledge, diet diversity attitudes, undernutrition knowledge, undernutrition attitudes, overnutrition knowledge, and overnutrition attitudes. The questions were in line with the Nutrition Knowledge, Attitudes, and Practices (KAP) manual published by FAO (Marias and Glasauer 2014). Each correct response was worth 1 point and an incorrect one was worth 0. The final score was the sum of these scores, which was then converted to a scale from 0 to 1, with 1 as the maximum total score.

Table 1 reports descriptive statistics on nutritional knowledge scores for the three districts. Sensitivity to nutrition-related problems, as measured by the KAP score,

Table 1 Description of the sample

Variable	Urban	Peri-Urban	Rural
Age of household representative	33.3 (5.8)	30.6 (6.0)	29.6 (6.8)
<i>Gender of household representative</i>			
Male (%)	49.08	50.00	49.77
Female (%)	50.92	50.00	50.23
<i>Education of household representative</i>			
Primary school or no formal education (%)	1.40	9.25	35.32
Secondary school (%)	4.67	21.39	35.32
High school (%)	11.21	32.37	24.31
University and college (%)	82.71	36.99	5.05
Household size	4.88 (1.49)	5.54 (1.35)	4.90 (1.22)
<i>Income (millions of VND)</i>			
Less than 7 (%)	7.34	26.63	72.4
From 7 to 11 (%)	22.48	32.07	22.62
From 11 to 20 (%)	31.65	22.28	3.62
Greater than 20 (%)	38.53	19.02	1.36
Nutrition knowledge score	0.7 (0.2)	0.6 (0.2)	0.5 (0.2)
Number of observations	214	184	221

Means and standard deviation (in parentheses) are reported for age, household size, and nutrition knowledge score

decreased along the urban to rural gradient. One-way analysis of variance and Tukey's range test show significant differences in nutritional knowledge scores between the three districts.¹¹

3.4 Food Sources

In addition to the specific food items and their quantities, the 24-h recall also documented the sources of food consumed, which include "supermarket," "convenience shop," "specialized shop," "wet market," "own production," and "other" sources (detailed description in Table 2). These food sources can be categorized in different ways, either between formal-licensed food retail business versus informal-self-organized unlicensed food retail business, or between modern, hybrid, and traditional outlets (Wertheim-Heck and Raneri 2019). Many low- and middle-income countries such as Vietnam are experiencing the development of a dual system, with supermarkets taking a larger share of household expenditures on non-staple and processed goods, while meat and fruits are mainly bought in traditional, smaller grocery stores and fresh markets (Food and Agriculture Organization of the United Nations 2013).

Although the collected data does not break down the share of food goods purchased by households from each food source, the share of each food source in household total calorie intake can be computed as an approximation.

Some households exhibit zero shares for some food sources, and the presence of zeros for food sources varies from district to district. Hence, zero shares are observed for most food sources with the exception of "wet market" in urban and peri-urban districts and "own production" in rural ones. The following question then arises: how should the observed zeros be interpreted? By construction, the zero shares capture the fact that no food coming from the corresponding food source has been eaten during the period covered by the 24-h recall survey. A food source share equal to zero does not indicate that the household never makes its purchases there, especially since the food source is present in the household food environment (for instance, there are even two supermarkets in the rural district) and since small, but non-zero, shares for the considered food source are also observed for many households in the same district. It therefore does not seem realistic to consider observed zero values as "true" zeros, reflecting either no potential access to the food source (food deserts) or a corner solution in the household's choice of food sources. Instead, we treat observed zeros as "rounded" zeros, and using the log-ratio *EM* algorithm proposed by Palarea-Albaladejo et al. (2007), we replaced values of zero with imputed values.

Table 3 reports the main characteristics of the distributions of food source shares after imputation. "Wet market" appears to be the most important food source in the urban district, followed by "specialized shop." While "wet market" is still the predominant food source in peri-urban districts, "own production" is now the second food source followed by "convenience shop." "Own production" is the main food

¹¹Results are available from the authors upon request.

Table 2 Food sources and their characteristics

Food source	Characteristics	Main food items provided by this source
Supermarket	Licensed food retail business	Processed foods and beverages
	Large variety of branded products	Imported and frozen food Fruits and vegetables
Convenience shop	Small grocery store	Processed foods and beverages
	Independently owned and operated	Rice, noodles, flour Eggs, cakes
Specialized shop	Small modern store	Fresh and/or organic produce
	Clear price tag and promotion	Fruits and vegetables
	Employees run the store	Meat and imported foods
Wet market	Managed by local authorities	Variety of fresh food products
	Rental of a tiny space (1-5m ²)	
	No need of a business license	
Own production (including gifts)	Own plot or garden	Mainly rice, maize, potatoes
	Own farm	Fruits and vegetables Pig, chicken, eggs
Other sources	Direct farm supply	Mainly vegetables, rice, meat
	Online shopping	Cakes
	Wild food	

source in the rural district, followed by “convenience shop.” “Supermarket” only appears as a marginal food source in the peri-urban and rural districts, while this food source is more present in the urban district.

3.5 Diet Quality

Diet quality is an important measure for the understanding of food security because of the synergistic nature of micro- and macronutrients and the association of healthy diet patterns with reduced risk of diet-related disease and illness. Diet quality can be measured using the Diet Quality Index International (DQI-I) (Kim et al. 2021). Their aim was to capture the fact that food intake patterns are likely to be more heterogeneous globally than nutrient intake patterns. Therefore, they proposed an index which incorporates both nutrient and food perspectives of the diet in the assessment, providing a means to better describe the diversity of consumption from country to country.

The DQI-I focuses on four major aspects of a high-quality, healthy diet: variety, adequacy, moderation, and overall balance, covering nutritional concerns of both

Table 3 Description of food source shares (percentage) after imputation

District	Minimum	Median	Mean	Maximum
<i>Urban district</i>				
Supermarket	0.10	2.30	1.60	92.80
Convenience shop	≤0.01	0.25	0.81	79.20
Specialized shop	0.10	10.70	4.47	93.10
Wet market	0.40	44.80	35.80	100
Own Production	0.03	0.10	0.64	87.90
Others	≤0.01	4.65	2.25	50.70
<i>Peri-urban district</i>				
Supermarket	≤0.01	≤0.01	≤ 0.01	77.50
Convenience shop	0.10	5.60	4.04	80.60
Specialized shop	0.02	2.30	1.09	94.40
Wet market	1.30	332.25	29.96	99.20
Own Production	0.10	32.00	10.17	88.30
Others	≤0.01	2.25	0.54	45.60
<i>Rural district</i>				
Supermarket	≤0.01	≤0.01	≤0.01	34.10
Convenience shop	0.10	10.20	6.41	99.50
Specialized shop	≤0.01	0.13	0.26	88.70
Wet market	≤0.01	2.50	1.34	93.30
Own Production	0.20	73.50	54.34	100
Others	≤0.01	≤0.01	≤0.01	24.10

Reported means are geometric ones

developed and developing countries. Specific components are assessed and scored for each aspect, as summarized in Table 4. The total DQI-I is the sum of these component scores, producing a total score between 0 and 100. A higher score indicates a higher quality diet. Macronutrient and micronutrient intakes required for the computation DQI-I were computed using the 2017 Vietnamese Food Composition Table (Viet Nam National Institute of Nutrition 2017) and scoring procedures were adapted to the Vietnamese context using Vietnamese dietary guidelines (Ministry of Health 2013).

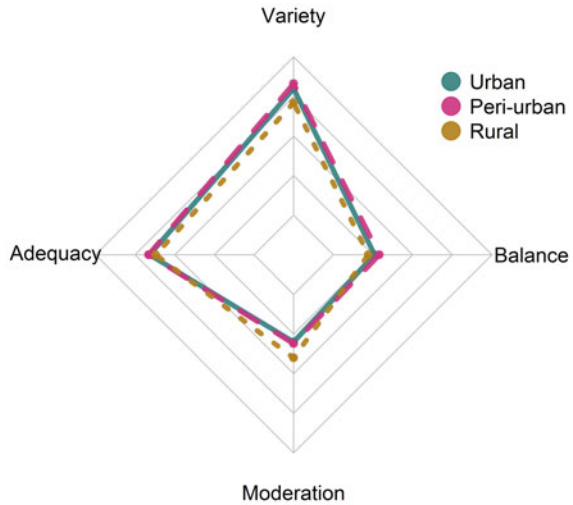
Quality of diet appears to be better in the peri-urban district, where the mean DQI-I score was 55.5% of the maximum possible score (with a standard deviation of 7.9) than in the urban and rural districts, with mean scores equal to 53.9% (9.7) and 53.2% (9.0), respectively. By way of comparison, the mean DQI-I score was approximately 60% of the maximum score in both China and in the US in the mid-1990s (Kim et al. 2021).

Table 4 DQI-I component definitions

DQI-I component	Grouping of diet quality component	Scoring criteria	Score
Variety-food groups	5 food groups: meat/poultry/ fish/egg, dairy/beans, grains, fruits, and vegetables	Each food group awarded 0 or 3 pts. 3 points awarded if at least 1 item from that group was consumed	0–15
Variety-protein sources	6 sources: meat, poultry, fish, dairy, beans, eggs	≥3 sources consumed: 5 pts 2 sources consumed: 3 pts 1 source consumed: 1 pts 0 sources consumed: 0 pts	0–5
Adequacy	8 groups: vegetables, fruit, grain, fiber, protein, iron, calcium, vitamin C	Between 0 and 5 points awarded for each of the 8 adequacy groups, depending on percentage of or Recommended Daily Allowances (RDA) met	0–40
Moderation	6 groups: total fat, saturated fat, cholesterol, sodium, empty calorie foods	Between 0 and 6 points awarded for each of the 5 moderation groups, depending on percentage of RDA met	0–30
Balance	2 groups: macronutrient ratio, fatty acid ratio	Between 0 and 6 points awarded, depending on ratio of macronutrients and between 0 and 4 points awarded depending on ratio of fatty acids	0–10
DQI-I		Grand total =	0–100

Mean scores for components of DQI-I by district are reported in Fig. 2, expressed as a percentage of the maximum values they can reach. Urban and peri-urban districts have similar profiles in terms of variety, adequacy, and moderation, with mean variety scores of 80% and 83%, mean adequacy scores of 66% and 67%, and mean moderation scores of 30% and 31%, of the corresponding maximum scores, respectively. Mean variety and adequacy scores of the rural district are lower: 71% and 62%, respectively, while higher in the case of moderation (40%). Balance is better fulfilled in the peri-urban district with a mean score of 29% (26% and 22% in urban and rural districts, respectively).

Fig. 2 Average DQI-I components by district



4 Results

4.1 Exploratory Analysis

Principal balances have emerged as a relevant tool for extracting compositional information to use in the statistical modeling of compositional data (Solans et al. 2019). The technique is used here to summarize the information provided by the data on food sources in the three districts (after imputation of zero values). Table 5 shows the total variance and percentages of explained variance by principal balances for each district. The structure of principal balances and their distributions are shown in Fig. 3. Two figures are thus reported for each district. The first figure on the left shows how some food sources are contrasted against others for each principal balance. Food sources appearing in the numerator (resp. denominator) of a principal balance are represented with points (resp. triangles). For instance, the first principal balance (z_1) shows how the “supermarket” and “other” sources are contrasted against the remaining food sources, for the rural district. The figure on the right shows the sample distributions of the principal balances, using boxplots. Note, for instance, that sample values of the first principal balance are always positive for the rural district, meaning that households living in that district always consume, on average, more calories from “wet market,” “specialized shop,” “convenience shop,” and “own production” sources than from “supermarket” and “other” sources.

The results show significant contrasts between the districts. Consider first the peri-urban district (panel (b) of Fig. 3). The first principal balance, which captures 86.77% of the total variance, compares “supermarket” with all other food sources. This result could be expected as “supermarket” appears to be the most marginal food source in this district. The importance of the contribution of the first principal balance

Table 5 Percentage of total variance explained by each principal balance

Principal balance	Urban	Peri-urban	Rural
z_1	31.78	86.77	67.91
z_2	22.23	7.84	18.72
z_3	18.42	3.20	6.84
z_4	17.08	1.35	3.80
z_5	10.49	0.84	2.73
Total variance	28.05	252.95	94.68

to total variance becomes clear when we consider the extent of its distribution. Values of the first principal balance are negative, with a few rare exceptions, and very large in absolute value, reflecting how small the “supermarket” share is when compared to the geometric mean of other food source shares. The second principal balance, whose contribution to total variance is 7.84%, compares “other” sources to the remaining four food sources: “specialized shop,” “own production,” “convenience shop,” and “wet market.” Its interpretation is similar to that of the first principal balance. “Other” food sources are marginal in the peri-urban district, but the contrast between this share and the geometric mean of the four food sources is less pronounced than the one for the first balance. The three remaining principal balances capture only small percentages of total variance, respectively. Their distributions are less and less sparse, while exhibiting mainly negative values. Food sources in the numerators of the considered principal balances are always smaller than the geometric mean of those in the denominators. For instance, the share of “wet market” is always greater than those of “convenience shop,” but their ratio is not very variable among peri-urban district households.

The first principal balance found for rural district compares the four food sources: “wet market,” “specialized shop,” “convenience shop,” and “own production,” with the marginal ones: “supermarket” and “other,” whose shares are negligible (panel (c) of Fig. 3). This principal balance captures a large part, 67.91%, of the total variance. As expected, the geometric mean of the four food sources is always larger than the geometric mean of the last two, and the distribution of their ratio is spread over the positive part of the real line. The second principal balance, which captures 18.72% of the total variance, compares “supermarket” with “other.” While negligible, the “other” share is always larger than the “supermarket” one, with few exceptions. The third, fourth, and fifth principal balances summarize the comparison between “wet market,” “specialized shop,” “convenience shop,” and “own production.” The “wet market” share appears to always be smaller than the geometric mean of the “specialized shop,” “convenience shop,” and “own production” shares. The “specialized shop” share is always smaller than the geometric mean of the “convenience shop” and “own production” shares. And, finally, the “convenience shop” share is always smaller than the “own production” share. As expected, the distributions of these three principal balances vary less and less, meaning that the ratios become nearly constant when moving from the third principal balance to the fifth one.

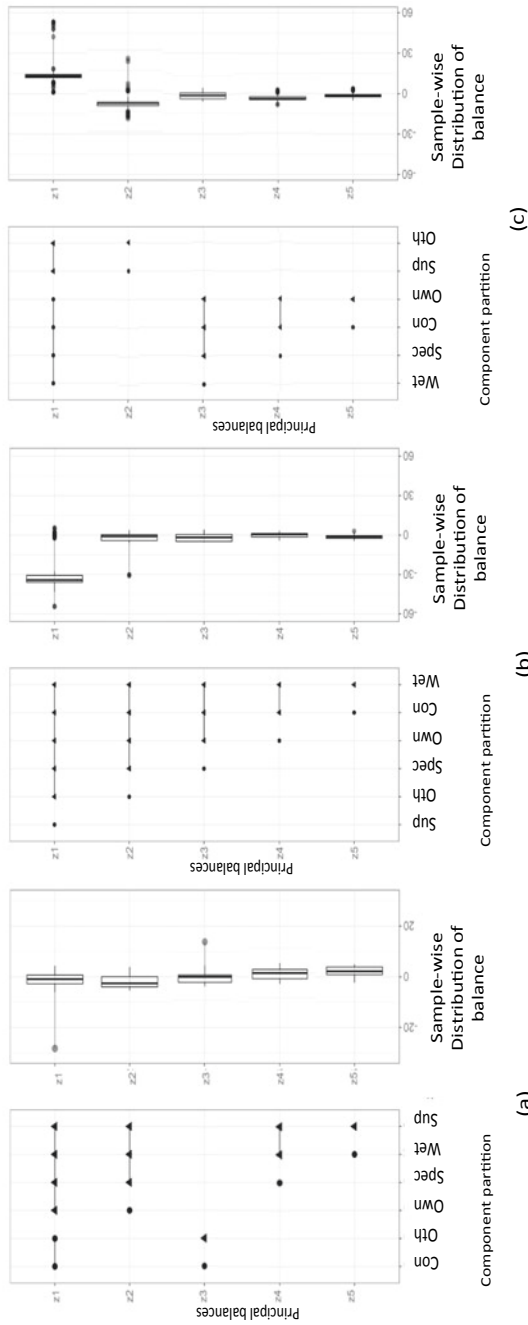


Fig. 3 Principal balances and their sample distributions for urban (panel (a)), peri-urban (panel (b)), and rural (panel (c)) districts. *Note:* *Sup:* Supermarket; *Con:* Convenience shop; *Spec:* Specialized shop; *Own:* Own production; *Wet:* Wet market; *Oth:* Others

The construction of principal balances in the urban district case does not exhibit one or two balances capturing a large percentage of total variation as observed with other districts (panel (a) of Fig. 3). The first principal balance for the urban district compares food sources for which many negligible shares are observed, mainly “convenience shop” and “other sources,” with “own production,” “specialized shop,” “wet market,” and “supermarket.” This principal balance captures only 31.78% of the total variance, and its distribution clearly indicates that the geometric mean of “convenience shop” and “other sources” shares is always smaller than the geometric mean of the remaining food sources. The second principal balance, which captures 22.23% of the total variance, compares “own production” with “specialized shop,” “wet market,” and “supermarket.” “Own production” share appears to always be smaller than the geometric mean of the three other food sources. The third balance shows the contrast between “convenience shop” and “other sources.” Most values of this principal balance are negative, meaning that the “other sources” share is larger than the “convenience shop” share for most households. The fourth principal balance compares “specialized shop” with “wet market” and “supermarket.” The distribution of this principal balance exhibits mainly positive values. The geometric mean of the “wet market” and “supermarket” shares are generally larger than the “specialized shop” share. Finally, the fifth principal balance contrasts “wet market” and “supermarket.” Here too, the principal balance takes mainly positive values. The “supermarket” share appears to be larger than the “wet market” one for most households.

4.2 Regression Analysis

The purpose of this section is to analyze associations between diet quality and food sources. Associations are assessed by regressing diet quality indicators, DQI-I or its components, on the previously built principal balances, and household socio-demographic characteristics as control variables. The main regression results are summarized in Table 6.

The empirical results give a contrasting picture of the associations within the three districts. We first consider the peri-urban district. Only a few associations are highlighted. The synthetic measure of diet quality, DQI-I, is positively associated with a relative increase of “own production” share when compared to the geometric mean of the “convenient shop” and “wet market” shares (z_4). Balance, which examines the overall balance of the diet in terms of the proportionality of energy sources and fatty acid composition, is positively associated with a relative increase in the ratio between “convenience shop” and “wet market” shares (z_5).

In the rural district, balance is positively associated with a relative increase in “wet market” share when compared to the geometric mean of “specialized shop,” “convenience shop,” and “own production” shares (z_3). Moderation is positively associated with a relative increase in the “specialized shop” share when compared to the geometric mean of the “convenience shop” and “own production” shares

Table 6 Summary of regression results by district

Principal balance	Variety	Adequacy	Moderation	Balance	DQI-I
<i>Urban district</i>					
z1: {Con, Oth} vs {Own, Spec, Wet, Sup}	0.177	0.196	-0.953**	-1.394***	-0.311
z2: {Own} vs {Spec, Wet, Sup}	1.251**	1.38***	0.218	0.885	0.956***
z3: {Con} vs {Oth}	-0.387	0.882*	-1.642***	-1.279**	-0.345
z4: {Spec} vs {Wet, Sup}	-1.025	-0.732	0.655	-0.174	-0.319
z5: {Wet} vs {Sup}	-1.435*	-1.828***	2.626***	0.154	-0.215
<i>Peri-urban district</i>					
z1: {Sup} vs {Oth, Spec, Own, Con, Wet}	-0.044	-0.001	-0.09	-0.08	-0.044
z2: {Oth} vs {Spec, Own, Con, Wet}	0.255	-0.005	0.053	-0.479	0.017
z3: {Spec} vs {Own, Con, Wet}	-0.145	-0.233	0.033	-1.05	-0.217
z4: {Own} vs {Con, Wet}	0.392	0.798	1.082	-0.327	0.690*
z5: {Con} vs {Wet}	0.287	-0.007	0.134	2.775**	0.372
<i>Rural district</i>					
z1: {Wet, Spec, Con, Own} vs {Sup, Oth}	0.214	-0.02	0.148	-0.183	0.061
z2: {Sup} vs {Oth}	-0.097	-0.109	-0.231	-0.144	-0.146
z3: {Wet} vs {Spec, Con, Own}	0.641	-0.022	-0.386	1.199*	0.123
z4: {Spec} vs {Con, Own}	-0.641	-0.15	0.923*	-0.375	0.051
z5: {Con} vs {Own}	-2.310***	-1.299*	0.203	0.870	-0.834**

Notes (1) DQI-I and its components are expressed as percentages of the maximum attainable scores. (2) In addition to the constant term and principal balances, regressions include gender, age, and education level of household head, with household income level and size, and nutrition knowledge, as control variables. (3) ***, ** and *: significant at 1%, 5%, and 10%, respectively. (4) *Sup*: Supermarket; *Con*: Convenience shop; *Spec*: Specialized shop; *Own*: Own production; *Wet*: Wet market; *Oth*: Others

(z_4). Negative associations are found between variety, which assesses whether intake comes from diverse sources both across and within food groups, adequacy, which evaluates the intake of dietary elements that must be supplied sufficiently to guarantee a healthy diet, and overall diet quality and a relative increase in “convenience shop” share relative to “wet market” share (z_5).

Results from the urban district highlight three important insights. First, moderation and balance are negatively associated with a relative increase in the geometric mean of “convenience shop” and “other sources” shares compared to the geometric mean of “own production,” “specialized shop,” “wet market,” and “supermarket” shares (z_1). A similar association arises with a relative increase in the “convenience shop” share compared to “other sources” share (z_3). Eating more food from a “convenience” shop than from other food sources appears to have a negative effect on diet quality in its dimensions linked to risk of chronic diseases or obesity. Second, variety, adequacy, and overall diet quality, as measured by the DQI-I score, are positively associated with a relative increase in relative “own production” share when compared to the average share of “supermarket,” “wet market,” and “specialized shop” (z_2). Thus, a household’s diet becomes less diverse and healthy when the relative “own production” share decreases in the urban site. Third, variety and adequacy are negatively associated with an increase in the ratio between “wet market” and “supermarket” shares, while it exhibits a positive association with moderation (z_5). Eating more food from supermarkets than from wet markets appears to have a negative effect on diet quality, which is linked to a higher risk of chronic diseases, despite being positively associated with a greater diversity in diet and better adequacy. This result is in line with those of the literature on the impact of supermarkets, and therefore of increased accessibility to highly processed food, on the quality of diet (Demmler et al. 2018).

5 Concluding Remarks

The aim of this chapter is to provide some initial evidence on the association between diet quality and its components (variety, adequacy, moderation, and balance) with food sources in lower-middle-income countries. By explicitly considering the compositional nature of food source data, this chapter shows the contribution of compositional data analysis to the treatment of this question, through the use of recent developments on principal balances and rounded zeros. Empirical implementation using a detailed survey on three Vietnamese districts highlights the contrast in the profiles of food sources of different areas and their association with diet quality.

The work presented in this chapter faces some limitations. First, due to dietary data limitations, this study focuses on Vietnamese households with children under 5 years of age. A more in-depth investigation is therefore necessary to see if the results obtained are also valid for households with children above the age of 5. Second, as this study is the first time the DQI-I was applied to assess diet quality in Vietnam, we were not able to evaluate its validity and credibility. Having said that, we customized

this international indicator to adapt to the Vietnamese Dietary Guidelines to improve its validity. Third, the application of compositional data analysis raises two important issues: interpretation of covariate impacts in compositional models, and the lack of consideration of total calorie intake in addition to its decomposition according to food sources as another explanatory variable in regression models. These two issues are the subject of numerous current works (Beal et al. 2018; Morais et al. 2018; Thomas-Agnan and Morais 2019). Their application to the empirical question treated in this chapter is left for future research.

Acknowledgements The authors would like to thank the editors of the book and two referees for their thoughtful comments and suggestions. This chapter has also benefited from valuable comments by Trang Nguyen. Trinh T. H. and Huynh T. T. T. contributions to this research were supported by the CGIAR Research Program on Agriculture for Nutrition and Health (CRP-A4NH) through its Flagship Program on Food Systems for Healthy Diets. Financial Support from the CIRAD—INRA GloFoodS program is fully acknowledged.

References

- Aitchinson, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70, 57–65.
- Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall.
- Amemiya, T. (1985). *Advanced econometrics*. Cambridge, MA: Harvard University Press.
- Beal, T., Le Danh, T., Nguyen, D. S., Simioni, M., Thomas-Agnan, C., Trinh, H. T. (2018). Macronutrient balances and body mass index: new insights using compositional data analysis with a total at various quantile orders. *Toulouse School of Economics* (WP 18-921). Toulouse, France
- Corrêa Leite, M. L. (2016). Applying compositional data methodology to nutritional epidemiology. *Statistical Methods in Medical Research*, 25(6), 3057–3065.
- Corrêa Leite, M. L. (2019). Compositional data analysis as an alternative paradigm for nutritional studies. *Clinical Nutrition ESPEN*, 33, 207–212.
- Corrêa Leite, M. L., & Prinelli, F. (2017). A compositional data perspective on studying the associations between macronutrient balances and diseases. *European Journal of Clinical Nutrition*, 71, 1365–1369.
- Demmler, K. M., Ecker, O., & Qaim, M. (2018). Supermarket shopping and nutritional outcomes: A panel data analysis for Urban Kenya. *World Development*, 102, 292–303.
- Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the royal Statistical Society Series B*, 39, 1–38.
- Egozcue, J. J., Pawłowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geosciences*, 35, 279–300.
- Food and Agriculture Organization of the United Nations (2013) *The state of food and agriculture 2013*. Rome, Italy: FAO
- Gibson, R. S., Charrondiere, U. R., & Bell, W. (2017). Measurement errors in dietary assessment using self-reported 24-h recalls in low-income countries and strategies for their prevention. *Advances in Nutrition*, 8(6), 980–991.
- HLPE: Nutrition and Food Systems. (2017). A report by the high level panel of experts on food security and nutrition of the committee on world food security, Rome.
- Huynh, T. T. T., Pham, T. H., Trinh, T. H., Duong, T. T., Nguyen, T. M., Hernandez, R., Lundy, M., Nguyen, T. K., Nguyen T. L., Nguyen, T. H., Vuong, T. V., Nguyen, T. H., Truong, T. M., Do, T. P. H., Raneri, J., Hoang, T. K., de Haan, S. Partial Food Systems Baseline Assessment at

- the Vietnam Benchmark Sites. A4NH report, The Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT), Hanoi, Vietnam.
- Kim, S., Haines, P. S., Siega-Riz, A. M., Popkin, B. M. (2003). The Diet Quality Index-International (DQI-I) provides an effective tool for cross-national comparison of diet quality as illustrated by China and the United States. *The Journal of Nutrition*, 133, 3476–3484
- Marias, Y. F., & Glasauer, P. (2014). *Guidelines for assessing nutrition-related knowledge, attitudes and practices*. Food and Agriculture Organization of the United Nations (FAO), Rome: Italy.
- Martín-Fernández, J. A., Palarea-Albaladejo, J., & Olea, R. A. (2011). Dealing with Zeros. In: V. Pawlowsky-Glahn & A. Buccianti (Eds.), *Compositional Data Analysis* (pp. 43–58). Hoboken: Wiley
- Martín-Fernández, J. A., Palarea-Albaladejo, J., & Gómez-García, J. (2003). Markov chain Monte Carlo method applied to rounding zeros of compositional data: First approach. In S. Thió-Henestrosa & J. A. Martín-Fernández (Eds.), *Compositional data analysis workshop*. Girona: University of Girona.
- Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics & Data Analysis*, 56(9), 2688–2704.
- Martín-Fernández, J. A., Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2018). Advances in principal balances for compositional data. *Mathematical Geosciences*, 50(3), 273–298.
- Ministry of Health. (2013). Food-based dietary guidelines—Viet Nam, 10 tips on proper nutrition for period 2013–2020, Hanoi, Vietnam. <http://www.fao.org/nutrition/education/food-dietary-guidelines/regions/countries/vietnam/en/>
- Monteiro, C. A., Moubarac, J. -C., Cannon, G., Ng, S. W., & Popkin, B. (2013). Ultra-processed products are becoming dominant in the global food system. *Obesity Reviews*, 14(52), 21–28.
- Morais, J., Thomas-Agnan, C., & Simioni, M. (2018). Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics*, 47(5), 1–25.
- Palarea-Albaladejo, J., Martín-Fernández, J. A. (2015) zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143, 85–96
- Palarea-Albaladejo, J., & Martín-Fernández, J. A. (2008). A modified EM algorithm for replacing rounded zeros in compositional data sets. *Computer & Geosciences*, 34, 902–907.
- Palarea-Albaladejo, J., Martín-Fernández, J. A., & Gómez-García, J. (2007). A parametric approach for dealing with compositional rounded zeros. *Mathematical Geosciences*, 39, 625–645.
- Pawlowsky-Glahn, V., Egozcue J. J., Tolosana-Delgado, R. (2011). Principal balances. In *The 4th International Workshop on Compositional Data Analysis CoDaWork2011* (pp. 1–10). Girona: University of Girona
- Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. Chichester: Wiley.
- Qaim, M. (2017). Globalisation of agrifood systems and sustainable nutrition. *Proceedings of the Nutrition Society*, 76(1), 12–21.
- Quinn, T. P. (2018). Visualizing balances of compositional data: A new alternative to balance dendrograms. *F1000Research*, 7, 1278. <https://doi.org/10.12688/f1000research.15858.1>
- Reardon, T., & Timmer, C. P. (2012). The economics of the food system revolution. *Annual Review of Resource Economics*, 4(1), 225–264.
- Solans, M., Coenders, G., Marcos-Gragera, R., Castelló, A., Gràcia-Lavedan, E., Benavente, Y., et al. (2019). Compositional analysis of dietary patterns. *Statistical Methods in Medical Research*, 28(9), 2834–2847.
- Thomas-Agnan, C., Morais, J. (2019). Covariates impacts in compositional models and simplicial derivatives. *Toulouse School of Economics* (WP 19-1057). Toulouse, France
- Trinh, H. T., Dhar, B. D., Simioni, M., de Haan, S., Huynh, T. T. T., Huynh, T. V., & Jones, A. D. (2020) Supermarkets and household food acquisition patterns in Vietnam in relation to population

- demographics and socioeconomic strata: Insights from public data. *Frontiers in Sustainable Food Systems*, 4. <https://doi.org/10.3389/fsufs.2020.00015>.
- Trinh, H. T., Morais, J., Thomas-Agnan, C., & Simioni, M. (2019). Relations between socioeconomic factors and nutritional diet in Vietnam from 2004 to 2014: New insights using compositional data analysis. *Statistical Methods in Medical Research*, 28(8), 2305–2325.
- Viet Nam National Institute of Nutrition. (2017). *Vietnamese Food Composition*. Hanoi, Vietnam: Medical Publishing House.
- Wertheim-Heck, S. C. O., & Raneri, J. E. (2019). A cross-disciplinary mixed-method approach to understand how food retail environment transformations influence food choice and intake among the urban poor: Experiences from Vietnam. *Appetite*, 142, 104370

Tools for Empirical Studies in Economics and Social Sciences

Mobility for Study and Professional Integration: An Empirical Overview of the Situation in France Based on the CÉREQ generational surveys



Bastien Bernela, Liliane Bonnal, and Pascal Favard

Abstract This chapter serves to elucidate the empirical reality of the phenomenon of geographical mobility among students and young graduates, based on data taken from five generational surveys conducted by CÉREQ. Our study shows that the degree of mobility among students' region of origin, region of education, and region of employment is relatively low: less than one in three high school graduates move to another region for their university studies, and less than one in three university graduates move to another region to find employment. The children of senior executives/Master's degrees are more likely to move to another region to pursue further education or find employment. Furthermore, more than half of such interregional movements correspond to people returning home. These results appear to demonstrate that individuals remain strongly geographically rooted: relatively few people move, and some of those movements correspond to people returning home.

1 Introduction

Geographical mobility has become a skill that students are encouraged to develop throughout their time in higher education. It is touted as essential for both high school graduates embarking on a new course of study and students considering post-graduate work. This mobility often corresponds to young people leaving home for the first time, and many will move residence several times during their time in higher education.

B. Bernela

Université de Poitiers CRIEF, UFR de sciences économiques: 2 rue Jean Carbonier TSA 81100, 86073 Poitiers cedex 09, France
e-mail: bastien.bernela@univ-poitiers.fr

L. Bonnal (✉)

Université de Poitiers CRIEF and TSE, UFR de sciences économiques: 2 rue Jean Carbonier TSA 81100, 86073 Poitiers cedex 09, France
e-mail: liliane.bonnal@univ-poitiers.fr

P. Favard

Université de Tours IRJI, Department of Economics, Tours Cedex, France
e-mail: pascal.favard@univ-tours.fr

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_29

573

The reasons behind such movements are many and varied: they may be a result of students moving to enroll on courses of study not available in the immediate vicinity of their family home, particularly for longer courses, and subsequently entering the workforce. As such, the geographical distribution of students depends primarily on the distribution of higher education options and their relative attractiveness.

Although the issue of student mobility has been a central concern of university policy since the expansion of access to higher education in the late 1990s, actual student mobility remains limited: each year around 6 students in every 100 move to a new town or region to pursue their studies, a phenomenon which has remained stable over the past decade (Perret 2007; Baron and Perret 2008). There is no obvious upward trend in this mobility, as one might have expected. The phenomenon is partly connected to the territorial distribution of educational options, which determines the geographical and social distances between high school graduates/students and centres of education.

The first priority of higher education institutions, and universities in particular, is to train the young people of today who will make up the workforce of tomorrow. This is a mission close to the hearts of local government, with a keen interest in developing a pool of qualified manpower with the potential to satisfy the labour requirements of the local economy; hence why local authorities (regions, in particular) are so closely involved with higher education policy within their territory. With the emergence of the knowledge economy, attracting students is now a development priority for local territories. Young graduates are considered to be an indispensable resource for the dynamic development of the local and regional job markets and are thus held to be a source of comparative advantage for the regions to which they move. The idea that territories influence the spatial allocation of resources for competitive considerations has been studied in some detail by fellow researchers, seeking to define the winners and losers of mobility dynamics at a macro-economic level (Baron 2009; Hoare and Corver 2010; Coquard 2019). Why are students and graduates so mobile? According to the standard economic theories of job hunting, this migration is the fruit of an individual, rational process of cost–benefit analysis. Put simply, young graduates looking for employment are prepared to move in order to hunt occupations that are not available locally or find jobs that offer more attractive salaries than those in the near vicinity. The academic literature on human capital (Sjaastad 1962) holds that the potential impact of such migration mirrors the level of qualifications of those concerned, i.e. university graduates are particularly concerned. Non-economic factors only came to be included in mobility analysis at a much later date. Factors such as the psychological burden of being separated from loved ones (Schwartz 1973; Dahl and Sorenson 2010) can help to explain the rate of non-mobility and the major influence of geographical proximity on those who do move. A better understanding of the mechanisms of mobility among students and recent graduates would enable local authorities (at the municipal and regional levels) to implement policies designed to attract and retain these groups. Indeed, local elected officials are frequently worried that graduates will move elsewhere after their studies, compromising the “return on investment” of funds invested in education.

The aim of this chapter is to examine the geographical trajectories of university graduates, specifically their movements in relation to their studies and first jobs. These two moments correspond to the first occasions on which young people make individual decisions about their geographical location which are not (or at least not entirely) dependent upon their family. On average, how many students pursue further studies (and then enter the labour market) in their region of origin, or else in another region? What are the defining characteristics of those who move?

We also propose to look at movements corresponding to students returning to their regions of origin upon completing their higher education. This phenomenon has received relatively little attention in the existing literature, although empirical studies have highlighted its importance in other geographical contexts (Niedomysl and Amcoff 2011; Rérat 2014).

Taken together, these geographical indications offer some empirical perspective on the true room for manoeuvre of territories seeking to boost their attractiveness.

This chapter utilizes data from five generational surveys conducted by CÉREQ (1998, 2001, 2004, 2007, and 2010). Section 2 is devoted to a brief explanation of these data. Section 3 presents a number of descriptive statistics and our empirical strategy. Section 4 is given over to the principal results of our econometric models.

2 The Data

Since the early 1990s, the CÉREQ¹ has been running a series of longitudinal studies focusing on graduates' first few years of professional life. These surveys examine the professional integration and progress of graduates leaving education, over a period of three years (or 5, for certain generations). One of the objectives of these surveys is to produce integration indicators (employment rate, unemployment rate, rate of graduates employed on permanent contracts, etc.) for different levels of education, sectors, etc. The surveys thus yield information that helps to improve our understanding of the different integration processes and paths experienced by graduates at the outset of their careers.

In order to be included in the survey for a given "generation" X (where X represents the year in which the subjects left education), subjects must satisfy all of the following criteria:

- must have been enrolled in an educational institution for the academic year corresponding to the year group X ,
- must have left education in this same academic year,
- must not have returned to education after a spell away in this academic year,
- must not have returned to education in the academic year $X+1$,
- must be less than 35 years old in the year the survey is conducted,

¹Centre for research on employment and qualifications. <http://www.cereq.fr/articles/Enquete-Generation/Presentation-detaillee-de-Generation>.

- must be resident in France in the year X. This condition excludes those who studied in France but are working overseas when the study is conducted, i.e. in the year X+3.

The sample strategy for generational surveys is based on random sampling across all courses and across the country as a whole. The data can thus be considered representative for different categories of qualification at the regional level.

The generational surveys employ the same set of questions, methodology, and analytical framework for all individuals surveyed, regardless of their academic background, level of qualification, and field or path of study. This makes it possible to compare and evaluate the impact of these different characteristics on the variations observed during the first post-graduation years, looking at parameters such as success in finding employment, type of job, salary, etc. Furthermore, the generational surveys are constructed with reference to the date at which respondents leave education, not their date of birth. As such, and regardless of their level of qualification, the young graduates surveyed enter a labour market which may be more or less favourable but which will at least be identical for all. It is of course possible that the effect of the general economic outlook on success in navigating the job market may be different for different types and levels of qualification, but in theory, it is still easier to make comparisons.

Thanks to their detailed questionnaire and substantial sample size (cf. Tables 1 and 6 in the appendix), these surveys contain, in addition to details of respondents' academic careers and qualifications, information regarding their gender, social background, nationality, place of residence, geographical mobility, family status, etc. Surveys are conducted, retrospectively, three years after respondents leave the education system. The questionnaire allows young graduates to describe systematically, month by month, the different situations they have encountered since leaving education. It is therefore possible to reconstruct their professional trajectories (employment, unemployment, inactivity, and return to education) and define different categories of integration which correspond to these trajectories (rapid employment, stable employment, stalling, downgrading, etc.).

The data used are taken from five generational surveys (1998, 2001, 2004, 2007, and 2010). We use abbreviations to indicate the successive generational surveys, with G1998 for those leaving higher education in 1998, G2001 for those leaving in 2001, and so on. A representative sample of French graduates are surveyed three years after completing their studies: for example, the Class of 2010 was surveyed in 2013. This system was established in order to better understand the educational and professional trajectories of young people. For this chapter, the data from these five generational surveys encompassing the whole population of university graduates were aggregated to distinguish between four different levels of qualification: BTS/DUT and equivalent (two years of higher education), Bachelor's degrees and equivalent (three years), Master's degrees and equivalent (five years), and Doctoral degrees (cf. Tables 1 and 6 in the appendix for weighted figures).

For each generation, the data include spatial variables which enable us to reconstruct individual geographical trajectories. We know where respondents started mid-

Table 1 Numbers in each generation and level of qualification

Generation class	Year of survey	BTS/DUT	Bachelor's degree	Master's degree	Doctorate	Total
G1998	2001	9 720 (46.2%)	5 710 (27.1%)	4 112 (19.5%)	1499 (7.1%)	21 041 (32.2%)
G2001	2004	3 262 (45.9%)	1 681 (23.6%)	975 (13.7%)	1 195 (16.8%)	7 113 (10.9%)
G2004	2007	5 999 (44.5%)	1 854 (13.7%)	4 167 (30.9%)	1 473 (10.9%)	13 493 (20.6%)
G2007	2010	4 789 (44.0%)	2 498 (22.9%)	2 627 (24.1%)	981 (9.0%)	10 895 (16.7%)
G2010	2013	3 114 (24.2%)	4 507 (35.1%)	3 539 (27.5%)	1 686 (13.1%)	12 846 (19.6%)
Total		26 885 (42.11%)	16 250 (24.9%)	15 420 (23.6%)	6 834 (10.4%)	65 389 (100%)

Source CÉREQ data (generational surveys G1998, G2001, G2004, G2007 and G2010). The numbers shown in this table are unweighted.

dle school,² the location of the higher education institution where they obtained their final qualification, and their location three years after graduating.³ This information allows us to analyze the interregional mobility of students, between their region of origin and the region in which they pursue further studies—we call this “educational mobility”—and between the region of study and the region in which they are employed three years later—we call this “employment mobility”. From a geographical perspective, the data derived from the generational surveys are representative on a regional scale, which is why we have decided to situate our analysis of mobility among students and young graduates at the regional level.⁴

Before analyzing the data in depth, it is important to highlight the heterogeneity within the student population with regard to mobility. Figure 1 illustrates the sizable variation in the mobility rates of students and graduates with different levels of qualification. On average, 18.8% of young people finish their studies in a region other

²The region in which respondents took the Baccalaureate is not available for one of the generations, which prevents us from making systematic use of this variable. Nonetheless, in the three other generations for which figures are available, it appears that more than 95% of high school graduates received their Baccalaureate in the same region in which they started middle school. This rate is so high that the location variables for middle school and school leaving can be used almost interchangeably. For the rest of this section, we will thus use the term “region of origin”.

³We focused on individuals in employment three years after graduating. The data therefore excludes graduates who are not in employment, registered unemployed, or who have taken up further studies. Table 8 in the appendix indicates that around 90% are employed. For those who are not, long high school graduates are unemployed while short high school graduates are returned to school.

⁴For this study, we used the old regional divisions. Metropolitan France was until recently divided into 22 regions and 26 educational academies: the Ile-de-France region is split into three academies (Créteil, Paris, and Versailles), the Rhône-Alpes, and Provence-Alpes-Côte d’Azur region each have two academies (Grenoble and Lyon; Marseille and Nice). Following the redrawing of the administrative map, France now has just 13 regions.

than their region of origin, and, three years after graduating, 19.7% of graduates are working in a region other than the region in which they gained their degree. The corresponding figures are 9.4% and 11.3% respectively for those with qualifications equivalent to the Baccalaureate or lower (\leq Baccalaureate: individuals that did not take the baccalaureate or that took it and failed it, or passed it and then left education, or passed it and embarked upon further studies which they did not complete), and 26.6% and 26.8% for those with qualifications higher than the Baccalaureate ($>$ Baccalaureate: individuals that passed the baccalaureate, continued with further studies, and graduated from an institute of higher education, whatever the grade or the level of qualification attained). There is therefore a clear disparity between the mobility behaviours of young people who do not pursue higher education and those who do, both during their studies and when it comes to finding work. This clear difference can be largely attributed to the geographical location of educational institutions and qualified jobs, with opportunities in the immediate local vicinity becoming scarcer, the higher one climbs on the qualification ladder. Zooming in on young people who do continue their studies after the Baccalaureate, once again we can observe a high degree of heterogeneity. 18.2% of those who undertake short courses of study (BTS/DUT, two years of study after the baccalaureate) complete their studies in a different region, with 22.0% changing region to find a job. These levels are close to the mean figures for the population as a whole. The rate of educational mobility rises to 27.4% for those with Bachelor's degrees or similar (baccalaureate plus three years), 49.0% for those with Master's degrees (baccalaureate plus five years), and 47.5% for doctoral graduates (baccalaureate plus eight years). The rate of employment mobility is 27.6% for those with Bachelor's degrees or similar, 45.3% for those with Master's degrees, and 33.2% for doctoral graduates. Master's graduates are therefore the most mobile. It is interesting to note that for both Master's and doctoral graduates, the rate of educational mobility is higher than the rate of employment mobility: those with superior qualifications are thus more likely to move to another region to complete their studies than they are to find a job in a new region once their studies are over.

3 The Micro-Economic Approach: Identifying the Factors that Determine Mobility Behaviour

In this chapter, we want to examine the determinants of migration patterns: migration to study and migration to work. More precisely we analyze the personal characteristics that impact educational and employment mobility decisions, but also return mobility.

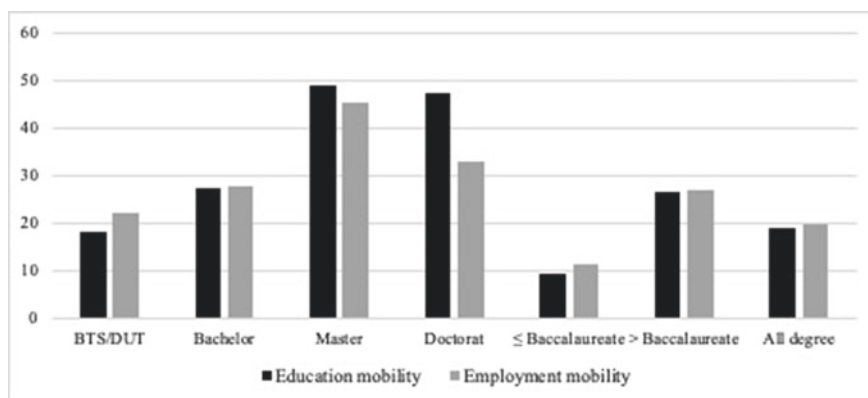


Fig. 1 Rate of mobility by level of qualification. *Source:* CÉREQ (*generational studies G1998, G2001, G2004, G2007, and G2010*).

3.1 Constructing Mobility Variables

We propose to examine three types of mobility: (1) educational mobility, corresponding to a change of region between the student's region of origin (i.e. where they started middle school) and the region where they obtained their final qualification, (2) employment mobility, corresponding to a change of region between the region where students obtained their final qualification and the region in which they were employed three years later, and (3) return mobility, which corresponds to those who return to their region of origin to find work after completing their higher education elsewhere. These mobility rates are detailed in Tables 2 and 3. Specifically, Table 2 shows the proportion of students who move regions for each level of qualification, generation, and type of studies. One in three students will change regions either for their studies or to find work. Among the third of students who do experience educational mobility, almost one in three will then return to their home region to enter the labour market. Educational mobility is more common among those who undertake longer courses of study (almost half of all Master's and doctoral students, and just one in five for those who undertake two-year programmes). These students, and particularly those with Master's degrees, are much more mobile (almost half) when they enter the labour market. Educational mobility has grown significantly, reaching 40% for the 2010 generation. The 2004 generation saw a sharp increase in employment mobility, which rose from 25.1% (2001 generation) to 34.4% before stabilizing at around 40% from the 2007 generation onwards. While barely one in five students from the 1998 and 2001 generations returned home after their studies, the rate of return doubled for the ensuing generations. Return mobility is most common among those students who enroll in short courses (over 40% for graduates of two-year programmes and over 35% for holders of Bachelor's degrees).

Table 2 Observed mobility rates

		Population	%	Mobility		
				Education (Ed)	Employment (Em)	Return (R)
Overall		65389	100	30.3	32.1	31.8
Qualification	BTSDUT	26885	41.1	20.1	24.4	42.1
	Bachelor's	16250	24.9	25.6	29.1	35.2
	Master's	15420	23.6	46.6	46.8	29.8
	Doctorate	6834	10.4	44.4	36.1	13.6
Generation	1998	21041	32.2	19.7	24.5	18.9
	2001	7114	10.9	25.7	25.1	20.0
	2004	13493	20.6	33.9	34.4	38.1
	2007	10895	16.7	37.7	40.9	39.5
	2010	12846	19.6	39.9	38.3	34.8
Discipline	Law, Economics & Management (LEM)	15928	24.4	25.6	29.7	31.1
	Languages & Literature (L&L)	3672	5.6	29.9	27.3	24.2
	Humanities & Social Sciences (HSS)	8933	13.7	28.0	27.0	31.2
	Basic & Applied Sciences (BAS)	27025	41.3	34.6	36.9	30.5
	Health	9831	15.0	27.8	28.9	41.0

Key: 41.1% of those leaving education graduated with a two-year diploma. Among this sub-population, 20.1% moved for their studies and 24.4% moved to find employment. Many of these individuals moved for both their studies and to find employment, with graduates returning to their native regions accounting for 42.1% of the total mobility. The numbers shown in this table are unweighted.

Table 3 presents the frequencies of the various possible combinations of such movements. Of the subjects in the sample, 56.6% did not move at all. This rate decreases as the level of qualification increases (68.1% for students on two-year courses but just 40% for Master's and doctoral students) and with successive generations (65% for the 1998 generation and 49% for the generations since 2007) but remains relatively stable within the different disciplines (around 60% for short courses, and just over 51% for science students). The proportion of young people concerned by employment mobility alone is stable across all of the sub-samples (between 11% and 15%) while 11.5% of students move only for their studies. This

Table 3 Descriptive statistics for different types of mobility

Mobility	Return (R)	NO	NO	NO	NO	YES
	Education (Ed)	NO	NO	NO	YES	YES
	Employment (Em)	NO	YES	NO	YES	YES
Overall		56.5	13.3	11.5	9.1	9.6
Degree	BAC+2	68.1	11.8	7.5	4.1	8.5
	Bachelor's	60.5	14.0	10.4	6.2	9.0
	Master's	39.0	14.4	14.3	18.4	13.9
	Doctorate	40.4	15.2	23.5	14.9	6.1
Generation	1998	64.6	15.7	10.9	5.1	3.7
	2001	60.5	13.9	14.4	6.1	5.1
	2004	55.0	11.1	10.6	10.4	12.9
	2007	49.1	13.2	10.0	12.9	14.9
	2010	48.7	11.5	13.0	13.0	13.9
Discipline	LEM	60.9	13.4	9.4	8.3	8.0
	L&L	56.7	13.4	15.9	6.7	7.2
	HSS	60.0	11.9	13.0	6.3	8.8
	BAS	51.2	14.2	11.9	12.2	10.6
	Health	60.4	11.8	10.8	5.6	11.4

General remarks: YES or NO answers for the three kinds of mobility (Ed,Em,R). For the first line: 56.5% of students did not move at all (NO,NO,NO); 13.3% moved only to find employment (NO,YES,NO); 11.5% moved only for educational purposes (YES,NO,NO); 9.1% moved for their studies and moved to find employment mobility but did not move back home (YES,YES,NO); and, finally, 9.6% moved for all three reasons (YES,YES,YES). The sum of each line is 100%.

rate is higher among those undertaking longer courses (Master's degrees, and even more so for doctorates) and for students of languages and literature or the arts. Almost 20% of students move twice. For half of them, the second move corresponds to a return to their home region (where they started middle school). The principal independent variables taken into consideration to account for these different types of mobility are educational variables (degree, discipline, and age), socio-economic variables (professional status and origin of parents, gender, family status, and whether or not they worked during their studies), and geographical variables (area of residence, region, and previous moves). The main descriptive statistics associated with all of the variables used in our empirical econometric analysis of the data are given in Table 7 in the appendix.

3.2 *Factors Determining Mobility*

3.2.1 **Educational Factors**

On the one hand, educational options are not the same in all regions (particularly for two-year courses and Master's degrees) and, on the other hand, the nature and number of jobs available vary from region to region. It is therefore important to take into account not only the highest qualification attained upon leaving education, but also the characteristics of those qualifications using indicators which reflect the level and discipline of studies.

3.2.2 **Socio-Economic Factors**

We begin by examining the impact of gender, bearing in mind that the results presented in the existing literature are not always convergent on this point. For example, Faggian et al. (2007b) shows that in the United Kingdom women are more likely than men to move for educational or employment reasons, whereas Kazakis and Faggian (2017) for the USA and Ciriaci (2014) for Italy find the opposite. Haussen and Uebelmesser (2018), meanwhile, do not consider gender to have any effect on employment mobility.

We then construct a variable for students required to resit school years (which serves as a proxy for age). For educational mobility, resitting is measured by an indicator identifying those students who were at least one year behind by the time they entered higher education. For employment mobility, we distinguish between individuals whose academic progress was slightly delayed (one or two years) and those who fall behind by three years or more. Age differences on this scale may correspond to changes of course or a decision to return to education. Late graduation (i.e. individuals completing their studies at a greater age) may be associated with a stronger sense of attachment to their home region, and thus with lower mobility rates. Once again, the existing literature diverges on this topic. While age appears to have a negative effect on educational mobility, it has no effect on employment mobility in Italy (Ciriaci 2014) or Germany (Haussen and Uebelmesser 2018); it has a positive influence on all types of mobility (educational, employment, and return) in the USA (Kazakis and Faggian 2017); and no significant effect in South Korea (Ma et al. 2017).

In terms of personal characteristics, we also include the social background and geographical location of students' parents. We might expect students from more modest backgrounds to have a lower rate of educational mobility than others. It is easier i) for students from privileged backgrounds to imagine themselves undertaking lengthy courses of study away from home and ii) for their families to bear the costs of studying in another region. Furthermore, since higher education is largely concentrated in big towns and cities, the probability of educational mobility should

be higher among students whose parents live in rural areas. An indicator for this information was taken into consideration.

Finally, conjugal status is an important factor in geographical mobility and we can safely assume that being in a relationship and having children will reinforce a respondent's sense of territorial and relational attachment, thus decreasing the likelihood that they will move.

3.2.3 Geographical Factors

Regional indicators were introduced to serve as control variables. In order to explain why individuals move to study or choose to return to their home regions, we took into account their region of origin (the place their parents were resident when they started middle school), and in order to account for employment mobility, we considered the region in which they completed their studies. The existing economic literature shows that probability of moving is higher among those who have previously moved for educational reasons (Faggian et al. 2007a; Ciriaci 2014; Ma et al. 2017; Haussen and Uebelmesser 2018) or moved internationally (Haussen and Uebelmesser 2018), hinting at the importance of prior experience with mobility. We thus include two indicators representing these two types of mobility and assume that they have a positive effect on employment mobility.

3.3 Empirical Strategy

Our primary objective is to analyze the determinant factors in educational and employment mobility decisions, as well as return mobility in those cases where individuals move both for their studies and to find work. With this goal in mind, we developed two models, both based on the simultaneous calculation of two Probit equations. The first equation in both models is designed to reflect educational mobility. In Model 1, the second equation covers employment mobility, while that in Model 2 covers return mobility for those individuals already affected by educational mobility.

The equation for educational mobility hinges on the dichotomous variable Ed , which assumes the value 1 if the student has experienced educational mobility and 0 if not.⁵ Specifically, educational mobility ($Ed=1$) is determined by the latent variable $Ed^* = X_{Ed}\beta_{Ed} + u_{Ed}$ which is positive. This variable depends on the observed and exogenous individual characteristics X_{Ed} , the vector of the parameters associated with these characteristics β_{Ed} , and the random measurement error u_{Ed} , which is assumed to follow a standard normal distribution.

Employment mobility, meanwhile, is represented by the dichotomous variable Em , which is 1 if the student has experienced employment mobility and 0 if not. For

⁵The index assigned to each student is omitted here to make the notations easier to read.

Model 1 in particular, students can be said to have experienced employment mobility if the associated latent variable $Em^* = Ed\gamma + X_{Em}\beta_{Em} + u_{Em}$ is positive, while if it is negative, the student did not move at the end of their studies. This latent variable depends on having previously experienced educational mobility (Ed), a variable that is potentially endogenous (γ is the parameter associated with this form of mobility), along with a set of exogenous individual characteristics X_{Em} (β_{Em} is the vector for the parameters to be calculated) and a random error term u_{Em} that is assumed to follow a standard normal distribution. Both measurement error terms are supposed to be correlated (σ_{EdEm}). The vector (u_{Ed}, u_{Em}) thus follows a bivariate normal distribution $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_{EdEm}\right)$ where $\Sigma_{EdEm} = \begin{pmatrix} 1 & \sigma_{EdEm} \\ \sigma_{EdEm} & 1 \end{pmatrix}$. The set of variables X_{Em} includes the majority of the variables in X_{Ed} . There are thus four contributions to likelihood:

$$\begin{aligned} P(Ed = 1, Em = 1) &= \Phi_2(X_{Ed}\beta_{Ed}, \gamma + X_{Em}\beta_{Em}, \sigma_{EdEm}) \\ P(Ed = 1, Em = 0) &= \Phi_2(X_{Ed}\beta_{Ed}, -\gamma - X_{Em}\beta_{Em}, -\sigma_{EdEm}) \\ P(Ed = 0, Em = 1) &= \Phi_2(-X_{Ed}\beta_{Ed}, X_{Em}\beta_{Em}, -\sigma_{EdEm}) \\ P(Ed = 0, Em = 0) &= \Phi_2(-X_{Ed}\beta_{Ed}, -X_{Em}\beta_{Em}, \sigma_{EdEm}) \end{aligned}$$

where $\Phi_2(\cdot, \cdot, \rho)$ is the distribution function of the bivariate normal distribution with mean 0, variance 1, and covariance ρ .

For Model 2, return mobility concerns students who have previously experienced educational mobility ($Ed = 1$). If a student did not attend an institution of higher education in the same region that she started middle school, she is considered an example of return mobility (i.e. $R = 1$) if the corresponding latent variable R^* is positive so that $R^* = X'_R\beta_R + u_R$. This underlying variable is dependent upon a set of explanatory variables X_R , the vector of the associated parameters β_R , and an error term u_R which is assumed to follow standard normal distribution. We also assume that the error terms associated with educational and return mobilities are correlated, noting their covariance σ_{EdR} . The vector (u_{Ed}, u_R) thus follows a bivariate normal distribution $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_{EdR}\right)$ where $\Sigma_{EdR} = \begin{pmatrix} 1 & \sigma_{EdR} \\ \sigma_{EdR} & 1 \end{pmatrix}$. X_R includes all explanatory variables of X_{Ed} , to which we add information concerning the conjugal situation of students. This model has three contributions to likelihood:

$$\begin{aligned} P(Ed = 0) &= \Phi(-X_{Ed}\beta_{Ed}) \\ P(Ed = 1, R = 0) &= \Phi_2(X_{Ed}\beta_{Ed}, -X_R\beta_R, -\sigma_{EdR}) \\ P(Ed = 1, R = 1) &= \Phi_2(X_{Ed}\beta_{Ed}, X_R\beta_R, \sigma_{EdR}) \end{aligned}$$

4 Results

The estimating coefficients associated with the probability of experiencing educational, employment, and return mobility are given in Table 4. These estimates are calculated for all graduates in employment three years after the end of their studies.

Our analysis of these results assumes all other factors to be equal. Before going into further detail regarding the coefficients associated with the explanatory variables, it is worth looking at the correlations between the error terms of the two equations. These coefficients are negative and statistically significant. As regards the coefficient connecting the equations for educational and employment mobility, it appears that those students who are highly likely to experience educational mobility are also unlikely to experience employment mobility: non-observable characteristics thus appear to have contrasting effects on educational and employment mobilities. We might, for example, invoke factors such as home town size or the properties of educational institutions. Ciriaci (2014) demonstrates that the quality of educational institutions (measured using ranking tables) and their size have opposing effects: they increase the probability of educational mobility and decrease the probability of employment mobility. Institution size may be considered a proxy for the local labour market, since the major universities are generally found in the largest cities, where the jobs on offer for young graduates are most numerous.

Nevertheless, when we introduce educational mobility in order to explain employment mobility, the coefficient associated with this variable is positive, significant, and relatively high. This result, which at first sight may appear to contradict those set out in the preceding paragraph, in fact, explains a certain dynamic of mobility, since students who have already moved once in order to pursue higher education are more likely to move again to find employment.

4.1 Educational Mobility

The probability of educational mobility is comparable for men and women. This probability increases significantly in successive generations. The duration and discipline of studies also have a significant impact on the probability of educational mobility. Students choosing short courses of study (two or three years post-Baccalaureate) show a below-average proclivity for educational mobility. With regard to different disciplines of study, those students with the lowest and highest probabilities of educational mobility are students of Law—Economics—Management and fundamental sciences, respectively.

Children of workers and salaried employees (professionals but not executives) have a below-average probability of educational mobility. This result is consistent with the existing literature, suggesting that students from lower-income households have lower mobility rates than others, essentially for financial reasons. Children of immigrants (at least one parent born overseas) have a higher rate of educational mobility.

Students whose parents live in rural areas are also more concerned about educational mobility. Region of origin also plays a role: growing up in one of France's larger regions, and particularly the Paris region, reduces the probability of students

Table 4 Estimating the probability of different mobilities

		Mobility					
		Education		Employment		Return	
		coef	s.d	coef	s.d	coef	s.d
Intercept		-1.481***	0.03	-1.262***	0.11	1.168***	0.12
Gender	Male	-0.003	0.01	0.036***	0.01	0.005	0.01
Degree	BTS/DUT	ref.		ref.		ref.	
	Bachelor	0.200***	0.02	0.152***	0.02	-0.212***	0.02
	Master	0.779***	0.02	0.368***	0.02	-0.744***	0.02
	Doctorate	0.653***	0.02	0.144***	0.03	-0.924***	0.04
Discipline	LEM	ref.		ref.		ref.	
	L&L	0.143***	0.03	-0.052*	0.03	-0.186***	0.03
	HSS	0.132***	0.02	-0.051**	0.02	-0.099***	0.02
	BAS	0.161***	0.02	0.086***	0.02	-0.139***	0.02
	Health	0.293***	0.02	0.093***	0.02	-0.129***	0.03
Generation	1998	ref.		ref.		ref.	
	2001	0.182***	0.02	-0.068***	0.02	-0.082***	0.03
	2004	0.394***	0.02	0.051**	0.02	0.022	0.06
	2007	0.465***	0.02	0.106***	0.02	-0.036	0.06
	2010	0.530***	0.02	0.065**	0.02	-0.104*	0.06
Mother	Born elsewhere	0.099***	0.03	0.005	0.03	-0.082**	0.03
	Prof./exec	ref.		ref.		ref.	
	Prof./non exec	-0.127***	0.01	-0.039***	0.01	0.133***	0.02
	Other	-0.030	0.02	-0.079***	0.02	0.018	0.03
Father	Born elsewhere	0.103***	0.03	0.029	0.03	-0.111***	0.03
	Prof./exec	ref.		ref.		ref.	
	Prof./non exec	-0.184***	0.03	-0.041***	0.01	0.223***	0.01
	Other	-0.094***	0.02	-0.094***	0.02	0.122***	0.03
Parents' residence	Rural area	0.063***	0.01			-0.011	0.02
Years behind (middle school)	Yes	-0.025	0.03				
Years behind (higher education)	1 or 2 years			0.011	0.01	-0.036***	0.01
	3 years or more			-0.125***	0.02	-0.095***	0.02

(continued)

Table 4 (continued)

		Mobility					
		Education		Employment		Return	
		coef	s.d	coef	s.d	coef	s.d
In a relationship	Yes			0.003	0.01	-0.058***	0.01
Children	Yes			-0.285***	0.02	0.021	0.02
Employment during studies	No			ref.		ref.	
	Occasionally			0.007	0.01	-0.017	0.01
	Regularly			-0.117***	0.02	-0.073***	0.02
Regions		yes		yes		yes	
housing rental costs				0.002	0.00	-0.007	0.01
International mobility during studies	Yes			0.125***	0.02	-0.032**	0.02
Education mobility	Yes			1.324***	0.07		
Correlation coefficient between education mobility and	Employment			-0.123***	0.04		
	Return					-0.941***	0.03

*Key: For each type of mobility, the first column shows the estimated coefficient, the second column indicates significance, and the third gives the standard deviation of the coefficient. The coefficient is significant if over: * 10%, ** 5%, and *** 1%. The coefficients associated with the various regions can be provided by the authors if required.*

moving to another region for their higher education.⁶ The effects of these two variables illustrate the importance of considering the density of education options on offer locally as a factor determining student mobility.

4.2 Employment Mobility

Unlike educational mobility, employment mobility is indeed influenced by gender. The probability is significantly higher for men than it is for women.

Generally speaking, graduates with a bachelor’s degree or higher are more likely to move to find employment than those who graduate from shorter courses (two-year programmes). Master’s graduates are the most mobile category. Graduates in the fields of healthcare and fundamental sciences are the most likely to move for

⁶The coefficients associated with the various regions can be provided by the authors if required.

employment reasons, whereas humanities graduates have the lowest probability of employment mobility.

As with educational mobility, children of workers and salaried employees are less likely to experience employment mobility. This probability is even smaller among children of families where the mother does not work or the father is not present.

Falling more than two years behind (i.e. being older than average when leaving education), having worked regularly during the period of study, and having children all reduce the probability of employment mobility. As discussed above, mobility breeds mobility: having travelled abroad during one's studies increases the probability of moving regions after graduation. Finally, with regard to geographical variables, completing one's higher education in the Paris region strongly reduces the probability of seeking work elsewhere: this is the only region with a significant negative effect, which may be attributed to the nature of the labour market in Paris where the concentration of executive-level jobs is much greater than that found elsewhere in the country.

4.3 Return Mobility

The probability of return mobility is not dependent upon either gender or generation.

Students who choose short courses of higher education (two years) are more likely than others to return to their home region, which seems to suggest that these students are more attached to their native regions and/or are more likely to find employment there.

The social background of students also has an impact on return mobility, since children of executives are less likely to return home. The country of origin of parents does not have a significant effect on the probability of this kind of mobility.

On average, falling behind in terms of academic years reduces the probability of return, and this probability decreases as the age gap rises.

While having a partner reduces the probability of return, having a child does not have a significant impact on this probability. Finally, regular employment and spending a spell abroad both have a slightly negative impact on the probability of returning home.

It also appears that those regions with the lowest probability of educational mobility are also the regions where the probability of return mobility is highest: to put it simply, students hailing from regions where few people move for their studies are more likely to return to those regions if they do indeed leave for a spell.

Building upon these results, the impact of spatial mobility is measured at the individual level. Following Faggian et al. (2007a, 2007b), it is possible to classify graduates on the basis of their discrete migratory choices: the initial migration of students moving to enter higher education and the subsequent migration of graduates entering the labour market, while also considering whether or not these movements involve a return to their region of origin. We can thus observe five categories:

Table 5 Relationship between migration and job quality

Employment conditions	Executive contract		Yes	Yes	No	No	All contracts
			Permanent contract	Yes	No	Yes	
Mobility behaviour	Non-Migrant	(1)	34.7	6.1	47.2	11.9	100.0
		(2)	1 904.1	1703.9	1 622.3	1 479.0	1 748.9
	University stayer	(1)	42.6	7.3	40.4	9.8	100
		(2)	2012.0***	1829.6***	1 645.7	1531.4***	1843.2***
	Late migrant	(1)	41.9	7.1	40.2	10.8	100
		(2)	1954.2*	1818.6***	1 636.9	1520.1**	1803.9***
	Repeat migrant	(1)	42.5	7.3	38.6	11.7	100
		(2)	2091.8***	1844.1***	1699.1**	1567.8***	1902.8***
	Return migrant	(1)	32.7	6.4	45.7	15.3	100
		(2)	2042.6***	1827.9***	1656.3*	1537.2**	1822.4***
	All	(1)	37.4	6.6	44.3	11.7	100
		(2)	1 958.4	1 764.8	1 633.9	1 504.8	1 794.1

(1) % of the column total; (2) mean wage. Significantly different from non-migrants? mean wage at * 10%; ** 5%; *** 1%.

- repeat migrants: moving for education and employment,
- returning migrants: moving for education and employment, returning to the region of origin to enter the job market,
- university stayers: moving for education only,
- late migrants: moving for employment only,
- non-migrants: no mobility.

One way of measuring the impact of mobility is to cross-compare employment conditions with these migration profiles (Table 5). We use three variables: type of contract (permanent versus fixed-term), job status (executive versus non-executive), and wages (including bonuses). Non-migrants and returning migrants are less likely than the other categories to be on executive contracts, but there is no real difference when it comes to permanent contracts. It appears that all of the other categories are significantly better off than non-migrants in terms of salary.

Table 6 Numbers in each generation and level of qualification

Generation class	Year of survey	BTS/DUT	Licence	Master	Doctorate	Total
G1998	2001	124 921 (50.7%)	63 959 (25.9%)	49 352 (20.0%)	8 370 (3.4%)	246 602 (20.0%)
G2001	2004	119 624 (48.1%)	63 797 (25.7%)	52 645 (21.2%)	12 499 (5.0%)	248 565 (20.1%)
G2004	2007	117 393 (44.1%)	43 564 (16.4%)	92 909 (34.9%)	12 208 (4.6%)	266 074 (21.5%)
G2007	2010	94 930 (38.5%)	70 139 (28.5%)	67 422 (27.4%)	13 841 (5.6%)	246 312 (19.9%)
G2010	2013	67 918 (29.7%)	62 547 (27.4%)	83 196 (36.4%)	14 756 (6.5%)	228 417 (18.5%)
Total		524 787 (42.4%)	304 006 (24.6%)	345 524 (28.0%)	61 654 (5.0%)	1 235 970 (100%)

Source: CÉREQ data (generational surveys G1998, G2001, G2004, G2007, and G2010). The numbers shown in this table are weighted. The weighting variables ensure that the data are representative at the regional and qualification levels. Percentages for each level and generation are given in parentheses.

5 Conclusion

The geographical mobility of students and graduates is a significant issue for regional governments, influencing the policies they adopt with a view to attracting and retaining talent. Our research casts new light upon the empirical reality of this phenomenon with the help of data taken from the CÉREQ generational surveys.

Based on data from five generational surveys (1998, 2001, 2004, 2007, and 2010), we propose an empirical analysis of the movements of students in higher education and recent graduates. Our study shows that mobility between the region of origin, region of study, and region of employment is relatively low: fewer than one in three high school graduates moves to a different region to pursue their studies, and fewer than one in three university graduates moves to a different region to find employment. Nevertheless, mobility has followed a strong upward trajectory during the period we observed: (1) educational mobility doubled between the 1998 generation and the 2010 generation, increasing from 20 to 40%, (2) employment mobility increased from 25 to almost 40% between these two generations, and (3) return mobility rose from 19 to 35%. The increase in educational and employment mobilities, primarily among children of executives and graduates of Master's programmes, in fact, conceals a higher proportion of students returning to their home regions. This serves to illustrate the strong sense of geographical attachment felt by many people, even those who are highly qualified.

The low rate of mobility observed among students and graduates nonetheless conceals considerable heterogeneity in individual experiences, determined by different socio-demographic backgrounds, regions, choice of studies, etc. The fact that the

Table 7 Descriptive statistics

		Mobility						
		Overall	Education		Employment		Return	
			no	yes	no	yes	no	yes
Overall		65389	45607	19782	44422	20967	13483	6299
%		100	69.7	30.3	67.9	32.1	68.2	31.8
Generation	1998	32.2	37.0	21.0	35.8	24.6	25.0	12.5
	2001	10.9	11.6	9.2	12.0	8.5	10.8	5.8
	2004	20.6	19.6	23.1	19.9	22.1	21.0	27.6
	2007	16.7	14.9	20.8	14.5	21.3	18.4	25.7
	2010	19.6	16.9	25.9	17.8	23.5	24.7	28.3
Degree	BTS/DUT	41.1	47.1	27.3	45.8	31.2	23.2	36.1
	Bachelor	24.9	26.5	21.0	25.9	22.6	20.0	23.3
	Master	23.6	18.1	36.3	18.5	34.4	37.4	34.0
	Doctorate	10.4	8.3	15.4	9.8	11.8	19.5	6.6
Discipline	LEM	24.4	26.0	20.6	25.2	22.6	20.9	20.2
	L&L	5.6	5.6	5.5	14.7	11.5	12.8	12.4
	HSS	13.7	14.1	12.7	38.4	47.6	48.2	45.4
	BAS	41.3	38.7	47.3	6.0	4.8	6.2	4.2
	Health	15.0	15.6	13.8	15.7	13.5	12.0	17.8
Gender	Male	43.2	42.2	45.5	41.7	46.4	46.8	42.6
	Female	56.8	57.8	54.5	58.3	53.6	53.2	57.4
Mother	Born in France	92.2	91.3	94.2	91.3	93.9	94.5	93.5
	Born elsewhere	7.8	8.7	5.8	8.7	6.1	5.5	6.5
	Prof./exec	22.3	20.4	26.5	21.4	24.0	28.7	21.9
	Prof./non exec	53.1	56.3	45.7	55.4	48.2	45.2	46.9
	Other	24.6	23.3	27.7	23.1	27.8	26.1	31.3
Father	Born in France	91.2	90.2	93.4	90.2	93.2	93.7	92.6
	Born elsewhere	8.8	9.8	6.6	9.8	6.8	6.3	7.4
	Prof./exec	42.4	39.0	50.3	40.2	47.0	53.1	44.1
	Prof./non exec	51.3	54.7	43.4	53.2	47.3	40.5	49.8
	Other	6.3	6.3	6.3	6.6	5.7	6.4	6.2
Parents' residence (middle school)	Rural area	19.9	18.7	22.7	18.8	22.3	21.5	25.2
	Urban area	80.1	81.3	77.3	81.2	77.7	78.5	74.8
Years behind (middle school)	Yes	4.2	4.7	3.1	4.7	3.2	2.9	3.4
	No	95.8	95.3	96.9	95.3	96.8	97.1	96.6
Years behind (higher education)	No	55.3	55.9	54.0	53.9	58.3	52.7	56.7
	Yes, 1 or 2 years	30.7	30.5	31.1	30.8	30.4	31.4	30.4
	Yes, 3 years or more	14.0	13.6	14.9	15.3	11.2	15.8	12.8
In a relationship	Yes	52.5	51.2	55.6	52.6	52.4	57.5	51.5
	No	47.5	48.8	44.4	47.4	47.6	42.5	48.5
Children	Yes	11.6	11.3	12.2	13.1	8.3	13.7	9.0
	No	88.4	88.7	87.8	86.9	91.7	86.3	91.0

(continued)

Table 7 (continued)

		Mobility						
		Overall	Education		Employment		Return	
			no	yes	no	yes	no	yes
Employment during studies	No	38.0	38.6	36.5	36.9	40.3	35.5	38.8
	Occasionally	47.9	46.7	50.8	47.5	48.8	50.4	51.5
	Regularly	14.1	14.7	12.7	15.6	10.8	14.1	9.7
International mobility during studies	Yes	11.4	8.8	17.5	9.3	15.8	18.0	16.4
	No	88.6	91.2	82.5	90.7	84.2	82.0	83.6

Key: 41.1% of those leaving higher education have completed two-year programmes. Among those students who have (and have not) experienced educational mobility, 47.1% (27.3%) have two-year qualifications. The percentages are 31.2 (45.8%) for employment mobility (or the absence thereof). For those who have experienced educational mobility, the percentages of those returning and not returning are 36.1% and 23.2%, respectively, for those with two-year qualifications.

Table 8 Status three years after leaving school by diploma

	G1998	G2001	G2004	G2007	G2010
Short higher education cycle (BTS/IUT—Bachelor)					
Employment	90.0	90.7	89.1	86.8	83.2
Unemployment	5.0	5.3	5.2	6.9	8.9
Out-of-labour force	2.0	1.2	1.5	1.8	1.9
Return to school	3.0	2.8	4.1	4.5	6.0
Long higher education cycle (Master—Doctorate)					
Employment	92.2	88.3	88.8	88.7	87.1
Unemployment	4.7	9.8	7.0	8.3	9.4
Out-of-labour force	1.8	0.7	1.6	1.5	1.4
Return to school	1.3	1.2	2.6	1.5	2.1

Source: Céreq data (generational surveys G1998, G2001, G2004, G2007, and G2010). The numbers shown in this table are unweighted.

mobility rate is relatively low, combined with the importance of proximity effects and return mobility, strongly limits the room for manoeuvre available to local authorities when it comes to attracting students and recent graduates. In our opinion, there are two major contradictions between the positions adopted and the actions taken by territorial decision-makers with regard to geographical mobility. First and foremost, there appears to be a clear consensus emerging in favour of greater geographical mobility; the European Union has provided the framework for this ambition, with a view to creating a pan-European community for research, higher education, and the circulation of knowledge, an ambition which has been widely adopted at the territorial level. But at the same time, local decision-makers are concerned about “brain drain” or the loss of graduates, reducing the return on investment of funds allocated to education. The maxim which holds that “mobility is to be encouraged” thus runs up against the injunction that “we don’t want to see the best talent leaving

our region”, as if the (desirable) goal of mobility could somehow leave the region of origin unaffected. The second contradiction arises from the disconnect between the idea that mobility is too low (aggravating unemployment by preventing the smooth alignment of demand for and supply of labour) and the development policies implemented in the 1990s and 2000s to expand access to higher education in the regions. The creation of new universities, university outposts in medium-sized towns, and a large number of technical training institutes and BTS vocational programmes has paved the way for a genuine democratization of higher education, expanding access for students from modest backgrounds who are geographically isolated from the major seats of learning.

6 Appendix

See Tables 6, 7, 8

References

- Baron, M. (2009). Villes et régions en concurrence pour comprendre l’offre de formations universitaires ? *Espaces et sociétés*, 1(136–137), 135–154.
- Baron, M., Perret, C. (2008). Comportements migratoires des étudiants et des jeunes diplômés. ce que révèle le niveau régional. *Géographie, Économie, Société*, 10(2), 223–242. <https://doi.org/10.3166/ges.10.223-242>
- Ciriaci, D. (2014). Does university quality influence the interregional mobility of students and graduates? the case of Italy. *Regional Studies*, 48(10), 1592–1608. <https://EconPapers.repec.org/RePEc:taf:regstd:v:48:y:2014:i:10:p:1592-1608>
- Coquard, B. (2019). Ceux qui restent: Faire sa vie dans les campagnes en déclin. Collection L’envers des faits, La Découverte. <https://books.google.fr/books?id=6C2fygEACAAJ>
- Dahl, M. S., & Sorenson, O. (2010). The social attachment to place. *Social Forces*, 89(2), 633–658. <https://doi.org/10.1353/sof.2010.0078>, <https://academic.oup.com/sf/article-pdf/89/2/633/6884732/89-2-633.pdf>.
- Faggian, A., McCann, P., & Sheppard, S. (2007a). Human capital, higher education and graduate migration: An analysis of Scottish and Welsh students. *Urban Studies*, 44(13), 2511–2528.
- Faggian, A., McCann, P., & Sheppard, S. (2007b). Some evidence that women are more mobile than men: Gender differences in UK graduate migration behavior. *Journal of Regional Science*, 47(3), 517–539.
- Haussen, T., & Uebelmesser, S. (2018). Job changes and interregional migration of graduates. *Regional Studies*, 52(10), 1346–1359.
- Hoare, A., & Corver, M. (2010). The regional geography of new young graduate labour in the UK. *Regional Studies*, 44(4), 477–494.
- Kazakis, P., & Faggian, A. (2017). Mobility, education and labor market outcomes for us graduates: Is selectivity important? *The Annals of Regional Science*, 59(3), 731–758.
- Ma, K. R., Kang, E. T., & Kwon, O. K. (2017). Migration behavior of students and graduates under prevailing regional dualism: The case of South Korea. *The Annals of Regional Science*, 58(1), 209–233.

- Niedomysl, T., & Amcoff, J. (2011). Why return migrants return: Survey evidence on motives for internal return migration in Sweden. *Population, Space and Place*, 17(5), 656–673.
- Perret, C. (2007). Note de recherche typologie de l'insertion professionnelle des diplômés de l'enseignement supérieur dans les régions françaises au regard des mobilités géographiques. *Revue d'Economie Régionale & Urbaine* juillet (2), 293. <https://doi.org/10.3917/reru.072.0293>
- Rérat, P. (2014). Highly qualified rural youth: Why do young graduates return to their home region? *Children's Geographies*, 12(1), 70–86.
- Schwartz, A. (1973). Interpreting the effect of distance on migration. *Journal of Political Economy*, 81(5), 1153–1169.
- Sjaastad, L. A. (1962). The costs and returns of human migration. *Journal of Political Economy*, 70(5, Part 2), 80–93

Toward a FAIR Reproducible Research



Christophe Bontemps and Valérie Orozco

Abstract Two major movements are actively at work to change the way research is done, shared, and reproduced. The first is the reproducible research (RR) approach, which has never been easier to implement given the current availability of tools and DIY manuals. The second is the FAIR (Findable, Accessible, Interoperable, and Reusable) approach, which aims to support the availability and sharing of research materials. We show here that despite the efforts made by researchers to improve the reproducibility of their research, the initial goals of RR remain mostly unmet. There is great demand, both within the scientific community and from the general public, for greater transparency and for trusted published results. As a scientific community, we need to reorganize the diffusion of all materials used in a study and to rethink the publication process. Researchers and journal reviewers should be able to easily use research materials for reproducibility, replicability, or reusability purposes or for exploration of new research paths. Here we present how the research process, from data collection to paper publication, could be reorganized and introduce some already available tools and initiatives. We show that even in cases in which data are confidential, journals and institutions can organize and promote “FAIR-like RR” solutions where not only the published paper but also all related materials can be used by any researcher.

1 The Need for Reproducible Research

During the last decade, a great number of papers have been published on the problem of irreproducibility of research (Nature 2013) and on the crisis in science due to errors (Reinhart and Rogoff 2010) or fraud (Ioannidis 2005), leading to a lack of trust in published results. One response to this credibility crisis has been a renewal of interest

C. Bontemps · V. Orozco (✉)

Toulouse School of Economics - INRAE, University of Toulouse Capitole, Toulouse, France
e-mail: valerie.orozco@inrae.fr

C. Bontemps

e-mail: christophe.bontemps@inrae.fr

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_30

595

in the “reproducible research” (RR) approach, as defined initially by geologist John Claerbout as the possibility of the “*replication [of a paper] by other scientists*” (Claerbout 1990).

However, despite an apparent consensus on the general problem, the publication of papers exhorting the scientific community to publish reproducible results, and the dissemination of tools, good practices, and courses, we still observe considerable weaknesses in both researcher and journal practices, leading to the scarce dissemination of raw scientific materials.¹ Many journals do not have a replication or data and code availability policy, and others have implemented a simple supplementary materials section on their website. Thus, a clear organization and precise guidelines on how to achieve the initial goals of reproducibility in science are still lacking.

Science needs verification and thus several conditions have to be fulfilled: First, the research has to be done in such a manner that it can easily be reproduced. Second, the materials used to produce the results have to be available to others. Finally, somebody, i.e., a referee for a journal or another researcher, has to reproduce and validate the published results using the materials available. These conditions may seem very strong, but we argue that they are necessary to prove the validity of any research. Our message here is that even in empirical work where we use data and code to produce a result, we have to prove our findings. We follow the idea of LeVeque (2009) that “*constructing a computer program isn’t so different from constructing a formal proof*” and claim that reproducing a result issued from a computer program should not be different from reproducing a formal proof.

This paper is devoted to two practical problems that have received little attention in economics and statistics so far: How, in practice, can we ensure that the results published in a paper have been reproduced and verified? How are all the materials used to produce the results of that paper shared with the community? These are very complex questions that can be even more complex when materials are confidential. Our goal here is thus to question the overall organization of research leading to the publication of a paper.

The paper is organized as follows. In the next section, we identify and illustrate the current problems that limit the reproducibility of research and survey some important initiatives that have been proposed. In Sect. 3, we introduce some recent initiatives and propose new schemes involving researchers, journals, and the research community. We illustrate the problem of sharing research materials when they are confidential in Sect. 4. Section 5 lists the incentives and impediments related to the proposed approach and concludes the paper.

¹We consider here that the research process starts once the data are collected and in possession of the researcher. We do not address here the issue of reproducibility for data collection in experimental economics or field experiments (Bowers et al. 2017).

2 Reproducible Research in Practice

The notion of reproducibility has been discussed lengthily in the literature, sometimes with conflicting terminologies (Benureau and Rougier 2018). We follow Barba (2018) who summarizes it by the equation “same data + same method = same results”.²

For Gentleman and Temple Lang (2007), this idea means that the “data+code” *compendium* used in a paper is made available to the readers so that they can first verify (reproduce) the results and second conduct alternative analyses of the work. The notion of reproducibility is thus related to the similar notion of verification or scientific proof.

In complement to RR, the Open Science movement emerged 10 years ago and aims at sharing data used in research as a patrimonial and cumulative goal. This movement has seen the involvement of many stakeholders, communities, and institutions (e.g., the Open Government Partnership, the Center for Open Science, and the Research Data Alliance).³ These initiatives have focused primarily on the big questions behind open science implementation such as the FAIR principles (findable, accessible, interoperable, and reusable) proposed by Wilkinson et al. (2016). FAIR does not mean open but, in brief, requires some accessibility to findable elements (most often datasets or at least metadata). The principles call for materials to be shared in a format that others can use and reuse.

These two movements are very active and influence the way institutions, research centers, and national statistical offices (NSOs) construct their infrastructures and data centers. To illustrate how these movements may affect the research publication landscape, we propose to reduce all the materials needed to produce a paper to only three key elements: the data, the code, and the workflow, even if elements such as the documentation and the computing environment are also of great importance for some papers. We use the pictograms presented in Fig. 1 throughout the paper. In view

Fig. 1 Pictograms of inputs and outputs and of main actors (simplified)



²We will not discuss here the question of the precise meaning of “same results”.

³At the European level, one should mention OpenAIRE and in France the “Plan national pour la science ouverte” (<https://www.ouvrirelascience.fr/>).

of the RR approach as defined by Claerbout (1990), all these elements, not only the research paper, should be shared with the scientific community.

We also illustrate the process leading to the publication of a paper by focusing on 3 major actors: the researcher(s), the journal that handles the publication process, and the scientific community that should benefit from a new publication. Later in the paper, we will see that other actors, such as research institutions, data providers, and funding agencies, may play an important role in the publication process and its outcome.

Since the publication of the seminal paper by Claerbout (1990), several authors have pointed out weaknesses in researchers' day-to-day practices and have proposed tools, solutions, and advice. Many of these proposals underline existing ways of improving our practices toward RR, focusing on technical solutions. Applied statisticians and econometricians can now enjoy tools developed in R, Python, Stata, SAS, and MATLAB (Orozco et al. 2020). Others (Baiocchi 2007) have identified possible organizational improvements at the researcher level and proposed principles to link a paper to its raw components, data and code through a clear workflow, following the original idea of *reproducible research documents* proposed by Knuth (1984, 1992) and his *literate programming* approach. Other principles include a clear organization of work and files, greater attention to versioning, good documentation of the research workflow, good writing practices for code using layouts, and naming conventions. Automating the whole workflow is also recommended and encouraged.

We have also seen the emergence of companion websites and “executable” papers allowing online code editing and execution (Hurlin et al. 2014; Gorp and Mazanek 2011). Platforms (e.g., Code Ocean, Exec&Share, and SHARE) have been created to allow code to be run online using materials stored by other researchers. With this technology, researchers from around the world are able to rerun the exact same code as the author, change parameter values to see the impact on the results, or even replicate the code with another dataset.

However, despite the efforts observed and all the tools and methods mentioned above, implementing RR has often been a challenge in practice. In a recent survey, Chang and Li (2017) attempted to update the seminal work of Dewald et al. (1988) and successfully replicated only 33% of 67 papers published in economics journals. Other examples in other disciplines exhibit similar features (Miyakawa 2020). This current situation, whereby journals do not ask for or check the raw materials used to produce a research paper, is represented in Fig. 2.⁴

There are many reasons for the overall “irreproducibility” (Nature 2013) of research. It is true that crafting reproducible papers may require more time and effort than that needed for papers with code that will be used only once. However, when researchers, especially young ones, invest in RR practices and tools, they generally become more efficient in their day-to-day practices. They are able to reuse their own materials, reproduce their own results, answer referees' questions more quickly, and test various specifications of their models with little effect on the time spent reporting

⁴See also Table 2 in Appendix 1, for a synthesis of the cases presented throughout the paper.

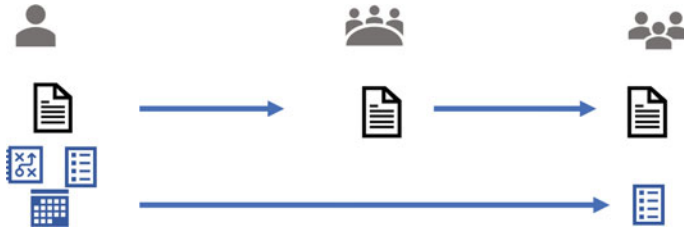


Fig. 2 The researcher (left) uses code and data following a (written) workflow and submits a (reproducible) paper to a journal (middle) that neither asks for nor checks the raw materials. The researcher then decides to share (or not) some materials (e.g., the code) with the community (right)

new results in a paper (McCullough 2009). There is also some evidence that papers that share their research materials are more likely to be cited (Christensen et al. 2019).

The way researchers apply RR methods and share their research materials may be another important reason for the prevalence of irreproducibility. When they share their materials, researchers do it in an unstructured manner, and many do not link the materials to each other or to the associated publication (Baker 2016). Even if available, the data may not be easily accessible. In their book, Christensen et al. (2019) remark that when shared, submitted data were also frequently an “unlabeled and undocumented mess”. Most of the time, researchers decide on their own initiative to share (or not) their materials, on their website, on GitHub, or “upon request”.

Data availability and data accessibility are thus ubiquitous problems.⁵ These problems are not the sole responsibility of researchers. Many scientists currently do share code and data, but this takes time and effort. We have to acknowledge that even if some researchers do share their research materials, others simply do not want to. In other disciplines, data collection represents a substantial effort. Researchers may not want to lose their investment and do not want to share this rent with others after a first publication. The lack of incentives to share and the lack of sharing solutions are thus two barriers to reproducibility. They exist either because no clear data and code-sharing policies are defined by journals or because the technical solutions proposed are not good ones.

In Figs. 2 and 3, we illustrate different unsatisfactory yet frequently observed situations in which the research community has no access to the original raw materials even when the paper was originally designed by the researcher to be reproducible. In Fig. 3, only the code is finally shared with the research community, even if the journal has access to all the materials. The materials may be “available upon request”, allowing the author the freedom to decide what to share. It is not guaranteed in either case that the paper was crafted to be reproducible nor that the referees were able to reproduce the results when reviewing and then accepting the paper for publication. Several examples of errors, such as the famous one in Reinhart and Rogoff (2010),

⁵Other issues that we do not address directly here include the digital preservation of research data (Akers and Doty 2013) or the preservation of software (Di Cosmo and Zacchiroli 2017).

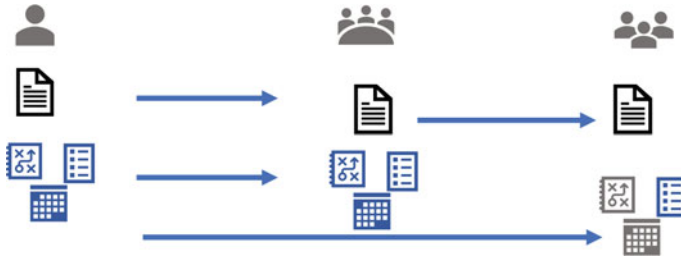


Fig. 3 The researcher (left) sends all materials to the journal (middle) but in the end shares some material (e.g. code) directly online (e.g. GitHub)

cast doubt on the role of referees and their ability or willingness to ask for and use submitted materials to reproduce results during the referee process.

This issue is not limited to applied research: Even theoretical mathematics papers display the same types of problems, with some analyses failing to be reproduced during the referee process (Gouëzel et al. 2019).

For journals, reproducing a submitted paper requires many combined conditions that are still rarely met. First, the journal's policy must require that all materials be sent before or during the referee process. This implies that the submitted paper must have been done in a reproducible manner by the researchers in the first place. Second, the referees must have the skills, willingness, and incentives to check the empirical proofs. Finally, the paper has to be reproduced, which implies some technical requirements, time, and resources. Even if a journal is willing to reproduce a paper, it may not be able to do so and such conditions are still rarely fulfilled. In fact, the situations depicted in Fig. 3 in which the researchers are the sole party responsible for the quality and reproducibility of the files shared and no one checks what is stored are still the most commonly observed.

In Figs. 4 and 5, we illustrate different types of observed situations where the community has access to a verified paper.⁶ These solutions assume that journals have a clear data and code availability policy that is enforced. This requires that the raw materials provided by the researchers closely follow the RR principles. The whole workflow, including the curation of the original (raw) dataset collected either by the researcher or through a data provider, should be written in a readable form for humans and computers, following the literate programming ideas proposed by Knuth (1992).

If these mandatory conditions are fulfilled by researchers, journals may implement two different strategies. They could try to reproduce the results and then signal or certify that the results are correct (Figs. 4 and 5). This is what is expected from any scientific journal willing to maintain a reputation as a trusted publication. Alternatively, if a journal lacks sufficient human or financial resources to achieve this task, it should at least organize the diffusion of the materials and leave the verification

⁶In these figures, for clarity reasons, we do not illustrate the fact that researchers may share their materials themselves.



Fig. 4 The researcher (left) sends all materials to the journal (middle). The journal shares only the code online for the community (right) and signals that the code has been carefully checked, either internally or by a trusted third-party agency

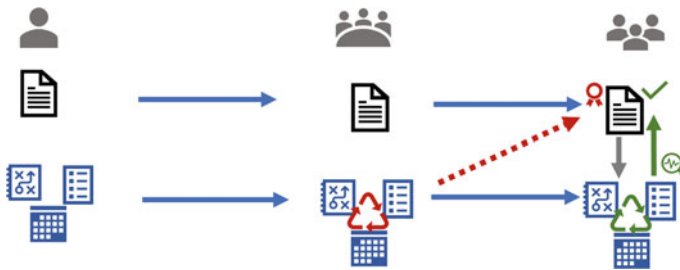


Fig. 5 The researcher (left) sends all materials to the journal (middle). The journal or a trusted third party certifies that the materials used in the paper actually reproduce the results. The journal also shares all the materials (code—data—workflow) online for the community (right), which can also check, reproduce, and reuse the materials

process to the community.⁷ The best solution is illustrated in Fig. 5, where the journal or a trusted third party reproduces and checks the paper’s results and then organizes the sharing of the materials they have used for the benefit of the community. Such a transparent organization may seem difficult to establish in practice, but it is in fact already implemented by some journals.

3 Implementing FAIR and RR Principles in Practice

In an empirical study examining 346 journals in economics and business studies, Vlaeminck and Herrmann (2015) showed that only 20% of the journals have a data policy. In Table 1, we compile the data and code availability policies for statistics

⁷In 2003, H. Pesaran announced the creation of a new section of the Journal of Applied Econometrics dedicated to the replication of published empirical papers (Pesaran 2003). Since then, some journals have followed this idea leading to an increase in the number of replication papers in economics (Mueller-Langer et al. 2019). The site PubPeer (<https://pubpeer.com/>) is also a way to allow users to discuss and review scientific research.

Table 1 Overview of statistics and economics journal policies in 2020

Journal	Policy	Platform used
<i>(In bold mandatory policies, in plain sharing is encouraged)</i>		
Statistics journals		
Annals of Statistics	–	–
Annals of Applied Statistics	(data+code) sharing	– (Supplementary materials archive)
Biometrics	(data+code) sharing	multiple (OSF, Dataverse)
Biometrika	–	–
Comput Stat & Data Analysis	(data+code) sharing	multiple (Mendeley, DRYAD, ICPSR, RunMyCode, ...)
Electronic Journal of Statistics	code sharing	Statlib
J of Business & Econ. Stat.	(data+code) sharing	Figshare
J of Comp. and Graph. Stat.	(data+code) sharing	Figshare
J of Multivariate Analysis	data citation	–
JASA	(data+code) sharing	JASA Dataverse, JASA GitHub
J of the Royal Statistical Society	(data+code) sharing	–
J of Statistical Software	(data+code) sharing	– (Supplementary materials archive)
Stat. Methods in Medical Research	(data+code) sharing	– (Supplementary materials online)
Statistics & Probability Letters	data citation	–
Statistics and Computing	data sharing	multiple (Figshare, Dryad, openICPSR, Dataverse)
Stoch. Proc. & their Applications	data citation	–
CSBIG	data sharing	–
Economics journals		
Am Econ Review	(data + code) sharing	OpenICPSR (AEA Data and code repository)
J Finance	code, data sharing	–
Q J Economics	(data + code) sharing	data repository (Dataverse) linked to the QJE website

Table 1 (continued)

Econometrica	(data + code) sharing	– (Supplementary materials webpage)
J Financial Econ	data + code sharing	multiple (Mendeley, DRYAD, ICPSR, RunMyCode)
J Political Econ	(data + code) sharing	– (JPE website)
Rev Financial Stud	–	–
J Econ Theory	(data + code) sharing	multiple (Mendeley, DRYAD, ICPSR, RunMyCode)
Rev Econ Studies	(data + code) sharing	Oxford Journals Review Archive
J Econometrics	(data + code) sharing	multiple (Mendeley, DRYAD, ICPSR, RunMyCode)
J Econ Literature	(data + code) sharing ^{AEA}	OpenICPSR ^{AEA}
J Monetary Econ	(data + code) sharing	multiple (Mendeley, DRYAD, ICPSR, RunMyCode)
J Econ Perspectives	(data + code) sharing ^{AEA}	OpenICPSR ^{AEA}
Rev Econ & Stat	(data + code) sharing	–
Eur Econ Review	(data + code) sharing	– (EER website)
Int Econ Review	–	–
J Int Econ	data, code sharing	Mendeley repository
Economic Journal	–	–
J Public Econ	(data + code) sharing	multiple (Mendeley, DRYAD, ICPSR, RunMyCode)
Game Econ Behav	(data + code) sharing	multiple (Mendeley, DRYAD, ICPSR, RunMyCode)
RAND J Economics	–	–
J Money Credit Bank	(data + code) sharing	web data archives
Economic Theory	data sharing	multiple repositories
J Bus & Econ Stat	(data + code) sharing	- (Supplementary online materials)
Economics Letters	data, code sharing	multiple (Mendeley, DRYAD, ICPSR, RunMyCode, ...)
J Appl Econometrics	data, code sharing	JAE data archive
A J Political Science	(data+code) sharing	AJPS <i>Dataverse</i>

^{AEA} indicates that the journal follows the strict AEA Data and Code Availability Policy. We use here a nonexhaustive list of journals. In statistics, we select the most important ones according to the Web of Science index. In economics, we use the journals listed in McCullough (2009). We choose to eliminate some specialized journals or journals publishing mainly theoretical work. (see <https://www.aeaweb.org/journals/policies/data-code/>)

and economics journals as published on their websites in 2020. We can observe a great heterogeneity of practices. From this nonexhaustive list, we confirm that many journals still do not publish any policy at all. Even if the publication of erroneous results has probably helped journals improve their referee process, the vast majority only “encourage” authors to share their materials, sometimes without any guidance on how or where to put the materials. Until very recently, very few (e.g., JSS and AJPS) checked that the code runs and reproduces the key results (Christian et al. 2018). Moreover, Duvendack et al. (2017) showed that less than 10% of economic journals (28 out of 333) have a majority of their empirical papers supplying data and code. This suggests that even when data and code policies are written, their enforcement is lax.

Nevertheless, there are some very good examples that should be inspiring. The JASA, for example, has an “associate editor for reproducibility”, responsible for the technical review of manuscripts before, during, or after the usual review process (Fuentes 2016). The Journal of Statistical Software (JSS) follows the scheme of Fig. 5 and asks for a standalone replication script that must enable reproducibility. Since 2019, all journals published by the American Economic Association require that authors share their data and code, which are systematically checked “within reasonable limits of time and computing resources”.⁸ These verifications can be costly for journals, but we may expect that the costs should decrease over time with the improvement of researchers’ and reviewers’ practices.⁹ This process can also be outsourced to specialists. An example is given by the American Journal of Political Science (AJPS), which contracted the Odum Institute for Research in Social Science to systematically check that research materials confirm the results of submissions (Crabtree 2011).

To succeed in organizing the way materials are shared while preserving the link with the results included in the paper, journals could extend the FAIR principles, developed primarily for datasets, to all the research materials. In practice, sharing data and other materials together can be quite complicated. Moreover, we believe that researchers should not organize the sharing themselves and that journals should align a strict mandatory policy with a clear organization and resources. In this regard, the Journal of Applied Econometrics (JAE) was a sort of pioneer, implementing its own data archive from the late 1980s. One satisfying technical solution is now offered by the *Dataverse Network*, developed at Harvard University. The network hosts collections of studies, embedding all materials for a paper in a single object, called a *dataverse* (King 2007; Leeper 2014). This solution is recommended by some journals (e.g., JASA, AJPS, QJE, PLOS, and Nature). Other journals, following

⁸Some useful resources facilitate the process (see https://social-science-data-editors.github.io/guidance/Verification_guidance.html). The Transparency and Openness Promotion (TOP) proposes also varying levels of replication policies for journals (Nosek et al. 2015).

⁹Jacoby William (2017) analyzed the AJPS verification policy and reported an average of 8 person-hours per manuscript to curate and replicate the analyses. The publication workflow, involving more rounds and resubmissions, is also much longer.

the recommendations of research institutions (NSF, ERC), use *Figshare*, *Zenodo*, *Mendeley*, or *ICPSR* (see Table 1).¹⁰

Following the FAIR principles implies not only that the materials should be shared or accessible, but also that they should be interoperable and findable. The interoperability principle can be interpreted here as the ability for any reader to have access to materials stored in a readable format.¹¹ It should also be easy to find the materials without any ambiguity. We agree with the American Sociological Review that “*citing datasets used in published research is just as important as citing journal articles, books, and other sources that contributed to the research*”. This means that the data and other materials should be identified in a consistent way. To precisely identify datasets and code, journals or institutions may request that a digital object identifier (DOI) be attached to each element described in the paper. DOIs are the backbones of the findable component of the FAIR principles. Even for data with constrained access (e.g., proprietary or confidential data), a DOI should provide enough elements, at least the metadata, to retrieve enough information describing the data (or other materials) used (Fenner et al. 2017).¹²

In the context of big data or machine learning analysis, sharing the materials and the analysis are important issues that can be challenging (Crosas et al. 2015). The R package *Zelig* automatically creates a workflow embedding all the procedures and algorithms used in the analysis into a single object that may then be exported and shared. The NSF-funded Whole Tale (<https://wholetale.org/>) may also be a scalable solution enabling the creation, publication, and execution of “tales” or executable objects embedding data (potentially big data), code, and the complete software environment used to produce research findings.

Another difficulty may be due to the length of the publication process, which can be quite extensive. The reviewing process often requires additional tests or modifications. Thus, the code as well as the datasets and even the workflow may change with the evolution of the paper under review. Researchers following the RR approach and writing RR papers are already familiar with version control tools such as GitHub for their code. They should easily integrate their data and workflow in the same spirit.

Another large step would be to question the access policies of publications, which are often paywalled, and to promote open-access publications such as PLOS (<https://plos.org/>) or arXiv (<https://arxiv.org/>) in our fields.

¹⁰A complete list of solutions is detailed in *The Registry of Research Data Repositories* (<http://re3data.org>) a service of *DataCite*. In addition, *CoreTrustSeal* provides certification to repositories and lists the certified ones.

¹¹For datasets, the FAIR interoperability principle suggests the use of open formats such as CSV files instead of proprietary formats (.xls). For code, open-source software should be preferred to avoid exclusive access (Vilhuber 2019). The metadata should also follow standards (Dublin core or DDI). References and links to related data should also be provided (Jones and Grootveld 2017).

¹²The *DataCite* project (Brase 2009) is a popular resource to locate and precisely identify data through a unique DOI.

4 Confidential Data

Confidential or proprietary data are often cited as an obstacle to reproducibility, mainly because of data accessibility restrictions such as those imposed by the European GDPR.¹³ In economics, Christensen and Miguel (2018) observed that there has been a small increase in empirical papers (using data) published coinciding with a significant increase in data exemptions. They found that nearly half of the papers using data are not reproducible because the data are not available. These exemptions may be due to the use of confidential data or to other restrictions limiting data availability.

Restrictions may also come from data providers' preservation policies or national statistical offices (NSOs) data management conditions allowing only remote and strictly controlled access to data through secure virtual terminals.¹⁴ Journals and reviewers are then unable to access the data and cannot check the results, having access to the paper and the code only.¹⁵ According to Lagoze and Vilhuber (2017), 50% of confidential data used in papers are from NSOs, and each NSO has its own data-sharing restriction and regulation. To handle the complexity of security-level restrictions and to allow third parties or reviewers to access confidential materials, Sweeney et al. (2015) proposed a system of *datatag* repositories. Each *datatag* repository documents the way data and other sensitive materials can be shared, reducing the complexity of the situation to a small number of tags.

Some authors propose altering original raw data into "safe data", potentially accessible by anybody, using blurring or aggregation techniques to remove sensitive details such as individual information (Alter and Gonzalez 2018). Other methods include adding random noise or swapping individual responses between otherwise similar respondents while maintaining the same likelihood distribution (Boker et al. 2015). In our view, these methods do not seem compatible with the principles of transparency used in the RR approach.

Nonetheless, solutions exist to preserve privacy while allowing other researchers to access and replicate results. One principle is based on a trusted third party having a secure access to confidential data and on an interactive platform for queries and answers (Dwork et al. 2009). The curator model (Crosas et al. 2015), depicted in Fig. 6, could be implemented using secured-sharing platforms such as *dataverses* or *datatags*.¹⁶ Referees could have access to materials securely and reproduce and check the validity of the results, even with confidential data. A reproducibility certification

¹³ There are many sources of confidential and nonshareable data (Christensen and Miguel 2018; Lagoze and Vilhuber 2017).

¹⁴ In France, the CASD (<https://www.casd.eu/>) is a single-access portal to many public data providers (INSEE, ministries, etc.). Researchers are not allowed to copy all the materials locally on their machine, and only some type of outputs can be extracted.

¹⁵ The code may also contain some confidential elements. In particular, the code used for the initial data curation may contain, e.g., brand or city names and addresses.

¹⁶ Some data providers, in particular NSOs, already perform RR on their confidential data, controlling output files and code, to check for confidentiality restrictions (Lagoze and Vilhuber 2017).



Fig. 6 The researcher (left) shares all materials (code—data—workflow) privately with the trusted third party. Access may be contracted with the journal (middle). A certification of reproducibility is sent to the journal together with the paper. The paper is published with the reproducibility certification, but only nonconfidential elements, such as code, are available to the community (right)

could be emitted during or even before the submission process to attest to the full reproducibility of the results.

The recent and promising Certification Agency for Scientific Code and Data (CASCAD), supported by the French National Science Foundation (CNRS), proposes the solution of prepublication reproducibility certification (Pérignon et al. 2019). Authors can ask CASCAD to review their materials, even when confidential data are used, and to certify the reproducibility of their paper’s results (Fig. 6). The replication process can be performed before publication, which facilitates the journal review process. Alternatively, journals may outsource this task to the certification agency during the referee process. Ex ante contracts between the trusted third party, the data provider, and the journal may facilitate the process. The code used and the certification report may then be hosted on an open-access repository, such as *Zenodo*, and be accessible by the research community.

It is possible to attest that a paper using confidential data has been reproduced so that no doubt remains over the validity of the results. However, only a few people, namely, the third-party agency or some reviewers, had access to the research materials, as only the certified code may be finally shared with the community.

Thus, working with confidential data does not necessarily mean having access to the dataset. Technical solutions, such as the one illustrated in Fig. 7 and based on the idea of “*data enclaves*”, exist. The Inter-university Consortium for Political and Social Research (ICPSR) has developed virtual and physical “*data enclaves*”, allowing online data analysis with strict restrictions on queries, data access, and downloads (Dunn and Austin 1998). At the time of its creation long ago, physical access to ICPSR resources and all the outputs created were humanly controlled. Currently, an online data analysis tool is used to evaluate output for disclosure risk prior to displaying for the end user. This is again a curator model accepting queries from any researcher and providing only approved analysis output, including the elements (estimations, tables, and figures) published in the paper.

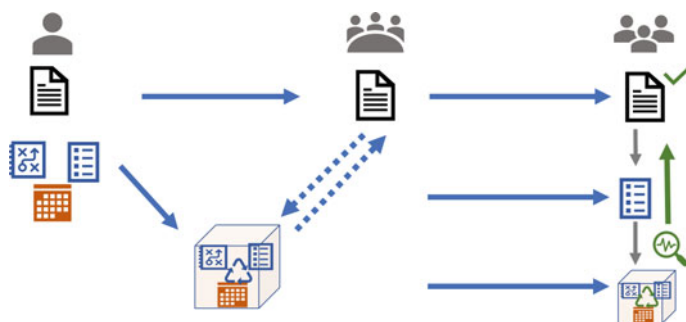


Fig. 7 The researcher (left) shares all materials (code—data—workflow), preferably as a *package* or *tale* that is securely accessible by the journal’s reviewers (middle) within a secured platform. The paper is published with access to a secured query platform (“*data enclave*”, such as the ICPSR). The community (right) has access to the secured query platform and can use the materials (e.g., code) without having access to the confidential elements

5 Conclusion

To many, the path toward findable, accessible, interoperable, and reproducible research may seem paved with obstacles. We argue here that this is the path to the future, considering the great challenges that we, as a scientific community, should overcome. Erroneous publications, unavailable research materials, and long and sometimes archaic publication processes have generated a research crisis within the research community and within society itself. Some publications (Science 2011) advocate for a great change in the individual and collective practices of scientists, journals, funders, institutions, and societies, acknowledging *Claerbout’s principles* that “*an article (· · ·) in a scientific publication is not the scholarship itself, (· · ·) the actual scholarship is the complete set of instructions which generated the figures*” (de Leeuw 2001). We argue that our community and the public at large will greatly benefit from a change toward greater transparency and better-organized research.

This change will only occur if all research actors agree to adhere to FAIR and reproducible research principles. Most of these principles can be gradually implemented as a growing process toward more reproducible practices.

Researchers should initiate these changes. They now have all the resources needed to improve their individual practices to create more reproducible papers by embedding code and data in a documented and written workflow understandable by others, including researchers’ own “future selves” (Gentzkow and Shapiro 2013). They are probably in the process of changing their habits already under the pressure of certain journals and research institutions such as the ERC and NSF.

Researchers are also reviewers, some of them even journal editors, and thus can promote many valuable actions. First, journals should reorganize the review process and the way the results of submitted papers are checked, including when some materials are confidential. This is probably one of the most challenging issues, requiring new skills for reviewers, additional resources, and a clear internal setup for sharing the submitted paper’s materials. Certified trusted third parties already exist if that process has to be outsourced. Having an “associate editor for reproducibility”, as the

JASA does, could also be a good idea to solve many practical questions such as when to check the validity of the submitted materials (before, during, or after the classical review) or to organize the verification and the relations with the author. Second, journals should have a clear data and code availability policy with a proper check of materials by reviewers. Journals should also organize the way research materials are shared.¹⁷ Contracting with public repositories such as *Zenodo*, *Mendeley*, or *Figshare* would reduce the constellation of individual and self-maintained repositories and the fragmentation of arbitrarily different, incompatible standards (Sansone et al. 2019). These platforms also guarantee perennial access to the core materials of science. Finally, if implemented, this process will provide strong incentives for researchers to only produce RR papers. Some journals (e.g., *AJPS* and *Biostatistics*) propose badges tagging reproducible research. Such a practice seems to increase the proportion of papers using open practices and to improve the preservation of research materials (Rowhani-Farid and Barnett 2018). Other challenges concern the free access to articles published in paywalled journals and the recognition of open-journals such as *PLOS* and *ArXiv*.

Data providers, whether they are private or public, could also facilitate the changes that we call for. Quite often, they are aware of these problems and have implemented some processes for their internal publications that could be inspiring (Lagoze and Vilhuber 2017). In the near future, they may pay increasing attention to the use of the data they provide, checking published results either to publicize their activity or to criticize a misuse.¹⁸ In the case of confidential data, providers and, in particular, NSOs may also find some interest in promoting and organizing the way their data are findable and accessible (Pérignon et al. 2019). Thus, data providers should encourage researchers, institutions, and journals to produce more reproducible and reproduced papers. It is therefore likely that partnerships among journals, data providers, and private or public third parties will increase in the future.

Research institutions have already started to impose some conditions on funded projects or grants by requiring researchers to follow a strict RR approach, by promoting the dissemination of the FAIR and RR approaches, and by financing public infrastructures hosting FAIR research materials repositories.¹⁹

If “*science is organized knowledge*” (Spencer 1854), then we should all work for better organization for better science. We believe that the FAIR and reproducible research movements are there to jointly provide organized resources, tools, and practices. Changing the publication workflow and our habits may be a long and probably costly journey. Not changing could be even more costly.

¹⁷Alter and Gonzalez (2018) suggested that to “protect” researchers who want to use their data first (before sharing), journals can propose an “embargo”.







¹⁸A recent lawsuit involving the popular training program CrossFit showed that a paper by Smith et al. (2013) erroneously showed an increased risk for injuries for its users. Although the paper was retracted later, the impacts on the researcher’s career were severe (for details, see <https://retractionwatch.com/>).

¹⁹The European Research Council (ERC) recommends “to all its funded researchers that they follow best practice by retaining files of all the research data they have used during the course of their work and that they be prepared to share this data with other researchers”.

Acknowledgements Christine Thomas-Agnan is our former teacher and a great colleague always available for helpful and interesting discussions for several years. We were very happy to write this paper as a demonstration of our gratitude. The authors wish also to thank the participants of the Banco de Portugal Reproducible Research Workshop in Porto (2019) for the stimulating discussions, which are at the origin of this paper. We are grateful to Virginie Piguet as well as the two anonymous referees for their careful reading and inspiring comments and suggestions.

Appendix 1: Synthesis of All Situations Illustrated on Figs. 2–7

Table 2 Comparison of the various situations presented in the paper

	Who shares? and What?			
	Researcher		Journal	
	Shares to journal	Shares to community	Shares to community	Verification & signal
 <p>(Fig. 2)</p>	–	Some materials (e.g. the code)	–	–
 <p>(Fig. 3)</p>	All materials	Some materials (e.g. the code)	–	–
 <p>(Fig. 4)</p>	All materials	(maybe)	Code	RR verification & signaling code check
 <p>(Fig. 5)</p>	All materials	(maybe)	All materials	RR verification & RR certification
 <p>(Fig. 6)</p>	All materials (to a third party)	–	Nonconfidential materials (such as code)	RR verification & RR certification
 <p>(Fig. 7)</p>	All materials (to a secured platform)	–	All materials (use only) (on a secured query platform) (no access to confidential elements)	(maybe)

References

- Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*, 8(2), 5–26.
- Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, 73(2), 146–156.
- Baiocchi, G. (2007). Reproducible research in computational economics: Guidelines, integrated approaches, and open source software. *Computational Economics*, 30(1), 19–40.
- Baker, M. (2016). Why scientists must share their research code. *Nature News*.
- Barba, L. A. (2018). Terminologies for reproducible research. arXiv preprint [arXiv:1802.03311](https://arxiv.org/abs/1802.03311).
- Benureau, F. C. Y., & Rougier, N. P. (2018). Re-run, repeat, reproduce, reuse, replicate: Transforming code into scientific contributions. *Frontiers in Neuroinformatics*, 11, 69.
- Boker, S. M., Brick, T. R., Pritikin, J. N., Wang, Y., von Oertzen, T., Brown, D., et al. (2015). Maintained individual data distributed likelihood estimation (middle). *Multivariate Behavioral Research*, 50(6), 706–720.
- Bowers, J., Higgins, N., Karlan, D., Tulman, S., & Zinman, J. (2017). Challenges to replication and iteration in field experiments: Evidence from two direct mail shots. *American Economic Review*, 107(5), 462–65.
- Brase, J. (2009). DataCite - A global registration agency for research data. In *2009 4th International Conference on Cooperation and Promotion of Information Resources in Science and Technology* (pp. 257–261).
- Chang, A. C., & Li, P. (2017). A preanalysis plan to replicate sixty economics research papers that worked half of the time. *American Economic Review*, 107(5), 60–64.
- Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3), 920–80.
- Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science*. Berkeley: University of California Press.
- Christian, T.-M., Lafferty-Hess, S., Jacoby, W., & Carsey, T. (2018). Operationalizing the replication standard: A case study of the data curation and verification workflow for scholarly journals. *International Journal of Digital Curation*, 13(1), 114–124.
- Claerbout, J. (1990). Active documents and reproducible results. *SEP*, 67, 139–144.
- Crabtree, J. D. (2011). Odum institute user study: Exploring the applicability of the dataverse network.
- Crosas, M., King, G., Honaker, J., & Sweeney, L. (2015). Automating open science for big data. *ANNALS of the American Academy of Political and Social Science*, 659(1), 260–273.
- de Leeuw, J. (2001). Reproducible research. The bottom line.
- Dewald, W. G., Thursby, J. G., & Anderson, R. G. (1988). Replication in empirical economics: The journal of money, credit and banking project: Reply. *American Economic Review*, 78(5), 1162–1163.
- Di Cosmo, R., & Zacchiroli, S. (2017). Software heritage: Why and how to preserve software source code.
- Dunn, C. S., & Austin, E. W. (1998). Protecting confidentiality in archival data resources. *IASSIST Quarterly*, 22(2), 16–16.
- Duwendack, M., Palmer-Jones, R., & Reed, W. R. (2017). What is meant by “replication” and why does it encounter resistance in economics? *American Economic Review*, 107(5), 46–51.
- Dwork, C., Naor, M., Reingold, O., Rothblum, G. N., & Vadhan, S. (2009). On the complexity of differentially private data release: Efficient algorithms and hardness results. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing* (pp. 381–390).
- Fenner, M., Crosas, M., Grethe, J., Kennedy, D., Hermjakob, H., Rocca-Serra, P., et al. (2017). A data citation roadmap for scholarly data repositories. bioRxiv.
- Fuentes, M. (2016). Reproducible research in JASA. *AMSTAT News: The Membership Magazine of the American Statistical Association*, 17.

- Gentleman, R., Temple Lang, D. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16(1), 1–23.
- Gentzkow, M., & Shapiro, J. (2013). Nuts and bolts: Computing with large data. In *Summer Institute 2013 Econometric Methods for High-Dimensional Data*.
- Van Gorp, P., & Mazanek, S. (2011). SHARE: A web portal for creating and sharing executable research papers. *Procedia Computer Science*, 4, 589–597.
- Gouëzel, S., & Shchur, V. (2019). A corrected quantitative version of the Morse lemma. *Journal of Functional Analysis*, 277(4), 1258–1268.
- Hurlin, C., Pérignon, C., & Stodden, V. (2014). RunMyCode.org: A novel dissemination and collaboration platform for executing published computational results. Open Science Framework.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Jacoby William G., Lafferty-Hess, S., & Christian, T.-M. (2017). Should journals be responsible for reproducibility?
- Jones, S., & Grootveld, M. (2017). How FAIR are your data?
- King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods & Research*, 36(2), 173–199.
- Knuth, D. E. (1984). Literate programming. *The Computer Journal*, 27, 97–111.
- Knuth, D. E. (1992). Literate programming. Center for the Study of Language and Information.
- Lagoze, C., & Vilhuber, L. (2017). O privacy, where art thou? Making confidential data part of reproducible research. *CHANCE*, 30(3), 68–72.
- Leeper, T. J. (2014). Archiving reproducible research with R and dataverse. *R Journal*, 6(1).
- LeVeque, R. J. (2009). Python tools for reproducible research on hyperbolic problems. *Computing in Science and Engineering (CiSE)*, 19–27. Special issue on Reproducible Research.
- McCullough, B. D. (2009). Open access economics journals and the market for reproducible economic research. *Economic Analysis and Policy*, 39(1), 117–126.
- Miyakawa, T. (2020). No raw data, no science: Another possible source of the reproducibility crisis.
- Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2019). Replication studies in economics—How many and which papers are chosen for replication, and why? *Research Policy*, 48(1), 62–83.
- Nature, Editor. (2013). Reducing our irreproducibility. *Nature*, 496, 398.
- Nosek, B. A., & Coauthors. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425.
- Orozco, V., Bontemps, C., Maigne, E., Piguët, V., Hofstetter, A., Lacroix, A., et al. (2020). How to make a pie: Reproducible research for empirical economics & econometrics. *Journal of Economic Surveys*, 34(5), 1134–1169.
- Pérignon, C., Gadouche, K., Hurlin, C., Silberman, R., & Debonnel, E. (2019). Certify reproducibility with confidential data. *Science*, 365(6449), 127–128.
- Pesaran, H. (2003). Introducing a replication section. *Journal of Applied Econometrics*, 18(1), 111.
- Reinhart, C. M., & Rogoff, K. S. (2010). Growth in a time of debt. *American Economic Review*, 100(2), 573–78.
- Rowhani-Farid, A., & Barnett, A. G. (2018). Badges for sharing data and code at biostatistics: An observational study [version 2; peer review: 2 approved]. *F1000Research*, 7(90).
- Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L., et al. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, 37(4), 358–367.
- Science, S. (2011). Challenges and opportunities. *Science*, 331(6018), 692–693.
- Smith, M. M., Sommer, A. J., Starkoff, B. E., Devor, S. T. (2013). Crossfit-based high-intensity power training improves maximal aerobic fitness and body composition. *The Journal of Strength and Conditioning Research*, 27(11), 3159–3172.
- Spencer, H. (1854). The art of education.
- Sweeney, L., Crosas, M., & Bar-Sinai, M. (2015). Sharing sensitive data with confidence: The datatags system. *Technology Science*.

- Vilhuber, L. (2019). Report by the AEA data editor. *AEA Papers and Proceedings*, 109, 718–729.
- Vlaeminck, S., & Herrmann, L.-K. (2015). Data policies and data archives: A new paradigm for academic publishing in economic sciences? In B. Schmidt, & M. Dobрева (Eds.), *New avenues for electronic publishing in the age of infinite collections and citizen science* (pp. 145–155). Amsterdam: IOS Press.
- Wilkinson, M., Dumontier, M., Aalbersber I., Appleton, G., Axton, M., Baak, A. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(160018).

“One Man, One Vote” Part 2: Measurement of Malapportionment and Disproportionality and the Lorenz Curve A: Introduction and Measurement Tools



Olivier de Mouzon, Thibault Laurent, and Michel Le Breton

Abstract The main objective of this paper is to explore and estimate the departure from the “One Man, One Vote” principle in the context of political representation and its consequences for distributive politics. To proceed to the measurement of the inequalities in the representation of territories (geographical under/over representation) or opinions/parties (ideological under/over representation), we import (with some important qualifications and adjustments) the Lorenz curve which is an important tool in the economics of income distribution. We consider subsequently some malapportionment and disproportionality indices. We provide several applications of these concepts in Chap. 32.

1 Introduction

This paper is in the continuation of our earlier paper (de Mouzon et al. 2020) dedicated to an analysis of the “one man, one vote” principle in the specific context of the U.S. Electoral College.

In that paper, the focus was on the degree of violation of the “one man, one vote” principle in the context of voting. It was postulated that the variable of interest that we wanted ideally to be the same for all voters was the probability of being decisive in

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-73249-3_31) contains supplementary material, which is available to authorized users.

O. de Mouzon

Toulouse School of Economics, INRAE, University of Toulouse Capitole, Toulouse, France
e-mail: olivier.de_mouzon@inrae.fr

T. Laurent (✉)

Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France
e-mail: thibault.laurent@tse-fr.eu

M. Le Breton

Institut Universitaire de France and Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France
e-mail: michel.lebreton@tse-fr.eu

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_31

an election. It was demonstrated that the “one man, one vote” principle was violated and that the identity of the beneficiaries was dependent upon the a priori probability model which was considered. However, for the three probability models which were investigated, the ratio between the most advantaged citizens and the less advantaged ones was always around 3.

In this second paper, we want to examine the “one man, one vote” principle when, instead of a binary ideological issue, the public decision consists in the distribution of resources among several territories like districts, counties, states or countries depending upon the context. It will be postulated that the actual distribution of resources among these territories is highly dependent upon how these territories are represented in the public body in charge of deciding the distribution of these resources.

The objective is to contrast the actual distribution with the distribution of resources that would arise as the solution of a social choice or welfare optimization problem postulating an equal treatment of individuals. We are interested in comparing the positive solution (which depends upon political representation of territories) and the normative solution (which only depends upon the population of the territories) as theory (in particular bargaining theory) suggests that any deviation from equality/proportionality in representation leads to a deviation from equality/proportionality in the sharing of resources.

This primary objective leads us to revisit an important issue in politics: how to measure malapportionment? Malapportionment¹ defines a situation where the allocation of seats/representatives across districts deviates from the allocation that would result from a strict application of population proportionality.²

This methodological issue spans a number of diverse and important situations including, in addition to legislative districting, the presidential Electoral College in the U.S. and the European Council of Ministers. For each of these situations, we can make an instantaneous photograph of how apportionment looks like. The photograph can consist of a single measure or a set of measures or even a curve as we will see. Collecting photographs at several points in time and/or for different territories paves the way for a study of the evolution of malapportionment along a time dimension (time series) or a spatial dimension (cross-section data).

With such measurement tools, we will be in a position to answer questions like: What has been the evolution of malapportionment in France over the last parliamentary elections? Could we say that, in the process of electing their “conseillers départementaux”, malapportionment is more severe in the Département “Morbihan” than it is in the Département “Creuse”? We could also, using the same tools, evaluate the impact on malapportionment of a particular redistricting plan like for instance the one which has been implemented in France in March 2015 for the election of

¹In this paper, we will be mostly interested in malapportionment in the context of districting. We assume that the partition of a territory (a union of countries, a country, a region or “départements” in France here, etc.) is given into districts (countries, states, congressional districts, “cantons” in the case of the French “Départements”, etc.). Each district is identified by its population size. The data on which measurement is based consists of the vector of seats and the vector of population sizes.

²Of course, the concepts introduced for the measurement of malapportionment can be (and in fact are) extensively applied also in the context of party representation.

local representatives: this plan included, among other things, a division by two of the number of “cantons”.

Malapportionment remains one of the key issues in political science. The “one man, one vote” principle is considered a pillar of any democratic system, and any violation of that principle is perceived as going against the democratic ideal. The adoption of the universal suffrage was certainly an important move towards this principle but it is well documented and recognized that in reality, the voice of some citizens may count more than the voice of others.

The books of Ansolabehere and Snyder (2008) and Balinski (2004) contain a lot of evidence indicating that this issue is not a secondary one. At that stage, it is important to say that this is not only a question of equality in political rights but, as we argued above, it is also a question of allocation of the resources/budgets which are under the control of the elected representatives.

As demonstrated forcefully by Ansolabehere and Snyder (2008),³ districts which are over (respectively under) represented tend to catch a larger (respectively smaller) share of the cake. Their cross-sectional analysis shows that counties with relatively more legislative seats per person prior to redistricting receive relatively more transfers from the state per person. They calculated that population equalization significantly altered the flow of state transfers to counties, diverting approximately \$7 billions annually from formerly over-represented to formerly under-represented counties. Clearly “the American experience provides clear evidence of the political consequences of unequal representation”.

Maaser and Stratmann (2016) reach a similar conclusion for Germany: they find that districts with a greater number of representatives receive more government funds. Kauppi and Widgrén (2004, 2007) and García-Valiñas et al. (2016) have also demonstrated that political representation within the E.U. council is a key driver of the distribution of the E.U. budgets.

In this paper, we import from economics some tools which have been developed to evaluate the intensity of inequality in income/wealth/health (or other continuous variable impacting the well being of individuals) distribution.⁴

In contrast to economics, here the variable under scrutiny is seats. We argue that the tools of economists, on top of which the *Lorenz curve* and the *Gini Index* are very much appropriate to handle the measurement of malapportionment once the right variables have been introduced. In doing so, we follow the steps of Van Puyenbroeck (2006) who already suggested the fruitfulness of that connection in his pioneering must-read paper.⁵

As forcefully demonstrated by Van Puyenbroeck (2006), it is important to be careful in importing these tools as measuring departures from the equality principle in politics calls for some important adjustments.

³See their documented Chap. 6 as well as their (2002) paper.

⁴See e.g. Lambert (2001) for a nice presentation of the main ideas and results in that area.

⁵It is also important to point out that concepts from the theory of majorization (Marshall et al. 2011) have also been used to compare different apportionment methods (Lauwers and Puyenbroeck 2006b, a; Marshall et al. 2002).

The paper is organized as follows. In the first section, we introduce the main concepts and notations together with a general framework to evaluate the distance between ideal and real public decisions. We then move to distributive politics and define the Lorenz curve and some important indices. A supplementary material divided into 5 parts contains additional material on malapportionment and voting, majorization with weights and other technical developments.

2 Descriptive Statistics and Measurement of Malapportionment/Disproportionality

The framework developed in this paper can accommodate two different measurement issues.

For both of them, we want to examine and compute the “distance” from the “one man, one vote” principle. In the first subsection, we present the two settings. Then, in the second subsection, we develop a framework explaining the connection between representation and public decision. The third subsection is the key subsection. It explains how to construct the Lorenz curve in our setting and argues against some alternative constructions of this Lorenz curve. The last subsection introduces some of the main indices and in particular the two main ones which are used in Chap. 32.

2.1 Two Settings

In the first set of applications, we consider a territory (a country, a region, a local government, etc.) divided into K electoral sub-territories (states, counties, electoral districts, etc.).

The representatives of the territory are all elected at the district level. Hereafter, we will denote by N_k the population size of district k and by R_k the number of representatives apportioned to district k for all $k = 1, 2, \dots, K$. The territory can be a local/regional territory (like a “département” or a region in France or a State in the U.S.) and the assembly of representatives a council in charge of the policies decided and implemented at the level of this territory.

The inputs of the measurement issue addressed in this first case consist of two vectors: the vector of populations $\mathbf{N} = (N_1, N_2, \dots, N_K)$ and the vector of representatives $\mathbf{R} = (R_1, R_2, \dots, R_K)$.⁶

Such a pair (\mathbf{N}, \mathbf{R}) will be called a *geographical pattern/situation*. In many applications, we will assume that $\mathbf{R} = (1, 1, \dots, 1)$. Let us finally denote by \mathbf{n} and \mathbf{r} the vectors of shares $\mathbf{n} = (\frac{N_1}{N}, \frac{N_2}{N}, \dots, \frac{N_K}{N})$ and $\mathbf{r} = (\frac{R_1}{R}, \frac{R_2}{R}, \dots, \frac{R_K}{R})$.

⁶Hereafter, N and R denote, respectively, the total number of voters and the total number of representatives.

While our notations seem to privilege the time series analysis of a pattern, we would like to point out that the same tools allow for a cross-sectional analysis (for instance, we can use the tools to compare different territories, at any given point in time, as done for instance in Ansolabehere and Snyder 2008) or a comparison between a pre-reform and a post-reform situation.

In the second type of application, the focus is on an election involving V voters, K parties and S seats. In this case, we will denote by V_k the number of people voting for party k ⁷ and by S_k the number of seats won by party k for all $k = 1, 2, \dots, K$.

The inputs of the measurement issue addressed in this case consist of two vectors: the vector of votes $\mathbf{V} = (V_1, V_2, \dots, V_K)$ and the vector of representatives $\mathbf{S} = (S_1, S_2, \dots, S_K)$. Such a pair (\mathbf{V}, \mathbf{S}) will be called an *ideological pattern/situation*. We will denote by \mathbf{v} and \mathbf{s} the vectors of shares $\mathbf{v} = (\frac{V_1}{V}, \frac{V_2}{V}, \dots, \frac{V_K}{V})$ and $\mathbf{s} = (\frac{S_1}{S}, \frac{S_2}{S}, \dots, \frac{S_K}{S})$.

Given either a *geographical pattern/situation* (\mathbf{N}, \mathbf{R}) (or (\mathbf{n}, \mathbf{r})) or an *ideological pattern/situation* (\mathbf{V}, \mathbf{S}) (or (\mathbf{v}, \mathbf{s})), we want to measure how far we are from the “one man, one vote” reference norm.

2.2 Mapping Representation into Public Decisions

To compare two different situations (\mathbf{N}, \mathbf{R}) and $(\mathbf{N}, \mathbf{R}')$, we introduce a set of feasible public decisions \mathcal{D} .

We assume that each citizen $i = 1, \dots, N$ has a utility function U_i on \mathcal{D} . Before exploring the influence of (\mathbf{N}, \mathbf{R}) in the positive decision making process, we define a normative reference that will be used as a benchmark in subsequent comparisons. Hereafter, we will focus on the utilitarian norm. From that perspective, the welfare attached to decision d is:

$$\sum_{i=1}^N U_i(d).$$

Let us denote by $d^*(\mathbf{N}, \mathbf{U})$ the decision which maximizes utilitarian welfare where \mathbf{U} denotes the profile (U_1, \dots, U_n) of utility function. Given the decision $d(\mathbf{U}, \mathbf{N}, \mathbf{R})$ undertaken by the council of representatives, we may evaluate the “distance” between the two in several ways. For instance, we could consider:

$$\sum_{i=1}^N U_i(d^*(\mathbf{N}, \mathbf{U})) - \sum_{i=1}^N U_i(d(\mathbf{U}, \mathbf{N}, \mathbf{R})),$$

⁷Of course, the expression “number of people voting for party k ” is possibly ambiguous if the electoral mechanism is complicated and/or if it involves several rounds. This framework only applies to elections where the ballots consist of lists of candidates (possibly one) with a party affiliation. In the case of several rounds, we retain the first-round votes.

or we could consider the departure from the perspective of each individual, i.e. :

$$(U_1(d^*(\mathbf{N}, \mathbf{U})) - U_1(d(\mathbf{U}, \mathbf{N}, \mathbf{R})), \dots, U_N(d^*(\mathbf{N}, \mathbf{U})) - U_N(d(\mathbf{U}, \mathbf{N}, \mathbf{R}))).$$

Since these measures depend upon the particular profile \mathbf{U} that is considered, we may prefer to consider ex ante evaluations where \mathbf{U} is drawn randomly from a set \mathcal{U} of admissible profiles according to a specific⁸ probability model λ .

Then for the two measures above, we move to expectations with respect to λ .

$$\Delta_\lambda^1(\mathbf{N}, \mathbf{R}) = E_\lambda \left[\sum_{i=1}^N U_i(d^*(\mathbf{N}, \mathbf{U})) - \sum_{i=1}^N U_i(d(\mathbf{U}, \mathbf{N}, \mathbf{R})) \right],$$

$$\begin{aligned} \Delta_\lambda^2(\mathbf{N}, \mathbf{R}) = & E_\lambda [(U_1(d^*(\mathbf{N}, \mathbf{U})) - U_1(d(\mathbf{U}, \mathbf{N}, \mathbf{R})), \dots, U_N(d^*(\mathbf{N}, \mathbf{U})) - U_N(d(\mathbf{U}, \mathbf{N}, \mathbf{R})))] = \\ & E_\lambda [(U_1(d^*(\mathbf{N}, \mathbf{U})), \dots, U_N(d^*(\mathbf{N}, \mathbf{U})))] - E_\lambda [(U_1(d(\mathbf{U}, \mathbf{N}, \mathbf{R})), \dots, U_N(d(\mathbf{U}, \mathbf{N}, \mathbf{R})))]. \end{aligned}$$

These two measures are derived from welfare foundations. Since this paper is about the measurement of the distance to the “one man, one vote” principles, we will modify later at the margin these measures to make sure that the values do not depend upon irrelevant factors. For instance, we do not want the size of the population per se to have an impact on the comparisons. In some other applications, we do not want some specific elements of the set \mathcal{D} to have an impact on the comparisons. We will explain in due time how to adjust the above measures to do so.

To proceed with these measures, we do need a detailed description of the derivation of $d(\mathbf{U}, \mathbf{N}, \mathbf{R})$. Depending upon the nature of the set \mathcal{D} , many alternative institutions can be considered. To study the behaviour of representatives within these institutions, we will need to model the objectives of the representatives and the nature of the game that they play among themselves. We limit our attention here to two canonical cases.

The first canonical case is the classical binary framework: $\mathcal{D} = \{0, 1\}$ and for each i , there are two possible utility functions: either $U_i(1) = 1$ and $U_i(0) = 0$ or $U_i(1) = 0$ and $U_i(0) = 1$. Here $d^*(\mathbf{N}, \mathbf{U})$ is the popular majority decision. If all the representatives of territory k endorse the majority opinion among voters in territory k , the decision $d(\mathbf{U}, \mathbf{N}, \mathbf{R})$ denotes the majority decision in the council. $d(\mathbf{U}, \mathbf{N}, \mathbf{R})$ need not to be equal to $d^*(\mathbf{U})$: an outcome such that $d(\mathbf{U}, \mathbf{N}, \mathbf{R}) \neq d^*(\mathbf{N}, \mathbf{U})$ is called in voting an election inversion.

For any probability model λ , we can (in principle) compute $\Delta_\lambda^i(\mathbf{N}, \mathbf{R})$ for $i = 1$ and 2. A large value indicates a large departure from the popular majority decision which is here the reference outcome to define at best “one man, one vote”.

Using Δ_λ^1 informs about the distance from a decision that reflects the “one man, one vote” principle. It is important to call attention to the fact that postulating that the

⁸Since the paper is about the “one man, one vote” principle, then the probability model itself must display symmetry across voters.

ideal “one man, one vote” is reflected at best by the majority mechanism demands more than the equal treatment of voters.⁹

Using Δ_λ^2 gives us a more detailed information about the decomposition of the aggregate difference Δ_λ^1 into its individual components. Indeed if $\Delta_\lambda^1(\mathbf{N}, \mathbf{R}) > 0$, then some of (maybe all) the coordinates of $\Delta_\lambda^2(\mathbf{N}, \mathbf{R})$ are positive. As for Δ_λ^1 , the reference to the majority outcome is important: having the vector $\Delta_\lambda^2(\mathbf{N}, \mathbf{R})$ on the diagonal of \mathbb{R}^N is not enough. If we compare several mechanisms (all different from the majority mechanism), on the basis of their respective vectors $\Delta_\lambda^2(\mathbf{N}, \mathbf{R})$, we may opt for a criterion different from the utilitarian one.

Instead of measuring the distance from the reference point through utilities, we could (as in de Mouzon et al. 2020, for the U.S. Electoral College) calculate for each state k , a number measuring the decisiveness of a voter from state k .¹⁰

A perfect application of the “one man, one vote” principle would require the perfect equality of these K numbers. In reality, these numbers differ among themselves. Appendix 1 (see supplementary material) contains a computation of $\Delta_\lambda^1(\mathbf{N}, \mathbf{R})$ and $\Delta_\lambda^2(\mathbf{N}, \mathbf{R})$ and a third measure in the case where $K = 3$ and $\mathbf{R} = (1, 1, 1)$. As advocated, we can of course normalize the above two measures in order to make them invariant with respect to the size N of the population to allow comparisons of situations where the population sizes are not the same. For instance, instead of Δ_λ^1 , we could consider $\widehat{\Delta}_\lambda^1$ defined as follows:

$$\widehat{\Delta}_\lambda^1(\mathbf{N}, \mathbf{R}) = \frac{\Delta_\lambda^1(\mathbf{N}, \mathbf{R})}{\phi(N, \lambda)}$$

where $\phi(N, \lambda)$ is a function taking care of the population scale factor.

In the second canonical case (often recorded under the headings “Distributive Politics” or “Divide the Dollar”), the set of public policies \mathcal{D} is a simplex:

$$\mathcal{D} = \mathcal{S} \equiv \left\{ X \in \mathbb{R}_+^K : \sum_{k=1}^K X_k = M \right\}, \tag{1}$$

where M is a positive number. The council decision consists in a distribution of the total budget M across the K territories. In such a case, it is natural to assume that U_i depends only upon X_k (where k is the territory where i lives) and is strictly increasing with respect to that variable.

If we assume further that the share of the budget received by territory k is divided equally among its residents, i.e. if the good which is considered is purely private (no

⁹In the paper, we restrict ourselves to two-step majority mechanisms (i.e. indirect majority elections through a set of representatives described by \mathbf{R}) but we could compute $\Delta_\lambda^1(N, V)$ for any voting mechanism mapping $\{0, 1\}^N$ onto $\{0, 1\}$. Any anonymous mechanism (even peculiar ones like selecting the minority candidate, drawing randomly the winner or drawing randomly a dictator) treats equally the voters.

¹⁰For some probability models λ , the two approaches are equivalent.

economies of scale), the benefit of a resident of territory k is $\frac{X_k}{N_k}$ and then the utility derived by i from decision d is:

$$U_i\left(\frac{X_k}{N_k}\right).$$

Further, if we postulate symmetry in inter-comparison of utilities, i.e. that U does not depend upon i , then the utilitarian welfare attached to decision d is:

$$\sum_{k=1}^K N_k U\left(\frac{X_k}{N_k}\right). \tag{2}$$

If U is strictly concave, maximization of (2) under constraint (1), i.e.

$$\max \sum_{k=1}^K N_k U\left(\frac{X_k}{N_k}\right),$$

under the constraints $X \in \mathcal{S}$ yields an unique interior solution:

$$X_k^*(\mathbf{N}, U) = \frac{N_k}{N} M \equiv n_k M \text{ for all } k = 1, \dots, K.$$

The reference point is perfect proportionality. According to the utilitarian principle, each territory should receive a share of the budget proportional to its population. We note that now, in contrast to the first canonical case, the reference point does not depend upon the profile U . There are several ways to transform (\mathbf{N}, \mathbf{R}) into $d(U, \mathbf{N}, \mathbf{R})$. Hereafter,¹¹ we will focus on the case where:

$$X_k(U, \mathbf{N}, \mathbf{R}) = \frac{R_k}{R} M \equiv r_k M \text{ for all } k = 1, \dots, K.$$

As before, $d(U, \mathbf{N}, \mathbf{R})$ does not depend upon U in that case. If the utility function¹² U is drawn randomly from a set \mathcal{U} of admissible utility functions according to a specific probability model λ , we obtain¹³:

$$\Delta_\lambda^1(\mathbf{N}, \mathbf{R}) = E_\lambda \left[\sum_{k=1}^K N_k \left(U\left(\frac{M}{N}\right) - U\left(\frac{MR_k}{RN_k}\right) \right) \right].$$

¹¹An alternative to the one considered in the text is described in Appendix 4 (see supplementary material).

¹²In contrast to the first canonical case, it is assumed that the profile is diagonal and summarized by a single increasing utility function.

¹³And similarly: $\Delta_\lambda^2(\mathbf{N}, \mathbf{R}) = E_\lambda \left[U\left(\frac{N_1}{N} M\right) - U\left(\frac{R_1}{R} M\right), \dots, U\left(\frac{N_K}{N} M\right) - U\left(\frac{R_K}{R} M\right) \right]$.

This measure calls for several comments. For any fixed U , we want this measure to be as small as possible. Ideally, we would like

$$\Delta_U^1(\mathbf{N}, \mathbf{R}) \equiv \sum_{k=1}^K N_k \left(U\left(\frac{M}{N}\right) - U\left(\frac{MR_k}{RN_k}\right) \right)$$

to be as small as possible pointwise, i.e. for all admissible U . For any given U , Δ_U^1 evaluates the distance between the “ideal” and the reality induced by the situation (\mathbf{N}, \mathbf{R}) . When we compare two situations (\mathbf{N}, \mathbf{R}) and $(\mathbf{N}', \mathbf{R}')$ such that $N = N'$, the difference $\Delta_U^1(\mathbf{N}', \mathbf{R}') - \Delta_U^1(\mathbf{N}, \mathbf{R})$ writes:

$$- \left[N' \sum_{k=1}^K n'_k U\left(\frac{M'r'_k}{N'n'_k}\right) - N \sum_{k=1}^K n_k U\left(\frac{Mr_k}{Nn_k}\right) \right] + NU\left(\frac{M}{N}\right) - N'U\left(\frac{M'}{N'}\right)$$

i.e. what matters are the vector of shares (\mathbf{n}, \mathbf{r}) and $(\mathbf{n}', \mathbf{r}')$ and the per capita budgets $\frac{M}{N}$ and $\frac{M'}{N'}$. When $\mathbf{N} \neq \mathbf{N}'$, this simplification does not hold unless we replace utilitarian welfare by average utilitarian welfare. Since we are interested in measuring deviation from proportionality per se, we don't want the size of the budget and the size of the population to have an effect on measurement¹⁴ and we change the measure $\Delta_\lambda^1(\mathbf{N}, \mathbf{R})$ above into the following one:

$$\widehat{\Delta}_\lambda^1(\mathbf{N}, \mathbf{R}) = -E_\lambda \left[\sum_{k=1}^K n_k U\left(\frac{r_k}{n_k}\right) \right] = E_\lambda \left[\sum_{k=1}^K n_k g\left(\frac{r_k}{n_k}\right) \right] \equiv E_\lambda \left[\widehat{\Delta}_g^1(\mathbf{N}, \mathbf{R}) \right]$$

where $g \equiv -U$. If we consider \mathcal{U} to be the set of increasing and concave functions on \mathbb{R} , the set \mathcal{G} of admissible g is the set of decreasing and convex functions on \mathbb{R} .

2.3 The Lorenz Order

By choosing a specific convex function g , we can order any two situations (\mathbf{N}, \mathbf{R}) and $(\mathbf{N}', \mathbf{R}')$ according to the measure $\widehat{\Delta}_g^1$:

¹⁴In inequality measurement, these two properties are called *scale invariance* (if the amount of resource received by each individual is multiplied by the same constant, then inequality remains unchanged) and *population invariance* (if the number of individuals in each of the K groups is multiplied by the same constant, then inequality is unchanged). In contrast, in welfare measurement, the resource scale and population scales matter. This is why some authors move from Lorenz curves to generalized Lorenz curves. These questions are discussed in Appendix 2.

$$(\mathbf{N}, \mathbf{R}) \text{ dominates } (\mathbf{N}', \mathbf{R}') \text{ iff } \sum_{k=1}^K n_k g\left(\frac{r_k}{n_k}\right) \leq \sum_{k=1}^K n'_k g\left(\frac{r'_k}{n'_k}\right)$$

This dominance reads as follows: Given g , the situation (\mathbf{N}, \mathbf{R}) is closer to the ideal “one man, one vote” than the situation $(\mathbf{N}', \mathbf{R}')$. This ordering is complete, i.e. any two situations can be compared. But it is also very sensitive to the choice of a specific convex function g in the class \mathcal{G} . Two different convex functions could lead to two opposite statements. One way to proceed is to move to expectation, i.e. to consider the measure $\widehat{\Delta}_\lambda^1$ defined before for some probability measure on \mathcal{G} . Since such measure may be sensitive to the probability model λ , an alternative road consists in considering the following *partial* ordering:

$$(\mathbf{N}, \mathbf{R}) \text{ unambiguously dominates } (\mathbf{N}', \mathbf{R}') \text{ iff } \sum_{k=1}^K n_k g\left(\frac{r_k}{n_k}\right) \leq \sum_{k=1}^K n'_k g\left(\frac{r'_k}{n'_k}\right)$$

for all convex functions g .

This partial ordering¹⁵ is presented in Appendix 2 (see supplementary material). It is an extension of the classical majorization ordering¹⁶ (Marshall et al. 2011) which is exclusively defined over the subclass of situations such that $n_k = n'_k = \frac{1}{K}$ for all $k = 1, \dots, K$.

How do we prove or disprove that (\mathbf{N}, \mathbf{R}) unambiguously dominates $(\mathbf{N}', \mathbf{R}')$? The main characterization theorems are also presented in Appendix 2 (see supplementary material). The most important one consists in introducing the Lorenz curve which is defined here as follows.

First, for any situation (\mathbf{N}, \mathbf{R}) , we consider the rearrangement of the coordinates of the vector $(\frac{r_1}{n_1}, \frac{r_2}{n_2}, \dots, \frac{r_K}{n_K})$ from the lowest to the largest. From this vector, denoted $(\frac{\tilde{r}_1}{\tilde{n}_1}, \frac{\tilde{r}_2}{\tilde{n}_2}, \dots, \frac{\tilde{r}_K}{\tilde{n}_K})$, we construct the following curve which contains (according to us) all the relevant statistical information on (\mathbf{N}, \mathbf{R}) .

We plot on the horizontal axis all the cumulative fractions: $0, \tilde{n}_1, \tilde{n}_1 + \tilde{n}_2, \tilde{n}_1 + \tilde{n}_2 + \tilde{n}_3, \dots, 1$ and on the vertical axis all the cumulative ordered fractions $0, \tilde{r}_1, \tilde{r}_1 + \tilde{r}_2, \tilde{r}_1 + \tilde{r}_2 + \tilde{r}_3, \dots, 1$.

This provides a sample of $K + 1$ points in the unit square $[0, 1]$ including $(0, 0)$ and $(1, 1)$. This sample is increasing and convex in the sense that:

$$\frac{\sum_{j=1}^k \tilde{r}_j}{\sum_{j=1}^k \tilde{n}_j} \geq \frac{\sum_{j=1}^{k-1} \tilde{r}_j}{\sum_{j=1}^{k-1} \tilde{n}_j} \text{ for all } k = 1, \dots, K.$$

¹⁵Strictly speaking, we should consider the more restricted family \mathcal{G} , i.e. assume that g is also decreasing. But since $\sum_{k=1}^K n_k \frac{r_k}{n_k} = \sum_{k=1}^K n'_k \frac{r'_k}{n'_k} = 1$, we can show that imposing that restriction does not change the partial ordering.

¹⁶This partial ordering is also known as second-order stochastic dominance.

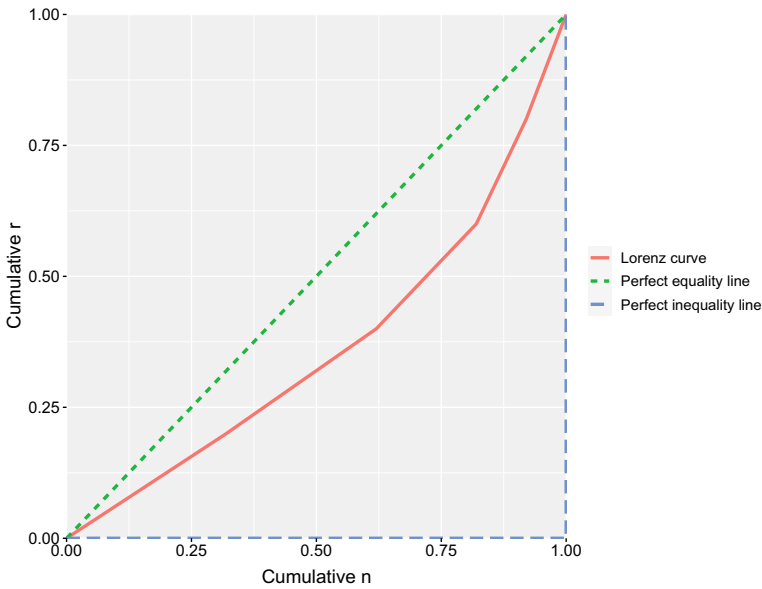


Fig. 1 Illustration of the Lorenz curve

For convenience, we identify this finite set of points to a curve by piece-wise linear interpolation between any pair of adjacent points. Let us denote by $L_{(\mathbf{N}, \mathbf{R})}(x)$ this curve defined for all $x \in [0, 1]$ and with values in $[0, 1]$. For the sake of illustration, the construction of such a curve is depicted in Fig. 1 when $K = 5$, $\mathbf{n} = (0.10, 0.32, 0.30, 0.20, 0.08)$ and $\mathbf{r} = (0.2, 0.2, 0.2, 0.2, 0.2)$.

We will say that the pattern (\mathbf{N}, \mathbf{R}) (strictly) Lorenz dominates the pattern $(\mathbf{N}', \mathbf{R}')$ iff $L_{(\mathbf{N}, \mathbf{R})}(x) \geq L_{(\mathbf{N}', \mathbf{R}')} (x)$ for all $x \in [0, 1]$ (with a strict inequality for at least one value of x).

Implicit in the above construction is the fact that the relevant units in our comparison are the individuals and what they ultimately receive through the redistribution of the resources.

This should be contrasted with alternative choices as those discussed and criticized by Van Puyenbroeck (2006).¹⁷ For instance, in many measures, the frequencies do not appear in the weighted sum and scholars look instead at an ordering like:

$$\sum_{k=1}^K g\left(\frac{r_k}{n_k}\right) \leq \sum_{k=1}^K g\left(\frac{r'_k}{n'_k}\right) \text{ for all convex functions } g.$$

¹⁷Van Puyenbroeck (2006) and also Goldenberg and Fisher (2019) contain a lot of developments including discussions about the rearrangement. In particular, they spend time contrasting the arrangement based on the ratios $\frac{r_k}{n_k}$ with the arrangement based on the differences $r_k - n_k$. Both papers agree that if k and k' are such that $\frac{r_k}{n_k} > 1$ and $\frac{r_{k'}}{n_{k'}} < 1$, then k' should be on the left of k but possibly disagree on k and k' when they are on the same side.

In doing so, we move from the voters to the territories or the parties as relevant recipients/units. This choice amounts to drawing a Lorenz curve where the relevant coordinates on the horizontal (vertical) axis are $0, \frac{1}{K}, \frac{2}{K}, \dots, \frac{K-1}{K}, 1$ ($\left(0, \frac{\tilde{r}_1}{\sum_{k=1}^K \frac{r_k}{n_k}, \frac{\tilde{r}_1 + \tilde{r}_2}{\sum_{k=1}^K \frac{r_k}{n_k}}, \dots, \frac{\sum_{k=1}^{K-1} \tilde{r}_k}{\sum_{k=1}^K \frac{r_k}{n_k}}, 1\right)$).

Van Puyenbroeck (2006) also discusses the possibilities offered by plotting respectively the coordinates $0, \frac{1}{K}, \frac{2}{K}, \dots, \frac{K-1}{K}, 1$ on the horizontal axis and the coordinates $0, n_1, n_1 + n_2, \dots, 1$ and $0, r_1, r_1 + r_2, \dots, 1$ on the vertical axis where the coordinates are rearranged according to the ordering attached to n and r (under the presumption that these two orderings are the same). As pointed out by Goldenberg and Fisher (2019) and Van Puyenbroeck (2006), this leads to problematic issues when we analyze seat transfers.¹⁸

In addition to Van Puyenbroeck (2006), the Lorenz curve that we use in our paper has also been used by Colignatus (2017c,b,a) in a series of applications to recent electoral data. In Chap. 32, we compute this Lorenz curve for several apportionment situations and one ideological situation. Let us conclude this section with three remarks.

First, note that when $M = 1$ and when we limit ourselves to vectors \mathbf{R} with integer coordinates in the unitary simplex, then the Lorenz curve has a very simple shape depicted in Fig. 2.

When the choice is the vector \mathbf{R} where the k th element is equal to 1, then the curve is flat until $1 - n_k$ and is linear after. This implies that if k and l are such that $n_k > n_l$, then the Lorenz curve attached to k is above the Lorenz curve attached to l . The Lorenz curve ordering is compatible with the absence of an election inversion. From the Lorenz perspective, it is always better to allocate the seat to the candidate with the highest number of votes.

Second, note that when we compare (\mathbf{N}, \mathbf{R}) and $(\mathbf{N}', \mathbf{R}')$ when $\mathbf{R} = \mathbf{R}' = (1, 1, \dots, 1)$, the ordering of the units on the horizontal axis amounts to the ordering of the units from the most populated to the less populated.

Third, note that the Lorenz criterion is also useful to compare situations where the map of the districts has been reshaped. For instance, we may consider two situations (\mathbf{N}, \mathbf{R}) and $(\mathbf{N}', \mathbf{R}')$ where $K' = \frac{K}{2}$, N' is deduced from N through a matching, i.e. according to a grouping of the old districts by pairs, $\mathbf{R} = (1, \dots, 1)$ and $\mathbf{R}' = (2, \dots, 2)$.¹⁹ If we move from the first situation to the second one, we may wonder what the best matchings from a Lorenz perspective are.²⁰

¹⁸The sum $\sum_{k=1}^K \frac{\tilde{r}_k}{n_k} = \sum_{k=1}^K \frac{r_k}{n_k}$ is not invariant under the transfer of seats between territories or parties. Van Puyenbroeck writes “Conversely, and equally unfortunately, it seems difficult to sustain that the latter construct, $\frac{1}{K} \sum_{k=1}^K \frac{r_k}{n_k}$, provides a reasonable benchmark of equality”.

¹⁹In the second situation, the number of districts has been reduced by one half while the total population and the total number of representatives have remained unchanged.

²⁰Clearly, grouping is always a good move from a Lorenz perspective. Note that this question is formally related to the issue of aggregation. When we move to a more aggregated level and average the values accordingly, we lose information and we ultimately underestimate malapportionment or disproportionality.

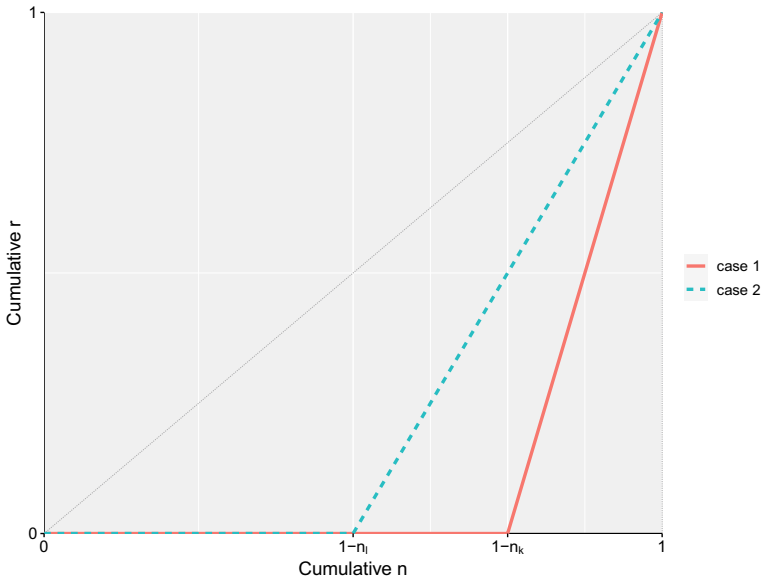


Fig. 2 Illustration of the Lorenz curve

2.4 Malapportionment and Disproportionality Indices

The Lorenz ordering defined in the preceding section is partial (we cannot compare (\mathbf{N}, \mathbf{R}) and $(\mathbf{N}', \mathbf{R}')$ when their Lorenz curves intersect). To overcome this difficulty when it arises, it is useful to complement the measurement analysis based on Lorenz by computing the value of some indices. A (relative) index is a function I which maps any situation (\mathbf{N}, \mathbf{R}) into a real number $I(\mathbf{N}, \mathbf{R})$ and satisfies the monotonicity property:

$$\text{If } L_{(\mathbf{N}, \mathbf{R})}(x) \geq L_{(\mathbf{N}', \mathbf{R}')} (x) \text{ for all } x \in [0, 1] \text{ then } I(\mathbf{N}, \mathbf{R}) \leq I(\mathbf{N}', \mathbf{R}').$$

as well as the scale and population invariance properties.

Among the most popular indices, the *Gini index* is defined as follows²¹:

$$G(\mathbf{N}, \mathbf{R}) = \frac{1}{2} \sum_{k=1}^K \sum_{j=1}^K n_k n_j \left| \frac{r_k}{n_k} - \frac{r_j}{n_j} \right|.$$

We could of course import from the inequality measurement literature other indices among which Atkinson–Kolm’s indices are defined as follows:

²¹It is defined alternatively as the surface of the area between the diagonal and the Lorenz curve $L_{(\mathbf{N}, \mathbf{R})}$.

$$AKT_\alpha(\mathbf{N}, \mathbf{R}) = \begin{cases} 1 - \left(\sum_{k=1}^K n_k \left(\frac{r_k}{n_k} \right)^{1-\alpha} \right)^{\frac{1}{1-\alpha}} & \text{if } \alpha \neq 1, \\ 1 - \left(\prod_{k=1}^K \left(\frac{r_k}{n_k} \right)^{n_k} \right) & \text{if } \alpha = 1. \end{cases}$$

The parameter α is a parameter of inequality aversion. The larger α is, the larger is the aversion to inequality. When α tends to $+\infty$, this index tends to

$$1 - \underset{1 \leq k \leq K}{\text{Min}} \frac{r_k}{n_k}.$$

We could alternatively²² consider the class of generalized entropy indices defined as follows:

$$GE_\alpha(\mathbf{N}, \mathbf{R}) = \begin{cases} \frac{1}{\alpha(\alpha-1)} \left(\sum_{k=1}^K n_k \left(\frac{r_k}{n_k} \right)^\alpha - 1 \right) & \text{if } \alpha \neq 0, 1, \\ \sum_{k=1}^K n_k \left(\frac{r_k}{n_k} \ln \frac{r_k}{n_k} \right) & \text{if } \alpha = 1, \\ - \sum_{k=1}^K n_k \ln \frac{r_k}{n_k} & \text{if } \alpha = 0. \end{cases}$$

The class of indices GE_α is part of the general class of indices I defined as follows:

$$I(\mathbf{N}, \mathbf{R}) = \sum_{k=1}^K n_k g\left(\frac{r_k}{n_k}\right) \text{ where } g \text{ is a convex function.}$$

To conclude this point,²³ let us mention the DK index (after Dauer and Kelsay 1955) which is advocated by Ansolabehere and Snyder (2008).

Let x^* be the unique value of x such that $L_{(\mathbf{N}, \mathbf{R})}(x) = 0.5$. From what precedes, x^* is larger than 0.5. The DK index attached to the pattern (\mathbf{N}, \mathbf{R}) , denoted $DK(\mathbf{N}, \mathbf{R})$ is the number $1 - x^*$.

It evaluates the smallest size of a population of citizens which control a majority of representatives in the assembly. For instance if $DK(\mathbf{N}, \mathbf{R}) = 0.32$, it means than in the context (\mathbf{N}, \mathbf{R}) , 32% of the electorate controls 50% of the seats/representatives. Here we prefer to have large values of DK which means that, strictly speaking, the index should be defined as being x^* itself.

All these indices are useful in the case where the Lorenz curves intersect. Drawing the Lorenz curves of situations is always important as when they do not intersect, it shows that the conclusion does not depend upon the choice of a particular index. In contrast, when they intersect, indices help to say something on the evolution of

²²The two classes of indices are ordinally equivalent since they deduce from each other through increasing transformations. See e.g. Lambert (2001).

²³We could also consider other measures like for instance the ratio between the largest coordinate and the smallest one but note that while popular in inequality measurement, this number is insensitive to changes in other parts of the vectors.

malapportionment/disproportionality. In Chap. 32, we will focus on the Gini and DK indices.

Let us remind the reader that the indices considered in this section are those which are monotonic with respect to the Lorenz ordering introduced before. This means for instance that the popular Gallagher index (Gallagher 1991) $GA(\mathbf{N}, \mathbf{R})$ defined as follows:

$$GA(\mathbf{N}, \mathbf{R}) = \sqrt{\frac{1}{2} \sum_{k=1}^K (r_k - n_k)^2},$$

is not an index as defined above since it is not always monotonic with respect to the Lorenz ordering.

This difficulty with Gallagher’s index is pointed out in Goldenberg and Fisher (2019) and Renwick (2015). There is an enormous literature²⁴ on the measurement of disproportionality. As emphasized by Van Puyenbroeck (2006), who refers to a “zoo of no fewer than 19 proposed indices” many of them, including among others some versions of Gini’s index,²⁵ are problematic if the concern is to examine how far we are from the “one man, one vote” principle.

Acknowledgements We are pleased to contribute to this volume in honour of Christine Thomas as an expression of our friendship and gratitude to her. The second author would like to thank her for her role as a mentor within the TSE research community and the long list of projects to which he participated under her leadership. We also express our gratitude to Philippe De Donder and Karine Van Der Straeten for their careful reading of an earlier version of the manuscript and their constructive criticisms. Finally, we acknowledge funding from ANR under grant ANR-17-EURE-0010 (Investissements d’Avenir program).

References

- Ansolabehere, S., & Snyder, J. M. (2008). *The End of Inequality: One Person, One Vote and the Transformation of American Politics*. New York: W. W. Norton.
- Balinski, M. (2004). *Le suffrage universel inachevé*. Belin.
- Bouyssou, D., Marchant, T., & Pirlot, M. (1947). A characterization of two disproportionality and malapportionment indices: The Duncan and Duncan index and the Lijphart index. *Annals of Operations Research*, 284, 147–163.

²⁴Karpov (2008) compares 18 indices. See also Chessa and Fragnelli (2012); Cox and Shugart (1991); Fry and McLean (1991); Monroe (1994); Pennisi (1998); Taagepera and Grofman (2003) out of many. There are also papers developments axiomatic analysis of some malapportionment and disproportionality indices (see e.g. Bouyssou et al. 1947; Koppel and Diskin 2009).

²⁵Many of these indices are simply defined as functions expressing (up to some normalization and/or ordinal transformation) a kind of “distance” between the population/vote shares and the seat shares which happens to be equal to 0 iff the two vectors coincide. Gallagher (1991) uses least squares but some other authors (Loosemore and Hanby 1971) uses absolute deviations.

- Chessa, M., & Fragnelli, V. (2012). A note on measurement of disproportionality in proportional representation systems. *Mathematical and Computer Modelling*, 55, 1655–1660.
- Colignatus, T. (2017a). Lorenz en Gini for the French elections of 2017. <https://boycottholland.wordpress.com/2017/07/19>.
- Colignatus, T. (2017b). Proportional representation. *Lorenz diagram and Gini measure*. <https://boycottholland.wordpress.com/2017/06/22>.
- Colignatus, T. (2017c). Two conditions for the application of Lorenz curve and Gini coefficient to voting and allocated seats. *MPRA Paper*, 80297.
- Cox, G. W., & Shugart, M. S. (1991). Comment on Gallagher's proportionality, disproportionality and electoral systems. *Electoral Studies*, 10, 348–352.
- Dauer, M. J., & Kelsay, R. G. (1955). Unrepresentative states. *National Municipal Review*, 46, 571–575.
- de Mouzon, O., Laurent, T., Le Breton, M., & Moyouwou, I. (2020). One man, one vote part 1: Electoral Justice in the U.S. Electoral College: Banzhaf and Shapley/Shubik versus May. In M. Diss & V. Merlin (Eds.), *Essays by and in honor of William Gehrlein and Dominique Lepelley*. Springer.
- Fry, V., & McLean, I. (1991). A note on Rose's proportionality index. *Electoral Studies*, 10, 52–59.
- Gallagher, M. (1991). Proportionality, disproportionality and electoral systems. *Electoral Studies*, 10, 33–51.
- García-Valiñas, M., Kurz, S., & Zaporozhets, V. (2016). Key-drivers of E.U. budget allocation: Does power matter? *European Journal of Political Economy*, 43, 57–70.
- Goldenberg, J., & Fisher, S. D. (2019). The Sainte-Lague index of disproportionality and Dalton's principle of transfers. *Party Politics*, 25, 203–207.
- Karpov, A. (2008). Measurement of disproportionality in proportional representation systems. *Mathematical and Computer Modelling*, 48, 1421–1438.
- Kauppi, H., & Widgrén, M. (2004). What determines E.U. decision making? needs, power or both? *Economic Policy*, 19, 221–266.
- Kauppi, H., & Widgrén, M. (2007). Voting rules and budget allocation in the enlarged E.U. *European Journal of Political Economy*, 23, 693–706.
- Koppel, M., & Diskin, A. (2009). Measuring disproportionality, volatility and malapportionment: Axiomatization and solutions. *Social Choice and Welfare*, 33, 281–286.
- Lambert, P. J. (2001). *The Distribution and Redistribution of Income*. Manchester University Press.
- Lauwers, L., & Puyenbroeck, T. V. (2006a). The Balinski-Young comparison of divisor methods is transitive. *Social Choice and Welfare*, 26, 603–606.
- Lauwers, L., & Puyenbroeck, T. V. (2006b). The Hamilton apportionment method is between the Adams method and the Jefferson method. *Mathematics of Operations Research*, 31, 390–397.
- Loosemore, J., & Hanby, V. J. (1971). The theoretical limits of maximum distortion: Some analytic expressions for electoral systems. *British Journal of Political Science*, 1, 467–477.
- Maaser, N., & Stratmann, T. (2016). Distributional consequences of political representation. *European Economic Review*, 82, 187–211.
- Marshall, A. W., Olkin, I., & Arnold, B. C. (2011). *Inequalities: Theory of Majorization and Its Applications* (2nd ed.). London: Academic Press.
- Marshall, A. W., Olkin, I., & Pukelsheim, F. (2002). A majorization comparison of apportionment methods in proportional representation. *Social Choice and Welfare*, 19, 885–900.
- Monroe, B. (1994). Disproportionality indexes and malapportionment: Measuring electoral inequity. *Electoral Studies*, 13, 132–149.
- Pennisi, A. (1998). Disproportionality indexes and robustness of proportional allocation methods. *Electoral Studies*, 17, 3–19.

- Renwick, A. (2015). Electoral disproportionality: What is it and how should we measure it?. <http://bit.ly/1RSUA4a>.
- Taagepera, R., & Grofman, B. (2003). Mapping the indices of seats-votes disproportionality and inter-election volatility. *Party Politics*, 9, 659–677.
- Van Puyenbroeck, T. (2006). Proportional representation, Gini coefficients, and the principle of transfers. *Journal of Theoretical Politics*, 20, 498–526.

“One Man, One Vote” Part 2: Measurement of Malapportionment and Disproportionality and the Lorenz Curve B: Applications



Olivier de Mouzon, Thibault Laurent, and Michel Le Breton

Abstract This chapter contains applications of the tools to the evaluation of malapportionment and disproportionality, already presented in Chap. 31. It is applied to the 2010 Electoral College and the French parliamentary and local elections with a special attention to the electoral reform of 2015. In these applications, the Lorenz curve ordering is almost conclusive, and consequently the Gini and *DK* indices are aligned and complement the almost complete ranking derived from Lorenz.

1 Introduction

This chapter applies to real-world cases the Lorenz curve and the Gini and *DK* indices presented in Chap. 31.

We apply these tools to several situations. First, we evaluate the Lorenz curve together with the Gini and the Dauer and Kelsay (*DK* hereafter) indices for the latest national legislatures. Second, we explore the Lorenz curve of each “département” in the Metropolitan part of France as well as two indices before and after the 2015 electoral reform. Third and last, we estimate the evolution of disproportionality over the last French parliamentary elections and the 2010 U.S. Electoral College.

The four real-world cases considered here are:

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-73249-3_32) contains supplementary material, which is available to authorized users.

O. de Mouzon

Toulouse School of Economics, INRAE, University of Toulouse Capitole, Toulouse, France
e-mail: olivier.de_mouzon@inrae.fr

T. Laurent (✉)

Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France
e-mail: thibault.laurent@tse-fr.eu

M. Le Breton

Institut Universitaire de France and Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France
e-mail: michel.lebreton@tse-fr.eu

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_32

633

1. The Evolution of the geographical Lorenz curve in the “Assemblée Nationale” of the French 5th Republic.
2. The Evolution of the ideological Lorenz curve in the “Assemblée Nationale” over the recent twenty five years of the French 5th Republic.
3. The Evolution of the geographical Lorenz curve in the “départements” before and after the 2015 electoral reform.
4. The 2010 Electoral College in the USA.

Supplementary material including original data and code can be found at <http://www.thibault.laurent.free.fr/code/4CT>.

2 The Evolution of the Geographical Lorenz Curve in the “Assemblée Nationale” of the French 5th Republic

The composition of the “Assemblée Nationale” results from the election of a single representative in each electoral district known as “circonscription électorale”.

The number of electoral districts has changed over time but since 1986, this number has remained unchanged at 577. Further, as pointed out by Sauger and Grofman (2016),¹ changes in the number and/or map of electoral districts, i.e. redistricting plans² in France³ have been very infrequent. They write:

“First redistricting plans in France have been very infrequent. During the more than five decades of the Fifth Republic, France has had only three censuses leading to redistricting: One linked to the initial district plan in 1958, then one in 1986 and, most recently, in 2009. Second, France has used the Adams method for apportionment. Of the set of standard apportionment methods the Adams method can be shown to be the one most favorable to small units by assuming even the smallest of them at least one seat. Third, [...], malapportionment in France at the constituency level is only partially constrained by constitutional rules.”

¹The factual and institutional informations reported in the beginning of this subsection are taken from Sauger and Grofman (2016). In the third section of their paper, Sauger and Grofman (2016) provide an assessment of the evolution of malapportionment from 1988 to 2012 for two different choices of units: “départements” on one hand and electoral districts on the other hand. Their results over the period from 1993 to 2012 are aligned with us.

²As noted by Sauger and Grofman (2016), “In France, we may think of redistricting as a two-step procedure. First, seats have to be allocated to geographically defined administrative units. In France, going back at least as far as the 3rd Republic, a divisor rule, called ‘méthode de la tranche’ (called the Adams rule in the U.S.) is used to allocate seats to ‘départements’. Second, within ‘départements’, single seat constituencies require that their boundaries be specified, and that rules be laid down about the degree of population equality needed across them. In France, the basis of apportionment and of evaluating population equality is namely persons (residents) rather than citizen population, registered voters, or something else”.

³On the U.S. history, we refer to Ansolabehere and Snyder (2008) and Cox and Katz (2002). Their books contain among other things a lively presentation of what is sometimes called the “reapportionment revolution” initiated by the Supreme Court decision on March 26, 1962 in the case *Baker v. Carr*.

The 1986 plan provided for mandatory redistricting to be conducted after each second general census but it was only on the basis of the third census, that of 2006 (results published in 2009), that a post-1986 districting plan was created. This failure to follow the law caused protest and the Constitutional council issued a first warning about the malapportionment issue. Sauger and Grofman (2016)’s appendix describes the timing of the 2009 redistricting plan. They point out that

“The 2008 constitutional reform introduced by the government places a ceiling on maximum assembly size of 577 and also required the creation of eleven new seats designated for French citizens living abroad. Since the imposition of both requirements could not be done while preserving the existing seat allocations, a new apportionment was a necessity... It also introduced a new element, an independent commission, the consultative council, which is to review and publish a public statement on any redistricting bill, although the option is only advisory”.

In spite of some continuity between 1986 and 2009, as noted by Sauger and Grofman (2016):

“There were changes in the requirement for minimal representation of each apportionment unit. The rule of a minimum of two seats for each metropolitan ‘département’ implemented in 1986 was rejected as unconstitutional by the Constitutional Council, and minimal representation was decreased to one seat per unit. That decision asserted that representation should be based mainly on population. [...] When we move from apportionment of seats to ‘départements’ to district lines within ‘départements’, a guiding principle was that the population of districts within any ‘département’ were not to be under or over 20 percent of the mean district population of the ‘département’ – except for special circumstances. Two other principles approved by the Constitutional Council are also very important. The first principle requires the territorial continuity of districts. The second principle is that cantons with fewer than 40,000 residents should be kept intact within a single constituency even if splitting the canton would have allowed for greater population equality across the districts within a ‘département’ (86-208DC). The third principle is that municipalities with less than 5000 inhabitants should be kept intact”.

This election occurs every 5 years which means there were 6 elections between 1993 and 2017 (1993, 1997, 2002, 2007, 2012, 2017). For each election, we have the results of the votes at the two rounds. Among the variables collected, we have the number of people who have the right to vote, the number of voters and the votes obtained by the different candidates.

It is important to point out that in the first application, only the data related to the number of people who have the right to vote is available for each election. Since the method to allocate the deputies is related to the number of inhabitants which is different from the number of people who have the right to vote, our conclusions are valid under the presumption that the ratio (voters/inhabitants) is sufficiently stable across time and space.

Figure 1 represents the boxplot and kernel density plot of the number of people who have the right to vote per “circonscription électorale”, with respect to the year of the election. On this figure, it is obvious that the distribution of the number of voters for each deputy has globally increased across the elections (meaning that the French electoral population has increased throughout the country over the years).

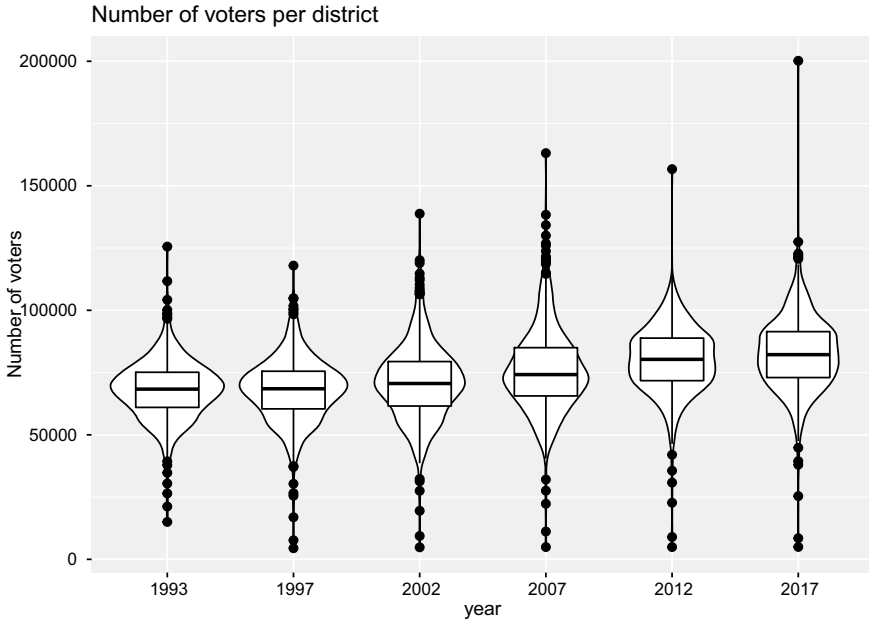


Fig. 1 Year by year boxplot (the boxplot presents three sample statistics—the lower quartile, the median and the upper quartile; it also presents the outliers) and kernel density plot of the number of people who have the right to vote per “circonscription électorale”

For any fixed election, we observe both outliers and a strong variance in the data, which seems to indicate that some circonscriptions are better (those with less voters) or worse (those with more voters) represented. Moreover, the distributions around the median seem uniform for the elections held in 2012 and 2017, while there are not for the elections held before.

In the next section, we try to better understand this distribution for a fixed year.

2.1 Analysis of the 2017 Election

We consider the population data in 2013 which is the one which is supposed to be used to settle the geographical boundaries of the circonscriptions.

Those geographical boundaries were used for the 2017 election. For this election, 10 districts were allocated to the French citizens living in foreign countries. We did not include these circonscriptions hereafter.

We look at the number of deputies observed per “département” to check if the rule “a ‘département’ has at least one deputy and an additional deputy is allocated every additional 125,000 inhabitants” is indeed followed.

Figure 2 plots the number of inhabitants per representative in each “département” with respect to its population. Note that “départements” ZS (Saint-Pierre-et-Miquelon), ZW (Wallis-et-Futuna) and ZX (Saint-Martin) have few inhabitants and

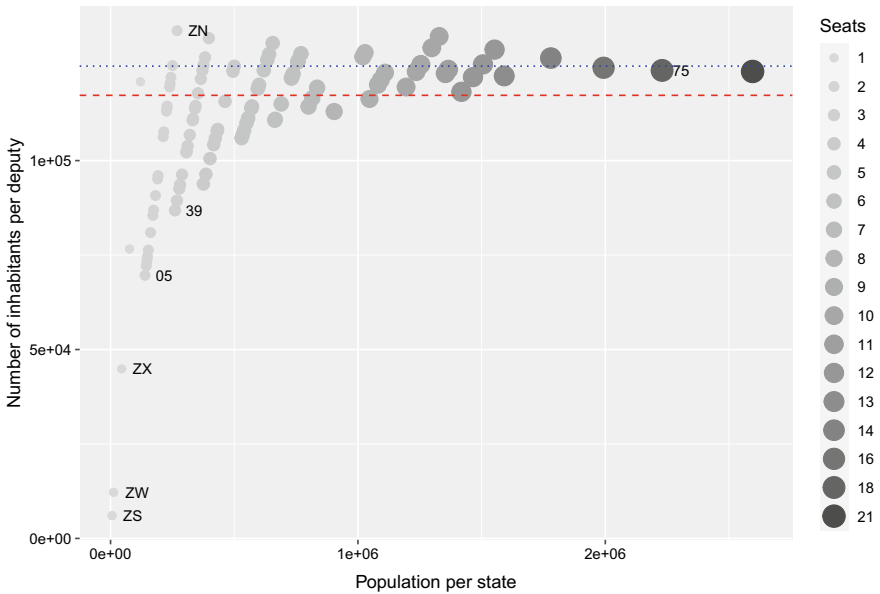


Fig. 2 Number of inhabitants per deputy for each “département” with respect to its population and dotted lines representing the average level (in dashed line) and the theoretical threshold leading to a supplementary deputy (in dotted line)

have 1 deputy each. It explains why the ratio (number of inhabitants)/(number of deputy) is very low for these circonscriptions.

Since the seat variable is integer valued, the rule is constant over intervals of populations and admits discontinuous jumps:

- 1 deputy if the number of inhabitants is lower than 125,000,
- 2 deputies if the number of inhabitants is between 125,000 and 250,000,
- etc.

The “départements” which are close to the lower bound are favoured and the “départements” which are close to the upper bound are disadvantaged. For example “départements” 05 (Hautes-Alpes) and ZN (Nouvelle Calédonie) have two deputies, but the first one has 139,279 inhabitants and the second has 268,767 inhabitants. In this case, it is interesting to notice that this last “département” should have three deputies like “département” 39 (Jura) which has 3 deputies and 260,502 inhabitants. The departure lies in the fact that the population data considered here is not the same as the one used to design the circonscription.

For the biggest “départements” (like 75-Paris), we observe that the ratio (number of inhabitants)/(number of deputies) is close to the theoretical blue line 125,000. The red line corresponds to the total number of inhabitants divided by the number of deputies and is equal to 117,274.

2.2 Lorenz Curve

Figure 3 plots the Lorenz curve for each year.

We define by n_k^j the number of voters and r_k^j the number of deputy in circonscription $k, k = 1, \dots, 577$ in election $j, j = 1997, 2002, 2007, 2012, 2017$ (the election of year 1993 was not kept because the data was incomplete). We plot on the horizontal axis all the cumulative fractions: $0, \tilde{n}_1^j, \tilde{n}_1^j + \tilde{n}_2^j, \tilde{n}_1^j + \tilde{n}_2^j + \tilde{n}_3^j, \dots, 1$ and on the vertical axis all the cumulative ordered fractions $0, \tilde{r}_1^j, \tilde{r}_1^j + \tilde{r}_2^j, \tilde{r}_1^j + \tilde{r}_2^j + \tilde{r}_3^j, \dots, 1$, where \tilde{n}_k^j and \tilde{r}_k^j have been defined in Chap. 31.

Zooming on this figure leads to the following observations:

1. 2012 seems always above the other curves except in two cases, where it is just under but still very close to the maximum curve (1997 when $\tilde{n} < 0.0075$ and mainly 2017 when $\tilde{n} > 0.9625$).
2. 2007 is below all the other curves when $\tilde{n} < 0.72$ except for $\tilde{n} < 0.007$ where it is just above but very close to 2017.
3. 2002 is below all the other curves when $\tilde{n} > 0.72$ except for $\tilde{n} > 0.985$ where it is just under but very close to 1997.

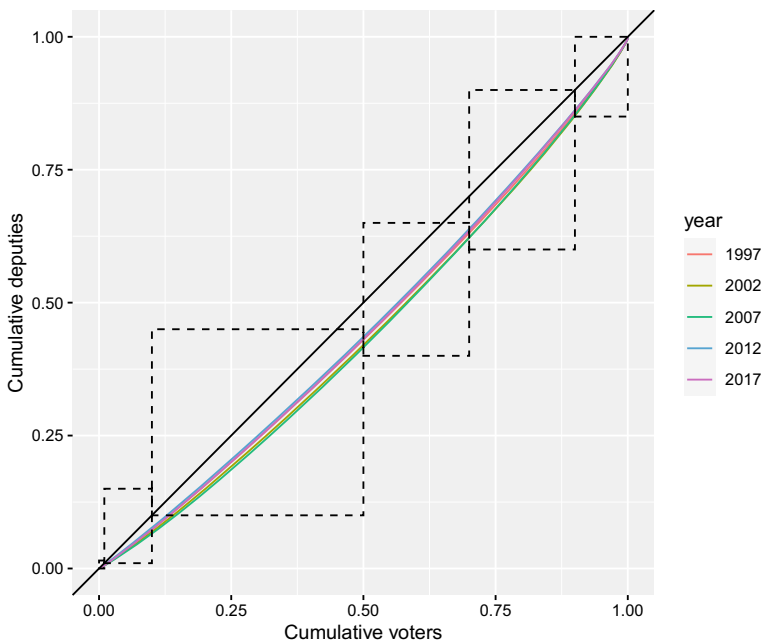


Fig. 3 Lorenz curve for the last five French “Assemblée Nationale” elections (zooms for the dotted rectangles can be found in the supplementary material)

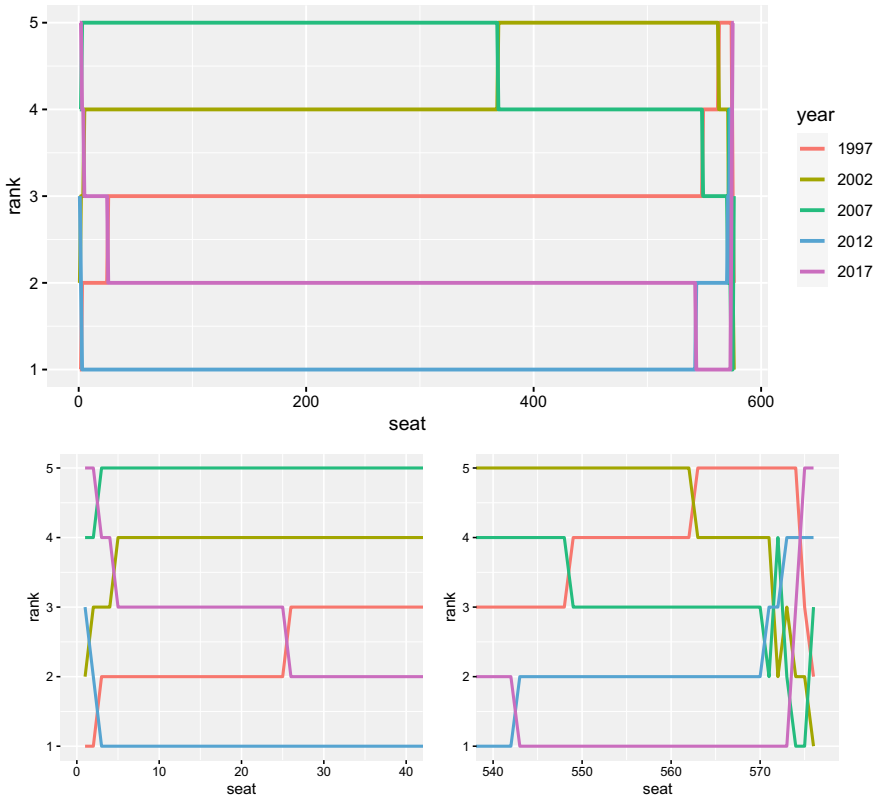


Fig. 4 Ranking of the closest Lorenz curve to the diagonal for each seat in the last five French “Assemblée Nationale” elections (and zooms for the first and last seats, where most crosses are observed)

4. At the beginning of the curve (i.e. $\tilde{n} < 0.5$), 1997 and 2017 are very close (except for $\tilde{n} < 0.05$ where 2017 is below 1997), then (when $\tilde{n} > 0.5$) 1997 is below 2017 (except for $\tilde{n} > 0.999$).

Moreover, we observe very few crossings between the curves. To check this, the rankings of the 5 studied elections were computed seat by seat (for each of the 577 seats) and are presented in Fig. 4. The link with the specific Lorenz curve of this application is the following: all the curves are based on 577 dots, which share the same y-coordinates (cumulative deputies). Hence, the ranking is easily obtained.

All the curves cross one another at least once. Yet, in all pairs of curves but one, there is always one curve that clearly is above the other one for most of the graph (at least around 94% of the graph). And the 6% or less of the graph where the situation is reversed is always at the very beginning or the very end of the graph.

The only pair of curves that does not match this trend is 2002 and 2007: in about 72% of the graph, the 2002 curve is above the 2007 one. Then, for the 28% or so

remaining part of the graph, the situation is reversed (except at the very end where the curves cross three times).

In spite of these few intersections, it seems that the elections in 2007 and 2002 were the least fair ones, then 1997, 2017 and finally the election in 2012 was the most fair ones (the one just after the application of the new rules).

2.3 Gini Index

The Gini index leads to the following results:

Best: 2012 ($G = 0.0464$) < 2017 ($G = 0.0497$) < 1997 ($G = 0.0517$) < 2002 ($G = 0.0589$) < Worst: 2007 ($G = 0.0613$)

Unsurprisingly, the Gini index confirms the conclusion derived from the Lorenz analysis: the elections in 2007 and 2002 were the least fair ones, then 1997, 2017 and finally the election in 2012 was the most fair ones.

2.4 DK Index

For the *DK* index, both versions (discrete⁴ and continuous⁵) lead to similar results.

Here are those for the continuous case:

Best: 2012 ($DK = 0.435$) < 2017 ($DK = 0.431$) < 1997 ($DK = 0.429$) < 2002 ($DK = 0.419$) < Worst: 2007 ($DK = 0.416$)

Again, the ranking is the same as with the Gini index and in line with what was conjectured from the Lorenz curve shapes.

> Conclusion on Application 1

In this application, the Lorenz curve ordering is almost conclusive, and consequently the Gini and *DK* indices are aligned with it for the fairness ranking of the studied elections. Moreover, the curves and indices are very close from one year to another, meaning that the fairness of the different elections seems quite stable in time. It is clear that the 2012 reform has designed circonscriptions fitting “at best” the population distribution of that year, leading to the fairest election. For the following election, the population had evolved a little, leading to a small decrease in the fairness of the year 2017 election. But its fairness seems very close to the year 1997. And year 1997 is two elections after the previous circonscription apportionment (which occurred in 1988). Then 2002 is one more election away as 2007. So it seems quite logical that the fairness

⁴In the discrete case: we search for the value of $x^* = \min(x_k)$, $k = 1, \dots, n$ so that $L(x_k) > 0.5$ and we get the *DK* with $1 - x_k^*$.

⁵More computation time is needed (due to the linear interpolation) for the continuous case, but the results are more accurate.

tends to decrease when moving away from the last apportionment, because population changes tend to follow a time trend.

3 The Evolution of the Ideological Lorenz Curve in the “Assemblée Nationale” of the French 5th Republic

In this section, we consider the same data used previously.

However, instead of considering the effect of apportionment (as in the previous section), we focus here on the differences between the vote shares and seat shares obtained for each competing party. For instance, Fig. 5 shows these differences for each of the 17 competing parties of the year 2012 election.

On this figure, the parties are ranked with respect to the Lorenz curve order: the first party is the one that was best off for this election and the last party is the one that was worst off (highest vote shares with no seat). The 4 first parties benefited from the electoral system (higher seat shares than vote shares) at the expense of the 11 others. Some parties with higher vote shares than others still get lower seat shares than the latter.

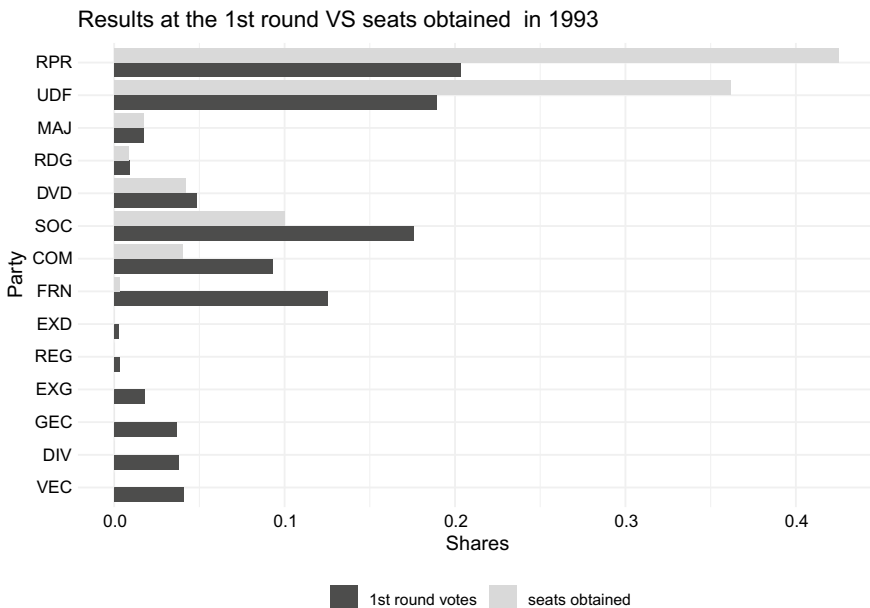


Fig. 5 Differences between the vote shares and seat shares obtained for each of the 17 competing parties of the year 2012 election

Hence, the correlation between the two shares is unclear. These huge differences cannot be explained by the small malapportionment studied in the previous section (especially in the case of the year 2012 where the malapportionment was the lowest). In fact, these differences are mainly due to the electoral system.⁶

Similar discrepancies are observed for the five other elections (years 1993, 1997, 2002, 2007 and 2017). They can be seen in the supplementary material.

3.1 Lorenz Curve

Figure 6 plots the different elections' Lorenz curves on the same graph.

Let us define by n_k^j the number of voters for party k , $k = 1, \dots, K^j$ and r_k^j the number of deputies obtained by party k in election j , $j = 1997, 2002, 2007, 2012, 2017$. We plot on the horizontal axis all the cumulative fractions: $0, \tilde{n}_1^j, \tilde{n}_1^j + \tilde{n}_2^j, \tilde{n}_1^j + \tilde{n}_2^j + \tilde{n}_3^j, \dots, 1$ and on the vertical axis all the cumulative ordered fractions $0, \tilde{r}_1^j, \tilde{r}_1^j + \tilde{r}_2^j, \tilde{r}_1^j + \tilde{r}_2^j + \tilde{r}_3^j, \dots, 1$, where \tilde{n}_k^j and \tilde{r}_k^j have been defined in Chap. 31.

Yet, as the party choice set K^j differs from one election to another (even in quantity) and also from one circonscription to another, it is difficult to explain the observed differences.

If we compare to Fig. 3, it appears that the Lorenz curves of Fig. 6 are much further away from the diagonal and with a much higher variability from one election to another. In fact, a deputy represents more or less the same number of voters throughout the country, but the seat shares are not necessarily in line with the vote shares.

Up to $\tilde{n} = 26\%$, the election in 2017 seems to be the most proportional one (closest to the diagonal). Then, after $\tilde{n} = 26\%$, it is 2007.

On the contrary, up to $\tilde{n} = 21\%$, the election in 2002 seems to be the least proportional one (farthest from the diagonal). Then, after $\tilde{n} = 21\%$, it is 1993. In fact, up to $\tilde{n} = 14\%$, the two curves overlap each other. So 2002 is a little worse than 1993 only for 5% of the vote shares.

Hence, 1993 seems the farthest from the proportional rule. Then, 1997, 2002 and 2017 seem close. Finally, 2007 seems the closest to the proportional rule, followed by 2012.

There are some curves crossing. Most of the crosses are on the first third of the vote shares. Then, after $\tilde{n} = 54\%$, the curves do not cross.

⁶Note that the vote shares are computed at the first round of the election and the seat shares are computed after the second round. The rules to be able to maintain candidacy between the two rounds, and the game of political alliances strongly differ from the proportional rule.

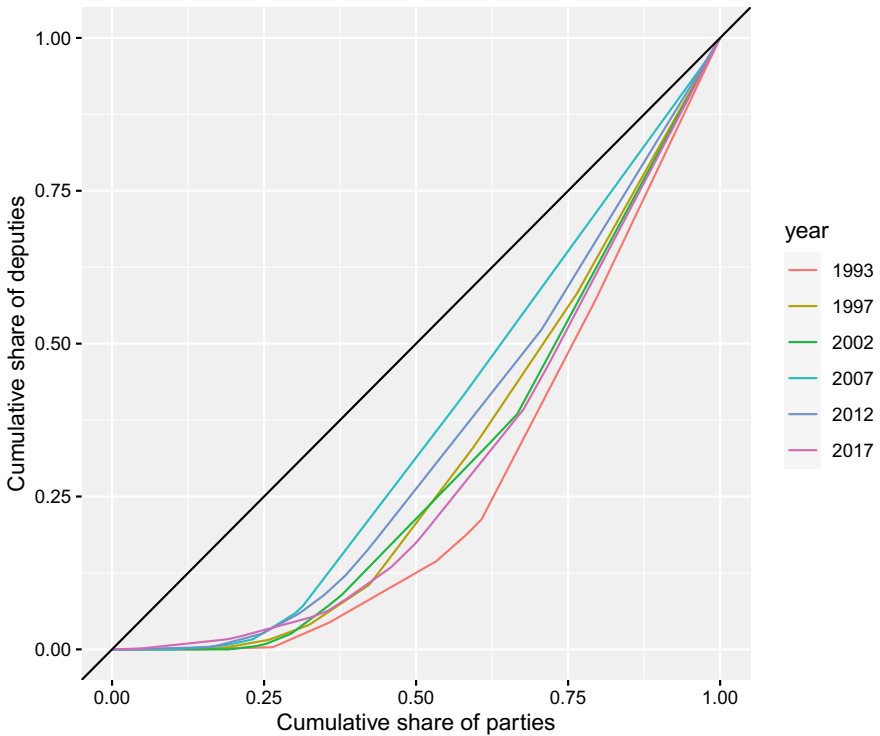


Fig. 6 Lorenz curves for 1993, 1997, 2002, 2007, 2012 and 2017 elections

3.2 Gini Index

The Gini index leads to the following results:

Best: 2007 ($G = 0.134$) < 2012 ($G = 0.162$) < 1997 ($G = 0.189$) < 2002 ($G = 0.195$) < 2017 ($G = 0.201$) < Worst: 1993 ($G = 0.233$)

As in the previous application, the Gini index is in line with the almost Lorenz ordering: the elections in 2007 and 2012 were the most proportional ones, then 1997, 2002 and 2017 and finally the election in 1993 was the least proportional one.

Specifically, the Gini index enables to break the ties, i.e. to rank the three years that were close but not straightforwardly ranked in terms of Lorenz curves.

3.3 DK Index

The continuous *DK* index leads to the following results which coincide with those derived from Gini:

Best: 2007 ($DK = 0.360$) < 2012 ($DK = 0.312$) < 1997 ($DK = 0.291$) < 2002 ($DK = 0.271$) < 2017 ($DK = 0.264$) < Worst: 1993 ($DK = 0.242$)

> **Conclusion on Application 2**

Like in the first application, the Gini and DK indices are aligned and complement the almost complete ranking derived from Lorenz. However, in this application, the curves and indices are far from the principle of proportionality mostly due to the electoral system. Moreover, we observe more variability from one election to another that could also be explained by the different party choice-sets across time and space.

4 The Evolution of the Geographical Lorenz Curve in the “départements” Before and After the 2015 Electoral Reform

The main objective of this section is to explore how the geographical Lorenz curve at the “département” level has changed as the result of an electoral reform simultaneous with some redistricting.

Each “département” elects a chamber of representatives. This legislative body is in charge of a number of local policies and redistributes resources across the territories within the perimeter of the “département”.⁷

This election proceeds from a division of the “département” into districts called cantons. Before 2015, the district magnitude was equal to 1: there was one seat per district and ballots consisted of a single candidate.

From 2015 on, several changes were implemented. First, the number of districts has been basically divided by two.⁸ Second, the district magnitude was increased from 1 to 2 with a very peculiar “winner-takes-all” electoral formula: each ballot consists of a ticket (not a list) of candidates (one male, one female).

The main objective of this reform was to guarantee the perfect equality of the two genders in the chamber. The electoral reform leaves unchanged the size of the

⁷This issue of malapportionment has an intrinsic interest but since one of our main motivations was rooted in distributive politics, it is legitimate to ask which fraction of the resources of a department could be considered discretionary enough to be modelled as a divide to the dollar game? Besides some anecdotal evidence that part of the budget falls into that category, we do not have, so far estimates of the fraction that can be classified under this heading. We thank Karine Van Der Straeten who has raised this question.

⁸In fact, the result of the division is rounded to the closest upper even number. Moreover, this number is at least 17 for “départements” with 500,000 or more inhabitants, and 13 with 150,000 or more.

chamber⁹ and has two components: a new map of the districts and a new electoral formula. It must also be pointed out that the reform has been exploited as an opportunity to solve at least partially the severe malapportionment problems of the historical electoral maps. This combination of multiple changes makes the problem quite complicated to analyze.

In the rest of this section, we will proceed to an evaluation of the 2015 reform from the perspective of the Lorenz curve. But before doing so, let us call the attention of the reader on the alternative evaluation, motivated by voting among two camps, that was presented in Sect. 2.2 in Chap. 31. We could indeed apply this method here with a focus on the colour (D or R) of the chamber. It may well happen that a majority of the voters of the “département” vote D and a majority of districts vote R. This is what we have called an election inversion. We could in particular compute how the measures Δ_λ^1 and Δ_λ^2 have changed under the reform, for some λ . This question is explored in Le Breton et al. (2017), where another index (called an index of disproportionality) is also introduced. As demonstrated there, if the principle “one man, one vote” is defined from that voting perspective, it is not clear that the reform led to an improvement.

In the context of distributive politics, things are different. Had the reform exclusively consisted in merging two old districts to create a new one, the post electoral reform Lorenz curve would have been closer to the diagonal than the pre-electoral reform one. This follows from a sequential application of the Pigou–Dalton principle. When two districts merge, the equal distribution within the new district dominates the unequal distribution prevailing in the union of the two old ones. We cannot apply without qualification this argument to the actual reform for many reasons, on top of which the fact that the redrawing of the map of districts was not as simple as a series of pairing. In this section, we look carefully at this question. As in Application 1, we focus here on the geographical distribution of the seats.¹⁰

4.1 Lorenz Curve

Figure 7 shows 100 graphs (one per “département” before the reform).

Each graph shows the Lorenz curve at the last “département” election before the reform (dotted line) and, for 98 of them, the one just after the reform (full line).¹¹

It is obvious that 96 “départements” out of 98 are better off after the reform: the full line is always closer to the diagonal than the dotted one. This means that

⁹In fact, the size of the chamber has slightly increased in some “départements”, as explained in Footnote 8.

¹⁰Again, we consider here the number of voters (available for the elections before and after the reform) and not the number of inhabitants (as such data was not easily available for each canton). So our conclusions are valid under the assumption that the ratio (voters/inhabitants) is sufficiently stable across time and space.

¹¹Note that the full line is present only for 98 “départements”: two small “départements” (overseas) disappear after the reform, so only the dotted line appears for them.

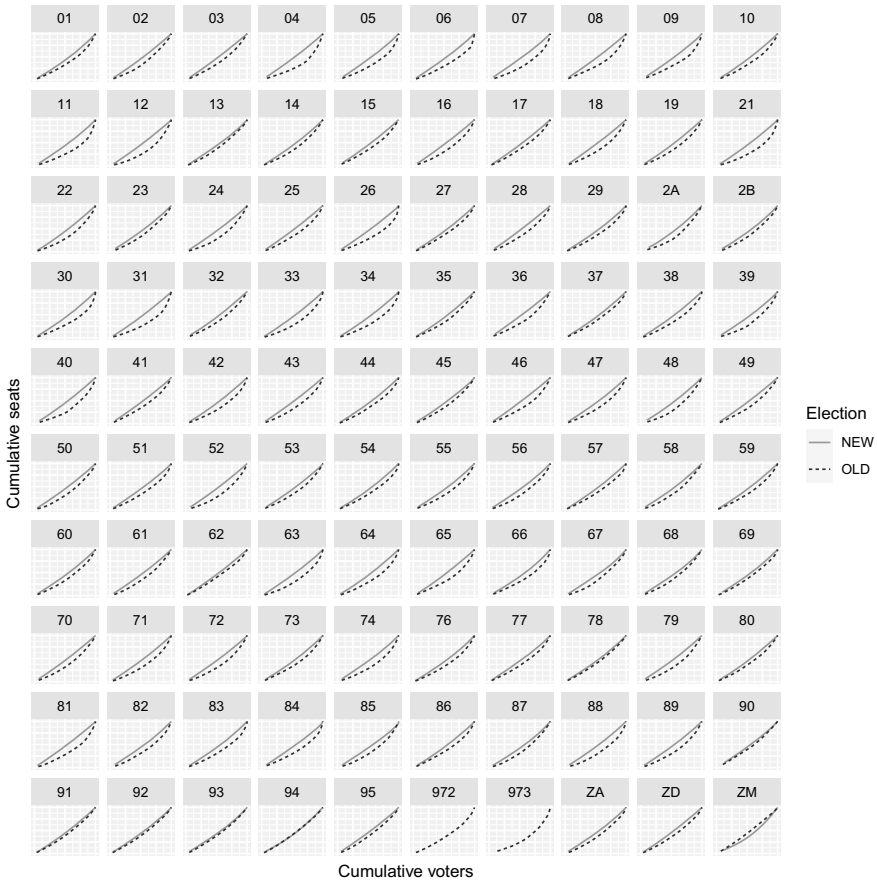


Fig. 7 Lorenz curves for each French “département” in the case of the 2015 election (full lines), after the reform, compared to the previous one (dotted lines, just before the reform)

the reform has enabled us to take into account the population changes that had occurred with time, similarly to the case of Application 1 (2012 election, right after the reform, was fairer than the previous one). There is a clear exception in the case of one “département” (Mayotte, last graph on the figure) which is worse off after the reform. This may occur when the actual number of voters and number of inhabitants are not so well correlated throughout that “département”.

The only questionable case is for “département” 94 (Val-de-Marne), but both dashed and full lines are very close to the diagonal and almost overlapping. In fact, the full line is closer to the diagonal at the beginning and at the end of the graph. The dotted line is only very slightly closer to the diagonal from $\tilde{n} = 34\%$ to $\tilde{n} = 74\%$. So, in the case of “département” 94, a few cantons are worse off, but overall “département” 94 seems better off.

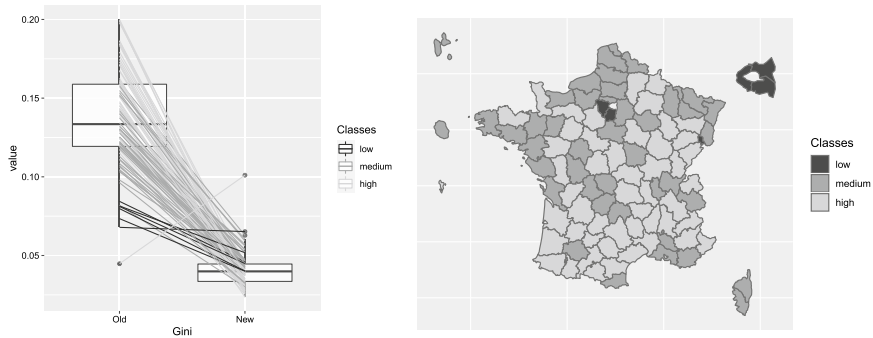


Fig. 8 Gini index before and after the reform. The 3 classes “high”, “medium” and “low” are defined with respect to the slopes and are represented on the map

It is interesting to notice that Fig. 7 before the reform shows dotted lines that can still be not so far from the diagonal (as all studied elections of Application 1) and others that are much further away (similarly to the worst case of Application 2). The graphs with a red line far from the diagonal correspond to “départements” where the population has probably most changed, so they most benefit from the reform, even if all of them reach a fairer situation.

4.2 Gini Index

Figure 8 on the left represents the Gini index for each “département” before and after the reform.

A line represents the evolution of the same “département”. The position of the boxes and the slopes of the lines indicate clearly a negative trend (except for two “départements”: Mayotte, clearly positive, and Val-de-Marne, flat but positive). We have represented the lines with different colours according to the absolute values of the slopes.

Our idea is to represent on a map those classes of “départements” and visualize¹² if there exists a spatial autocorrelation. It appears that there exists a spatial autocorrelation and a trend North/South. The “départements” with the largest changes are mostly located in the South of France. The “départements” nearby Paris seem the ones with the smallest changes. Finally, the “départements” with medium changes are mostly located in the North. It would be interesting to use a spatial econometric approach and model the Gini index by some socio-economic factors, to explain these differences in behaviour.

As before, and unsurprisingly, the Gini index corroborates the judgments based on the Lorenz curves.

¹²This kind of representation has been studied by Laurent et al. (2012).

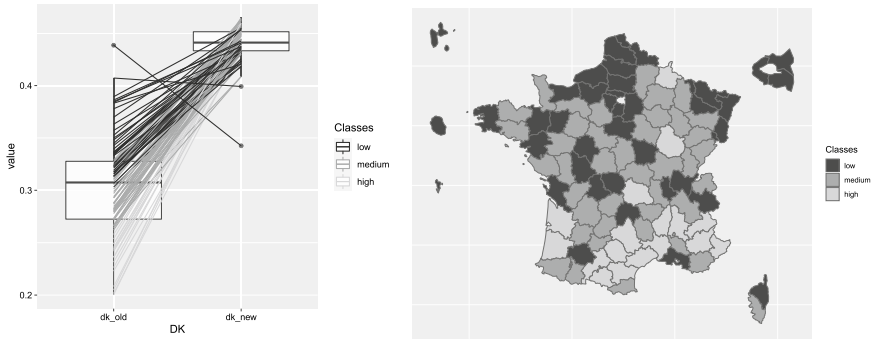


Fig. 9 *DK* index before and after the reform. The 3 classes “high”, “medium” and “low”, are defined with respect to the slopes and are represented on the map

4.3 *DK* Index

The left part of Fig. 9 represents the *DK* index for each “département” before and after the reform.

A line represents the evolution of the same “département”. We have represented the lines with different colours with respect to the absolute values of the slopes, as in the previous section (cf. Gini index).

The positions of the boxes and the slopes of the lines seem reversed compared to those obtained with the Gini index, so the conclusions are very similar (fairness is decreasing with the Gini index, and increasing with *DK*). The only noticeable change is for Val-de-Marne (“département” 94) where the more or less flat curve shows a small reduction in fairness, which is not in line with what the Gini index shows (we focus more on the area under the diagonal, which coincides with the Gini index).

> Conclusion on Application 3

We have shown that, except in the case of “département” Mayotte, all the new Lorenz curves are closer to the diagonal. As explained, the case of Mayotte could be explained by a change in the percentage of people who have the right to vote. In 96 out of 98 “départements”, the Lorenz Curves do not cross, so all indices (*G* or *DK*) confirm an improvement also. The interesting part of this application is for “département” Val-de-Marne (94): the Lorenz curves cross twice, meaning that the cantons having fewer seats per inhabitants and those that have most seats per inhabitants are better off, whereas the intermediary ones are worse off. This part is interesting because it shows that indices might not be aligned when the curves cross. Here, the Gini index concludes that the situation overall is better: the population update benefits to a majority of inhabitants.

On the contrary, the *DK* index concludes that the situation is worse because a smaller number of people may have half of the seats. To avoid misinterpretation, note that the inhabitants are not necessarily at the same abscissa before and after the reform, especially as the cantons are not the same in the two cases. So it might be the case that cantons that have the most population are better off after the reform, but they may concern totally different people.

Spatial analysis is required if we want to look at which areas are better/worse off. In fact, in this example, the situation might even not have occurred at all, had we access to the population of each canton (and not only to the number of voters): the two curves with population might not cross.

Also, we have looked at the curves as if they were continuous, but in reality, it might be difficult to obtain a set of cantons corresponding to at least half of the cantons, where the total population is under half of the total population. So the unfair situation captured by the *DK* index might in fact never occur. Finally, both curves are very close to the diagonal, so the situation is not so different before and after the reform. For all these reasons, no strong conclusion should be based on this result, but it is interesting to stress that the Gini and *DK* indices might not always be aligned, as in this case.

5 Electoral College

In this section, we consider the presidential U.S. elections during the 2010–2019 time period (based on the 2010 census).

The number of electoral votes (called hereafter “seats”) of a state is the sum of its number of representatives and number of senators (which is 2 for all states). The District of Columbia is allocated 3 seats. This data is fully presented in Table 1 of de Mouzon et al. (2020). The aim of this section is to compare the malapportionment when considering the allocation of the seats or the allocation of the representatives.

5.1 Lorenz Curve

Figure 10 presents the Lorenz curve when considering the number of seats (full line curve) and the number of representatives (dotted curve), based on the 2010 census.

The dotted curve is very close to the diagonal which shows that the representatives are allocated proportionally to the population of the State. However, the full line curve is always further away from the diagonal which indicates that the fact to allocate automatically 2 senators per state creates malapportionment.

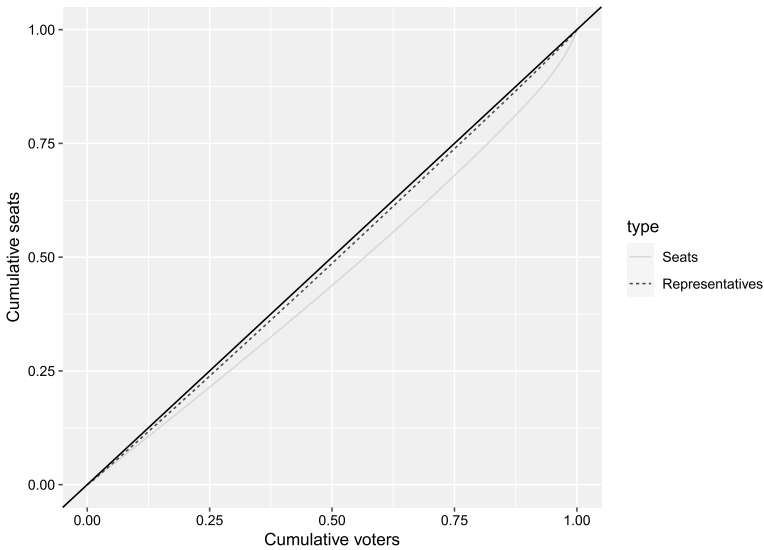


Fig. 10 Lorenz curve for the Electoral College in the U.S. elections based on the year 2010 census

5.2 Gini Index

The Gini index is equal to $G = 0.0484$ for the number of seats and $G = 0.0106$ when considering the number of representatives.

As the Lorenz curves do not cross, the Gini index is of course in line with what was observed in the previous section.

In the ideal situation corresponding to $G = 0$ (Lorenz curve aligned with the diagonal), it is interesting to observe that the number of seats has a Gini index 4.6 times higher than the number of representatives.

5.3 DK Index

The DK index is equal to $DK = 0.433$ for the number of seats and $DK = 0.486$ when considering the number of representatives.

Again, as expected, the DK index is in line with what was observed in the two previous sections.

It is interesting to observe that the distance to the ideal situation ($DK = 0.5$ when the Lorenz curve is aligned with the diagonal) is 4.8 times higher for the number of seats than the number of representatives.

Hence, both indices, Gini and DK , give a very similar relative difference to the ideal situation between the number of seats and the number of representatives.

> Conclusion on Application 4

This fourth application enables to show, in a simple setting, that both indices, Gini and DK , can sometimes be aligned even up to the relative difference to the ideal situation between two settings. This is of course not a general rule (e.g. in the previous application, we even had totally opposite outcomes in “Département” 94: with the Gini index finding an improvement after the reform, whereas the DK index finds a worse off situation).

Of course, this result is straightforward: the representatives are allocated on a proportional basis (and only suffer from the curse of rounding to integers their numbers). Adding two senators, whatever the population of the state, necessarily moves the curve further from proportionality of seats to populations of the states. And obviously, this result is not dependant on the census year: an equivalent result is obtained for the year 2000 census and any other.

Depending on the population distribution throughout the states in the different census years, it could be the case that some green (resp. red) curves are closer to the diagonal than others. But the green curves are always very close to the diagonal and the red ones always a little further away (although they are still close to the diagonal, as the curves of Application 1).

A more pragmatic question is to know whether the two “senatorial” seats really give a bonus to the small states in the presidential elections or whether they more or less correct some other unfairness (due to the fact that the biggest states have more representatives and thus more power in deciding who will be president). This question has been studied in the light of the three main voting probability models in de Mouzon et al. (2020).

Acknowledgements We acknowledge funding from ANR under grant ANR-17-EURE-0010 (Investissements d’Avenir program). Material from Sauger and Grofman (2016), quoted in Sect. 2, is used with permission from the publisher: Copyright ©2016 Elsevier Ltd. All rights reserved.

References

- Ansolabehere, S., & Snyder, J. M. (2008). *The end of inequality: One Person, One Vote and the transformation of American politics*. New York: W. W. Norton.
- Cox, G. W., & Katz, J. N. (2002). *Elbridge Gerry’s Salamander: The Electoral Consequences of the Reapportionment Revolution*. Cambridge: Cambridge university Press.
- de Mouzon, O., Laurent, T., Le Breton, M., & Moyouwou, I. (2020). One man, one vote part 1: Electoral Justice in the U.S. Electoral College: Banzhaf and Shapley/Shubik versus May. In M. Diss & V. Merlin (Eds.), *Essays by and in honor of William Gehrlein and Dominique Lepellety*. Springer.
- Laurent, T., Ruiz-Gazen, A., & Thomas-Agnan, C. (2012). GeoXp: An R package for exploratory spatial data analysis. *Journal of Statistical Software*, 47(2), 1–23.

- Le Breton, M., Lepelley, D., Merlin, V., & Sauger, N. (2017). Le scrutin binominal paritaire : un regard d'ingénierie électorale. *Revue économique*, *68*, 965–1004.
- Sauger, N., & Grofman, B. (2016). Partisan bias and redistricting in France. *Electoral Studies*, *44*, 388–396.

Visualizing France with Cartograms



Jonathan Haughton and Dominique Haughton

Abstract France has a long tradition of using statistical (choropleth) maps, which use shading to represent the spatial distribution of a variable, such as population, by department. Such maps lead the observer to underestimate the importance of urban areas, especially Paris. A solution that complements the choropleth map is to create a cartogram, which deliberately distorts each department so that the area is in proportion to the variable (such as population). Shading can then be used to show a second variable, typically representing density, on the same map. We illustrate the use of cartograms for the case of metropolitan France, with maps that show the spatial distribution of social housing, unemployment, immigration, suicides, election patterns, and the advance of COVID-19. The maps are relatively straightforward to construct, using ArcMap, but attention is needed to the use of colors and classifications. The cartograms reveal patterns that would not be clear based solely on traditional statistical maps.

1 Introduction

For almost two centuries, maps have been used to convey social, economic, and demographic information. In France, this is often done at the level of the department, which provides enough granularity to show spatial differences, but without overwhelming the observer with too much detail.

J. Haughton (✉)

Department of Economics, Suffolk University, 73 Tremont St., Boston, MA 02108, USA

e-mail: jhaughton@suffolk.edu

D. Haughton

Department of Mathematical Sciences, Bentley University, 175 Forest St.,

Waltham, MA 02452, USA

e-mail: dhaughton@bentley.edu

Université Paris I (SAMM), Paris, France

Université Toulouse I (TSE-R), Toulouse, France

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_33

There is a strong French tradition of innovative map-making. As far back as 1826, Charles Dupin produced a choropleth (statistical) map of France where departments with higher levels of illiteracy are colored in darker shades of gray.¹ The high levels of literacy in the Basque country, and in the northeast, emerge clearly. In 1858, Charles Joseph Minard superimposed pie charts on a map of departments indicating the number of cattle sent to the abattoirs of Paris. His most famous map depicts the losses suffered by Napoleon's army during the Russian campaign.² Among the many statistical maps produced by Emile Cheysson, one of the most interesting is a stack of maps of France designed to dramatize the reduction in travel time over the two centuries prior to its publication in 1888.³ While the term "cartogram" was sometimes applied to such maps, most are best thought of as choropleth maps, or imaginative extensions to them; Tobler (2004) reviews a number of these early examples of visual analytics.

Modern choropleth maps use shading that is in proportion to an underlying statistical variable and are a popular form of thematic map. The top row of Fig. 1 shows two such maps. The first shades departments in 2020 according to population density—a form of spatially intensive data; the dense areas around Paris, Lille, Lyon, and Marseille stand out, in contrast to the emptiness of the Alps and of the Massif Central. The top-right-hand map in Fig. 1 colors departments by total population—spatially extensive information—and draws the eye somewhat better toward the centers of population in France.

Yet neither map is entirely satisfactory, because the true size of the population, by area, is not apparent. One solution is to superimpose symbols—such as the circles in the lower left panel of Fig. 1—that are in proportion to population, creating a graduated circles map. This map is designed to mimic those produced by INSEE on its excellent website.⁴ The concentrations of population are certainly clearer here, but the circles overlap, so it is still hard to get a good sense of how the population is distributed.

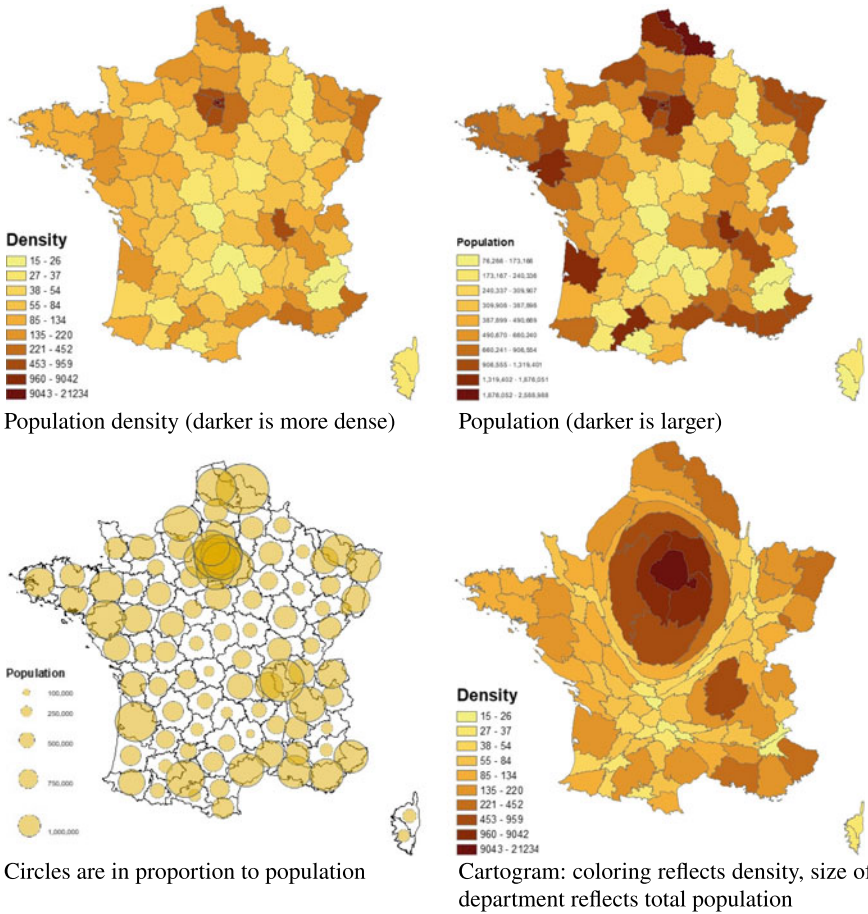
In this paper we argue that it is typically useful to complement the standard choropleth or graduated circles maps with cartograms, such as that shown in the bottom-right panel of Fig. 1. In what follows we explain what cartograms are, how they can be constructed, what makes a good cartogram, why they are illuminating, and the potential limits to their use. This essay is inspired by the work on visual analytics of Christine Thomas-Agnan and her colleagues.

¹Dupin's map may be found at http://math.yorku.ca/SCS/Gallery/images/dupin1826-map_200.jpg.

²Minard's map of pie charts may be found here: https://en.wikipedia.org/wiki/Charles_Joseph_Minard/media/File:Minard-carte-viande-1858.png and his celebrated "Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813" is available here: https://en.wikipedia.org/wiki/Charles_Joseph_Minard/char%0023/media/File:Minard.png.

³"Accélération des voyages en France depuis 200 ans", Plate 8a in *Album de Statistique Graphique* by Emile Cheysson. The map is available here: http://www.sci.utah.edu/~kpotter/Library/Papers/friendly:2008:GASS/friendly_2008_GASS_14.png.

⁴For instance, <https://www.insee.fr/fr/statistiques/2012713>.



Why does this deliberately distorted map work so well? It is probably because most viewers have a clear mental picture of France, and the cartogram provides a contrast, almost a shock, as they realize how unevenly the population is distributed spatially. If the population were evenly spread throughout Metropolitan France, the cartogram would look just like a “normal” map, and be uninteresting. Thus, the real power of the cartogram is that it “distorts the geography to overcome some of the problems of heterogeneous reality” (Field 2017), and the larger those differences, the more dramatic the visual effect. In practice, cartograms work best when shown side by side with standard choropleth maps, so that the differences become obvious.

Cartograms come in a wide variety of styles and types, as the recent review by Nusrat and Kobourov (2016) makes clear. Among the most effective, and the only ones we use here, are diffusion-based cartograms, constructed using the algorithm developed by Gastner and Newman (2004). These maps are statistically accurate, in that the size of the areas reflect the underlying variable; they are contiguous, so that areas are glued to each other, preserving the principle of adjacency; but they are geographically distorted.

Some of the most dramatic cartograms are those available on the WorldMapper website Worldmapper (2020), which offers a number of visualizations of the world as a whole, with areas reflecting such variables as population, GDP, HIV/AIDS cases, CO2 emissions, and the like. Hennig (2019) provides some background to the WorldMapper project. In the United States, cartograms have been widely used to visualize the national breakdown of votes in Presidential elections, where Flanagan (2016) has written about the “battle of the maps.” Low-density states are more likely to vote Republican, so an undistorted map shows a lot of red area; but when states are scaled by the number of votes, the country looks more evenly divided between red (Republican) and blue (Democrat).

Curiously, these types of cartograms have not been widely used for France, so in this paper we provide several examples, which are interesting in their own right, and may also inspire others to use the tool more frequently.

Below, we present a selection of cartograms related to housing, unemployment, immigration, elections, and COVID-19, with brief commentaries on the content, in the spirit of computer scientist Ben Shneiderman’s remark that “the purpose of visualization is insight, not pictures.”

What makes a useful map? In his classic book *The Visual Display of Quantitative Information*, Tufte (2001) argues that “there are two goals when presenting data: convey your story and establish credibility.” He emphasizes the importance of good design, argues for simplicity and the removal of graphical clutter (including extraneous “chartjunk”), and favors maximizing the “data to ink” ratio. These principles of parsimony are broadly applicable to cartograms, but maps (including cartograms) are typically more complex than graphs, and require attention to other principles. Buckley (2012) argues that maps, among other features, need to provide a clear visual contrast, and must be legible. Geographer Keith Clarke (2020) asserts that “good design makes a map more effective and interpretable”, and notes that the eye picks up similarities, the proximity of phenomena, and continuities when it scans

a map. This helps explain why the key patterns shown in a cartogram can be grasped almost at a glance.

In what follows, we show examples of cartograms that work well, as well as some that are not worth the trouble. Three generalizations, to which we return below, seem particularly important:

1. Cartograms are typically most effective when paired with choropleth maps.
2. The coloration of cartograms works best when it measures variables related to density (such as population density, or death rates) rather than totals (such as population, or deaths).
3. Cartograms are most helpful when the geographic distortions they produce are greatest.

Mechanically, it is relatively straightforward to create cartograms with ArcMap, which was used for all the cartograms shown here, but we have also achieved adequate results with ScapeToad (2008). A “cartogram tool for ArcGIS” (ESRI 2015) needs to be downloaded and installed; the instructions are clear, and, with a little practice, cartograms can be created as easily as new maps.

3 Housing and Camping

Of the 34.6 million residences in France, ten percent are secondary residences. The top-left panel of Fig. 2 shows that they are concentrated along the Atlantic seaboard, Mediterranean, and Alps. The cartogram below uses the same shading, but shapes the departments so that they reflect the number of secondary residences. The large number of secondary residences along the Mediterranean is striking and is in contrast to their dearth in the northeast.

The right-hand side of Fig. 2 shows another form of leisure facility, the campsite. The pattern is comparable to that of secondary residences, with the difference that there are essentially no campsites in or close to Paris—indeed this is one of the few cartograms for which the Parisian region remains unimportant—and they are relatively more common in Brittany and in the southwest. While secondary residences are popular in the mountainous east, this is not as true of campsites. This is a straightforward case where the cartograms quickly provide a more complete picture than the statistical maps above them. They also contain more information, because they illustrate two variables—the density of secondary residences (or campsites), and their total number.

The next maps (Fig. 3) show the distribution of social housing, which accounts for 14% of all residential units in the country. The density of social housing is highest in the north, but the cartogram shows more clearly how much social housing

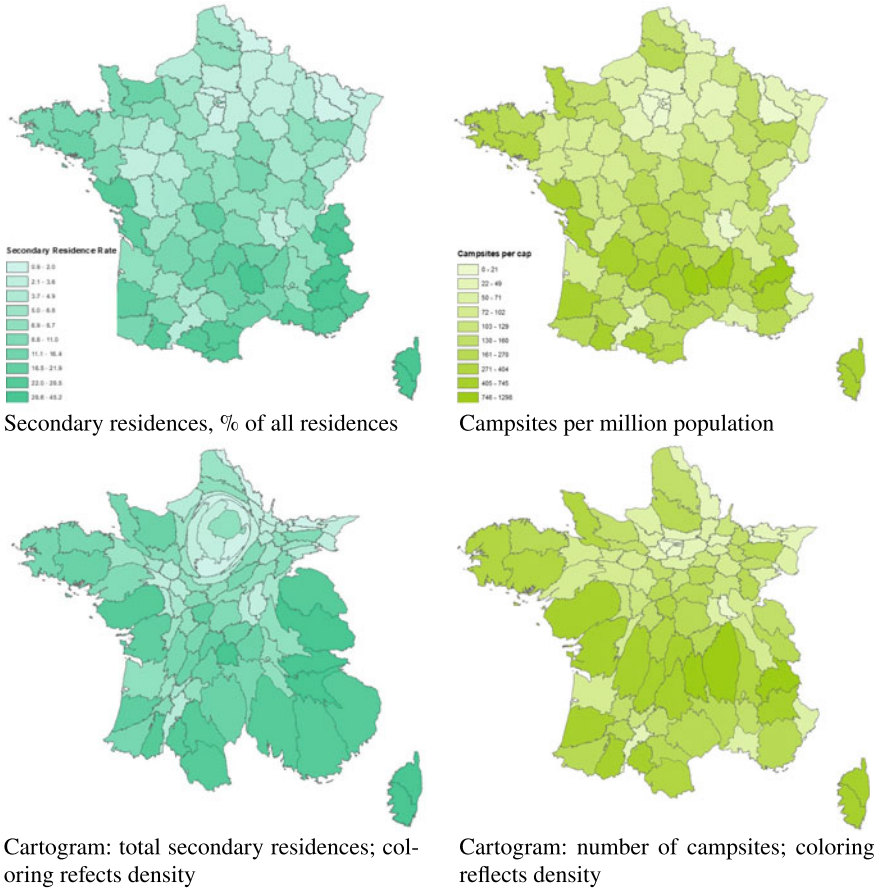


Fig. 2 Secondary Housing and Campsites by Department, 2015. *Data source:* INSEE. *Notes:* Secondary residences sharing reflects secondary residents as a percentage of all residents, and deciles range from 0.9-2.0% to 28.6-45.2%. Campsite shading reflects campsites per million population, and deciles range from 0-21 to 746-1298

there also is in and around Paris, and how little in the south and southwest. When compared to the population cartogram in Fig. 1, there is an evident geographical inequality in the availability of social housing. When this cartogram is set beside that for unemployment (Fig. 4), it is interesting to note that while both the north and southeast have high unemployment rates, only the north has a high incidence of social housing.

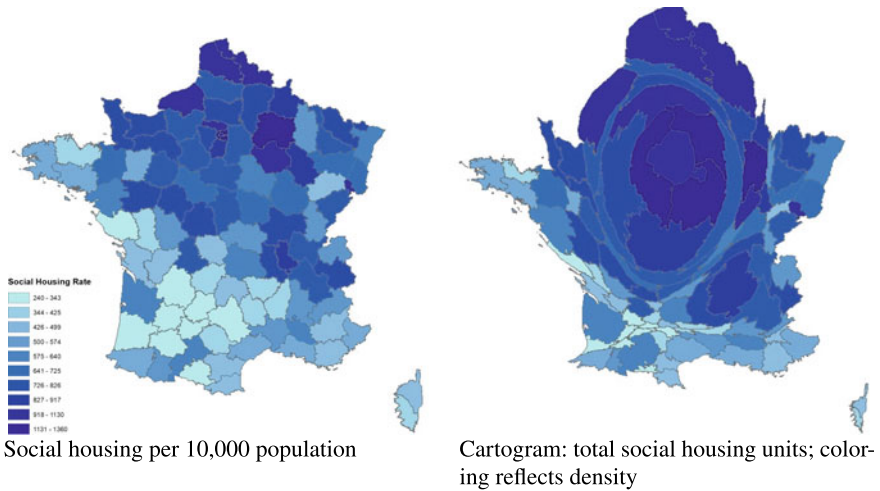
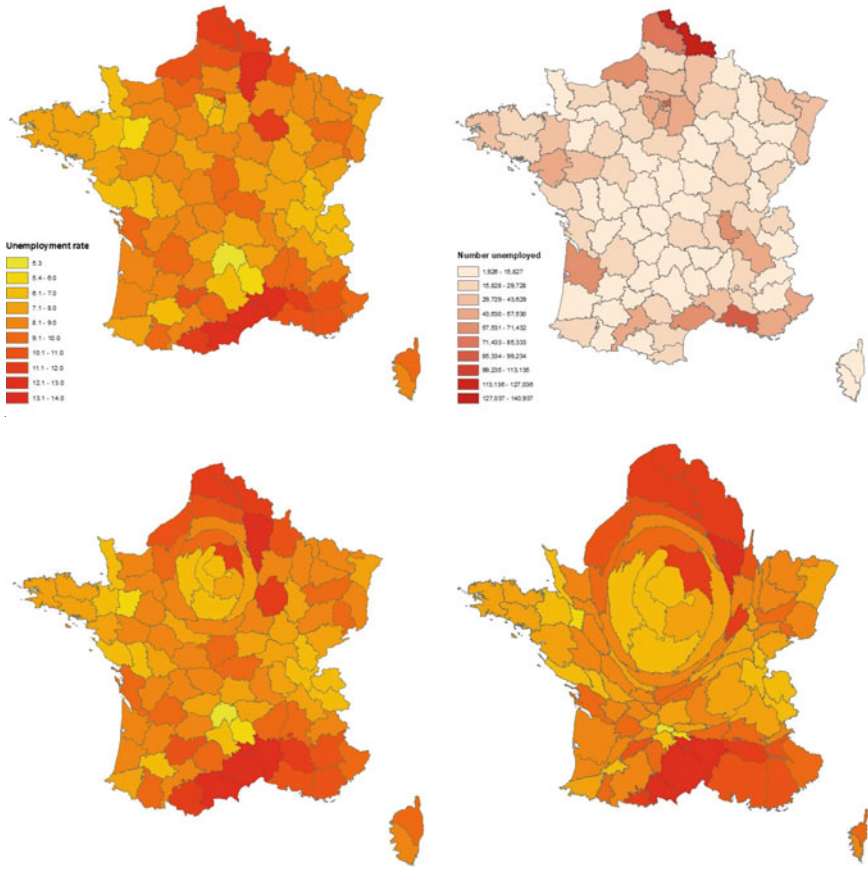


Fig. 3 Social Housing by Department 2015. *Data source:* INSEE. *Notes:* Shading reflects social housing per 10,000 population, and declines range from 240-343 to 1,131-1,360

4 Unemployment

The high rate of unemployment in France over the past several years is a matter of ongoing concern. There was also wide variation across departments in 2018, from 5.3% in Cantal to 14.4% in Pyrénées-Orientales. The top-left graph in Fig. 4 is a conventional map of the unemployment rate by department. The high rates at the northern and southern edges of France are notable. It is also difficult to see what is happening in the Paris area. The top-right graph shades, in beige, the departments with more unemployed people, and this changes the picture considerably: Ariège, for instance, has a high unemployment rate, but given its small population, has relatively few unemployed people. On the other hand, this choropleth graph is not very convincing, because it does not give sufficiently precise visual cues as to the extent of unemployment.

The bottom graphs in Fig. 4 are cartograms. On the left, the areas reflect the unemployment rates, but this is hardly an improvement over the map above it, and indeed is not an appropriate use of the cartogram: it is included here to make precisely that point. On the other hand, the cartogram on the bottom-right shows where the unemployed are to be found—in the greater Paris area, the north, and along the Mediterranean littoral. The persistence of differentials in unemployment rates is discussed in some detail in Aragon et al. (2003), who examined this issue in the context of the (then) Midi-Pyrénées Region: it is not always easy for the unemployed to move to where jobs are more plentiful, and there may also be reluctance to move from attractive parts of the country.



Cartogram: both areas and shading show the unemployment rate

Cartogram: areas reflect number of unemployed, shading shows unemployment rate

Fig. 4 Unemployment by Department, 2018. *Data source:* INSEE. *Notes:* Shading reflects unemployment rate (for red/yellow maps) and groups range from 5.3 to 13.1-14.0. Shading in tan map reflects total number of unemployed, and groups range from 1,925-15,827 to 127,037-140,937

5 Immigration

Immigration is a politically and economically sensitive topic in France, as elsewhere. In 2016 there were 6.1 million immigrants in Metropolitan France, out of a total population of 64 million. Figure 5 shows the spatial distribution of immigrants in a number of ways. The top-left panel is a choropleth map that shows the number of immigrants and classifies departments into deciles, so there are approximately equal numbers of departments with each degree of shading. The pattern of immigration is not obvious, although urban areas (Paris, Lille, Lyon, Toulouse, Bordeaux, and Marseille) appear to have relatively more immigrants.

The top-right map shows the same data, but uses a different shading protocol, with a set of fixed rather than relative classes. It shows that in most departments, the absolute number of immigrants is small—typically less than 30,000. A somewhat different picture emerges from the bottom-left panel of Fig. 5, where the shading reflects the density of immigrants, as measured by the number of immigrants per 1,000 population. The low density of immigrants in Brittany, indeed in the areas west of Paris, is striking and was not evident from the maps in the top row. The most useful map is probably the cartogram, which shows how strong the pull of the Île-de-France region is for immigrants: the maw of the greater Paris area accounts for over a third of the total.

It is clear that the choice of shading is important to the story that is to be told. This is true of all good maps, which emphasize features of interest and downplay elements that are not germane to the matter at hand. But the fact that map-making is as much art as science is also a call for the reader to be vigilant, because it means that maps can be, in effect, manipulated by their makers. Unless otherwise noted, we use a consistent shading scale for the maps within any figure, to allow for proper comparisons. So the high immigration density in the Bouches-du-Rhône department is shaded the same way in both panels at the bottom of Fig. 5.

6 Suicides

The geographic distribution of medical and health outcomes lends itself naturally to cartograms. To illustrate this, we consider the case of male suicides. In 2015, almost nine thousand people committed suicide in France, and just over three-quarters of these were men. At 12.1 suicides per 100,000 people in 2016, France has one of the highest suicide rates in Western Europe, more than double that of Spain and Italy, and somewhat higher than the global average of 10.6 (Organization 2017).

There were wide differences in male suicide rates from department to department in 2015, from an astonishing 343 per 100,000 males in Nord to just 9 per 100,000 in Belfort. The cartogram on the right-hand side of Fig. 6 gives a somewhat better sense of where the number of male suicides is large than the statistical map on the left: there are almost as many suicides in Brittany, or in the North, as in the Parisian Basin. High suicide rates appear to be a characteristic of rural and small-town France.

7 Elections

Cartograms have been widely used to show the spatial distribution of votes for Democrats and Republicans in the United States. In France, political competition does not divide so neatly along two-party lines: In the local elections of 2015 there were at least 17 significant political parties. Nguyen et al. (2018) classify these into three groups—Left, Right, and Far Right—and generate a ternary diagram, recreated

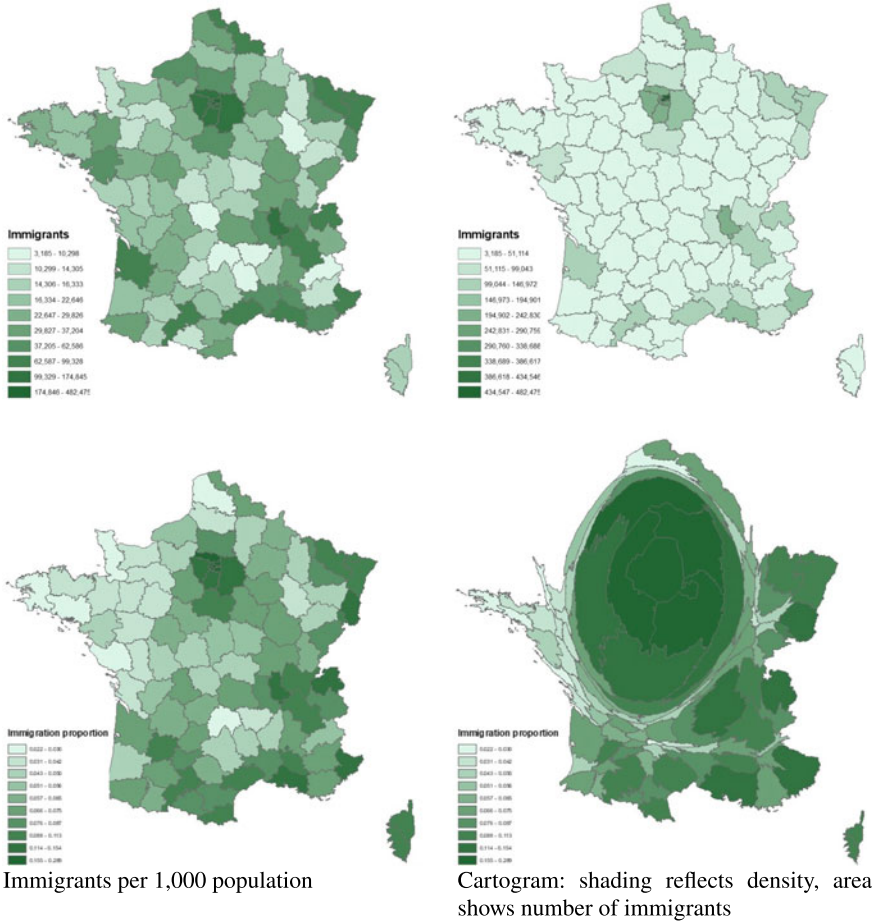


Fig. 5 Immigration by Department, 2016. Data source: INSEE

here in Fig. 7, that locates each department (except for Paris and Lyon, which have a separate electoral regime) on the relevant scales. For instance, the department of Aube is marked with a red X: 14% of its votes went to parties on the left, 49% to parties on the right, and 37% to parties on the far right. While the diagram does show the heterogeneity in voting patterns, it is not designed to reflect the spatial pattern of political preferences.

The spatial pattern of voting can be seen in the choropleth maps and cartograms shown in Fig. 8. The strength of left-leaning parties in the west and southwest is well-established and clear from both the map and the cartogram, although the latter does show the substantial number of left-leaning votes in the north and near Paris. The regional nature of the votes going to the Far Right is clear in the cartogram on the bottom-right of Fig. 8, perhaps to a greater extent than one would understand

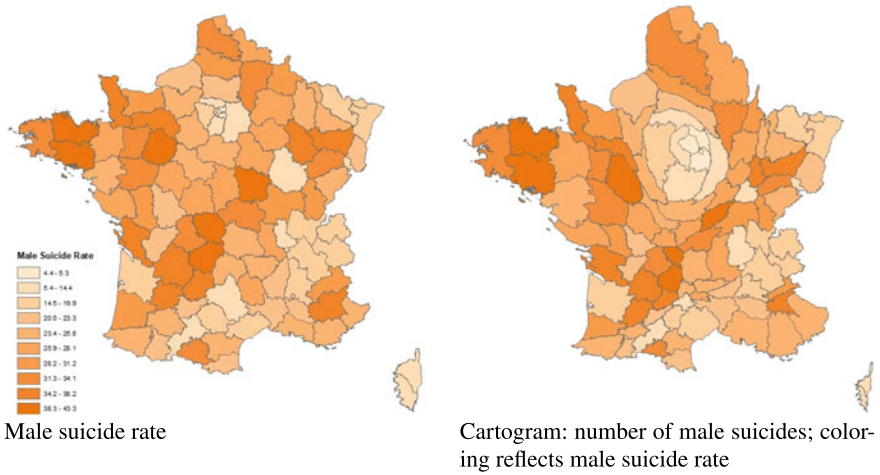


Fig. 6 Male Suicides by Department 2015. *Data source:* INSEE. *Notes:* Shading reflects suicide rate per 100,000 population, and declines range from 4.4-5.3 to 38.3-43.3

from the map above it. Despite these differences, the cartograms do not add a great deal to what may be seen from the conventional choropleth maps, in part because of the greater complexity of dealing with multiple political groupings rather than a binary choice, but also because the spatial variation in voting preferences is in fact relatively modest.

8 Covid-19

The devastating novel coronavirus that reached France in early 2020 had killed 516 people by March 20, and 16,643 by May 20, according to official statistics. The geographic distribution of deaths related to Covid-19 has been extremely uneven, with the great bulk of the cases occurring in the northeast, and in Paris and the surrounding areas. The pattern is dramatized in Fig. 9. The top row of choropleth graphs shows the evolution of the cumulative death rate from the virus between March 20 and May 20, during the period of its most rapid spread, and then the distribution of cumulative deaths as of October 4, 2020. Initially, deaths were mainly contained to the northeast, but gradually spread westwards.

The bottom row of Fig. 9 shows a series of cartograms that reflect the total number of deaths. The pattern is remarkable: through May 20, the great bulk of deaths continued to be in the greater Paris area, and in the northeast, with very few elsewhere, although by October the pattern of deaths spread more widely around the country. The early geographic confinement of the virus in France mirrors that of China, where the great bulk of cases were in Hubei province; Gao et al. (2020) represent

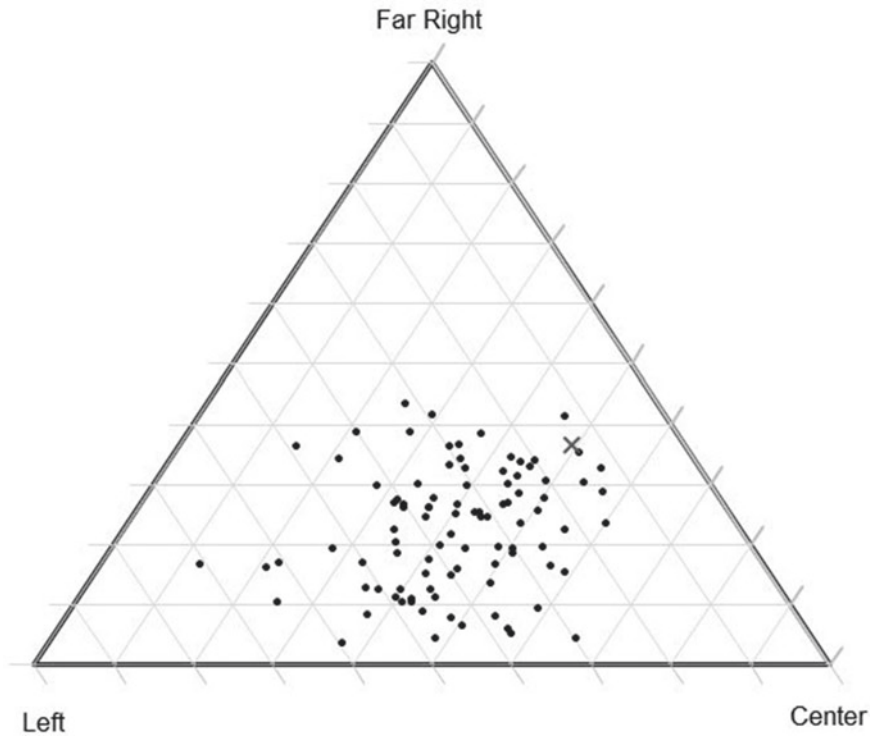


Fig. 7 Ternary Diagram of Voting by Departments in French Local Elections, 2015. *Data source:* [French Ministry of Interior](#)

this very effectively with a Dorling cartogram, which represents provinces by circles of different size rather than by maintaining the pattern of geographic contiguity as we have done. Worldmapper (2020) has also made good use of cartograms to map the worldwide evolution of COVID-19.

Cartograms show relativities, so the bottom-left cartogram that shows the distribution of 516 deaths has the same area as the bottom-right cartogram that reflects the geographic spread of 21,212 deaths. Perhaps the cartograms themselves should be scaled. The tiny boxed cartogram in the bottom-right panel of Fig. 9 is the cartogram for March 20 drawn to scale. A lot changed in six months.

9 Conclusions

Cartograms, of the type we have presented in the paper, can effectively complement the more traditional choropleth maps that have, up to now, been the main tool for conveying spatially distributed statistical data. They can be produced relatively easily

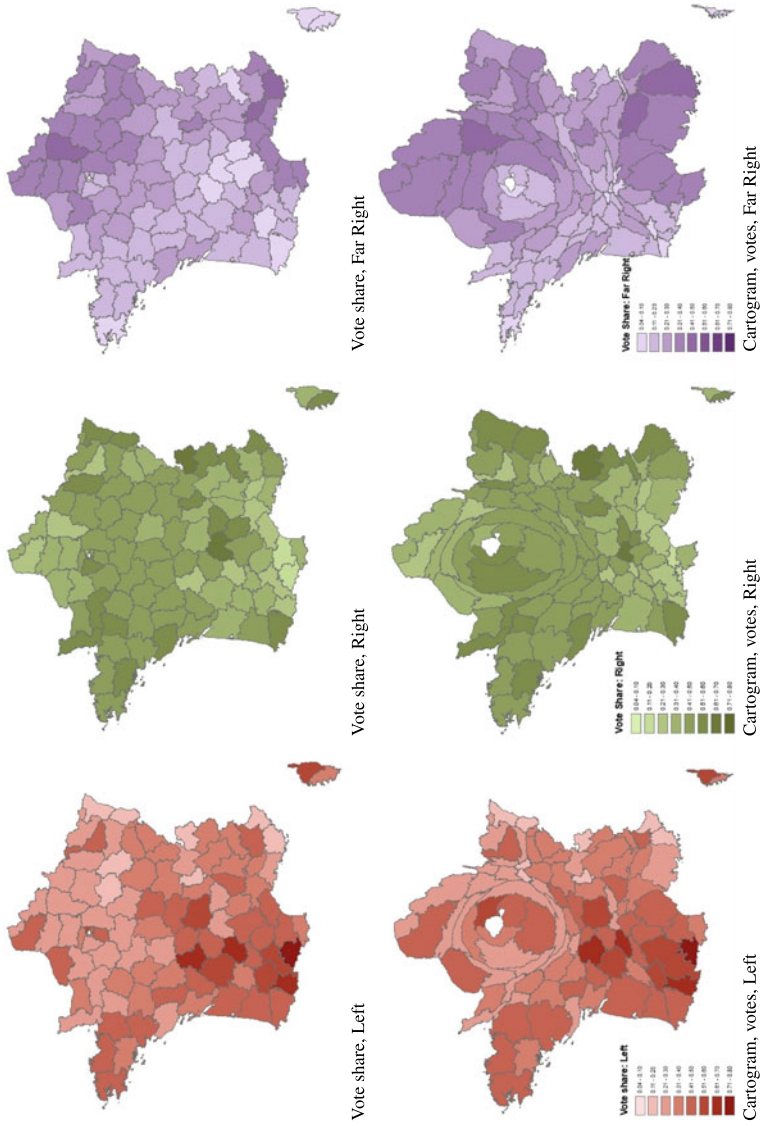
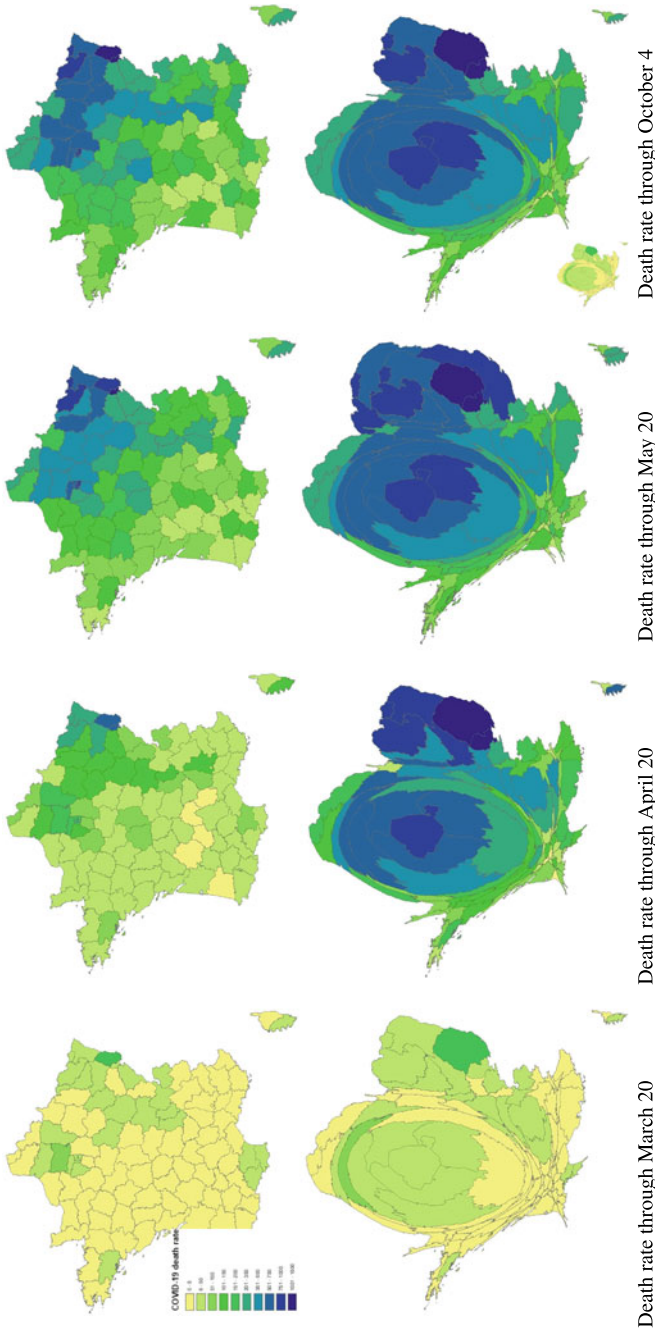


Fig. 8 Vote Shares of Round 1 in Local Elections, 2015. *Data Source:* French Ministère de l'Intérieur. *Notes:* Color scales are same for choropleth graphs and cartograms, and for each pair of graphs. Elections in Paris and Lyon were held at a different time and are not included here



Death rate through October 4

Death rate through May 20

Death rate through April 20

Death rate through March 20

Fig. 9 Number of Deaths Due to COVID-19, March-October 2020, by Department. *Data Source:* Coronavirus statistics (2020) *Notes:* In all graphs, shading reflects cumulative number of deaths per million population; in the cartograms (bottom row), area reflects number of deaths

and quickly using ArcMap, and they work very well when applied to France at the departmental level. By European standards, France is a large and heterogeneous country. Most standard maps do not adequately convey the importance of the Paris region, while cartograms make this absolutely clear. Cartograms of France are also effective because readers begin with a clear mental picture of what France looks like geographically and can appreciate the deviations from this baseline that are represented in a cartogram.

There are a few basic recommendations for anyone planning to create and present cartograms. First, they work best when coupled with choropleth maps, so the two can be compared side by side. Second, it is important to choose the shading carefully, as illustrated in Fig. 5; poor choices can give a misleading impression, or obscure an interesting pattern. Moreover, the shading should reflect a variable that measures density; the area of each unit in the cartogram measures the (relative) numbers. Third, cartograms are most revealing when the distribution of some attribute—population, campsites, suicides—differs substantially from the distribution of land area. When this is not the case, as with the election results shown in Fig. 9, the cartograms do not add a lot of insight.

We got interested in the topic of cartograms when we began to think about the work of Christine Thomas-Agnan, who has long had an interest in the spatial dimensions of statistics. Her paper on elections (Nguyen et al. 2018) led us to create cartograms for the same electoral data (Figs. 7 and 8). Her work on regional unemployment (Aragon et al. 2003) prompted us to construct cartograms for unemployment rates and levels (Fig. 4). And her development of GeoXp (Laurent et al. 2012) has been an inspiration: it is a package for exploratory spatial data analysis that allows one to create maps with statistical information at the same time as graphs with associated distributional data, such as densities, histograms, Lorenz curves, and the like. It would be interesting indeed if that tool could be expanded to draw cartograms as well.

References

- Aragon, Y., Haughton, D., Haughton, J., Leconte, E., Malin, E., Ruiz-Gazen, A., et al. (2003). Explaining the pattern of regional unemployment: The case of the midi-pyrénées region. *Papers in Regional Science*, 82, 155–174.
- Buckley, A. (2012). Make maps people want to look at. ESRI.com Special Section. <https://www.esri.com/news/arcuser/0112/files/design-principles.pdf>.
- Clarke, K. (2020). Lecture 14: Visual analytics and data exploration. <http://www.geog.ucsb.edu/~kclarke/Geography183/Lecture14.pdf>.
- Coronavirus statistiques. (2020). <https://www.coronavirus-statistiques.com/stats-globale/nombre-de-cas-coronavirus-par-region-par-departement/>.
- Field, K. (2017). Cartograms. In Wilson, J. P. (ed.) *The Geographic Information Science and Technology Body of Knowledge* (3rd Quarter 2017 Edition). <https://doi.org/10.22224/gistbok/2017.3.8>
- Flanagan, B. (2016). U.s. election 2016: Battle of the maps. <https://communityhub.esri.com/geoexchange/2016/11/1/us-election-2016-battle-of-the-maps>.

- Gao, P., Zhang, H., Wu, Z., & Wang, J. (2020). Visualising the expansion and spread of coronavirus disease 2019 by cartograms. *Environment and Planning A Economy and Space*, 52(4). <https://doi.org/10.1177/0308518X20910162>
- Gastner, M. T., & Newman, M. E. (2004). Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences*, 101(20), 7499–7504.
- Hennig, B. D. (2019). Geography: using cartograms to change our view of the world. *Geography* 104 (Part 2).
- Laurent, T., Ruiz-Gazen, A., & Thomas-Agnan, C. (2012). Geoxp: An r package for exploratory spatial data analysis. *Journal of Statistical Software*, 47(2), 1–23.
- Nguyen, T. H. A., Laurent, T., Thomas-Agnan, C., & Ruiz-Gazen, A. (2018). Analyzing the impacts of socio-economic factors on french departmental elections with coda methods. Tech. Rep. Working Paper 961, Toulouse School of Economics.
- Nusrat, S., & Kobourov, S. (2016). The state of the art in cartograms. *Computer Graphics Forum*, 35(3), 619–642.
- Organization, W. H. (2017). Suicide rate estimates, crude. <https://apps.who.int/gho/data/view.main.MHSUICIDEREgv?lang=en>.
- Scapetoad. (2008). <http://scapetoad.choros.place/>.
- Tobler, W. (2004). Thirty-five years of computer cartograms. *Annals of the Association of American Geographers*, 94(1), 58–73.
- Tufte, E. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press.
- Worldmapper. (2020). <https://worldmapper.org/>.

Kernel and Dissimilarity Methods for Exploratory Analysis in a Social Context



Jérôme Mariette, Madalina Olteanu, and Nathalie Vialaneix

Abstract While most of the statistical methods for prediction or data mining have been built for data made of independent observations of a common set of p numerical variables, many real-world applications do not fit in this framework. A more common and general situation is the case where a relevant similarity or dissimilarity can be computed between the observations, providing a summary of their relations to each other. This setting is related to the *kernel* framework that has allowed to extend most of standard statistical supervised and unsupervised methods to any type of data for which a relevant such kernel can be obtained. The present chapter aims at presenting kernel methods in general, with a specific focus on the less studied unsupervised framework. We illustrate its usefulness by describing the extension of self-organizing maps and by proposing an approach to combine kernels in an efficient way. The overall approach is illustrated on categorical time series in a social-science context and allows to illustrate how the choice of a given type of dissimilarity or group of dissimilarities can influence the output of the exploratory analysis.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-73249-3_34) contains supplementary material, which is available to authorized users.

J. Mariette · N. Vialaneix (✉)
Université de Toulouse, INRAE, UR MIAT, F-31320 Castanet-Tolosan, France
e-mail: Nathalie.Vialaneix@inrae.fr

J. Mariette
e-mail: Jerome.Mariette@inrae.fr

M. Olteanu
SAMM, Université Paris 1, F-75005 Paris, France
e-mail: madalina.olteanu@univ-paris1.fr

1 Introduction

While most of the statistical methods for prediction or data mining have been built for data made of independent observations of a common set of p numerical variables, many real-world applications do not fit in this framework. Typical such examples include categorical variables, relations between entities (e.g., a graph or network) or even more complex frameworks such as categorical time series. A particularly useful simplification of these more general situations is the case where a relevant similarity or dissimilarity can be computed between the observations, providing a summary of their relations to each other. In addition, when this similarity has some mild additional properties, it is called a *kernel* and provides a strong mathematical framework (Berlinet and Thomas-Agnan 2004) for extending most of standard statistical supervised and unsupervised methods to any type of data for which a relevant such kernel can be obtained (Cristianini and Shawe-Taylor 2000; Shawe-Taylor 2004). This approach has already proven useful in computational biology (Schölkopf et al. 2004) or in social sciences and humanities (Boulet et al. 2008; Massoni et al. 2013).

Nevertheless, the choice of a relevant kernel is still an open problem. Some authors have proposed to combine all candidate kernels into a “meta-kernel” which is an “optimal” linear or convex combination of the individual kernels. This approach is known as the “multiple kernel learning problem” and has been widely studied in the supervised framework (Gönen and Alpaydın 2011). The present chapter aims at presenting the less addressed unsupervised framework. More precisely, after a brief introduction to kernels and their relation with the more general similarity/dissimilarity settings (Sect. 2), we describe how statistical methods can be extended to the kernel framework by using the so-called “kernel trick” (Sect. 3). Section 4 focuses more precisely on the extension of an exploratory method, called Self-Organizing Maps (Kohonen 2001), to the kernel framework and discusses the issue of complexity and how it can be solved in this particular setting. Section 5 explains how kernels can be combined in an unsupervised setting, as a processing prior to the unsupervised methods presented before. The overall approach is illustrated in Sect. 6 on categorical time series in a social-science context: originally developed in bioinformatics, sequence analysis is indeed increasingly used in social sciences for the study of life-course processes. In this section, we discuss how the choice of a given type of dissimilarity or group of dissimilarities influences the output of the exploratory analysis and allows to extract relevant patterns from this particular kind of data.

2 Kernels and More General Proximity Data

2.1 Kernels and RKHS

Kernel methods consider the case where data are described by a *kernel* obtained from a Reproducing Kernel Hilbert Space (RKHS; Berlinet and Thomas-Agnan 2004). Usually, the sample of interest takes values in an arbitrary space, \mathcal{X} that encompasses a variety of data types. This sample is then described by a function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, which is symmetric ($\forall x, x' \in \mathcal{X}, K(x, x') = K(x', x)$) and positive ($\forall N \in \mathbb{N}, \forall (\alpha_i)_{i=1, \dots, N} \subset \mathbb{R}$ and $\forall (x_i)_{i=1, \dots, N} \subset \mathcal{X}, \sum_{i, i'=1}^N \alpha_i \alpha_{i'} K(x_i, x_{i'}) \geq 0$) and is called the *kernel*. Indeed, in this case, it is known (Aronszajn 1950; Berlinet and Thomas-Agnan 2004) that there exists a unique Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ and a unique application $\phi : \mathcal{X} \rightarrow \mathcal{H}$, such that

$$\forall x, x' \in \mathcal{X}, \quad \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = K(x, x').$$

$(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is the RKHS of K and is also often called *feature space*; ϕ is the *feature map* of K .

In statistics and machine learning, this framework is often used to deal with observations that are not just multidimensional vectors (e.g., categorical time series or graphs, among others) or to incorporate expert knowledge in the analysis (see Examples 1 and 2 below with standard examples of kernels often used in practice). The sample $(x_i)_{i=1, \dots, n}$ is then described by pairwise relations between observations, as measured by the kernel. This leads to the computation of the *kernel matrix* $\mathbf{K} = (k_{i i'})_{i, i'=1, \dots, n}$, with $k_{i i'} = K(x_i, x_{i'})$, which is symmetric and semi-definite positive, by definition of the kernel K .

The idea of kernel methods is to perform standard linear statistical analyses in the feature space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$. Since the only operations involved in these analyses are related to the computation of dot products and norms, the Hilbert space \mathcal{H} and the feature map ϕ are usually not explicitly given but used implicitly through the kernel K instead. This principle, which we illustrate below, is called the *kernel trick*.

Example 1 Some useful kernels

Kernels in \mathbb{R}^p . Kernel methods are often used for standard multidimensional data to provide more flexibility and non-linearity in the analyses. In these spaces, a trivial kernel is given by using the standard dot product of \mathbb{R}^p : $K(x, x') = (x')^\top x$, which leads to the trivial feature map $\phi = \text{Id}$. The feature space is then unchanged as compared to the original space ($\mathcal{X} = \mathcal{H} = \mathbb{R}^p$) and the performed statistical analysis is thus still linear. Among more interesting kernels for \mathbb{R}^p , one of the most popular is the Gaussian kernel (also called Radial Basis Function—RBF—kernel)

$K_\gamma(x, x') = e^{-\gamma\|x-x'\|^2}$, which shape is controlled by a hyper-parameter $\gamma > 0$. This kernel is of special importance since it is continuous and *universal* for every compact set of \mathcal{X}, \mathcal{C} (meaning that the set of all functions induced by $x \in \mathcal{C} \rightarrow K(x, \cdot)$ is dense in the set of all continuous functions $\mathcal{C} \subset \mathcal{X} \rightarrow \mathbb{R}$). This property allowed Steinwart (Steinwart 2001, 2002) to demonstrate the consistency of kernel classification and regression methods in the statistical sense (when the sample size grows to infinity and in terms of convergence of the error loss to its optimum). The polynomial kernel ($K(x, x') = (1 - (x')^\top x)^\gamma$ for $\gamma > 0$) and the exponential kernel ($K(x, x') = e^{(x')^\top x}$) are also universal kernels.

Kernels on graphs. In many application fields including social sciences and biology, graphs (also called networks) are widely used to represent pairwise relations between entities (friendship, professional contacts, regulation between genes, ...). A number of kernels for graphs have been proposed to provide a similarity measure between nodes based on the graph structure. Most of them are derived from regularized version of the Laplacian of the graph (Kondor and Lafferty 2002; Smola and Kondor 2003) and have been used in prediction or exploratory analyses in biology (e.g., for introducing known relations between genes (Vert and Kanehisa 2003; Rapaport et al. 2007) or in social sciences (e.g., to extract information from a medieval social network Boulet et al. 2008).

2.2 From General Similarities to Kernels

In practice, data are often described by similarities (or dissimilarities) that are not necessarily definite positive (see Example 2). This situation is addressed either by generalizing kernel methods to the “pseudo-Euclidean” framework (Goldfarb 1984; Ong et al. 2004), by embedding the sample directly into a Euclidean space whose dot product resembles the original similarity (Multidimensional Scaling—MDS—is one of these approaches Cox and Cox 2001), or by using a proper definite kernel instead of the original indefinite similarity. In the latter case, the chosen kernel is often obtained by a simple transformation of its spectrum meant to obtain only positive eigenvalues. Chen et al. (2009), Schleif and Tino (2015) are two reviews describing the topic of general similarity learning and its relation with kernel methods.

Example 2 Some more general similarities and dissimilarities

Categorical sequences or time series. Categorical sequences are naturally used in biology to represent the DNA sequences or proteins (with the categories being the amino acids). Among the many proposals for quantifying the similarities between two sequences, edit distances (also known as “Levenshtein distances” or “optimal matching dissimilarities”) (Needleman and Wunsch 1970) are one of the most famous. Their main idea is to quantify the minimum number of transformations needed to obtain a sequence from another one. A cost is associated with insertion, deletion, and substitution transformations to allow a flexible customization of these dissimilarities. These measures have been increasingly used in social sciences as well, for studying life-course processes (Abbott and Tsay 2000; Massoni et al. 2013). Section 6 describe in further details those dissimilarities.

Dissimilarities based on phylogeny. As already mentioned, kernels and dissimilarities can also embed prior expert information in their computation. A typical example is the case where variables are the abundances of different species for which a phylogeny information (a parental information between those species) is given. Such frameworks are met when studying the biodiversity of different places or in metagenomics for instance. In these applications, data are described by vectors of counts that represent the number of times given species or Operational Taxonomic Units (OTUs) have been found for a given individual. For these data, computing a measure of proximity between observations that accounts for the distances between the species has been shown to provide a more relevant information than the simple Euclidean distance between counts (Lozupone and Knight 2005; Lozupone et al. 2007). Such distances include the (weighted) UniFrac distance or the generalized UniFrac distance (Chen et al. 2012).

3 Basics of Statistical Learning with Kernels

3.1 Supervised Setting

A simple example of a supervised learning method that has been extended to kernels is the ridge regression. More precisely, when given a training sample $\{(\mathbf{x}_i, y_i)_{i=1, \dots, n}\}$ for which $\mathbf{x}_i \in \mathbb{R}^p$ and y_i is a real number, the ridge regression finds the best linear predictor for $(y_i)_i$ based on $(\mathbf{x}_i)_i$ that minimizes the squared loss plus a regularization term based on the ℓ_2 norm:

$$\beta^* = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \beta^\top \mathbf{x}_i)^2 + \lambda \|\beta\|^2 \tag{1}$$

for a given $\lambda > 0$, called regularization parameter, which is usually tuned by a cross validation approach. The solution of Eq. (1) is given by

$$\beta^* = \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top + \lambda \mathbb{I}_p \right)^{-1} \left(\sum_{i=1}^n y_i \mathbf{x}_i \right),$$

that can also be written as

$$\beta^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I}_p) \mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top (\mathbf{X} \mathbf{X}^\top + \lambda \mathbb{I}_n) \mathbf{y},$$

with $\mathbf{y} = (y_1, \dots, y_n)^\top$ and \mathbf{X} being the $(n \times p)$ -matrix with rows containing the \mathbf{x}_i so that the matrix $\mathbf{X} \mathbf{X}^\top$ is the $(n \times n)$ -matrix with entries the pairwise dot products $\mathbf{x}_i^\top \mathbf{x}_{i'}$ for all $i, i' = 1, \dots, n$. In summary, the solution writes

$$\beta^* = \sum_{i=1}^n \alpha_i^* \mathbf{x}_i \quad \text{with } \alpha^* = (\mathbf{X} \mathbf{X}^\top + \lambda \mathbb{I}_n)^{-1} \mathbf{y}. \tag{2}$$

The extension of this approach to samples $(x_i)_i$ taking values in an arbitrary space \mathcal{X} through the use of kernels is called *kernel ridge regression* (Saunders et al. 1998). The idea is simply to search for a linear predictor in the feature space induced by the kernel, \mathcal{H} , which transforms the optimization criterion of Eq. (1) into:

$$w^* = \operatorname{argmin}_{w \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (y_i - \langle w, \phi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|w\|_{\mathcal{H}}^2. \tag{3}$$

The best linear predictor in \mathcal{H} is thus given similarly as the solution of Eq. (2) but in the feature space, replacing \mathbf{x}_i by $\phi(x_i)$ and the \mathbb{R}^p dot product by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. In particular, this means that the matrix $\mathbf{X} \mathbf{X}^\top$ is replaced by a matrix with entries equal to $\langle \phi(x_i), \phi(x_{i'}) \rangle_{\mathcal{H}}$, which, by the so-called *kernel tricks*, turns out to simply be equal to \mathbf{K} . We thus have that

$$w^* = \sum_{i=1}^n \alpha_i^* \phi(x_i) \quad \text{with } \alpha^* = (\mathbf{K} + \lambda \mathbb{I}_n)^{-1} \mathbf{y}.$$

This result can also be found as a consequence of the Representer Theorem (Kimeldorf and Wahba 1970; Schölkopf et al. 2001) or directly solving the dual of Eq. (3):

$$\alpha^* = \operatorname{argmin}_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n (y_i - \alpha^\top \mathbf{K}_i)^2 + \lambda \|\alpha\|_{\mathbf{K}}^2, \tag{4}$$

in which \mathbf{K}_i is the i -th row of the kernel matrix \mathbf{K} and $\|\cdot\|_{\mathbf{K}}$ is the ℓ_2 norm induced by this matrix in \mathbb{R}^n : $\|\alpha\|_{\mathbf{K}}^2 = \alpha^\top \mathbf{K} \alpha$.

Variants of this framework include Support Vector Machines (SVM, Boser et al. 1992), for the classification case, or ϵ -SVM, for the regression case. In both cases, the main difference with the kernel ridge regression lies in the loss function but the main principle of the approach remains identical: the Representer Theorem allows to express the solution as a linear combination of the images by ϕ of the observations and the solution is obtained by solving a dual optimization problem obtained thanks to the use of the kernel trick.

3.2 Unsupervised Setting

Kernel methods have also been developed for the unsupervised setting. Among the most direct of these extensions, the generalization of PCA (Schölkopf et al. 1998) and that of k -means (Dhillon et al. 2004) are probably the most known and used. They both use approaches similar to the supervised case described in the previous section, and more precisely:

- computations related to the original method (*i.e.*, standard PCA and k -means) are performed in the feature space;
- to do so, the *kernel trick* is used instead of the standard computation of dot products or norms.

Kernel PCA.

Standard PCA is often presented as the eigendecomposition of the variance/covariance matrix associated to the $(n \times p)$ -matrix of sample measures, \mathbf{X} . Assuming without loss of generality that \mathbf{X} is centered, this eigendecomposition is equivalent to the dual eigendecomposition of $\mathbf{X}\mathbf{X}^\top$, that provides the coordinates (or scores) of the projection of \mathbf{X} on the different principal components. More precisely, if \mathbf{T} is the $(n \times k)$ column matrix with the first k eigenvectors of $\mathbf{X}\mathbf{X}^\top$, orthogonal and with a norm equal to $\frac{1}{\sqrt{\lambda_j}}$, then the $(p \times k)$ column matrix of the unit-scaled loadings (orthogonal and with a norm equal to 1) is $\mathbf{X}^\top \mathbf{T}$.

Kernel PCA uses a similar approach taking advantage of the analogy between $\mathbf{X}\mathbf{X}^\top$ and \mathbf{K} and between the i -th row of \mathbf{X} and $\phi(x_i)$. More precisely,

1. assuming that \mathbf{K} is centered in the feature space¹, the eigendecomposition of \mathbf{K} is obtained. It gives $(\lambda_j)_{j=1,\dots,k}$, the first k eigenvalues of \mathbf{K} , and $(t_j)_{j=1,\dots,k}$, the associated first k orthogonal eigenvectors with a norm equal to $\frac{1}{\sqrt{\lambda_j}}$;
2. the first k (orthogonal) unit-scaled loadings are thus obtained as

¹If \mathbf{K} is not centered, the centering operation is simply $\mathbf{K} - \frac{1}{n} \mathbf{1}_n \mathbf{K} - \frac{1}{n} \mathbf{K} \mathbf{1}_n + \frac{1}{n^2} \mathbf{1}_n \mathbf{K} \mathbf{1}_n$, in which $\mathbf{1}_n$ is an $n \times n$ matrix with all entries equal to 1.

$$w_j = \sum_{i=1}^n t_{ji} \phi(x_i),$$

and have a norm equal to 1 in \mathcal{H} . The coordinate of $\phi(x_i)$ on the j -th axis is thus $\langle w_j, \phi(x_i) \rangle_{\mathcal{H}} = \lambda_j t_{ji}$.

Kernel k -means.

Similarly, kernel k -means performs a standard k -means algorithm in the feature space \mathcal{H} . To do so, in addition to computing dot products and norms using the kernel trick, it is necessary to obtain a representation of the cluster barycenters. More precisely, if $(x_i)_{i \in C}$ are the observations assigned to a given cluster C , then, the barycenter is given by

$$\bar{x}_C = \frac{1}{|C|} \sum_{i \in C} \phi(x_i)$$

and its (squared) distance to any other observation, x_i , in the sample is obtained by

$$\begin{aligned} \|\phi(x_i) - \bar{x}_C\|_{\mathcal{H}}^2 &= \left\| \phi(x_i) - \frac{1}{|C|} \sum_{i' \in C} \phi(x_{i'}) \right\|_{\mathcal{H}}^2 \\ &= \|\phi(x_i)\|_{\mathcal{H}}^2 - \frac{2}{|C|} \sum_{i' \in C} \langle \phi(x_i), \phi(x_{i'}) \rangle_{\mathcal{H}} + \frac{1}{|C|^2} \sum_{i', i'' \in C} \langle \phi(x_{i'}), \phi(x_{i''}) \rangle_{\mathcal{H}} \\ &= k_{ii} - \frac{2}{|C|} \sum_{i' \in C} k_{ii'} + \frac{1}{|C|^2} \sum_{i', i'' \in C} k_{i'i''}. \end{aligned}$$

In both situations (kernel PCA and kernel k -means), the adaptation of the algorithms to kernel data is made by their direct rewriting in the feature space. New data points that were not previously in the feature space (principal components or barycenters) are represented by linear combinations of the images by the feature map, ϕ , of observations. In addition, distances to these new elements can be expressed in function of the kernel, using the kernel trick. These adaptations are thus very similar to the supervised case situations.

4 Kernel Self-Organizing Maps and Complexity Reduction

In this section, we present an extension of kernel k -means to a more general method, which simultaneously performs clustering and dimensionality reduction for visualization, namely, the *self-organizing map* (SOM) algorithm. Originally designed for unsupervised exploration of standard numerical datasets (Kohonen 2001), the method has been extended to handle non-numeric data by using approaches based on Multiple Correspondence Analysis (Cottrell and Letrémy 2005) or by relying on an algorithm that represents all the clusters by a prototype chosen among the

data (*median SOM*, Kohonen and Somervuo 1998). Even if very general, the latter approach is very restrictive and generates representation issues, with associated biases in the obtained maps. In the present section, we present the extension of SOM to kernels that has been introduced by several authors for batch and online versions (Mac Donald and Fyfe 2000; Boulet et al. 2008) and has also been generalized to data represented by general dissimilarities (rather than kernels) (Hammer and Hasenfuss 2010). The second part of this section will discuss associated complexity issues when the sample size is large and review the different strategies that can be implemented to overcome them.

4.1 Kernel Self-Organizing Maps

For the standard case of a dataset $(\mathbf{x}_i)_{i=1,\dots,n}$ of multidimensional observations $\mathbf{x}_i \in \mathbb{R}^p$, SOM algorithm is close to k -means algorithm, except that the clusters are organized on a map equipped with a distance, d . More precisely, a map (also sometimes called a grid) is a set of U clusters (also sometimes called units or neurons) associated to physical locations in a low dimensional space. The clusters are frequently positioned in \mathbb{R}^2 at coordinates $(a, b)_{a=1,\dots,A, b=1,\dots,B}$ with $AB = U$. Clusters are related to each other using pairwise distances, that can be, for instance, the Euclidean distances between their coordinates in \mathbb{R}^2 . In addition, every cluster, u , is summarized by a *prototype*, p_u that takes its values in the input space \mathbb{R}^p .

When fixing the number of neurons U , one should take into account the fact that SOM is more intended as a method for non-linear mapping and dimensionality reduction—in the sense of vector quantization—than as a method for clustering the data into a small number of clusters. Some authors (Ultsch and Siemon 1990) suggest to build very large SOMs, with a number of units U larger than the sample size, n . In this context, SOM essentially reduces to non-linear mapping and to mining the underlying distribution of the data. A second option, which is more commonly used in practice, consists in building medium size maps that are smaller than the sample size, but still large enough to have a few input observations representing each unit (Kohonen 2001). This strategy is a good trade-off between mapping and clustering, and a heuristic suggests to set U close to $\sqrt{n/10}$ (Villa-Vialaneix 2017).

The method aims at assigning every observation in the dataset to one of the clusters, while minimizing the distortion of the topology between the original space (here, \mathbb{R}^p) and the map (as seen through the distance d). The prototypes are thus expected to be representative of the observations assigned to their cluster, as the barycenter is representative of its cluster in kernel k -means. To do so, the stochastic version of the method iterates over two steps:

- an *assignment step* in which an observation, \mathbf{x}_i is randomly chosen and assigned to the unit with the closest prototype:

$$f(\mathbf{x}_i) = \operatorname{argmin}_{u=1,\dots,U} \|\mathbf{x}_i - p_u\|^2$$

where $f(\mathbf{x}_i)$ is the cluster to which observation \mathbf{x}_i is assigned;

- a *representation step* in which all prototypes are updated according to the new assignment:

$$\forall u = 1, \dots, U, \quad p_u \leftarrow p_u + \mu H(d(f(\mathbf{x}_i), u))(\mathbf{x}_i - p_u)$$

where $\mu > 0$ is chosen so as to vanish during the training process and H is a decreasing function, generally chosen such that $H(0) = 1$ and $\lim_{z \rightarrow +\infty} H(z) = 0$.

The method is usually initialized with a random choice of the prototypes in \mathbb{R}^p . Different heuristics can be used to choose T : either it is defined proportionally to the sample size (typically, $T = 5n$) or it is not fixed in advance and the algorithm stops when the iterations no longer modify the solution.

The extension of SOM to kernel data is based on the same key tools as the ones described for kernel PCA and kernel k -means, and which allow to re-write the algorithm in the feature space \mathcal{H} :

- the prototypes are expressed as convex combinations of the images by ϕ of the observations, and the assignment step is written in terms of coefficients related to each image $\phi(x_i)$;
- the representation step is expressed with \mathbf{K} by means of the kernel trick.

The full version of the method is provided in Algorithm 1.

Algorithm 1 Stochastic kernel SOM

- 1: $\forall u = 1, \dots, U$ and $\forall i = 1, \dots, n$, random initialization of the prototypes: $p_u^1 = \sum_{i=1}^n \beta_{ui}^1 \phi(x_i)$ with $\beta_{ui}^1 \in [0, 1]$ and $\sum_i \beta_{ui}^1 = 1$
- 2: **for** $t = 1$ to T **do**
- 3: Select randomly one observation $i \in \{1, \dots, n\}$ ▷ Assignment step

$$\begin{aligned} f^{t+1}(x_i) &= \operatorname{argmin}_{u=1, \dots, U} \|\phi(x_i) - p_u^t\|_{\mathcal{H}}^2 \\ &= \operatorname{argmin}_{u=1, \dots, U} \left(k_{ii} - 2 \sum_{l=1}^n \beta_{ul}^t k_{il} + \sum_{l, l'=1}^n \beta_{ul}^t \beta_{ul'}^t k_{ll'} \right) \end{aligned}$$

- 4: For all $u = 1, \dots, U$, ▷ Representation step

$$\begin{aligned} p_u^{t+1} &= p_u^t + \mu_t H^t(d(f^{t+1}(x_i), u))(\phi(x_i) - p_u^t) \\ \Leftrightarrow \beta_u^{t+1} &= \beta_u^t + \mu_t H^t(d(f^{t+1}(x_i), u)) (\mathbf{1}_i^n - \beta_u^t), \end{aligned}$$

where $\mathbf{1}_i^n$ is a vector of length n with all entries equal to 0 except for the i th, which is equal to 1.

- 5: **end for**
 - 6: **return** $(p_u^{T+1})_u$ (prototypes) and $(f^{T+1}(x_i))_i$ (clustering)
-

4.2 Complexity of Kernel SOM

Kernel methods are generally considered efficient to deal with large dimensional data (when the original space is a standard multidimensional space, \mathbb{R}^p , with p large) but often encounter scalability issues when the sample size, n , becomes large. As noted by Rossi (2014), the complexity of kernel SOM is $\mathcal{O}(n^2UT)$ whereas the complexity of the numeric SOM in \mathbb{R}^p is $\mathcal{O}(npUT)$. When $p \ll n$ and n is large, this cost can be very prohibitive since, typically, T is of order $\mathcal{O}(n)$. Different strategies have been developed to overcome this difficulty, among which approximations with dimensionality reduction or sparse representations (Hofmann et al. 2015; Mariette et al. 2017a) or exact approaches using a storage of intermediate results (Mariette et al. 2017b). Other alternatives to reduce the complexity of kernel methods directly use accelerated computations of the kernel dot products (used in most kernel methods) with tiled reduction schemes on GPU, without even storing the kernel itself (see KeOps, <https://www.kernel-operations.io/keops/index.html>). They would be a practicable approach to accelerate the assignment step of kernel SOM when the full kernel matrix itself does not fit in memory but we will restrict to the simpler case where the kernel is already computed and stored, in the remaining of this section.

Low rank and sparse approximations.

The first type of approaches relies on a simpler representation of the prototypes, with a reduced number of (non-zero) coefficients. Mariette et al. (2017a) proposes two types of solutions. The first one is very similar to the strategies developed in Hofmann et al. (2015) and uses a direct sparse approach in which an additional step is added to each iterations, aiming at thresholding the smallest coefficients $(\beta_{ui})_i$ for every prototype p_u . The second method relies on a prior step (a kernel PCA) to provide inputs that are the coordinates of the original observations on the first k principal components of the kernel PCA. The SOM algorithm then used is a simple numeric SOM with a complexity of $\mathcal{O}(nkTU)$ with $k \ll n$. However, this prior step has a high computational cost itself: the full eigendecomposition of \mathbf{K} has a computational cost of $\mathcal{O}(n^3)$ but it can be reduced with the Nyström approximation (Williams and Seeger 2000). This method allows to obtain an approximation of the eigendecomposition of \mathbf{K} using an eigendecomposition of a submatrix $\mathbf{K}^{(m)}$ based on m observations chosen at random in the original sample. The eigendecomposition approximation is even exact if the rank of \mathbf{K} is smaller than m . The complexity of the approach is reduced to $\mathcal{O}(nm^2)$ where m is usually chosen $\ll n$.

Exact approaches.

Most of the complexity of the kernel SOM comes from the assignment step, which is $\mathcal{O}(n^2U)$. Re-formulating this step and transforming it into the update of stored results, we reduced it to $\mathcal{O}(U)$. More precisely, the assignment step writes

$$f^{t+1}(x_i) = \operatorname{argmin}_{u=1,\dots,U} A_u^t - 2B_{ui}^t$$

with $A_u^t = \sum_{j,j'=1}^n \beta_{uj}^t \beta_{uj'}^t k_{jj'}$ and $B_{ui}^t = \sum_{j=1}^n \beta_{uj}^t k_{ij}$. Storing these quantities in memory (a vector of size U and a $(U \times n)$ -matrix), the representation step reduces to an update of A^t and B^t :

$$\begin{aligned} A_u^{t+1} &= (1 - \lambda_u(t))^2 A_u^t + \lambda_u(t)^2 k_{ii} + 2\lambda_u(t)(1 - \lambda_u(t)) B_{ui}^t \\ B_{ui'}^{t+1} &= (1 - \lambda_u(t)) B_{ui'}^t + \lambda_u(t) k_{i'i}, \end{aligned}$$

with $\lambda_u(t) = \mu_t H^t(d(f^{t+1}(x_i), u))$. The representation step thus has a complexity of $\mathcal{O}(nU)$ (update of B^t) and the total complexity of the approach is reduced to $\mathcal{O}(nUT)$. This reduction of the computational time is thus obtained at the cost of storing operations with a memory cost of $\mathcal{O}(U)$ and $\mathcal{O}(nU)$ for A^t and B^t , respectively.

5 Combining Kernels

Kernel methods have proven to be particularly efficient when data are described by multi-source and multi-type information obtained on the same n observations. In this case, each source of data, of a given particular type (numerical, graph data, factors,...), can be passed through a kernel, K^m ($m = 1, \dots, M$): this kernel provides the similarity information between observations, seen from the point of view of the source m . The advantage of such an approach is that it provides a common representation of the different sources that can be easily combined. A similar framework is the one where multiple kernels can be obtained from a single dataset, each capturing a specific feature. Combining these kernels avoids having to choose between them, and also benefits of the information coming from different aspects of the data. Among the combination approaches (Gönen and Alpaydin 2011), one that has been widely developed is the computation of a convex combination of the M kernels into a single meta-kernel:

$$\mathbf{K}^\gamma = \sum_{m=1}^M \gamma_m \mathbf{K}^m, \quad \text{st} \begin{cases} \gamma_m \geq 0, \forall m = 1, \dots, M \\ \sum_{m=1}^M \gamma_m = 1 \end{cases}.$$

In the context of supervised methods, the choice of $(\gamma_m)_m$ is usually done by solving a global optimization problem that aims at minimizing a prediction loss, with respect to the parameter of a given method (SVM for instance) and to the value of $(\gamma_m)_m$ (Zhao et al. 2009; Yu et al. 2012; Huang et al. 2012; Gönen and Margolin 2014). In the unsupervised setting, choosing relevant $(\gamma_m)_m$ is harder since the objective function might not be as easily designed or because, as it is the case for kernel PCA, its joint optimization to estimate the principal components and the $(\gamma_m)_m$ is degenerate (Speicher and Pfeifer 2017).

Several propositions have thus been made (Lin et al. 2010; Zhuang et al. 2011; Speicher and Pfeifer 2015; Wang et al. 2017; Mariette et al. 2018) to overcome

this issue and, in the latter, we proposed two solutions that can cope with non-numerical observations, contrary to the others. The first method, named **STATIS-UMKL** (where “UMKL” stands for Unsupervised Multiple Kernel Learning), is based on the STATIS method (L’Hermier des Plantes 1976; Lavit et al. 1994) and aims at searching for a consensual meta-kernel. More precisely, the method searches for the kernel that is the most similar, on average, to all the kernels to be combined, $(K^m)_{m=1,\dots,M}$:

$$\max_{\mathbf{v}} \sum_{m=1}^M \left\langle \mathbf{K}^{\mathbf{v}}, \frac{\mathbf{K}^m}{\|\mathbf{K}^m\|_F} \right\rangle_F \quad \text{for } \mathbf{K}^{\mathbf{v}} = \sum_{m=1}^M v_m \mathbf{K}^m,$$

$$\text{and } \mathbf{v} \in \mathbb{R}^M \text{ such that } \|\mathbf{v}\|_{\mathbb{R}^M}^2 = 1,$$

where $\langle \cdot, \cdot \rangle_F$ and $\|\cdot\|_F$ stand for the Frobenius dot product and norm. It is easy to show that the solution is given by the spectral decomposition of a $M \times M$ -matrix, \mathbf{C} , such that $C_{mm'} = \frac{\langle \mathbf{K}^m, \mathbf{K}^{m'} \rangle_F}{\|\mathbf{K}^m\|_F \|\mathbf{K}^{m'}\|_F}$ and γ is thus chosen as $\frac{\mathbf{v}}{\sum_m v_m}$.

The second method first creates a proxy of the local geometry induced by each kernel K^m using a k nearest neighbor graph and the global adjacency matrix of these M graphs, \mathbf{W} is then used in a global criterion. This criterion is designed to preserve at best the local geometry measured by \mathbf{W} in the feature space induced by the meta-kernel \mathbf{K}^γ :

$$\text{argmin}_{\gamma \in \mathbb{R}^M} \sum_{i,i'=1}^n W_{ii'} \|C_i(\gamma) - C_{i'}(\gamma)\|_{\mathbb{R}^n}^2,$$

$$\text{st } \gamma_m \geq 0 \text{ and } \sum_{m=1}^M \gamma_m = 1,$$

with

$$C_i(\gamma) = \left\langle \phi^\gamma(x_i), \begin{pmatrix} \phi^\gamma(x_1) \\ \vdots \\ \phi^\gamma(x_n) \end{pmatrix} \right\rangle_{\mathcal{H}^\gamma} = \begin{pmatrix} K^\gamma(x_i, x_1) \\ \vdots \\ K^\gamma(x_i, x_n) \end{pmatrix}.$$

This problem has a sparse solution that performs a selection of the kernels (some of the entries of $(\gamma_m)_m$ are forced toward 0) because of the convexity constraint $\sum_{m=1}^M \gamma_m = 1$ but this can be relaxed by replacing the ℓ_1 constraint with a constraint on the ℓ_2 norm instead (the two versions are called **sparse-UMKL** and **full-UMKL**).

Once the kernel is obtained, it can be used as input to kernel-based algorithms, like kernel PCA, kernel k -means, or kernel SOM for exploratory purpose.

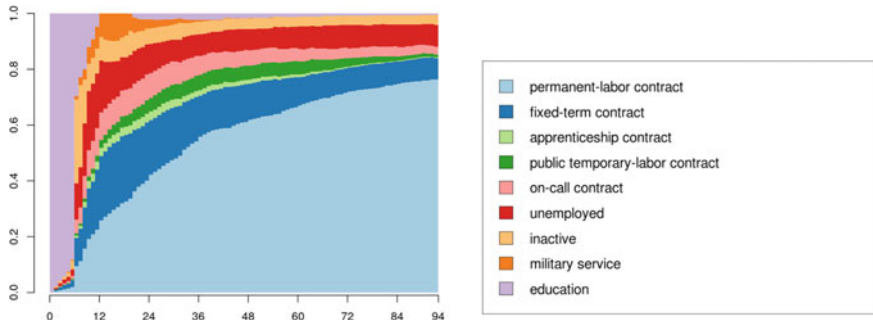


Fig. 1 Chronogram of the labor market structure as illustrated by the “Génération 98” dataset. *x*-axis is time (in month) and *y*-axis is the proportion of each type of contract

6 Application

The combined-kernel SOM algorithm is illustrated on data related to school-to-work transitions, extracted from the survey “Generation 98”². The dataset contains information on 16,040 young people having graduated in 1998 and monitored during 94 months after having left school. The labor-market status has nine categories, labeled as follows: permanent contract, fixed-term contract, apprenticeship, public temporary contract, on-call contract, unemployed, inactive, military service, education. The following stylized facts are highlighted by a first descriptive analysis of the data, as shown in Fig. 1³:

- permanent contracts represent more than 20% of all status after 1 year and their ratio continues to increase up to 50% after 3 years and almost 75% after 7 years;
- the ratio of fixed-term contracts is more than 20% after 1 year on the labor market, but it is decreasing to 15% after 3 years and then seems to converge to 8%;
- almost 30% of the young graduates are unemployed after 1 year. This ratio is decreasing and becomes constant, 10%, after the fourth year.

Some trajectories were duplicated (some people had exactly the same job trajectories) so, to reduce redundancy and computational time, we used only the 12,471 unique trajectories.

Three optimal combined kernels were computed from three sets of dissimilarities with different features, and each of these kernels was then used as input to the kernel SOM. The three sets of dissimilarities are described with more details in the next section. For each set of dissimilarities, the optimal combined kernel was obtained as follows:

² Available thanks to Génération 1998 à 7 ans - 2005, [producer] CEREQ, [diffusion] Centre Maurice Halbwachs (CMH).

³ The graphical illustrations were carried out using the `TraMineR` package (Gabadinho et al. 2011).

- first, each dissimilarity matrix, \mathbf{D} , was transformed into a (centered) similarity matrix by computing

$$\forall i, j = 1, \dots, n, s(x_i, x_j) = -\frac{1}{2} \left(\mathbf{D}_{ij}^2 - \frac{1}{n} \sum_{l=1}^n \mathbf{D}_{il}^2 - \frac{1}{n} \sum_{l=1}^n \mathbf{D}_{jl}^2 + \frac{1}{n^2} \sum_{l,l'=1}^n \mathbf{D}_{ll'}^2 \right).$$

Each of these resulting similarity matrices was used as a kernel, even though a small part of their spectra were non-positive;

- second, the resulting kernel matrices were optimally combined using STATIS-UMKL, as described in Sect. 5.

All dissimilarities were computed using the R package **TraMineR** (Gabadinho et al. 2011) and the combination of kernels was obtained using the R package **mixKernel**. Each combined kernel was processed through a kernel self-organizing map using the implementation provided in the R package **SOMbrero**. For each map to be trained, a 10×10 configuration was selected and default values of the package were chosen for the initialization step, the topology of the map (choice of H^t in Algorithm 1) and the decreasing value μ_t (also as in Algorithm 1). We decided to use 10×10 maps as a trade-off between having a meaningful visualization and a reduced number of meaningful typical trajectories. Since many trajectories are redundant—the permanent contracts are overrepresented—a map with a number of units equal to about a tenth of the number of inputs was a reasonable choice. Final results were represented as chronograms: more precisely, each unit of the map was featured by a rectangle containing the chronogram of the subsample of trajectories assigned to this unit.

6.1 Three Sets of Dissimilarities and Relations Between Them

There is currently a vast literature devoted to measuring similarities for longitudinal data, and the community agrees that the differences between the various criteria focus on three different aspects of sociological importance: the sequencing or the order in which the states appear, the timing, and the duration of the states. A recent and detailed review of these methods, focusing on these different aspects, and also introducing some new and versatile criteria, is available in Studer and Ritschard (2016). Starting from these considerations, three sets of distances were selected for the present study: the first aimed at focusing on the sequencing and possibly the duration, the second on the timing, and the third on the duration only.

The first group of dissimilarities contains one criterion based on the number of matching subsequences, and two based on generalizations of the classical OM metric (Needleman and Wunsch 1970; Abbott and Forrest 1986):

- The **SVRspell** distance, proposed by Elzinga and Studer (2015), is computed using the number of matching subsequences within the distinct sequences of states,

where the durations of the spells are weighted by some parameter b . This method is built to be sensitive to sequencing, and, depending on b , it may be also made sensitive to durations. In the following, we set $b = 0$, so that the sensitivity to the order only is put forward.

- The **OMspell** distance, introduced in Studer and Ritschard (2016), generalizes the OM distance to sequences of spells. The increasing size of the alphabet and of the costs to be specified are controlled through a linear function depending on a parameter δ , representing the cost of extending or compressing a spell by one unit of time. For small values of δ , the method favors the expansion or the compression of existing spells. For $\delta = 0$, **OMspell** reduces to the usual OM distance between the distinct sequences of states. Let us remark here that the smaller δ , the less sensitive the criterion is to the duration of the spells and the more it is to the sequencing.
- The **OMstran** distance, also introduced in Studer and Ritschard (2016), adapts the OM distance to sequences of transitions. Here also, the alphabet and the number of costs to be specified are much larger than in the usual setting, but the dimension of the parameters is reduced by considering a convex combination between the costs of the spells and the transition costs, controlled by some w . In the following, the value of w was fixed so as to favor a criterion sensitive to differences in sequencing.

The second set of dissimilarities was focused on highlighting timing. Four distances were tested: the Hamming distance (**HAM**), based on the number of non-matching states, the Euclidean distance (**EUCLID**), which, in this case, is the squared root of the Hamming, the χ^2 (**CHI2**), which is similar to the two previous dissimilarities except that it gives more weights to the infrequent states, and the Dynamic Hamming distance (**DHD**). For both **EUCLID** and **CHI2** distances, the sensitivity between duration and timing is controlled through a parameter, L . When $L = 94$ (*i.e.*, in our case the trajectory length), scores are similar to those of the Hamming family regarding timing, when $L = 1$, dissimilarity measures are more sensitive to duration. **DHD**, proposed by Lesnard (2010), generalizes the Hamming distance by considering an OM dissimilarity without insertions or deletions, and with the substitution costs defined at each temporal instant from the corresponding transition matrices. While taking the position in the sequence and time into account, **DHD** has often been criticized for its risk of over-parameterization. In the following, distance parameters used are specified between parentheses.

The third group of dissimilarities was defined to be sensitive to duration of states. Again, four distances were selected: Euclidean and χ^2 distances between state distributions in the whole trajectories, **OMspell** with $\delta = 1$, which is sensitive both to sequencing and duration, and a distance based on the length of the longest common subsequence of two trajectories, **LCS**, as described in Bergroth et al. (2000). Whereas the Euclidean distance is more sensitive to differences between states with a high duration, the χ^2 gives more importance to rare states.

Figure S1 of Supplementary material illustrates the relations between the different distances on a cosine matrix (computed from the Frobenius dot product) for the school-to-work transition dataset. Within the first group of distances, **SVRspell** with

$b = 0$ appears to behave very specifically, and is even very different from the other distances of its own group, except for the **OMspell** with $\delta = 0$. This is explained by the fact that **SVRspell** is sensitive to the sequencing only, and not at all to timing or duration. Furthermore, since it is based on the number of common subsequences only, its principle and computation are quite different from its OM-based counterparts.

The only distance highly correlated to **SVRspell** is the **OMspell** with $\delta = 0$, which is also sensitive to sequencing only. The **OMspell** with $\delta = 0$ is also very different from the distances in its own group, except for **OMspell** with $\delta = 0.1$ and **OMstran**. These two distances introduce some sensitivity with respect to the duration of the spells, while still mainly favoring sequencing. These results are consistent with the conclusions in Studer and Ritschard (2016), based on simulated data.

In the second group, the four distances are all very similar, and more particularly **HAM**, **Euclid** and **DHD**. The χ^2 -distance stands out because of its particular weighting. We can also note that all four distances in Group 2 are also very similar to **OMstran** and **OMspell** with $\delta = 0.1$, which favor sequencing and duration, and very different from **SVRspell** and **OMspell** with $\delta = 0$, which favor sequencing only.

Within the third group of distances, all distances are also very similar, even if the value of the cosine is a bit lower than within distances in the second group. The distances in the third group are also similar to the ones in the second group, to **OMspell** with $\delta = 0.1$ and to **OMstran** in the first group.

In conclusion, **SVRspell** and **OMspell** with $\delta = 0$ are very different from all other distances, which are globally similar. This indicates that, for this dataset, criteria favoring sequencing are quite opposite to those favoring timing or duration.

6.2 Results of the Clusterings

The map obtained with the first four distances is provided in Fig. 2. For the sake of conciseness, the other two maps are available in Figures S2 and S3 and in Section S2 of the Supplementary material.

As one may easily see, most of the clusters show smooth sigmoidal transitions between states, and in some cases “sandwich”-like representations. These patterns are inherent to the fact that the first set of distances was built to point out similarities in terms of sequencing and not in terms of timing or duration. A chronogram representation is not a well-suited representation in this case because it allows for temporal shifts. Despite that, some clusters of particular interest can be identified, such as the lower left corner of the map, showing the outcomes of public fixed-term contracts.

The final convex combination for the first group of distances was:

$$0.27 \times \mathbf{OMstran} + 0.26 \times \mathbf{OMspell}(\delta = 0) + 0.28 \times \mathbf{OMspell}(\delta = 0.1) + 0.19 \times \mathbf{SVRspell}.$$

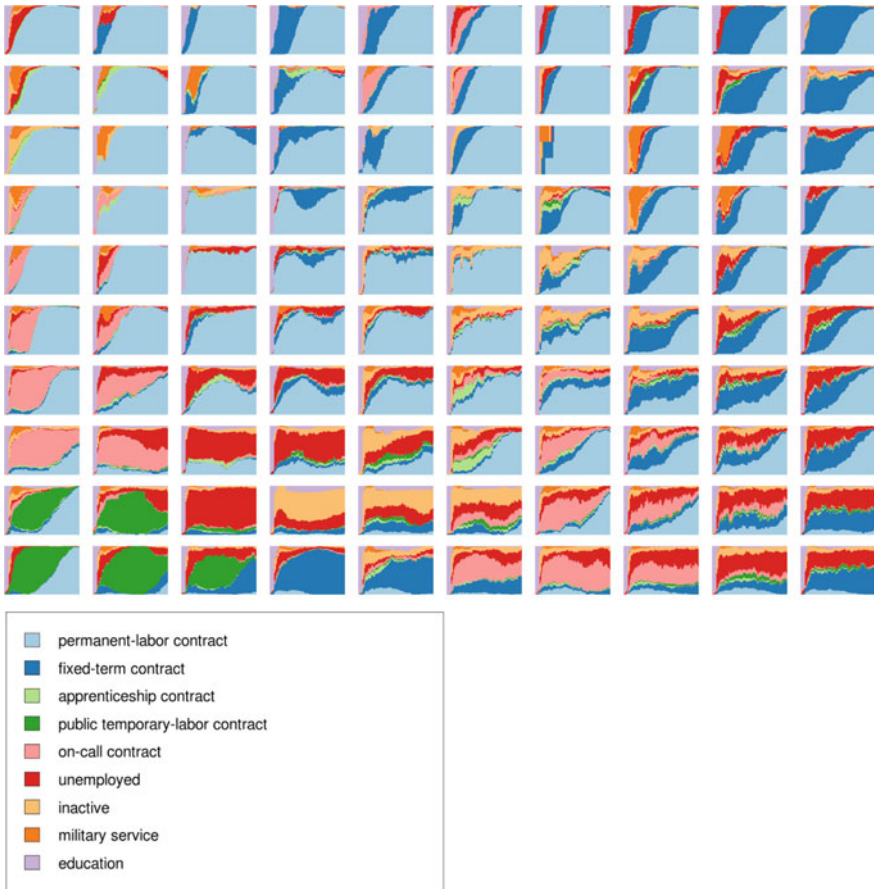


Fig. 2 Final map obtained with the first group of distances

It gives more weight to the OM-inspired distances, which appear to be more discriminant than the **SVRspell**. As it has already been stressed in the comparison between distances (Figure S1 of the Supplementary material), the **SVRspell** distance behaves very differently from all the other distances and, according to Studer and Ritschard (2016), it is very sensitive to sequencing and small random perturbations. The latter may be a reason for which the other dissimilarities appear to be fitter for clustering trajectories based on their sequencing properties.

6.3 Concluding Remarks

This chapter has presented a global approach to perform exploratory analysis in the presence of multiple sources of data or of multiple kernels describing different features of the data. The approach has been illustrated on a dataset of categorical time series representing labor-market status of recently graduated people. We have shown the effectiveness of the approach to identify a relevant typography of the dataset, with contrasting results depending on which features the focus is put on. Different dissimilarities led to highlight different characteristics of the trajectories, some less suited to chronogram representations than others. To fully exploit that diversity, alternative representations would be needed, which could be able to better represent similar duration of states or similar global distributions of the trajectories and thus to highlight the distance features.

Acknowledgements Nathalie Vialaneix would like to sincerely thank Professor Christine Thomas-Agnan for so many interesting discussions and for making her participate to the organization of “Journées de Statistique de la SFdS 2013” (an unforgettable adventure!). From a scientific point of view, Christine’s book (written with Alain Berlinet) has been her main starting point for her work on kernel methods and use of spline-based regularizations in functional data analysis. For all this, I am more than grateful!

References

- Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 16, 471–494.
- Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology. Review and Prospect. *Sociological Methods and Research*, 29(1), 3–33.
- Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3), 337–404.
- Bergroth, L., Hakonen, H., & Raita, T. (2000). A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000* (pp. 39–48). <https://doi.org/10.1109/SPIRE.2000.878178>.
- Berlinet, A., & Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Boston, Norwell, MA, USA / Dordrecht, The Netherlands: Kluwer Academic Publisher.
- Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *5th annual ACM Workshop on COLT* (pp. 144–152). ACM Press.
- Boulet, R., Jouve, B., Rossi, F., & Villa, N. (2008). Batch kernel SOM and related Laplacian methods for social network analysis. *Neurocomputing*, 71(7–9), 1257–1273. <https://doi.org/10.1016/j.neucom.2007.12.026>.
- Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., et al. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16), 2106–2113. <https://doi.org/10.1093/bioinformatics/bts342>.
- Chen, Y., Garcia, E., Gupta, M., Rahimi, A., & Cazzanti, L. (2009). Similarity-based classification: concepts and algorithm. *Journal of Machine Learning Research*, 10, 747–776.
- Cottrell, M., & Letrémy, P. (2005). How to use the Kohonen algorithm to simultaneously analyse individuals in a survey. *Neurocomputing*, 63, 193–207.

- Cox, T., & Cox, M. (2001). *Multidimensional Scaling*. Boca Raton, Florida, USA: Chapman and Hall/CRC.
- Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press.
- Dhillon, I. S., Guan, Y., & Kulis, B. (2004). Kernel k-means, spectral clustering and normalized cuts. In W. Kim, R. Kohavi, J. Gehrke, & W. DuMouchel (Eds.), *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD 2004)* (pp. 551–556). New York, NY, USA, Seattle, WA, USA: ACM. <https://doi.org/10.1145/1014052.1014118>.
- Elzinga, C. H., & Studer, M. (2015). Spell sequences, state proximities, and distance metrics. *Sociological Methods & Research*, *44*(1), 3–47.
- Gabardin, A., Ritschard, G., Müller, N., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, *40*(4).
- Goldfarb, L. (1984). A unified approach to pattern recognition. *Pattern Recognition*, *17*(5), 575–582. [https://doi.org/10.1016/0031-3203\(84\)90056-6](https://doi.org/10.1016/0031-3203(84)90056-6).
- Gönen, M., & Alpaydin, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, *12*, 2211–2268.
- Gönen, M., & Margolin, A. A. (2014). Localized data fusion for kernel k-means clustering with application to cancer biology. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Proceedings of Advances in Neural Information Processing Systems 27 (NIPS 2014)* (Vol. 27, pp. 1305–1313). Curran Associates, Inc.
- Hammer, B., & Hasenfuss, A. (2010). Topographic mapping of large dissimilarity data sets. *Neural Computation*, *22*(9), 2229–2284.
- Hofmann, D., Gisbrecht, A., & Hammer, B. (2015). Efficient approximations of robust soft learning vector quantization for non-vectorial data. *Neurocomputing*, *147*, 96–106. <https://doi.org/10.1016/j.neucom.2013.11.044>
- Huang, H. C., Chuang, Y. Y., & Chen, C. S. (2012). Multiple kernel fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, *20*(1), 120–134. <https://doi.org/10.1109/TFUZZ.2011.2170175>.
- Kimeldorf, G. S., & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, *41*(2), 495–502. <https://doi.org/10.1214/aoms/1177697089>.
- Kohonen, T. (2001). *Self-Organizing Maps* (3rd ed., Vol. 30). Berlin, Heidelberg, New York: Springer.
- Kohonen, T., & Somervuo, P. (1998). Self-organizing maps of symbol strings. *Neurocomputing*, *21*, 19–30.
- Kondor, R., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete structures. In C. Sammut & A. Hoffmann (Eds.), *Proceedings of the 19th International Conference on Machine Learning* (pp. 315–322). San Francisco, CA, USA, Sydney, Australia: Morgan Kaufmann Publishers Inc. 10.1.1.57.7612.
- Lavit, C., Escoufier, Y., Sabatier, R., & Traissac, P. (1994). The ACT (STATIS method). *Computational Statistics and Data Analysis*, *18*(1), 97–119. [https://doi.org/10.1016/0167-9473\(94\)90134-1](https://doi.org/10.1016/0167-9473(94)90134-1).
- Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research*, *38*(3), 389–419.
- L'Hermier des Plantes, H. (1976). Structuration des tableaux à trois indices de la statistique. Ph.D. thesis, Université de Montpellier. Thèse de troisième cycle
- Lin, Y., Liu, T., & CS, F. (2010). Multiple kernel learning for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*, 1147–1160.
- Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, *71*(12), 8228–8235. <https://doi.org/10.1128/AEM.71.12.8228-8235.2005>.
- Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative β eiversity measures lead to different insights into factors that structure microbial communi-

- ties. *Applied and Environmental Microbiology*, 73(5), 1576–1585. <https://doi.org/10.1128/AEM.01996-06>.
- Mac Donald, D., & Fyfe, C. (2000). The kernel self organising map. In *Proceedings of 4th International Conference on knowledge-based Intelligence Engineering Systems and Applied Technologies* (pp. 317–320).
- Mariette, J., Olteanu, M., & Villa-Vialaneix, N. (2017a). Efficient interpretable variants of online SOM for large dissimilarity data. *Neurocomputing*, 225, 31–48. <https://doi.org/10.1016/j.neucom.2016.11.014>.
- Mariette, J., Rossi, F., Olteanu, M., & Villa-Vialaneix, N. (2017b). Accelerating stochastic kernel som. In M. Verleysen (Ed.), *XXVth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2017)* (pp. 269–274). Bruges, Belgium: i6doc.
- Mariette, J., & Villa-Vialaneix, N. (2018). Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6), 1009–1015. <https://doi.org/10.1093/bioinformatics/btx682>.
- Massoni, S., Olteanu, M., & Villa-Vialaneix, N. (2013). Which distance use when extracting typologies in sequence analysis? An application to school to work transitions. In *International Work Conference on Artificial Neural Networks (IWANN 2013)*. Puerto de la Cruz, Tenerife.
- Needleman, S., & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453.
- Ong, C. S., Mary, X., Canu, S., & Smola, A. J. (2004). Learning with non-positive kernels. In C. Brodley (Ed.), *Proceedings of the XX1st International Conference on Machine Learning (ICML 2004)* (p. 81). New York, NY, USA, Banff, AB, Canada: ACM. <https://doi.org/10.1145/1015330.1015443>.
- Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., & Vert, J. (2007). Classification of microarray data using gene networks. *BMC Bioinformatics*, 8, 35. <https://doi.org/10.1186/1471-2105-8-35>.
- Rossi, F. (2014). How many dissimilarity/kernel self organizing map variants do we need? In T. Villmann, F. Schleif, M. Kaden, & M. Lange (Eds.), *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)* (Vol. 295, pp. 3–23). Advances in Intelligent Systems and Computing. Berlin, Heidelberg, Mittweida, Germany: Springer. https://doi.org/10.1007/978-3-319-07695-9_1.
- Saunders, G., Gammernan, A., & Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML'98)* (pp. 515–521). Madison, Wisconsin, USA.
- Schleif, F. M., & Tino, P. (2015). Indefinite proximity learning: a review. *Neural Computation*, 27(10), 2039–2096. https://doi.org/10.1162/neco_a_00770.
- Schölkopf, B., Herbrich, R., & Smola, A. (2001). A generalized representer theorem. In D. Heimbald & B. Williamson (Eds.), *Proceedings of the 14th Conference on Computational Learning Theory (COLT)* (Vol. 2111, pp. 416–426). Lecture Notes in Computer Science. Berlin Heidelberg: Springer. https://doi.org/10.1007/3-540-44581-1_27.
- Schölkopf, B., Smola, A., & Müller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5), 1299–1319. <https://doi.org/10.1162/089976698300017467>.
- Schölkopf, B., Tsuda, K., & Vert, J. (2004). *Kernel Methods in Computational Biology*. London, UK: MIT Press.
- Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge, UK: Cambridge University Press.
- Smola, A., & Kondor, R. (2003). Kernels and regularization on graphs. In M. Warmuth & B. Schölkopf (Eds.), *Proceedings of the Conference on Learning Theory (COLT) and Kernel Workshop* (pp. 144–158). Lecture Notes in Computer Science. Berlin Heidelberg, Washington, DC, USA: Springer. https://doi.org/10.1007/978-3-540-45167-9_12.

- Speicher, N. K., & Pfeifer, N. (2015). Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, *31*(12), i268–i275. <https://doi.org/10.1093/bioinformatics/btv244>.
- Speicher, N. K., & Pfeifer, N. (2017). Towards multiple kernel principal component analysis for integrative analysis of tumor samples. *Journal of Integrative Bioinformatics*, *14*(2), 20170019. <https://doi.org/10.1515/jib-2017-0019>.
- Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, *2*, 67–93.
- Steinwart, I. (2002). Support vector machines are universally consistent. *Journal of Complexity*, *18*, 768–791.
- Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *179*(2), 481–511. <https://doi.org/10.1111/rssa.12125>.
- Ultsch, A., & Siemon, H. P. (1990). Kohonen's self organizing feature maps for exploratory data analysis. In *Proceedings of International Neural Network Conference (INNC'90)* (pp. 305–308). Dordrecht, The Netherlands: Kluwer Academic Press.
- Vert, J., & Kanehisa, M. (2003). Extracting active pathways from gene expression data. *Bioinformatics*, *19*(Suppl. 2), ii238–ii244. <https://doi.org/10.1093/bioinformatics/btg1084>.
- Villa-Vialaneix, N. (2017). Stochastic self-organizing map variants with the R package SOMbrero. In J. Lamirel, M. Cottrell, & M. Olteanu (Eds.), *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (Proceedings of WSOM 2017)*. Nancy, France: IEEE. <https://doi.org/10.1109/WSOM.2017.8020014>.
- Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., & Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods*, *14*, 414–416. <https://doi.org/10.1038/nmeth.4207>.
- Williams, C., & Seeger, M. (2000). Using the Nyström method to speed up kernel machines. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems (Proceedings of NIPS 2000)* (Vol. 13). Denver, CO, USA: Neural Information Processing Systems Foundation.
- Yu, S., Tranchevent, L., Liu, X., Glanzel, W., Suykens, J. A., de Moor, B., et al. (2012). Optimized data fusion for kernel k-means clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(5), 1031–1039. <https://doi.org/10.1109/TPAMI.2011.255>.
- Zhao, B., Kwok, J., & Zhang, C. (2009). Multiple kernel clustering. In C. Apte, H. Park, K. Wang, & M. Zaki (Eds.), *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM)* (pp. 638–649). Philadelphia, PA: SIAM. <https://doi.org/10.1137/1.9781611972795.55>.
- Zhuang, J., Wang, J., Hoi, S., & Lan, X. (2011). Unsupervised multiple kernel clustering. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, *20*, 129–144.

Of Particles and Molecules: Application of Particle Filtering to Irrigated Agriculture in Punjab, India



Alban Thomas

Abstract We present an estimation method for agricultural crop yield functions, when unobserved productivity depends on water availability that is only partially observed. Using the setting of Bayesian non-linear filtering for estimating Hidden Markov Models, we discuss joint estimation of state variables and parameters in a structural production model with potentially endogenous regressors. An extension to particle filtering with resampling, convolution filter based on kernel regularization, is then discussed. We apply this non-parametric method to estimate a system of structural equations for rice crop yield and unobserved productivity on panel data for 10 districts in Punjab, India. Results based on computer-intensive resampling steps illustrate the interest of convolution particle filtering techniques, with low interquartile range of time-varying estimates. We compare fertilizer elasticity estimates with and without accounting for unobserved productivity, and we find a significant relationship between unobserved productivity and nitrogen fertilizer input, when the former is conditioned on district-level climate variables (summer rainfall, potential evapotranspiration).

1 Introduction

Irrigation has been a major driver of the tremendous increase in crop yield of major crops, in many developing countries over half a century. Yet, the dynamics of water productivity is not so easy to evaluate, because contrary to other production inputs such as labor, fertilizer, and pesticide, water used on crops is in most cases not observed by the practitioner.

While water is a major agricultural input in many settings due to the need to complement rainfall, its use is difficult to record for several reasons. One major reason is that irrigation water is seldom charged in proportion to its actual use. Second, even if water abstracted from surface or ground resources is observed, the ultimate water

A. Thomas (✉)

Toulouse School of Economics-Research, INRAE, University of Toulouse,
31000 Toulouse, France

e-mail: Alban.Thomas@inrae.fr

© Springer Nature Switzerland AG 2021

A. Daouia and A. Ruiz-Gazen (eds.), *Advances in Contemporary Statistics and Econometrics*, https://doi.org/10.1007/978-3-030-73249-3_35

691

volume available to plants is difficult to observe accurately, because of heterogeneous local soil and climate conditions. Third, proxies for water availability such as model-based agronomic simulations or data on irrigation facilities are not always available and are often of poor precision. Fourth, irrigation water is often a complement to rainfall, the latter being in most cases more and more insufficient given targeted crop yields. The contribution of rainfall to water availability for crops in irrigated agriculture is therefore a process with multiple channels, involving direct water use from rainfall, water stocking in reservoirs, canals and wells, as well as groundwater recharge (Lapworth et al. 2014).¹ As a consequence, evaluating agricultural input productivity in settings where water is a major input, from irrigation and/or rainfall, is a major empirical challenge.

Beside available water for crops, there are many more determinants of agricultural production that are also unobserved and that condition both crop yield and input use, resulting in possible simultaneity and/or selection bias in estimation. Production economics has provided the analyst with consistent methods to deal with the estimation of production functions under productivity that is unobserved by the econometrician but observed by the producer (the firm). When unobserved productivity is correlated with regressors such as inputs, least squares estimates of the production function are affected by a simultaneity bias (Sickles and Zelenyuk 2019). Major advances to deal with such bias have been Olley and Pakes (1996) and Levinsohn and Petrin (2003), who suggest controlling for unobserved productivity by introducing inverse demand functions for investment or intermediate inputs (labor), see Chap. 14 (Sect. 14.5) of Sickles and Zelenyuk (2019) for details. However, when only limited information on such inputs are available, other estimation methods are needed than this control function approach. Kutlu and Sickles (2012) have proposed a structural estimation procedure based on a Kalman filter technique applied to random coefficients. This procedure is derived from Kim (2006) and is adapted to structural models with endogenous regressors of the form $y_t = f(x_t, \beta) + w_t + \varepsilon_t$, where y_t is output, x_t is an observed input, ε_t is an i.i.d. error term, and w_t is unobserved heterogeneity, specified as a dynamic process $w_t = \rho w_{t-1} + z_t \gamma + u_t$, with z_t a vector of exogenous variables.

We propose to generalize the structural estimation approach above, described in Sickles and Zelenyuk (2019), by allowing unobserved heterogeneity w_t to interact with observed input x_t , and by relaxing distributional assumptions on the error terms of the model. To do this, we consider a more flexible procedure than the Kalman filter used in Kutlu and Sickles (2012), and we specify a non-linear production function with quadratic and interaction terms. The structural model we propose is dedicated to deal with potential endogeneity caused by unobserved productivity terms that are

¹When the objective is to evaluate water productivity in agriculture, these multiple contributions have to be accounted for, given that the relevant indicator in biophysical terms is water availability for the crop (for most crops, in the soil below the root zone). The latter can be related to available groundwater resources, irrigation facilities, and observed ambient climate variables. However, each of these components is not sufficient to represent the full volume of water available to grow crops (i.e., below the root zone of most plants), which in most cases forms a combination of direct rainfall on crops, groundwater recharge, and/or surface reservoirs used for irrigation.

likely to be correlated with observed inputs. It is of the form: $y_t = g(x_t, w_t, \beta) + \varepsilon_t$, and $w_t = \rho w_{t-1} + z_t \gamma + u_t$, where $g(\cdot)$ can be a flexible functional form that satisfy regularity and concavity conditions.

Our application to agriculture consists in specifying crop yield (production in tons of a crop such as rice, per unit of surface such as hectares) as an increasing and concave function of both observed fertilizer application rate and unobserved productivity. By specifying a dynamic process to represent the latter, a structural model can be considered with crop yield as the observed dependent variable, and productivity as an unobserved state variable, depending on local variables that condition its dynamics. Our model admits the specification discussed in Sickles and Zelenyuk (2019) as a special case and is justified by the need to represent agricultural production as a concave function in a primal framework, where a major input such as water is not directly observed. Although water availability can be associated with observed climate and soil conditions, irrigation water and actual water available to plants remain in most cases unobserved. For this reason, it is recognized in our framework as a major component, although not the only one, of total unobserved productivity. We emphasize at this stage that our purpose is not to estimate unobserved water availability or productivity, but we aim at providing a consistent estimation procedure for agricultural production (crop yield function) with unobserved heterogeneity (including water availability to crops). We use the theoretical setting of Hidden Markov Models (HMM) to specify and estimate such system of equations on time series data for Punjab, one of the most important Indian states in terms of agricultural production (in particular, rice and wheat). We use district-level data from the Icrisat Village Dynamics of South Asia project on rice crop yield, rainfall during the hot season (*Kharif*) and fertilizer use, complemented by data on potential evapotranspiration (ETP) to control the process of unobserved rice productivity.

To our knowledge, this is the first time a method of particle non-linear filtering is applied to the estimation of agricultural production (crop yield function), to account for unobserved productivity, partly associated with water availability. Although the paper does not propose any new estimation method, its major contributions are, first to extend the basic production model with unobserved productivity discussed in Sickles and Zelenyuk (2019), with a more flexible form interacting the latter with observed regressors; and second, to propose an original estimation method for such model, with less distributional restrictions than the Kalman filter approach to Hidden Markov Chains.

The chapter is organized as follows. In Sect. 2, we briefly present the HMM and discuss early estimation methods, before moving to a presentation of particle filtering as a general Bayesian non-linear filtering method. Joint estimation of state variables and parameters is also discussed. Section 3 presents an interesting extension to particle filtering with resampling: the convolution filter technique based on kernel regularization. The empirical application is presented in Sect. 4, in which we present the data, a first estimation of the crop yield function and model calibration, and estimation results. Section 5 concludes.

2 Estimation of Hidden Markov Models

Consider a dynamic state-space system of equations of the form

$$y_t = f(w_t; \theta; \varepsilon_t), \quad (1)$$

$$w_t = g(w_{t-1}; \theta; u_t), \quad (2)$$

where w_t is a scalar unobserved (hidden) state variable, y_t is the observed dependent variable, ε_t and u_t are independent (white noise) random terms. Functions $f(\cdot)$ and $g(\cdot)$ represent the probability density functions (pdf) of observation and state variables respectively. Parameters are gathered in vector θ in the observation and state equations, with possibly a common subset.

A wide variety of economic models can be specified as Equations (1) and (2), including production models dedicated to efficiency analysis of industries, firms, farms, etc. Of particular interest in empirical applications are stochastic frontier (SF) models, which measure the maximum amount of output that can be obtained from a given level of inputs. In the SF framework, measurement errors (two-sided error terms) and (in)efficiency indexes (one-sided error terms) are not observed, making the production frontier stochastic, and they may be time-varying as well as producer-specific. SF models are therefore interesting applications of Bayesian estimation methods because, as discussed in Koop and Steel (2001), such methods are useful in providing more accurate representation of parameter uncertainty through the use of prior knowledge (and possibly, imposing regularity conditions on the production frontier). Bayesian econometric methods may also provide more precise inference on (in)efficiency indexes in small samples, i.e., the (in)efficiency index can be retrieved directly through the computation of its predictive posterior.²

Bayesian econometric methods are particularly useful in cases where inference on individual (in)efficiency indexes is likely to be more precise when unobserved efficiency is explicitly considered as separated from unobserved heterogeneity. For example, Griffiths and Hajargasht (2016) consider a SF model in a Bayesian context with endogeneity of Type I and Type II (i.e., the two-sided error term is correlated with some regressors of the production frontier, or with the one-sided efficiency error term, respectively). Atkinson and Tsionas (2016) use Bayesian techniques to estimate optimal firm-specific directions for a set of inputs and outputs, estimating jointly the directional distance with first-order conditions associated with profit maximization. Gallan et al. (2014) use a Bayesian approach to estimate a SF model with a dynamic process for unobserved inefficiency and random parameters for the observed inefficiency component. Estimating SF models with a Bayesian approach typically requires computer-intensive numerical methods, in order to construct individual contributions to the likelihood function. Markov Chain Monte Carlo (MCMC) algorithms such as Gibbs sampling are generally used in empirical applications, see, e.g., Griffin and Steel (2007).

²See Chap. 7, pp. 168–177 in Koop (2003) on Bayesian econometric methods applied to SFM.

In this chapter, we consider a Bayesian approach that can be applied to a wide variety of models including SF models, which is based on non-linear filtering and which does not require the evaluation of the model likelihood function. We assume that the conditional density $p_t(y_t|w_t)$ exists and is bounded. For inference purposes, and in particular to estimate parameters θ , we need to form the conditional density of the state variable with respect to observations up to time t , $p_t(w_t|y_1, \dots, y_t) = p_t(w_t|y_{1:t})$. We will discuss the issue of parameter estimation below, and consider for the moment the objective of deriving an estimate of the state variable, w_t , conditional on fixed parameters θ and observations $\{y_1, \dots, y_T\}$. The Bayesian sequential approach would evaluate the posterior probability density function $p_t(w_t|y_{1:t})$ recursively, from the following relationship:

$$p_t(w_t|y_{1:t}) = p_t(y_t|w_t) \times \frac{\int p_t(w_t|w_{t-1})p_{t-1}(w_{t-1}|y_{1:t-1})dw_{t-1}}{p_t(y_t|y_{1:t-1})}, \quad (3)$$

which cannot be in general determined analytically. In the linear case with Gaussian random terms, the Kalman filter (Kalman 1960) can be used to produce a consistent estimate of the state variable, based on a recursive algorithm involving first and second Gaussian conditional moments. If the model is non-linear, the Extended Kalman Filter (EKF) can be considered, based on a linearization around previous estimates and applying Kalman filter recursion rules. In practice, however, the EKF can be unstable and produce biased estimates if the assumption of linearity (locally) is violated. Moreover, deriving the Jacobian matrices may be non-trivial in many applications, in particular when the model is highly non-linear. This is mainly because the EKF relies on a Taylor expansion that neglects higher order terms, leading to under-estimation of the covariance of the state variable. Corrections have been proposed, such as the Unscented Kalman Filter (UKF), which adds to the current state and its covariance estimate, expected values of second-order terms. Sample points denoted *sigma points* are selected around the mean of the state variable and are propagated through the non-linear model equations, to form updated mean and covariance estimates. Another advantage of the UKF is that closed-form expressions of the Jacobian matrices are not necessary. Difficulties such as convergence issues remain in practice with EKF or UKF, especially regarding the choice of coordinates for the state variable w_t and when the posterior distribution $p_t(w_t|y_{1:t})$ is unimodal.

2.1 Particle Filtering for Bayesian Non-Linear Filtering

Particle filtering is an extension of the Point Mass Filter (PMF), which was proposed for estimating non-linear and non-Gaussian models by approximating the posterior distribution over a deterministic grid of points, but with the original system of equations (not an approximation as in the EKF or the UKF). By contrast, the Particle Filter considers an adaptive stochastic grid for the state space (Murphy and Russell 2001). The principle of particle filtering is based on sequential importance sampling,

which consists in approximating the posterior pdf:

$$p_t(w_t|y_t) \approx \sum_{i=1}^N \omega_t^{(i)} \delta(w_t - w_t^{(i)}), \quad (4)$$

where N is the number of *particles* and $\{\omega_t^{(i)}, w_t^{(i)}\}_{i=1,2,\dots,N}$ defines a set of weighted particles and δ is the Dirac function. $\omega_t^{(i)}$ are normalized discrete weights such that $\sum_i^N \omega_k^{(i)} = 1, \forall k$ and can be computed recursively. In practice, the standard particle filter approach proceeds with the following steps:

- (1) Draw N particles from a proposal (prior) distribution $w_t^{(i)} \sim \pi(w_t|w_{t-1}^{(i)}, y_{0:t})$
- (2) Update particle weights according to the new observation:

$$\omega_t^{(i)} = \omega_{t-1}^{(i)} \frac{p_t(y_t|w_t^{(i)})p_t(w_t|w_{t-1}^{(i)})}{\pi(w_t|w_{t-1}^{(i)}, y_{0:t})}.$$

- (3) Compute normalized weights $\omega_t^{*(i)} = \omega_t^{(i)} / \sum_i^N \omega_t^{(i)}$.
- (4) Resample if particle degeneracy occurs.

A major drawback with the standard sequential importance sampling procedure above is degeneracy, very few particles being not zero as the algorithm proceeds. The solution is to include a resampling step to the sequential importance sampling. Let weight $\omega_t^{*(i)}$ represent the probability to draw the particle i associated with the set $\{w_t^{(i)}, i = 1, \dots, N\}$. We can draw N such “samples” from the discrete distribution of $\omega_t^{*(i)}$ and replace the old sample with a new one formed by N resampled indexes. Each weight is then set to $1/N$. Resampling can be performed at every step of the algorithm above, but this increases the variance of weights. Other options are to perform resampling at a lower frequency, e.g., every M -th step, or use adaptive resampling, based on the effective number of samples:

$$N^* \approx \left[\sum_{i=1}^N \left(\omega_t^{*(i)} \right)^2 \right]^{-1}. \quad (5)$$

The performance of the particle filter depends on a series of factors, in particular the choice of the proposal distribution $\pi(w_t|w_{t-1}^{(i)}, y_{0:t})$. The optimal importance density is $p(w_t|w_{t-1}^{(i)}, y_t)$ but it is not possible in general to draw from such distribution, except in very special cases, such as a finite state space. Another possibility is to draw from the state transition priori $p_t(w_t|w_{t-1})$, but this does not provide accurate state estimates, as the current observation y_t is overlooked. Refinements of the standard or sequential importance sampling particle filter with resampling have been proposed in

the literature, to produce a better choice of proposal distribution, which significantly improves the performance of the method.³

2.2 Joint Parameter Estimation

We now turn to the problem of non-linear filtering with parameter estimation. There are two main approaches to this problem: either use a non-Bayesian method to optimize some distance function with respect to parameter vector θ , or use a Bayesian method by augmenting the state space with parameters θ , treated the same way as the state variable w_t . Consider the non-Bayesian approach first. A conditional least squares estimator can be constructed as

$$Q_T(\theta) = \frac{1}{T} \sum_{t=1}^T [y_t - E_{\theta}(y_t|y_{1:t-1})]^2, \quad (6)$$

where $E_{\theta}(y_t|y_{1:t-1})$ is the conditional expectation of y_t depending on parameter θ .

It can be approximated with particle filtering as $E_{\theta}(y_t|y_{1:t-1}) \approx \sum_{i=1}^N \omega_t^{(i)} y_t^{(i)}$, where

$\omega_t^{(i)}$, $i = 1, \dots, N$, are particle weights and $y_t^{(i)}$ is sampled from the observation equation $f(w_t^{(i)}; \theta)$. The resulting conditional least squares estimator is then $\hat{\theta}_T^N = \operatorname{argmin}_{\theta} \hat{Q}_T^N(\theta)$, where

$$\hat{Q}_T^N(\theta) = \frac{1}{T} \sum_{t=1}^T \left[y_t - \sum_{i=1}^N \omega_t^{(i)} y_t^{(i)} \right]^2. \quad (7)$$

More general M-estimators can be considered, in particular the Maximum Likelihood estimator that can be derived from sampled contributions $y_t^{(i)}$ (as above) to the likelihood of observation y_t , $p_t(y_t, \theta)$. See Fernandez-Villaverde and Rubio-Ramirez (2007) for an application of particle filtering to macroeconomic models.

Consider then the Bayesian approach, where the vector of parameters is considered a latent variable, of the same nature as w_t , and which can be time-dependent. Methods based on Rao-Blackwellised particle filters have been proposed in the literature, as discussed in, e.g., Lindsten et al. (2012). They consist in most cases of a Gaussian conditional model, with normally distributed state and observation w_t and y_t with conditional moments depending on time-varying parameters θ_t . The problem discussed above without parameter estimation is now to evaluate the posterior distribution $p_t(w_{1:t}, \theta_{1:t}|y_{1:t})$, which can be factored as $p_t(w_{1:t}, \theta_{1:t})p_t(\theta_{1:t}|y_{1:t})$. The first

³A popular extension of the standard particle filter is the Rao-Blackwellized Particle Filter, which performs better than the standard particle filter when the number of state variables becomes large (see, e.g., Liu and Chen 1998).

term is Gaussian and can be evaluated with Kalman filter recursion rules. The second term is proportional to $p_t(y_t|\theta_{1:k}, y_{1:k-1})p_t(\theta_t|\theta_{t-1})p_t(\theta_{1:k-1}|y_{1:k-1})$ and can also be evaluated with a Kalman filter or from recursively drawing from importance distributions. Static parameters can also be accommodated by assuming that the posterior distribution of θ depends on sufficient statistics $T_t(w_{1:t}, y_{1:t})$, for which a recursion rule exists (Fearnhead 2002).

Particle filtering of both state and parameters has been applied while trying to deal with the problem of sample degeneracy, by adding random disturbances (*roughening penalties*) to parameter particles between updating steps. Model parameters are therefore artificially viewed as time-varying and are replaced by θ_t in an augmented state vector (Storvij 2002). Such artificial evolution allows one to generate updated parameter values at each step of the filtering algorithm, including a resampling step to take care of the sample attrition issue. The drawback with such method is that it introduces an artificial loss of information, so that approximated posterior distributions are too diffuse with respect to their theoretical counterparts (associated with fixed value parameters).

A solution is to regularize the model equations by, e.g., smoothing the empirical measure of the posterior distribution of θ with a Gaussian distribution, as proposed by Liu and West (2001). An important restriction is that, on top of the variances or random terms being non-zero (and the likelihood function to be tractable for some of the inference approaches), both the state and the observation variables must be regularized to avoid the problem of degeneracy of particle filters. The next section proposes an interesting alternative that deals with issues of joint estimation of state and parameters in non-linear filtering problems, namely, the regularization with convolution kernels.

3 The Convolution Particle Filter

We consider here a particle filter estimation procedure that avoids numerical issues such as impoverishment and degeneracy. The convolution particle filtering approach has been developed in tracking and shape recognition problems. It is based on a convolution kernel density estimation and a regularization of the distribution of state and observation variables. When the likelihood associated with the model and the sample does not admit an analytical solution, complex problems can nevertheless be solved.

Filters based on sequential importance sampling with resampling have however remaining drawbacks in practical use, although they have interesting theoretical properties. First, systems with non-noisy observations or with a very low noise to signal ratio may hamper convergence of such filters, because noise density is used to weight the particles. Second, as discussed in Hurzeler and Kunsch (1998), a degree of regularization on the distribution of the state variable is necessary to stabilize the signal to noise ratio, because of the discrete nature of the distribution approximations. Third and more importantly, regularized filters rely on the availability of the obser-

vation likelihood $p_t(y_t|w_t)$. If only a limited knowledge of such likelihood function is available, it may not be possible to update weights in step (2) of the algorithm above.

We consider here a regularization approach for the state and the observation distributions, which does not require analytical knowledge of $p(y_t|w_t)$, only the ability to simulate observation y_t from state variable w_t using the structural model (1)-(2). This approach is based on convolution kernel density estimation for the estimation, weight updating, and resampling steps, and it avoids the problem of no or small observation noise (Del Moral et al. 2001). We first consider the problem of non-linear filtering with fixed parameter values that are not estimated. See Vila (2012) for an analysis of the theoretical properties (almost sure convergence) of the algorithm.

Suppose that we can draw random samples of size N , $\{w_t^{(i)}, y_t^{(i)}\}$, from the state and observation pdfs $g(\cdot|w_{t-1})$ and $f(\cdot|w_t)$. A sample can then be drawn from their joint distribution by sequential simulation, starting from a proposal distribution $\pi_0(w_t)$, where an estimate of the joint density is

$$p_t(w_t, y_{1:t}) \approx \frac{1}{N} \sum_i^N \delta(w_t - w_t^{(i)}, y_t - y_{1:t}^{(i)}) \tag{8}$$

This is a direct generalization of the density approximation in (4), and its convolution can be computed to provide a kernel estimate of the true joint density:

$$p_t^N(w_t, y_{1:t}) = \frac{1}{N} \sum_i^N K_h^w(w_t - w_t^{(i)}) \times K_h^{\bar{y}}(y_{1:t} - y_{1:t}^{(i)}), \tag{9}$$

where $K_h^{\bar{y}} = \prod_{j=1}^t K_h^y(y_j - y_j^{(i)})$ and K_h^w and K_h^y and Parzen-Rosenblatt kernels with appropriate dimensions. The posterior conditional distribution of the state variable can then be estimated by

$$p_t^N(w_t|y_{1:t}) = \frac{\sum_i^N K_h^w(w_t - w_t^{(i)}) \times K_h^{\bar{y}}(y_{1:t} - y_{1:t}^{(i)})}{\sum_i^N K_h^{\bar{y}}(y_{1:t} - y_{1:t}^{(i)})} \tag{10}$$

A major aspect of particle filtering, which also applies to the convolution particle filter, is that numerical (quadratic) complexity involved in computing the term $K_h^{\bar{y}}(y_{1:t} - y_{1:t}^{(i)})$ (over t periods) can be avoided. This is because resampling weights associated with particles takes care of the dependence of density $p_t^N(w_t|y_{1:t})$ with respect to past observations $y_{1:t}$ (see Rossi and Vila (2006)). In effect, one can form the weight $\omega_t^{(i)}$ with the recursion $\omega_t^{(i)} = \omega_{t-1}^{(i)} \times K_h^y(y_t - y_t^{(i)})$, and consider instead

$p_t^N(w_t|y_{1:t}) = \sum_i^N \omega_t^{(i)} K_h^w(w_t - w_t^{(i)})$, where weights $\omega_t^{(i)}$ sum to 1 and are resampled.

A resampled convolution filter would (Rossi and Vila 2006) draw N particles at $t = 0$ from a proposal distribution $w_1^{(i)} \sim \pi_0$, and for $t = 1, \dots, T$, resample $(w_{t-1}^{(1)}, \dots, w_{t-1}^{(N)})$ from distribution p_{t-1} , and sample state and observation $w_t^{(i)} \sim g(w_{t-1}^{(i)})$ and $y_t^{(i)} \sim f(w_t^{(i)})$. The filter is finally updated: $p_t(w_t|y_{1:t}) = \sum_i^N \omega_t^{(i)} K_h^w(w_t - w_t^{(i)})$, where $\omega_t^{(i)} = K_h^y(y_t - y_t^{(i)}) / \sum_i^N K_h^y(y_t - y_t^{(i)})$.

Let us now turn to the joint estimation of state and parameter by convolution filter. As discussed in the previous section, sequential importance sampling schemes with resampling can accommodate parameter estimation, provided some regularization and resampling of parameter particles are performed. The performance of the convolution particle filter crucially depends on the design parameters, in particular kernel functions and their associated bandwidth parameters, the number of particles, and the proposal distribution function. The full algorithm for kernel convolution particle filter with parameter estimation is as follows:

- (1) $t = 0$. Initialization. Generate N particles from proposal distributions of state and parameters:
 $w_0^{(i)} \sim p_0(w)$, $\theta_0^{(i)} \sim p_0(\theta)$, and initialize weights $\omega_0^{(i)} = 1/N$.
- (2) Iterate for $t = 1, \dots, T$
 - if $t = 1$: Prediction, for $i = 1$ to N
 - (i) Sample state $w_1^{(i)} \sim g(w_0^{(i)}, \theta_0^{(i)})$, parameter $\theta_1^{(i)}$ from $\theta_0^{(i)}$ and observation $y_1^{(i)} \sim f(w_0^{(i)}, \theta_0^{(i)})$.
 - (ii) Go to step 3)
 - if $t > 1$:
 - (i) Resampling, for $i = 1$ to N :
 $(w_{t-1}^{(i)}, \theta_{t-1}^{(i)}) \sim p_{t-1}^N(w_{t-1}, \theta_{t-1}|y_{1:t-1})$, set weights $\omega_{t-1}^{(i)} = 1/N$
 - (ii) Prediction, for $i = 1$ to N : Sample state $w_t^{(i)} \sim g(\cdot|w_{t-1}^{(i)}, \theta_{t-1}^{(i)})$, parameter $\theta_t^{(i)}$ from $\theta_{t-1}^{(i)}$ and observation $y_t^{(i)} \sim f(w_t^{(i)}, \theta_t^{(i)})$
 - (iii) Update weights, for $i = 1$ to N : $\omega_t^{(i)} = \omega_{t-1}^{(i)} \times K_h^y(y_t - y_t^{(i)})$
 - (iv) Estimate conditional densities of w_t and θ_t :

$$p_t^N(w_t, \theta_t|y_{1:t}) = \left(\sum_{i=1}^N \omega_t^{(i)} \right)^{-1} \sum_{i=1}^N \omega_t^{(i)} K_h^\theta(\theta_t - \theta_t^{(i)}) K_h^w(w_t - w_t^{(i)}).$$

$$p_t^N(w_t|y_{1:t}) = \left(\sum_{i=1}^N \omega_t^{(i)} \right)^{-1} \sum_{i=1}^N \omega_t^{(i)} K_h^w(w_t - w_t^{(i)}).$$

$$p_t^N(\theta_t|y_{1:t}) = \left(\sum_{i=1}^N \omega_t^{(i)} \right)^{-1} \sum_{i=1}^N \omega_t^{(i)} K_h^\theta(\theta_t - \theta_t^{(i)}).$$

(3) Estimate final state and parameter values:

$$\hat{w}_t = \sum_{i=1}^N \omega_t^{(i)} w_t^{(i)}, \quad \hat{\theta}_t = \sum_{i=1}^N \omega_t^{(i)} \theta_t^{(i)}.$$

4 Empirical Application

Punjab has been a central state in India's green revolution, thanks to fertile soils and abundant surface and groundwater resources. Punjab produces currently about 20 percent and 11 percent, respectively, of India's wheat and rice production (Lapworth et al. 2014). Between 1970 and 2010, the area planted with rice increased from 390,000 ha to 2,826,000 ha (Punjab 1971, 1981, 2000, 2001, 2010). Punjab is characterized by a significantly higher irrigated rice yield (4,010 kg/ha) than other Indian states, with an average applied irrigation water of 180 cm, and an applied irrigation water productivity of 0.22 kg/m³ (Sharma et al. 2018). Rice and wheat, two of India's most important food crops, are also the most water-intensive: producing a kilogram of rice requires an average of 2,800 liters of water, while a kilogram of wheat takes 1,654 liters. As a consequence, ever growing withdrawal of groundwater for rice (paddy) cultivation has resulted in a rapid decline of the water table in Punjab districts.

The purpose of the present empirical application is to estimate a crop yield function for rice, with observed fertilizer input use but accounting for unobserved productivity. The latter is expected to depend on unobserved water availability, originating from direct rainfall and rainfall harvesting and irrigation facilities, and we assume that unobserved productivity may interact with fertilizer input use. To account for such interaction and non-linearities (concavity) in the crop yield function, we consider a quadratic production function for rice crop yield, depending on observed fertilizer application rate and unobserved productivity (assumed to depend on unobserved water availability, which depends upon observed ambient climate variables). We use district-level time series on rainfall during the hot season (*Kharif*) and on potential evapotranspiration (ETP) to control the process of unobserved productivity. Evapotranspiration is a widely used indicator of the pressure put by cropping systems on water availability and is negatively correlated with the latter, while rainfall is a positive contributor to it. The Penman-Monteith equation is currently used by

the food and Agricultural Organization to estimate a reference evapotranspiration (ET_0), which is multiplied if needed by crop coefficients K_c (see, e.g., Burman and Pochop 1994). It is important to note that care should be taken when interpreting results from a system of equations with unobserved productivity (with water availability as a major component) specified as a hidden Markov process. This is because many other unobserved factors, even aggregated at the district level, may influence crop yield, and water availability may not be easily singled out from unobserved productivity. Conditioning unobserved productivity on climate variables (rainfall, ETP) is justified because of its expected relationship with water availability for the crop. Checking whether such conditioning is relevant in a statistical sense is the only option we have at this stage, without imposing additional assumptions on the relationship between unobserved productivity, water availability for crops, and observed variables in state and observation equations.

4.1 Data and Model Specification

The sample used in the application is obtained from Indian district-level data collected from ICRISAT over the period 1966–2007, as part of the Village Dynamics in South Asia (VDSA) data collection effort. See Icrisat (2012) for a description of the VDSA data set. We have data on 11 districts followed from 1966–2007 in Punjab, but in order to maintain a balanced sample, we drop districts with too many missing observations over the 42 years. The final sample consists of 10 districts (Amritsar, Bathinda, Firozpur, Gurdaspur, Hoshiarpur, Jalandhar, Kapurthala, Ludhiana, Patiala, and Sangrur) and contains 420 observations in total.

We specify a quadratic crop yield function and an autoregressive AR(1) process for unobserved productivity w :

$$y_{it} = \beta_0 + \beta_1 x_{it} + \frac{1}{2} \beta_2 x_{it}^2 + w_{it} + \frac{1}{2} \beta_3 w_{it}^2 + \beta_4 x_{it} w_{it} + \sigma_\varepsilon \varepsilon_{it},$$

$$w_{it} = \rho w_{i,t-1} + \gamma_1 z_{it} + \gamma_2 r_{it} + u_{it},$$

where y_{it} is crop yield (ton/ha), x_{it} is fertilizer application rate (kg/ha), w_t is unobserved productivity (depending on water availability), z_{it} is the reciprocal of ETP and r_{it} is average summer (*kharif* rainfall, in mm/ha). We specify $z = 1/ETP$ for ease of interpretation without loss of generality, to have both exogenous explanatory variables in the state equation with the same expected (positive) sign. Quadratic terms in unobserved productivity and observed fertilizer application rate allow us to explore the degree of concavity in the crop yield function, while the interaction term $x_{it} \times w_{it}$ is useful to evaluate the direction of the relationship between productivity and actual fertilizer input use.

Table 1 presents descriptive statistics on the sample.

Table 1 Descriptive statistics

Variable	Unit	Mean	Std. deviation	Min.	Max.
Rice yield	(ton/ha)	2.9240	0.8167	0.9090	4.6540
Rice fertilizer	(kg/ha)	171.8289	77.4493	10.9879	325.2017
Summer rainfall	(mm/ha)	332.34	179.7496	0	994
1/ETP		0.1601	0.0168	0.1467	0.2188

Notes. 420 observations: 42 years (1966–2007), 10 Punjab districts. Icrisat VDSA Survey (Icrisat 2012), except ETP (evapotranspiration) obtained from the Indian Water Portal

Table 2 Crop yield function estimates

Parameter	Estimate	Standard error	<i>p</i> -value
1/ETP	23.5359***	3.0321	0.000
0.5(1/ETP) ²	−10.1237***	1.3500	0.000
Fertilizer	1.1724***	0.1694	0.000
Fertilizer ² /2	−0.3403***	0.0305	0.000
Fertilizer × (1/ETP)	−0.2163	0.1592	0.175
Intercept	−13.0280***	1.6796	0.000
σ_u^2	0.1412		
σ_e^2	0.1137		

Notes. 420 observations: 42 years (1966–2007), 10 Punjab districts. All variables are divided by their sample mean

**p* < 0.1

***p* < 0.05

****p* < 0.01. $R^2 = 0.5843$. Fisher test for individual effects $F(9,405) = 29.47$ (*p*-value = 0.000). ETP is potential evapotranspiration at district level (in mm/ha). σ_u^2 and σ_e^2 respectively denote the variance estimate of individual and i.i.d. error terms

Let θ denote the vector of parameters to be estimated, $\theta = \{\beta_0, \dots, \beta_5, \sigma_\varepsilon, \rho, \gamma_1, \gamma_2\}$. Initial values for some parameters in θ are obtained from a panel-data estimation with fixed effects. In the observation and state equations, we replace productivity w (not observed) by $1/ETP$; in the state equation, we do not have any proxy for the original $z = 1/ETP$, and we arbitrarily set initial parameter value γ_1 to 1. All observed variables are divided by their sample mean to have normalized variables with unit mean. Estimation results of the fixed-effect model for quadratic crop yield are presented in Table 2.

Most parameters are significant at the 5 percent level, except the interaction term between fertilizer and $1/ETP$. The crop function is increasing and concave in its arguments (fertilizer and $1/ETP$), but no significant substitutability pattern is found between both regressors. We present in Table 3 parameter estimates for the modified state equation. The autoregressive parameter of $1/ETP$ is significantly different from 1 so that a panel-data unit root test is not deemed necessary. Summer (July–August) rainfall is significant and positive as expected, indicating that when a year

Table 3 Dynamic panel-data equation for 1/ETP

Parameter	Estimate	Standard error	<i>p</i> -value
$(1/ETP)_{t-1}$	0.6216***	0.0357	0.000
Summer rainfall	0.0018***	0.0006	0.006
Constant	0.3760***	0.0372	0.000

Notes. 404 observations

* $p < 0.1$

** $p < 0.05$

*** $p < 0.01$. Sargan overidentifying restrictions test $\chi^2(39) = 9.8554$ (p -value = 1.000) with two-step GMM estimates

Arellano-Bond test for zero autocorrelation in first-differenced errors: Order 1 -1.4666 (p -value 0.1425) ; Order 2 -1.1619 (p -value 0.2453)

is relatively less dry, crop evapotranspiration decreases as plants demand less water for their growth (see Burman and Pochop 1994 for details).

As for random terms in the system of equations, we estimate the standard deviation of ε , σ_ε , and we impose a variance of 1 for u_t to reduce the number of parameters to estimate. We consider a model where all parameters are district-dependent, and all 10 districts are treated as independent time series with 42 time periods each. A final estimation step is also implemented on a single time series constructed from average values over the districts of all observed variables (y , x , z , and r) as well as values used for building the proposal distribution for w .

The initial distribution $\pi_0(w)$ is specified from groundwater availability (per ha) estimates for selected years (from 1998 to 2010) obtained from the Indian Ministry of Water. This is justified because, as discussed above, we assume a strong relationship between unobserved productivity w and unobserved water availability to grow crops. We select the highest value from these figures in each district, assuming that water available for crops has been diminishing over the 1966–2007 period. We are of course aware that water available for crops (in the soil) does not correspond to groundwater availability, however, it is the closest indicator we could find for our empirical application. We considered the minimum groundwater availability per hectare for the mean of the distribution, but we multiplied the standard deviation by a factor of 4.0 to account for a time-dependent drift in the distribution.

4.2 Estimation Results

The convolution particle filter procedure is implemented with various values of N , from 20,000 to 2,000,000 (2 million). We present here results obtained with 1,000,000 random draws, as increasing N above that value did not yield significantly different results. At the end of the estimation process, accuracy of parameter estimates can be examined from the interquartile range of the empirical distribution of θ . For the implementation of the convolution filter, we specify Gaussian kernels with bandwidth computed from the following rule of thumb: $h_A = 0.79 I Q_A \times N^{-1/5}$, where $A = w, y, \beta_i, i = 1, \dots, \dim \theta$ and $I Q_A$ is the Interquartile Range of A .

Table 4 Parameter median and interquantile range-By district

District	β_0	β_1	β_2	β_3	β_4	σ_ϵ	ρ	γ_1	γ_2
Gurdaspur	-6.3813 (0.8792)	0.3162 (0.0344)	-0.2423 (0.0263)	-7.6800 (0.8339)	0.2330 (0.0253)	0.9639 (0.1046)	0.4180 (0.0454)	1.9675 (0.2132)	0.0000 (0.0000)
Amritsar	-7.8846 (0.9110)	1.5137 (0.1643)	-0.4833 (0.0525)	-20.2338 (2.1986)	0.2068 (0.0225)	1.0991 (0.1194)	0.6819 (0.0740)	1.6532 (0.1796)	0.0023 (0.0003)
Kapurthala	-7.0148 (1.8837)	1.2162 (0.1320)	-0.1682 (0.0183)	-6.9980 (0.7601)	-0.6099 (0.0662)	1.1553 (0.1255)	0.5766 (0.0626)	-0.6526 (0.0709)	0.0045 (0.0005)
Jalandhar	-13.9783 (1.5422)	0.5888 (0.0638)	-0.3810 (0.0414)	-8.8282 (0.9588)	0.7067 (0.0768)	0.7563 (0.0823)	0.6123 (0.0664)	-0.3770 (0.0410)	0.0001 (0.0000)
Hoshiarpur	-21.0912 (2.3938)	1.2877 (0.1398)	-0.2703 (0.0293)	-6.4334 (0.6983)	0.0330 (0.0036)	0.8262 (0.0897)	0.4353 (0.0473)	0.4975 (0.0540)	0.0009 (0.0001)
Ludhiana	-4.9550 (0.5792)	1.2355 (0.1341)	-0.5004 (0.0543)	-6.2212 (0.6754)	-0.3398 (0.0369)	1.2028 (0.1307)	0.7762 (0.0843)	0.2745 (0.0298)	0.0046 (0.0005)
Firozpur	-13.4434 (1.4602)	1.0754 (0.1169)	-0.3089 (0.0336)	4.0754 (0.4423)	0.1629 (0.0177)	0.3602 (0.0391)	0.5885 (0.0638)	0.9630 (0.1047)	0.0018 (0.0002)
Bathinda	-12.8676 (1.3992)	1.3284 (0.1443)	-0.1291 (0.0140)	-9.4259 (1.0230)	-0.2667 (0.0290)	0.7735 (0.0840)	0.6147 (0.0669)	0.0974 (0.0105)	0.0024 (0.0002)
Sangrur	-2.5567 (0.3104)	1.7216 (0.1871)	-0.8333 (0.0905)	-12.1006 (1.3138)	-0.6772 (0.0734)	1.1328 (0.1231)	0.6425 (0.0697)	0.4390 (0.0477)	0.0006 (0.0001)
Patiala	-12.5266 (1.4551)	1.8910 (0.2053)	-0.2827 (0.0307)	-10.2794 (1.1148)	-0.8341 (0.0906)	1.3977 (0.1518)	0.7766 (0.0843)	0.5944 (0.0644)	0.0046 (0.0005)
All districts	-15.2154 (1.6947)	1.6301 (0.1769)	-0.0771 (0.0084)	-6.0257 (0.6542)	-0.8622 (0.0935)	0.4409 (0.0479)	0.4465 (0.0485)	1.3526 (0.1469)	0.0022 (0.0002)

Notes. 42 years (1966–2007), 10 Punjab districts. Median ($Q_{0.5}$) and Interquantile Range ($Q_{0.75} - Q_{0.25}$) are computed from the last period estimates ($t = 42$). The Interquantile Range is between parentheses. Number of draws 1,000,000

Table 5 Fertilizer elasticities of rice crop yield

District	OLS estimates	Convolution filter estimates
Gurdaspur	0.7702 (0.1444)	0.3069 (0.0860)
Amritsar	0.7664 (0.1253)	1.2372 (0.2393)
Kapurthala	0.7335 (0.1388)	0.4381 (0.2165)
Jalandhar	0.7138 (0.1483)	0.9145 (0.1820)
Hoshiarpur	0.8460 (0.1300)	1.0504 (0.1727)
Ludhiana	0.6466 (0.1530)	0.3953 (0.2253)
Firozpur	0.7025 (0.1745)	0.9294 (0.1682)
Bathinda	0.7110 (0.1842)	0.9326 (0.1873)
Sangrur	0.8004 (0.1241)	0.2111 (0.3510)
Patiala	0.7620 (0.1434)	0.7742 (0.3266)
All districts	0.7460 (0.1404)	0.6908 (0.2788)

Notes. 42 years (1966–2007), 10 Punjab districts. Standard errors for OLS and interquartile range (IQ) for convolution filter estimates are in parentheses

Estimation results from the convolution particle filter estimator are presented in Table 4, with Interquartile Range ($Q_{.75} - Q_{.25}$) to represent the precision of the algorithm. Based on such indicator, the convolution particle filter provides remarkably precise parameter estimates. Of particular interest are parameters β_3 and β_4 , associated with squared unobserved productivity and its interaction with fertilizer input, in the observation equation (y). The former is negative in all cases, indicating some degree of concavity in the unobserved productivity. As for β_4 , it is negative in five districts out of 10 (as for the sample average of all districts), leading us to reject a systematic form of complementarity (or substitutability) in the crop yield production, between productivity and nitrogen fertilizer input. Moreover, the autoregressive parameter in unobserved productivity is also estimated accurately and lies between 0.4180 (Gurdaspur district) and 0.7766 (Patiala district). Parameters γ_1 and γ_2 inform about the correlation between unobserved productivity on the one hand, and inverse ETP and rainfall on the other. Both are significant and, while rainfall is positively correlated with productivity in all districts, inverse ETP has a positive correlation with w in eight cases out of 10.

Finally in our empirical application, we compare the effect of fertilizer application rate on rice yield, when unobserved productivity is accounted for and when it is not. To do this, we compute the relative, marginal fertilizer productivity as the elasticity of rice yield with respect to fertilizer input use based on our convolution filter estimates, which we compare with OLS estimates (with only x and $0.5x^2$ as regressors). Results in Table 5 illustrate the consequence of introducing a latent productivity process, resulting in a much stronger heterogeneity in elasticity estimates than in the OLS case. However, for the average sample on all districts, the elasticity estimate is fairly similar between both estimation methods (0.7460 for OLS compared with 0.6908 for convolution filter).

5 Conclusion

This chapter uses the framework of Hidden Markov Models (HMM) to estimate a system of structural equations for agricultural production with unobserved productivity, in a context where the latter is expected to depend on unobserved water availability. We generalize the structural approach of production function estimation with endogenous regressors discussed in Chap. 14 of Sickles and Zelenyuk (2019), by considering a non-linear production function that accommodates concavity and interaction with observed inputs for the unobserved productivity process. To our knowledge, this is the first time a method of particle non-linear filtering is applied to the estimation of agricultural production (crop yield function), to account for unobserved productivity, partly associated with water availability. This work contributes to empirical literature by extending the production model with unobserved productivity discussed in Sickles and Zelenyuk (2019), with a more flexible form interacting the latter with observed regressors, and by proposing an original estimation method with less distributional restrictions than the Kalman filter approach to Hidden Markov Chains.

We use ICRISAT data for Punjab districts in India on rice crop yield, fertilizer use, rainfall during the *Kharif* season, and potential evapotranspiration for unobserved water availability (with their relationship with productivity as the main motivation). We discuss estimation methods for HMM, focusing on particle filter techniques for sequential importance sampling algorithms. In the framework of Bayesian non-linear filtering, we are particularly interested in joint estimation of state variables and parameters, for which we present the convolution filter technique based on kernel regularization as an interesting extension to standard particle filtering.

The estimation procedure proposed in this chapter can prove useful for practitioners, in particular those considering production models dedicated to efficiency analysis and prediction of productivity in agriculture and other economic activities. The general framework presented here allows for a wide range of possible specifications on unobserved inefficiency and heterogeneity random terms, making specification checks possible when investigating dynamic components of, e.g., stochastic frontiers. Moreover, the Bayesian non-linear filtering approach discussed in the chapter

is particularly interesting to consider in empirical settings where data are difficult to collect, or simply not available. This is for example often the case of productivity components such as available water in agriculture, or other inputs for which price and quantity data are not systematically recorded. As far as policy guidance is concerned, we stress that our approach should, whenever possible, be considered jointly with other data or information sources. However, the approach proposed here can be useful in providing information for helping decision-makers to design, e.g., better targeted policies given limited available information.

Regarding the empirical application, there are obvious caveats to note, in particular the fact that the state process we consider (unobserved productivity depending on water availability for crops) may well depend on other unobserved, time-varying components for which control variables are needed. The only way to ensure our interpretation of the hidden process corresponds effectively to productivity depending on unobserved water availability is to examine the statistical significance of variables used to condition the state process. In our case, available water for crops is assumed to (positively) depend on actual summer rainfall and potential evapotranspiration, and our estimates confirm that this is valid for rainfall in all cases, and in a majority of cases (eight districts out of 10) for inverse ETP. Our parameter estimates are fairly precise, based on final interquantile range, and confirm that this conditioning is relevant. Our model is estimated for rice at this time, and a possibility is to estimate the model for more crops, but keeping the same process for unobserved productivity, w_t . Concerning the choice of proxies to estimate initial parameter values and start the convolution filter algorithm, other choices than inverse ETP are possible, and robustness checks could be used with other proxies for unobserved heterogeneity.

Econometric methods used in the application can be extended and improved in several directions. First, block sampling and resampling can be considered, to better exploit the panel structure of the data. Another possibility would be to obtain w and θ in each time period for all districts successively, and then move to the next time period, hence allowing to draw a “block” sample accounting for a covariance matrix with error components and individual effects. Second, more efficient procedures for selecting the bandwidth parameters in kernel density can be considered, along the lines, e.g., of Botev et al. (2010). This is also true of the choice of the Parzen-Rosenblatt kernel $K(\cdot)$ used for convolution particle filter, other specifications than the Gaussian kernels (Epanechnikov, etc.) being likely to produce more efficient results. This is left for future research.

Acknowledgements The author would like to thank the editors of this Festschrift for the initiative, two anonymous reviewers for helpful comments and Christine Thomas-Agnan for providing the inspiration to finish this work at last. Alban Thomas acknowledges funding from Agence Nationale de la Recherche under grants ANR-16-CE03-0006 (project ATCHA-India) and ANR-17-EURE-0010 (Investissements d’Avenir program).

References

- Atkinson, S., & Tsionas, M. (2016). Directional distance functions: Optimal endogenous directions. *Journal of Econometrics*, *190*, 301–314.
- Botev, Z., Grotowski, J., & Kroese, D. (2010). Kernel density estimation via diffusion. *Annals of Statistics*, *38*, 2916–2957.
- Burman, R., & Pochop, L. (1994). *Evaporation, Evapotranspiration and Climatic Data*. Amsterdam: Elsevier Science B.V.
- Del Moral, P., Jacod, J., & Protter, P. (2001). The monte-carlo method for filtering with discrete-time observations. *Probability Theory and Related Fields*, *120*, 346–368.
- Fearnhead, P. (2002). Mcmc, sufficient statistics and particle filter. *Journal of Computational and Graphical Statistics*, *11*, 848–862.
- Fernandez-Villaverde, J., & Rubio-Ramirez, J. (2007). Estimating macroeconomic models: A likelihood approach. *Review of Economic Studies*, *74*, 1059–1087.
- Gallan, J., Veiga, H., & Wiper, M. (2014). Bayesian estimation of inefficiency heterogeneity in stochastic frontier models. *Journal of Productivity Analysis*, *42*, 85–101.
- Griffin, J., & Steel, M. (2007). Bayesian stochastic frontier analysis using winbugs. *Journal of Productivity Analysis*, *27*, 163–176.
- Griffiths, W., & Hajargasht, G. (2016). Some models for stochastic frontiers with endogeneity. *Journal of Econometrics*, *190*, 341–348.
- Hurzeler, M., & Kunsch, H. (1998). Monte carlo approximations for general state-space models. *Journal of Computational and Graphical Statistics*, *7*, 175–193.
- Icrisat. (2012). *Village Dynamics in South Asia, 2012*. ICRISAT-ICAR-IRRI, Hyderabad, India: Tech. rep.
- Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*, 35–45.
- Kim, C. (2006). Time-varying parameter models with endogenous regressors. *Economics Letters*, *91*, 21–26.
- Koop, G. (2003). *Bayesian Econometrics*. Chichester: Wiley.
- Koop, G., & Steel, M. (2001). Bayesian analysis of stochastic frontier models. In B. Baltagi (Ed.), *A Companion to Theoretical Econometrics*. Oxford: Blackwell Publisher.
- Kutlu, L., & Sickles, R. (2012). Estimation of market power in the presence of firm level inefficiencies. *Journal of Econometrics*, *168*, 141–155.
- Lapworth, K. D. J., Gopal, M. R., & MacDonald, A. (2014). Intensive Groundwater Exploitation in the Punjab – an Evaluation of Resource and Quality Trends. Tech. Rep. Open Report, OR/14/068. 45pp., British Geological Survey, Nottingham, UK.
- Levinsohn, J., & Petrin, A. (2003). Estimating production function using inputs to control for unobservables. *Review of Economic Studies*, *70*, 317–341.
- Lindsten, F., Schon, T., & Svensson, L. (2012). A non-degenerate rao-blackwellised particle filter for estimating static parameters in dynamical models. *IFAC Proceedings Volumes*, *45*, 1149–1154.
- Liu, J., & West, M. (2001). Combined parameter and state estimation in simulation-based filtering. In A. Doucet, N. Freitas, & N. de and Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. New York, NY: Springer.
- Liu, J. S., & Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, *93*, 1032–1044.
- Murphy, K., & Russell, S. (2001). Rao-blackwellised particle filtering for dynamic bayesian networks. In A. Doucet, N. Freitas, & N. de and Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*. Statistics for Engineering and Information Science. New York, NY: Springer.
- Olley, G., & Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, *64*, 1263–1297.
- Punjab, G. (1971, 1981, 2000, 2001, 2010). Statistical Abstract of Punjab. Tech. rep. Government of Punjab, Chandigarh, India.

- Rossi, V., & Vila, J. P. (2006). Nonlinear filtering in discrete time: A particle convolution approach. *Annals I.S.U.P.*, 50, 71–102.
- Sharma, B., Gulati, A., Mohan, G., Manchanda, S., Ray, I., & Amarasinghe, U. (2018). Water productivity mapping of major Indian crops. Tech. rep., National Bank for Agriculture and Rural Development (NABARD) and ICRIER (Indian Council for Research on International Economic Relations) report, New Delhi, India.
- Sickles, R., & Zelenyuk, V. (2019). *Measurement of productivity and efficiency. Theory and practice*. Cambridge: Cambridge University Press.
- Storvij, G. (2002). Particle filters in state space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing*, 50, 281–289.
- Vila, J. (2012). Enhanced consistency of the resampled convolution particle filter. *Statistics and Probability Letters*, 82, 786–797.