CLADAG 2021

BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by

Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE
UNIVERSITY
PRESS

PROCEEDINGS E REPORT

ISSN 2704-601X (PRINT) - ISSN 2704-5846 (ONLINE)

– 128 –

SCIENTIFIC PROGRAM COMMITTEE

Giovanni C. Porzio (chair) (University of Cassino and Southern Lazio - Italy)

Silvia Bianconcini (University of Bologna - Italy)
Christophe Biernacki (University of Lille - France)
Paula Brito (University of Porto - Portugal)
Francesca Marta Lilja Di Lascio (Free University of Bozen-Bolzano - Italy)
Marco Di Marzio ("Gabriele d'Annunzio" University of Chieti-Pescara - Italy)
Alessio Farcomeni ("Tor Vergata" University of Rome - Italy)
Luca Frigau (University of Cagliari - Italy)
Luis Ángel García Escudero (University of Valladolid - Spain)
Bettina Grün (Vienna University of Economics and Business - Austria)
Salvatore Ingrassia (University of Catania - Italy)
Volodymyr Melnykov (University of Alabama - USA)
Brendan Murphy (University College Dublin -Ireland)
Maria Lucia Parrella (University of Salerno - Italy)
Carla Rampichini (University of Florence - Italy)
Monia Ranalli (Sapienza University of Rome - Italy)
J. Sunil Rao (University of Miami - USA)
Marco Riani (University of di Parma - Italy)
Nicola Salvati (University of Pisa - Italy)
Laura Maria Sangalli (Polytechnic University of Milan - Italy)
Bruno Scarpa (University of Padua - Italy)
Mariangela Sciandra (University of Palermo - Italy)
Luca Scrucca (University of Perugia - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)
Mariangela Zenga (University of Milan-Bicocca - Italy)

LOCAL PROGRAM COMMITTEE

Carla Rampichini (chair) (University of Florence - Italy)

Chiara Bocci (University of Florence - Italy)
Anna Gottard (University of Florence - Italy)
Leonardo Grilli (University of Florence - Italy)
Monia Lupparelli (University of Florence - Italy)
Maria Francesca Marino (University of Florence - Italy)
Agnese Panzera (University of Florence - Italy)
Emilia Rocco (University of Florence - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)

CLADAG 2021
BOOK OF ABSTRACTS
AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE UNIVERSITY PRESS
2021

CLADAG 2021 BOOK OF ABSTRACTS AND SHORT PAPERS : 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021/ edited by Giovanni C. Porzio, Carla Rampichini, Chiara Bocci. — Firenze : Firenze University Press, 2021.
(Proceedings e report ; 128)

https://www.fupress.com/isbn/9788855183406

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 978-88-5518-340-6 (PDF)
ISBN 978-88-5518-341-3 (XML)
DOI 10.36253/978-88-5518-340-6

Graphic design: Alberto Pizarro Fernández, Lettera Meccanica SRLs
Front cover: Illustration of the statue by Giambologna, *Appennino* (1579-1580) by Anna Gottard

CLAssification and Data
Analysis Group (CLADAG)
of the Italian Statistical
Society (SIS)

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

This book is printed on acid-free paper
Printed in Italy

INDEX

Keynote Speakers

Plenary Session

Invited Papers

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

Contributed Papers

Preface

This book collects the abstracts and short papers presented at CLADAG 2021, the 13th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS). The meeting has been organized by the Department of Statistics, Computer Science, Applications 'G. Parenti' of the University of Florence, under the auspices of the University of Florence, the SIS and the International Federation of Classification Societies (IFCS).

CLADAG is a member of the IFCS, a federation of national, regional, and linguistically-based classification societies. It is a non-profit, non-political scientific organization, whose aims are to further classification research.

Every two years, CLADAG organizes a scientific meeting, devoted to the presentation of theoretical and applied papers on classification and related methods of data analysis in the broad sense. This includes advanced methodological research in multivariate statistics, mathematical and statistical investigations, survey papers on the state of the art, real case studies, papers on numerical and algorithmic aspects, applications in special fields of interest, and the interface between classification and data science. The conference aims at encouraging the interchange of ideas in the above-mentioned fields of research, as well as the dissemination of new findings. CLADAG conferences, initiated in 1997 in Pescara (Italy), were soon considered as an attractive information exchange market and became an important meeting point for people interested in classification and data analysis. A selection of the presented papers is regularly published in (post-conference) proceedings, typically by Springer Verlag.

The Scientific Committee of CLADAG 2021 conceived the Plenary and Invited Sessions to provide a fresh perspective on the state of the art of knowledge and research in the field. The scientific program of CLADAG 2021 is particularly rich. All in all, it comprises 5 Keynote Lectures, 26 Invited Sessions promoted by the members of the Scientific Program Committee, 10 Contributed Sessions, and a Plenary Session on *Statistical Issues in the COVID-19 Pandemic*. We thank all the session organizers for inviting renowned speakers, coming from many different countries. We are greatly indebted to the referees, for the time spent in a careful review of the abstracts and short papers collected in this book. The Conference was planned as an in presence event; unfortunately due to the persistent uncertainty of the COVID-19 epidemic condition, CLADAG 2021 will be completely online.

Special thanks are finally due to the members of the Local Organizing Committee and all the people who collaborated for CLADAG 2021. Last but not least, we thank all the authors and participants, without whom the conference would not have been possible.

Giovanni Camillo Porzio
Carla Rampichini
Chiara Bocci

Florence, September 2021

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

Keynote Speakers

Giovanni C. Porzio, University of Cassino and Southern Lazio, Italy, porzio@unicas.it, 0000-0003-1208-6991
Carla Rampichini, University of Florence, Italy, carla.rampichini@unifi.it, 0000-0002-8519-083X
Chiara Bocci, University of Florence, Italy, chiara.bocci@unifi.it, 0000-0001-8189-4445

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

OPTIMAL TRANSPORT METHODS FOR FAIRNESS IN MACHINE LEARNING

Jean-Michel Loubes[1]

[1] Université Toulouse Paul Sabatier, FRANCE
(e-mail: loubes@math.univ-toulouse.fr)

ABSTRACT: The principle of Supervised Machine Learning is to build a decision rule from a set of labeled examples called the learning sample, that fits the data. This rule becomes a model or a decision algorithm that will be used for all the population. Mathematical guarantees can be provided in certain cases to control the generalization error of the algorithm which corresponds to the approximation done by building the model based on the observations and not knowing the true model that actually generated the data set. More precisely, the data are assumed to follow an unknown distribution while only its empirical distribution is at hand. Yet potential existing bias, present in the learning sample, will be implicitly learnt and incorporated in the prediction. This leads to a potential amplification or generalization of bias that may create unfair decision rules. In this presentation we will present how optimal transport methods can be used to control the bias from machine learning algorithms. From a global point of view, group discrimination can be quantified by looking at the behaviour of the algorithm for different groups of individuals. This enables to measure the trade-off between the accuracy of the algorithm and the level of fairness using the notion of Wasserstein's barycenter. From an individual point, optimal transport methods provide an alternative way to define counterfactual worlds that explain how changes in some attributes of the individual may affect the decisions of an algorithm. This enables to recast the problem of training individually fair algorithms to ensuring regularity assumptions in both normal and counterfactual world.

CLASS OF MAPS FOR VISUALIZING CLASSIFICATION RESULTS

Peter J. Rousseeuw[1], Jacob Raymaekers[1] and Mia Hubert[1]

[1] Section of Statistics and Data Science, Dept of Mathematics, KU Leuven, Belgium, (e-mail: peter@rouseeuw.net, jakob.raymaekers@kuleuven.be, mia.hubert@kuleuven.be)

ABSTRACT: Classification is a major tool of statistics and machine learning. A classification method first processes a training set of objects with given classes (labels), with the goal of afterward assigning new objects to one of these classes. When running the resulting prediction method on the training data or on test data, it can happen that an object is predicted to lie in a class that differs from its given label. This is sometimes called label bias, and raises the question whether the object was mislabeled.

The proposed class map reflects how well an object lies within its class, by comparing to an alternative class as done in Rousseeuw (1987) for unsupervised classification. The class map also shows how far the object is from the other objects in its class, and whether some objects lie far from all classes. The goal is to visualize aspects of the classification results to obtain insight in the data.

The display is constructed for discriminant analysis, the k-nearest neighbor classifier, support vector machines, logistic regression, and coupling pairwise classifications. It is illustrated on several benchmark datasets, including some consisting of images and texts.

KEYWORDS: discriminant analysis, k-nearest neighbors, mislabeling, pairwise coupling, support vector machines.

References

RAYMAEKERS, J, & ROUSEEUW, P J. Transforming variables to central normality. *Machine Learning*.

RAYMAEKERS, J, ROUSEEUW, P J, & HUBERT, M. 2021. Class maps for visualizing classification results. *arXiv:2007.14495*.

ROUSEEUW, P J. 1987. Silhouettes: a graphical aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

UNDERSTANDING CROSS-VALIDATION AND PREDICTION ERROR

Robert Tibshirani[1], Stephen Bates[2] and Trevor Hastie[1]

[1] Departments of Statistics, and Biomedical Data Science, Stanford University, (e-mail: tibs@stanford.edu, hastie@stanford.edu)

[2] Departments of Statistics, and Electrical Engineering and Computer Sciences, UC Berkeley, (e-mail: stephenbates@cs.berkeley.edu)

ABSTRACT: Cross-validation is a widely-used technique to estimate prediction accuracy. However its properties are not that well understood. First, it is not clear exactly what form of prediction error is being estimated by cross-validation: one would like to think that cross-validation estimates the prediction error for the model and the data at hand. Surprisingly, we show here that this is not the case, (at least for the special case of linear models) and derive the actual estimand(s). This phenomenon occurs for most popular estimates of prediction error including data splitting, bootstrapping,C_p and AIC. Second, the standard (naïve) confidence intervals for prediction accuracy that are derived from cross-validation may fail to cover at the nominal rate, because each data point is used for both training and testing, inducing correlations among the measured accuracy for each fold. As a result, the variance of the CV estimate of error is larger than suggested by naïve estimators, which leads to confidence intervals for prediction accuracy that can have coverage far below the desired level. We introduce a nested cross-validation scheme to estimate the standard error of the cross-validation estimate of prediction error, showing empirically that this modification leads to intervals with approximately correct coverage in many examples where traditional cross-validation intervals fail.

QUANTILE-BASED CLASSIFICATION

Cinzia Viroli[1]

[1] Department of Statistical Sciences, University of Bologna,
(e-mail: `cinzia.viroli@unibo.it`)

ABSTRACT: The idea of using quantiles in classification is relatively recent. The median classifier for high-dimensional problems (Hall *et al.*, 2009), the quantile classifier (Hennig & Viroli, 2016); the ensemble and the directional quantile classifiers (Lai & McLeod, 2020; Farcomeni *et al.*, 2021) represent main relevant proposals for supervised classification. These ideas proved to perform well for high dimensional and skewed data compared to other classical classification strategies. For clustering purposes quantiles lead to analogues appealing advantages. In this context, K-quantiles have been recently introduced (Hennig *et al.*, 2019). In this talk the main quantile-based strategies for supervised and unsupervised classification will be presented and discussed, both from the theoretical and empirical points of view.

KEYWORDS: L1 distance, supervised and unsupervised classification, k-means, skewness, high-dimensional data

References

FARCOMENI, A., GERACI, M., & VIROLI, C. 2021. *Directional quantile classifiers*.

HALL, P., TITTERINGTON, D. M., & XUE, J.-H. 2009. Median-Based Classifiers for High-Dimensional Data. *Journal of the American Statistical Association*, **104**(488), 1597–1608.

HENNIG, C., & VIROLI, C. 2016. Quantile-based classifiers. *Biometrika*, **103**(2), 435–446.

HENNIG, C., VIROLI, C., & ANDERLUCCI, L. 2019. Quantile-based clustering. *Electronic Journal of Statistics*, **13**(2), 4849–4883.

LAI, Y., & MCLEOD, I. 2020. Ensemble quantile classifier. *Computational Statistics & Data Analysis*, **144**, 106849.

VERIDICAL DATA SCIENCE FOR RESPONSIBLE AI: CHARACTERIZING V4 NEURONS THROUGH DEEPTUNE

Bin Yu[1]

[1] Departments of Statistics, and Electrical Engineering and Computer Sciences, UC Berkeley (e-mail: binyu@berkeley.edu)

> "A.I. is like nuclear energy – both promising and dangerous"
>
> Bill Gates, 2019

ABSTRACT: Data Science is a pillar of A.I. and has driven most of recent cutting-edge discoveries in biomedical research. In practice, Data Science has a life cycle (DSLC) that includes problem formulation, data collection, data cleaning, modeling, result interpretation and the drawing of conclusions. Human judgement calls :wq:ware ubiquitous at every step of this process, e.g., in choosing data cleaning methods, predictive algorithms and data perturbations. Such judgment calls are often responsible for the "dangers" of A.I. To maximally mitigate these dangers, we developed a framework based on three core principles: Predictability, Computability and Stability (PCS). Through a workflow and documentation (in R Markdown or Jupyter Notebook) that allows one to manage the whole DSLC, the PCS framework unifies, streamlines and expands on the best practices of machine learning and statistics – bringing us a step forward towards veridical Data Science.

The PCS framework will be illustrated through the development of the DeepTune framework for characterizing V4 neurons. DeepTune builds predictive models using DNNs and linear regression and applies the stability principle to obtain stable interpretations of 18 predictive models.

Finally, a general DNN interpretation method based on contextual decomposition (CD) will be discussed with applications to sentiment analysis and cosmological parameter estimation.

References

ABBASI-ASL, R., CHEN, Y., BLONIARZ, A., OLIVER, M., WILLMORE, B.

D. B., Gallant, J. L., & Yu, B. 2018. The DeepTune framework for modeling and characterizing neurons in visual cortex area V4. *bioRxiv*.

Yu, B., & Kumbier, K. 2020. Veridical data science. *Proceedings of the National Academy of Sciences*, **117**(8), 3920–3929.

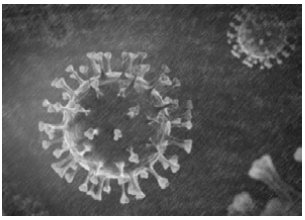
Plenary Session

Giovanni C. Porzio, University of Cassino and Southern Lazio, Italy, porzio@unicas.it, 0000-0003-1208-6991
Carla Rampichini, University of Florence, Italy, carla.rampichini@unifi.it, 0000-0002-8519-083X
Chiara Bocci, University of Florence, Italy, chiara.bocci@unifi.it, 0000-0001-8189-4445

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

Statistical Issues in the COVID-19 Pandemic



ORGANIZER
J. Sunil Rao
Chair
Division of Biostatistics,
University of Miami, USA

ABSTRACT: COVID-19 has become a pandemic of epic proportion, calling on scientific enquiry from a broad range of disciplines, including biology, chemistry, pharmacology, epidemiology, mathematics, statistics, and data science, among others. As a result, potential solutions to this problem have become highly interdisciplinary. Notwithstanding, statistics and data science have become paramount in the quest for providing evidentiary based answers to a host of scientific problems associated with this novel virus. Among these problems are the issues of vaccine development, development of therapeutics, testing, contract tracing, forecasting, and inferential analysis.

The effects of the virus have varied greatly from country to country reflecting differences in data reporting, public health infrastructure, politics, economics, social contexts and the role of civil society. This session will discuss specific statistical issues related to the COVID-19 pandemic and will bring together prominent researchers who will share their experiences from Israel to India to the US.

SPEAKERS

Daniel Diaz
Research Assistant Professor
Division of Biostatistics, University of Miami, USA

Jeffrey S. Morris
Chair
Division of Biostatistics, University of Pennsylvania, USA

Bhramar Mukherjee
Chair
Department of Biostatistics, University of Michigan, USA

Danny Pfeffermann
Chief Statistician of Israel and Professor of Statistics
Hebrew University of Jerusalem, ISRAEL & University of Southampton, UK

A SIMPLE CORRECTION FOR COVID-19 SAMPLING BIAS

Daniel Diaz[1]

[1] Division of Biostatistics, University of Miami, USA, (e-mail: ddiaz3@miami.edu)

ABSTRACT: COVID-19 testing has become a standard approach for estimating prevalence which then assist in public health decision making to contain and mitigate the spread of the disease. The sampling designs used are often biased in that they do not reflect the true underlying populations. For instance, individuals with strong symptoms are more likely to be tested than those with no symptoms. This results in biased estimates of prevalence (too high). Typical post-sampling corrections are not always possible. Here we present a simple bias correction methodology derived and adapted from a correction for publication bias in meta analysis studies. The methodology is general enough to allow a wide variety of customization making it more useful in practice. Implementation is easily done using already collected information. Via a simulation and two real datasets, we show that the bias corrections can provide dramatic reductions in estimation error.

A Seat at the Table: The Key Role of Biostatistics and Data Science in the COVID-19 Pandemic

Jeffrey Morris[1]

[1] Division of Biostatistics, University of Pennsylvania, USA,
(e-mail: jeffrey.morris@pennmedicine.upenn.edu)

Abstract: The novel virus SARS-CoV-2 has produced a global pandemic, forcing doctors and policymakers to "fly blind", trying to deal with a virus and disease they knew virtually nothing about. Sorting through the information in real time has been a daunting process—processing data, media reports, commentaries, and research articles. In the USA this is exacerbated by an ideologically divided society that has difficulty with mutual trust, or even agreement on common facts. The skills underlying our statistical profession are central to this knowledge discovery process, filtering out biases, aggregating disparate data sources together, dealing with measurement error and missing data, identifying key insights while quantifying the uncertainty in these insights, and then communicating the results in an accessible balanced way. As a result, we have had a central role to play in society to bring our perspective and expertise to bear on the pandemic to help ensure knowledge is efficiently discovered and put into practice. Unfortunately, our profession is often shy about asserting its perspective in broader societal ventures, perhaps not realizing the central importance of our perspective and mindset. I have authored a website and blog covid-datascience.com that represents my own person efforts to disseminate information I have found reliable and insightful regarding the pandemic, accounting for subtle scientific and data analytical issues and uncertainties about our current knowledge, and seeking to filter out political and other subjective biases.

Using experiences with the covid-datascience blog as a backdrop, I will highlight how statistical and data scientific issues have been central in understanding the emerging knowledge in the pandemic. I will discuss various broad issues I have seen impede the knowledge discovery process, including subjective bias causing individuals to ignore some information and magnify others, viral misinformation spread on social media platforms, danger of rushed and inadequately reviewed scientific studies, conflating of political concerns and scientific messaging, and incomplete and messaging from scientific leaders to

the broader community. I will discuss these concepts in various specific contexts, including identification of key modes of spread and effective mitigation strategies, vaccine safety and efficacy, durability of immune protection and risk of reinfections or breakthrough infections, and the emergence of variants of concern and how this affects the pandemic moving forward. I will finish with a call to urge statisticians to seek greater visibility and engagement with the media and policymakers to ensure our understanding of quantitative nuances is reflected in important societal-level decisions and dissemination of emerging scientific knowledge.

PREDICTIONS, ROLE OF INTERVENTIONS AND THE CRISIS OF VIRUS IN INDIA: A DATA SCIENCE CALL TO ARMS

Bhramar Mukherjee[1]

[1] Department of Biostatistics, University of Michigan, USA,
(e-mail: bhramar@umich.edu)

ABSTRACT: India, the world's largest democracy with 1.38 billion people, underwent five phases of national lockdown from March 25-June 30, 2020 and several phases of unlocking in Wave 1 of the COVID-19 pandemic. The virus curve turned the corner in mid-September of 2020 and it appeared that India could avoid a second resurgence in the Winter. Normalcy returned to the life of Indian people and vaccination had a sluggish start nationwide. Several hypotheses were being postulated for this miraculous recovery of India including that of herd immunity as implied by some serosurveys. Then came an astronomic wave 2 for India, where the daily case counts reached more than 400000 and daily death counts peaked around 4500. In this presentation, we provide a brief chronicle of the modeling experience of our study team over the last one year trying to understand the pandemic in India and explain what caused this devastating second wave, including the role of the Delta variant. We discuss methodological innovations by incorporating imperfect viral testing when using case-counts in an extended SEIR model for COVID-19. We use this model to estimate the unobserved infections and deaths leading to an estimate of the infection fatality rates in India for Waves 1 and 2 . This is joint work with many, with all supporting research materials and products available at covind19.org.

CONTRIBUTIONS OF ISRAEL'S CBS TO ROUT COVID-19

Danny Pfeffermann[1]

[1] Central Bureau of Statistics and Hebrew University of Jerusalem, Israel; University of Southampton, UK, (e-mail: D.Pfeffermann@soton.ac.uk)

ABSTRACT: In this presentation, I shall describe the major problems that the Central Bureau of Statistics in Israel (ICBS) had faced during the pandemic, discuss the methodological issues involved and how we dealt with them. Issues considered are lack of health data; performing special household, business and serological surveys; accounting for NMAR nonresponse; publication of flash estimates; estimation of excess mortality; seasonal adjustment, trend estimation and weighting of CPI items in a year of pandemic.

Invited Papers

Giovanni C. Porzio, University of Cassino and Southern Lazio, Italy, porzio@unicas.it, 0000-0003-1208-6991
Carla Rampichini, University of Florence, Italy, carla.rampichini@unifi.it, 0000-0002-8519-083X
Chiara Bocci, University of Florence, Italy, chiara.bocci@unifi.it, 0000-0001-8189-4445

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

ROBUST ISSUES IN ESTIMATING MODELS FOR MULTIVARIATE TORUS DATA

Claudio Agostinelli [1], Giovanni Saraceno [1] and Luca Greco [2]

[1] Department of Mathematics, University of Trento, (e-mail: claudio.agostinelli@unitn.it, giovanni.saraceno@unitn.it)

[2] University Giustino Fortunato, Benevento (e-mail: l.greco@unifortunato.eu)

ABSTRACT: We consider the problem of robust fitting for statistical models applied to multivariate torus data, e.g., data which are multivariate angles. We discuss two different definitions of outliers, "geometric" and "probabilistic" outliers, and the proposed robust methods to cope with them. We mainly focus on multivariate wrapped models together with some computational aspects.

KEYWORDS: circular data, multivariate torus data, outlier detection, robust estimation, wrapped models

1 Introduction

Multivariate circular data arise commonly in many different fields. Depending on the situation, observations can be thought as points on the surface of a hyper-sphere ($\mathbb{S}^{p-1}$) or as points on the surface of a torus ($\mathbb{T}^p = [0, 2\pi)^p$). While the first problem is well studied in literature, the latter received much less attention, even though it is more common. Here, we review some aspects of robust fitting of torus data according to wrapped models. The peculiarity of multivariate torus data is periodicity, that reflects in the boundedness of the sample space and often of the parametric space. Indeed, it is challenging to introduce the *geometric* concept of outliers, as points that are far from the bulk of the data. However, it is always possible to define circular outliers from a *probabilistic* point of view, as points that are unlikely to occur under the assumed model. Notice that outliers are model dependent, since they are defined with respect to the specified model. A first general attempt to develop a robust parametric technique for multivariate torus data can be found in Saraceno *et al.*, 2021 where a weighted likelihood estimator is introduced and outliers are defined using the probabilistic point of view. In contrast, Greco *et al.*, 2021 develop robust estimators based on S/M/MM-estimators as well as weighted likelihood estimators considering the geometric approach.

2 Wrapped models

Let $\mathbf{X}$ be a multivariate random variable with model density $m(\mathbf{x};\theta)$ on $\mathbb{R}^p$ parameterized by $\theta \in \Theta$. We can construct a wrapped model by $\mathbf{Y} = \mathbf{X} \bmod 2\pi$ where the mod operator is performed component-wise. The density function of $\mathbf{Y}$ takes the form of an infinite sum over $\mathbb{Z}^p$ given by

$$m^\circ(\mathbf{y};\theta) = \sum_{\mathbf{j}\in\mathbb{Z}^p} m(\mathbf{y}+2\pi\mathbf{j};\theta) \ .$$

A good approximation, denoted as m°_J, can be obtained, in most cases, with only few terms of the summation, so that $\mathbb{Z}^p$ is replaced by $\mathcal{C}_\mathcal{J} = \otimes_{s=1}^p \mathcal{J}$ where $\mathcal{J} = (-J, -J+1, \ldots, 0, \ldots, J-1, J)$ for some fixed J. The support of $\mathbf{Y}$ is bounded and given by $[0,2\pi)^p$, for convenience, and the parametric space Θ might be restricted as well to ensure identifiability. The $p-$dimensional vector $\mathbf{j}$ represents the wrapping coefficients vector, that is, it indicates how many times each component of the $p-$toroidal data point has been wrapped. Given a sample $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$, the approximated log-likelihood function is given by

$$\ell(\theta) = \sum_{i=1}^n \log m^\circ_J(\mathbf{y}_i;\theta) = \sum_{i=1}^n \log \sum_{\mathbf{j}\in\mathcal{C}_\mathcal{J}} m(\mathbf{y}_i+2\pi\mathbf{j};\theta) \ .$$

Assuming that we could observe the vectors $\mathbf{j}_i$ $(i = 1,\ldots,n)$, then we would have access to the unwrapped and unobserved sample $\hat{\mathbf{x}}_i = \mathbf{y}_i + 2\pi\mathbf{j}_i$. This leads to the following log-likelihood

$$\ell_C(\theta) = \sum_{i=1}^n \log m(\hat{\mathbf{x}}_i;\theta) = \sum_{i=1}^n \log m(\mathbf{y}_i+2\pi\mathbf{j}_i;\theta) = \sum_{i=1}^n \sum_{\mathbf{j}\in\mathcal{C}_\mathcal{J}} v_{i\mathbf{j}} \log m(\mathbf{y}_i+2\pi\mathbf{j};\theta) \ ,$$

where $\mathbf{v}_{i\mathbf{j}} = 1$ or $\mathbf{v}_{i\mathbf{j}} = 0$ according to whether $\mathbf{y}_i$ has $\mathbf{j} \in \mathcal{C}_\mathcal{J}$ as the wrapping coefficient vector and now the $\mathbf{j}_i$s are additional unknown parameters needed to be estimated. Optimization of the above log-likelihood can be performed naturally through a Classification-Expectation-Maximization algorithm, see Nodehi *et al.*, 2021 for more details. Hereafter, we concentrate on unimodal and elliptically symmetric densities m, i.e., given a strictly decreasing and non-negative function h and set $\theta = (\mu, \Sigma)$ for a location vector parameter μ and dispersion matrix Σ, then $m(\mathbf{x};\theta) \propto |\Sigma|^{-1/2} h\left((\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$.

3 Outliers in multivariate torus data

Consider $0 \leq \varepsilon < 0.5$ and an arbitrary distribution $g(\mathbf{x})$ in $\mathbb{R}^p$. According to the usual gross error model, the true density $f(\mathbf{x})$ of the data is given by

$f(\mathbf{x}) = (1-\varepsilon)m(\mathbf{x};\mu,\Sigma) + \varepsilon g(\mathbf{x})$ and hence the corresponding wrapped density would have the form

$$\begin{aligned} f^{\circ}(\mathbf{y}) &= (1-\varepsilon)\sum_{\mathbf{j}\in\mathbb{Z}^p} m(\mathbf{y}+2\pi\mathbf{j};\mu,\Sigma) + \varepsilon \sum_{\mathbf{j}\in\mathbb{Z}^p} g(\mathbf{y}+2\pi\mathbf{j}) \qquad (1) \\ &= (1-\varepsilon)m^{\circ}(\mathbf{y};\mu,\Sigma) + \varepsilon g^{\circ}(\mathbf{y}). \qquad (2) \end{aligned}$$

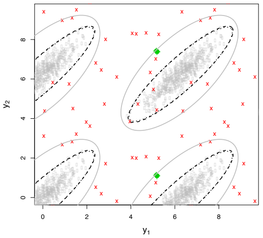
If we instead consider the approach leading to $\ell_C(\mu,\Sigma)$ and equation (1), for a given observation $\mathbf{y}_i$ we have

$$f^{\circ}(\mathbf{y}_i) \approx (1-\varepsilon)m(\mathbf{y}_i+2\pi\mathbf{j}_i;\mu,\Sigma) + \varepsilon g(\mathbf{y}_i+2\pi\mathbf{j}_i)$$

which suggests the classical geometric definition of outliers. In such cases, the degree of outlyingness of an observation is based on some "geometric" distance, e.g., the squared Mahalanobis distance. In contrast, we can define outliers directly on the torus, that is, according to equation (2), based on a "probabilistic" distance [Markatou *et al.*, 1998 and Agostinelli, 2007] where we compare the *true* density $f^{\circ}(\mathbf{y}_i)$ with the model density $m^{\circ}(\mathbf{y}_i;\mu,\Sigma)$. A measure of the agreement is provided by the finite sample Pearson residual function [Lindsay, 1994 and Markatou *et al.*, 1998], defined as $\delta_n(\mathbf{y}) = \frac{\hat{f}_n(\mathbf{y})}{\hat{m}(\mathbf{y};\theta)} - 1$ where $\hat{f}_n(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} k(\mathbf{y};\mathbf{y}_i,h)$ is a non-parametric kernel density estimate (with kernel function k and bandwidth h) of the true density $f(\mathbf{y})$ and $\hat{m}(\mathbf{y};\mu,\Sigma) = \int k(\mathbf{y};\mathbf{t},h)m(\mathbf{t};\mu,\Sigma)\, d\mathbf{t}$ is a smoothed version of the model density.

4 Example

Here, we illustrate the behavior of the robust estimators introduced in Saraceno *et al.*, 2021 and Greco *et al.*, 2021 using a simulated example. We point the reader to the cited papers for full details. The bulk of data has been drawn from a bivariate wrapped normal distribution with $\mu = 0$, $\Sigma = D^{1/2}RD^{1/2}$ where R is a random correlation matrix and $D = diag(\sigma\mathbf{1}_2)$ with $\sigma = \pi/4$. The sample size is $n = 500$ with 10% of contamination. Two types of outlying observations are considered: scattered and point-mass. It is suggested to represent circular data points after they have been unwrapped on a "flat" torus in the form $\mathbf{x} = \mathbf{y} + 2\pi\mathbf{j}$ for $\mathbf{j} \in \mathcal{C}_{\mathcal{J}}$. The figure shows the unwrapped bivariate points (grey points), the scattered (red crosses) and the point-mass (green plus) outliers. The bivariate fitted models are given in the form of ellipses based on the 0.99-level quantile of a χ^2_2 distribution. We show the results obtained using maximum likelihood estimator (grey line) and the proposed robust estimators. In particular, we



	$AS(\hat{\mu})$	$\Delta(\hat{\Sigma})$
MLE	0.001478	2.096517
probabilistic	0.000491	0.002610
geometric	0.000571	0.005987

consider the robust estimators based on the weighted likelihood technique, implemented according to geometric (dotted line) and probabilistic (dashed line) outliers. Finally, the table gives some measures of fitting accuracy.

References

Agostinelli, C. 2007. Robust Estimation for Circular Data. *Computational Statistics and Data Analysis*, **51**(12), 5867–5875.

Greco, L., Saraceno, G., & Agostinelli, C. 2021. Robust Fitting of a Wrapped Normal Model to Multivariate Circular Data and Outlier Detection. *Stats*, **4**(2), 454–471.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. 1986. *Robust Statistics: The Approach based on Influence Functions*. Wiley.

Hawkins, D. 1980. *Identification of Outliers*. Chapman & Hall.

He, X. 1992. Robust Statistics of Directional Data: A Survey. *Nonparametric Statistics and Related Topics*, 87–95.

Lindsay, B.G. 1994. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *The Annals of Statistics*, **22**, 1018–1114.

Markatou, M., Basu, A., & Lindsay, B. G. 1998. Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, **93**(442), 740–750.

Nodehi, A., Golalizadeh, M., Maadooliat, M., & Agostinelli, C. 2021. Estimation of parameters in multivariate wrapped models for data on a p-torus. *Computational Statistics*, **36**, 193–215.

Saraceno, G., Agostinelli, C., & Greco, L. 2021. Robust Estimation for Multivariate Wrapped Models. *Metron*. To appear.

BAYESIAN NONPARAMETRIC DYNAMIC MODELING OF PSYCHOLOGICAL TRAITS

Emanuele Aliverti [1]

[1] Department of Economics, University Ca' Foscari Venezia, (e-mail: emanuele.aliverti@unive.it)

ABSTRACT: This work focuses on investigating the evolution of different traits of psychosis during the COVID-19 pandemic. We develop a Bayesian nonparametric mixture model for multivariate categorical data, which characterizes the population' psychosis via a set of latent psychological profiles. Leveraging a time- and covariate-dependent stick-breaking construction for the mixture weights, the proposed specification characterizes the dynamic evolution of such latent traits across the pandemic, measuring the effect of subject-specific demographic information such as sex and age of the individuals.

KEYWORDS: Bayesian nonparametrics, categorical data, dynamic modeling, stick-breaking.

1 Introduction

Multivariate categorical data are routinely collected in a variety of applications (e.g., Agresti, 2003). Some common examples include surveys on opinions and feelings, where individuals are asked to fill in questionnaires reporting their level of agreement with different categorical items. This abundance of data has motivated a large literature on statistical models for high-dimensional categorical data, with penalized log-linear models (Nardi et al., 2012) and latent-structures (Lazarsfeld, 1950) being particularly popular in the literature (Aliverti and Dunson, 2020).

When the number of categorical variables increases, the number of free cells in the resulting contingency tables becomes extremely sparse, motivating novel approaches to provide compact representation of the observed structures. Bayesian nonparametric models are particularly appealing for this goal, leveraging on flexible specifications which adapt to the complexity of the observed data, characterizing uncertainty in a rigorous way (e.g., Dunson and Xing, 2009; Müller et al., 2015).

In this talk, we illustrate a Bayesian nonparametric dynamic model for the evolution of the population' psychosis during the COVID-19 pandemic.

According to the proposed model, a set of latent profiles characterizes the population-specific response patterns, while the individual propensity toward a specific profile is allowed to change in time and with subject-specific covariates, leveraging on a dependent stick-breaking construction for the mixture weights.

We illustrate the details of the proposed methodology and its application on the Italian population. Our empirical findings focus on the evolution of the psychosis across the pandemic and on the estimated sub-regional differences in terms of the impact of COVID-19 pandemic on the individual's psychology.

References

Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

Aliverti, E. and Dunson, D. B. (2020). Composite mixture of log-linear models for categorical data. *arXiv preprint arXiv:2004.01462*.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, pages 362–412.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer.

Nardi, Y., Rinaldo, A., et al. (2012). The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli*, 18(3):945–974.

Clustering financial time series using generalized cross correlations

Andrés M. Alonso[1], Carolina Gamboa[1] and Daniel Peña[1]

[1] Department of Statistics, Universidad Carlos III de Madrid, Spain.
(email: andres.alonso@uc3m.es, 100312917@alumnos.uc3m.es), daniel.pena@uc3m.es)

Abstract: In this paper we propose a procedure for clustering financial time series using the generalized cross correlations (GCC) between the estimated volatilities and the squared residuals of ARMA(p,q) models. Monte Carlo experiments are carried out to analyze the performance of the proposed procedure. We show that the procedure is able to recover the original clustering structures in all cases studied. Finally, the methodology is applied to a set of real data.

Keywords: unsupervised classification, dependence measure, conditional variance.

1 Introduction

A variety of methods have been proposed in the literature to cluster time series (see Caiado *et al.*, 2015 and the references cited there). In those methods the clustering problem is solved using two different strategies: the first one works directly on the original time series by defining an appropriate metric; in the second one, time series are projected in a smaller space of features or parameters. These methods are useful when the time series are independent, however, in many applications the assumption of independence does not hold. Few articles have proposed method for clustering by dependency. Zhang & An, 2018 proposed a distance measure based on copulas to measure general dependence of the time series. Alonso & Peña, 2019 introduced the generalized cross correlation metric based on all the cross correlations between two time series until a certain lag, k.

These two methods assumed that the dependency among the time series is on the levels, and do not consider the case in which the dependency is on the conditional variances. This fact is important in many fields. For example, in financial time series where asset returns do not present a strong structure in the levels but do present it in the volatility. Some studies have taken into account the similarity of the evolution of the conditional variances (see Otranto, 2008 and D'Urso *et al.*, 2013 for GARCH models).

In this work we study a procedure to cluster time series for dependency on the conditional variability, integrating the concepts of dependency and heteroscedasticity of a set of time series. We extend the results presented with Alonso & Peña, 2019 in the search for dependencies between the squares of two time series or between their estimated volatilities. In Section 2, we present the new methodology and Section 3 we illustrate its use in a real data example. Some Monte Carlo experiments are available upon request to the authors.

2 Clustering time series by volatility dependency

Let w_t and z_t be two stationary time series and let $x_t = w_t^2$, $y_t = z_t^2$ be their corresponding squares, that will also be stationary. Using the results given in Alonso & Peña, 2019, we are going to define a linear dependence measure between (x_t, y_t). We calculate the autocorrelations of x_t and y_t, $\rho_x(h)$ and $\rho_y(h)$, and the cross correlations between x_t and y_t, $\rho_{xy}(h)$, for lags $h = 0, \pm 1, \cdots, \pm k.$ The linear dependency between the two time series of squares can be summarized in the matrix

$$\mathbf{R}_k = \begin{pmatrix} \mathbf{R}(0) & \mathbf{R}(1) & \dots & \mathbf{R}(k) \\ \mathbf{R}(-1) & \mathbf{R}(0) & \dots & \mathbf{R}(k-1) \\ \vdots & \vdots & \dots & \vdots \\ \mathbf{R}(-k) & \mathbf{R}(-k+1) & \dots & \mathbf{R}(0) \end{pmatrix}, \tag{1}$$

where $\mathbf{R}(h) = \begin{pmatrix} \rho_x(h) & \rho_{xy}(h) \\ \rho_{yx}(h) & \rho_x(h) \end{pmatrix}$. Matrix $\mathbf{R}_k$ corresponds to the correlation matrix of the stationary process $(x_t, y_t, x_{t-1}, y_{t-1}, \dots, x_{t-k}, y_{t-k})^T$

The *generalized correlation coefficient* is defined using matrix $\mathbf{R}_k$ by

$$GCC(x_t, y_t) = 1 - \left(\frac{\det(\mathbf{R}_{yx,k})}{\det(\mathbf{R}_{xx,k}) \det(\mathbf{R}_{yy,k})} \right)^{1/(k+1)}, \tag{2}$$

where $\mathbf{R}_{xx,k}$ and $\mathbf{R}_{yy,k}$ are the correlation matrices for the $X_{t,k}$ and $Y_{t,k}$, respectively, and $\mathbf{C}_{xy,k}$ the matrix of cross-correlations between these two vectors.

This similarity measure $GCC(x_t, y_t)$ satisfies the following properties: (1) $GCC(x_t, y_t) = GCC(y_t, x_t)$; (2) $0 \leq GCC(y_t, x_t) \leq 1$, it takes the zero value in the case that the dependence between both variables is perfectly linear, and take the value one in the case that all cross correlations are zero. Based on this measure we define the dissimilarity between x_y and y_t as $d(x_t, y_t) = 1 - GCC(x_t, y_t)$, in that way, high dissimilarity values are associated to weak dependence and values close to zero will be related to strong dependence. Once the pairwise dissimilarities between time series are obtained, we can apply any clustering method that uses dissimilarity matrices as input.

3 Real data example

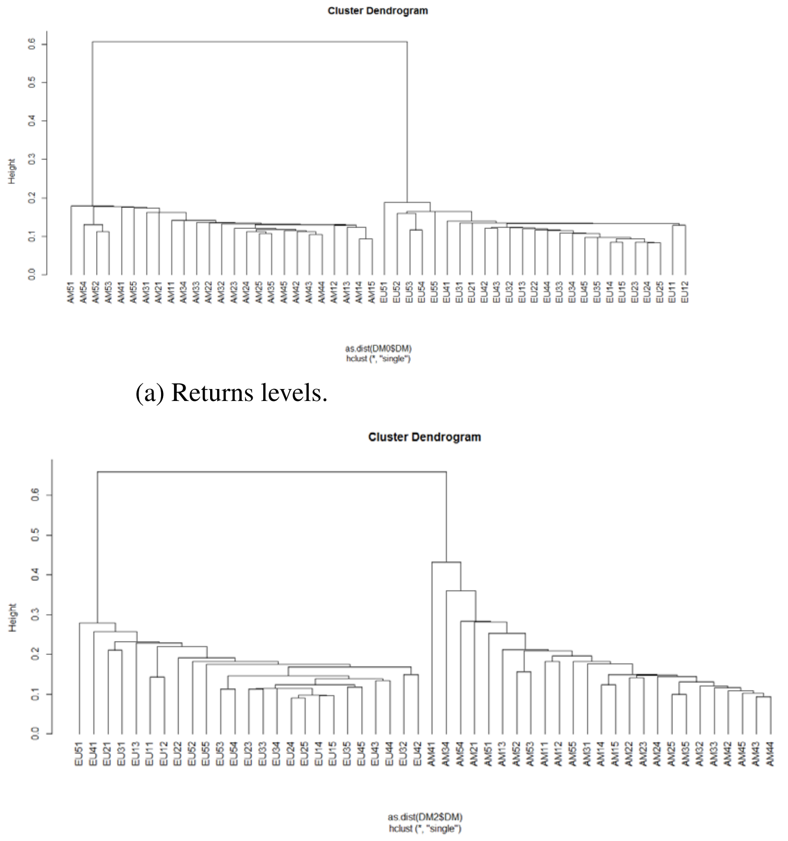
In this section we are going use the set of the portfolios designed by Kenneth R. French, which contains daily and equal-weighted returns of firms listed on the NYSE, AMEX, or NASDAQ. The portfolios are constructed based on different criteria such as companies size, book/market ratio, company capitalization and/or industry classification. See at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. We will analyze 100 time series that contains 25 portfolios based market equity (ME) and the ratio of book equity to market equity (BE/ME) for European (UE), Japanese (JAP), Pacific Asian (PA) -except Japan- and North American (AM) markets.

First, we obtain the dendrogram using single linkage and dissimilarity, $d(w_t, z_t)$, for the levels of daily returns. Silhouette statistic suggests four clusters which corresponds to the four regions analyzed. In addition, the series that belong to each clusters present a strong dependence between them except the Asian time series. When we use the squares of daily returns for clustering, Silhouette statistic finds five clusters: the same four groups as in the levels and a fifth group with a single time series belongs to Pacific Asia. Also it is observed that the group of American time series presents weaker dependencies than those observed in levels. In Figure 1, we show that the dependency structures based on levels differs from the ones based on squared returns. In particular, it is remarkable that portfolios AM12, AM13, AM14, AM15 make up a group of dependent series on the levels, however, this group is divided when squares are taken into account, the same is true for the group of portfolios AM52, AM53, AM54.

Acknowledgements: The authors gratefully acknowledge the financial support from the Spanish government Agencia Estatal de Investigación (PID2019-108311GB-I00 / AEI / 10.13039/501100011033).

References

ALONSO, A.M., & PEÑA, D. 2019. Clustering time series by linear dependency. *Statistics and Computing*, **29**, 655–676.



(a) Returns levels.

(b) Squared returns levels.

Figure 1: Dendrograms for American and European portfolios.

CAIADO, J., MAHARAJ, E.A., & D'URSO, P. 2015. Time-Series Clustering. *Pages 262–285 of: Handbook of Cluster Analysis*. Chapman and Hall/CRC.

D'URSO, P., CAPPELLI, C., DI LALLO, D., & MASSARI, R. 2013. Clustering of financial time series. *Physica A: Statistical Mechanics and its Applications*, **392**, 2114–2129.

OTRANTO, E. 2008. Clustering heteroskedastic time series by model-based procedures. *Computational Statistics & Data Analysis*, **52**, 4685–4698.

ZHANG, B., & AN, B. 2018. Clustering time series based on dependence structure. *PloS One*, **13**, e0206753.

Model-based clustering for categorical data via Hamming distance

Raffaele Argiento[1], Edoardo Filippi-Mazzola[2] and Lucia Paci[1]

[1] Università Cattolica del Sacro Cuore,
(e-mail: raffaele.argiento@unicatt.it, lucia.paci@unicatt.it)

[2] Università della Svizzera Italiana, (e-mail: edoardo.filippi-mazzola@usi.ch)

Abstract: In this work a model-based approach for clustering categorical data with no natural ordering is introduced. The proposed method exploits the Hamming distance to define a family of probability mass functions to model categorical data. The elements of this family are considered as kernels of a finite mixture model with unknown number of components. Fully Bayesian inference is provided using a sampling strategy based on a trans-dimensional blocked Gibbs-sampler, facilitating computation with respect to the customary reversible-jump algorithm. Model performances are assessed via a simulation study, showing improvements both in terms of prediction and estimation, with respect to existing approaches. Finally, our method is illustrated with application to reference datasets.

Keywords: Hamming distribution, mixture modelling, categorical data analysis, blocked Gibbs Sampling

Mining multiple time sequences through Co-clustering algorithms for Distributional data

Balzanella Antonio [1], Irpino Antonio[1] and Francisco T. de A. de Carvalho[2]

[1] Department of Mathematics and Physics, University of Campania "Luigi Vanvitelli", (e-mail: antonio.balzanella@unicampania.it, antonio.irpino@unicampania.it)

[2] CIN-UFPE, Av. Jornalista Anibal Fernandes, s/n - Cidade Universitria 50.740-560, Recife, PE, Brasil, (e-mail: fatc@cin.ufpe.br)

Abstract: This paper deals with the co-clustering of distributional data applied to multiple time sequences. The aims are: to get a double-partition of data into clusters of units and variables; to summarize the main concepts in the data through histogram prototypes; to overview the evolution over time of the monitored phenomenon. We extend the double k-means algorithm to handle distributional data by using the L_2 Wasserstein distance for comparing distributions. Moreover, we adapt double k-means algorithm to compute optimal relevance weights associated with the variables.

Keywords: co-clustering, distribution data, Wasserstein distance.

1 Introduction

In recent years, several authors (Arroyo & Maté, 2009; Balzanella & Irpino, 2020) have proposed summarizing temporal sequences by a set of distributions. In particular, they assume that the time domain of the sequences is split into non-overlapping time windows, and the distribution of the records, framed by each window, is estimated through histograms or kernel density estimators. This summarization allows us to retain most of the information regarding the monitored phenomenon, and to perform dimensionality reduction.

In this framework, we consider co-clustering of distribution data with the following objectives: 1) To summarize the main concepts in the data through histogram prototypes; 2) To reorganize the initial matrix into a block matrix; 3) to overview the evolution over time of the monitored phenomenon through the partition of the variables; 4) To evaluate the contribution of various periods (intervals of time) to the optimal partitioning by considering the weights of the variables; 5) to obtain a partition of the series so that groups of series that record similar data over time can be discovered.

We use a co-clustering approach (de A.T. De Carvalho et al. , 2021) that extends the classic alternated double k-means. It performs double partition of objects and variables to simultaneously discover blocks of subsets of the rows and columns of a data table according to a homogeneity criterion. We use two variants of this algorithm: the distributional double k-means (DDK) and the adaptive distributional double k-means (ADDK). The main difference between the two algorithm is that only ADDK computes the relevance weight for each variable. In both the variants, the internal variability of clusters or co-clusters is measured by the Wasserstein-based sum of squared errors (Irpino & Verde, 2015).

2 Distributional Double k-means (DDK) and the Adaptive Distributional Double k-means (ADDK)

Let us consider a set of N objects observed on P Distributional variables (DV). A DV takes as values one-dimensional theoretical or empirical (i.e., histograms) probability density functions.

The objects are indexed by i (with $i = 1, \dots, N$), the P variables are denoted by Y_j (with $j = 1, \dots, P$), and the $i-th$ one-dimensional distribution data (DD) of the Y_j variable is denoted by y_{ij}. The vector $\mathbf{y}_i = [y_{i1}, \dots, y_{iP}]$ contains the description of the $i-th$ object on the P DVs. Considering y_{ij} an empirical probability density function, we refer to Q_{ij} as the quantile function (qf), that is, the inverse of the cdf.

We use the the squared L_2 Wasserstein metric d_W^2 between the DD y_{ij} and $y_{i'j}$, with support in $\Re$, defined as: $d_W(y_{ij}, y_{i'j}) = \sqrt{\int\limits_0^1 \left[Q_{ij}(t) - Q_{i'j}(t)\right]^2 dt}$

In order to consider the relevance of each variable we use the following notion of adaptive distances based on the squared L_2 Wasserstein distance. Let us consider a vector of positive weights $\Lambda = [\lambda_1, \dots, \lambda_P]$. According to (De Carvalho & Lechevallier, 2009), a general expression for the adaptive (squared) L_2 Wasserstein distance is:

$$d_W^2(\mathbf{y}_i, \mathbf{y}_{i'} | \Lambda) = \sum_{j=1}^{P} \lambda_j d_W^2\left(y_{ij}, y_{i'j}\right) \tag{1}$$

with $\lambda_j > 0 \, \forall j$ and $\prod_{j=1}^{P} \lambda_j = 1$.

The objective is to obtain a co-clustering of the input data, that is a double partition of the data table into $C \times H$ blocks such that $\mathcal{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_C\}$ is a partition of the set of N objects into C clusters, and $\mathcal{Q} = \{\mathcal{Q}_1, \ldots, \mathcal{Q}_H\}$ is a partition of the set of P distributional-valued variables into H clusters.

Given the number of desired object clusters C and variable clusters H, the co-clustering returns the matrix $\mathbf{G}$ of prototypes, the partition $\mathcal{P}$ of the objects, and the partition $\mathcal{Q}$ of the variables. These are iteratively obtained by minimizing the following error function, denoted here as J_{DDK}:

$$J_{DDK}(\mathbf{G}, \mathcal{P}, \mathcal{Q}) = \sum_{k=1}^{C} \sum_{h=1}^{H} \sum_{e_i \in \mathcal{P}_k} \sum_{Y_j \in \mathcal{Q}_h} d_W^2(y_{ij}, g_{kh}), \tag{2}$$

where g_{kh} is the prototype of the co-cluster $\mathbf{Y}_{kh}$.

In most applications, variables may have a different relevance. We propose to obtain relevance weights by minimizing an objective function denoted by J_{ADDK}:

$$J_{ADDK}(\mathbf{G}, \Lambda, \mathcal{P}, \mathcal{Q}) = \sum_{k=1}^{C} \sum_{h=1}^{H} \sum_{e_i \in \mathcal{P}_k} \sum_{Y_j \in \mathcal{Q}_h} d_W^2(y_{ij}, g_{kh} | \Lambda), \tag{3}$$

where $d_W^2(.|\Lambda)$ is the adaptive (squared) L_2 Wasserstein distance computed between the generic y_{ij} and the prototype g_{kh} of the belonging co-cluster $\mathbf{Y}_{kh}$, weighted by the elements of Λ.

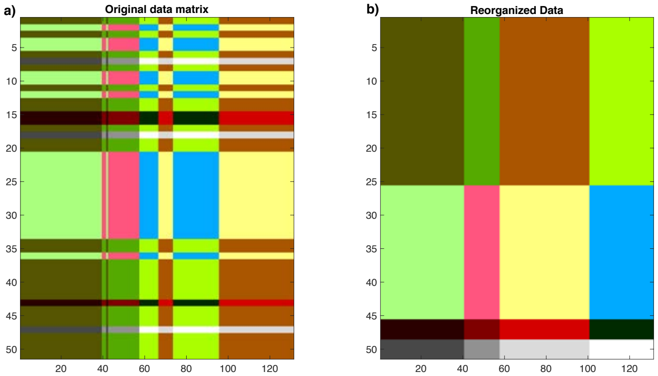
The basic scheme of the DDK and ADDK co-clustering algorithms is the following: from an initial random partitioning of the objects, into clusters of objects, and variables, the algorithms perform a sequence of alternating steps (three for DDK and four for ADDK) until the algorithms converge to a stationary value of the objective function:

i) *representation* step, in which the optimal representative (prototype) of each cluster is computed;
ii) *weighting* step (ADDK), in which the relevance weights for each variable and/or each component are computed;
iii) *object assignment* step, in which the optimal assignment of the objects to clusters is obtained;
iv) *variable assignment* step, in which the optimal assignment of the variables to clusters is obtained.

3 Conclusions

In this paper we propose to use two co-clustering algorithms for the analysis of time sequences. We tested the method on a real world dataset available at *http://db.csail.mit.edu/labdata/labdata.html* which collects some environmental variables inside a laboratory. We show in Fig. 1 the double partition for DDK. The left side shows the obtained co-clusters while the right side provides a reorganized version to highlight the main blocks.

Figure 1. *DDK algorithm: co-clustering structure.*



References

ARROYO, J., & MATÉ, C. 2009. Forecasting histogram time series with k-nearest neighbours methods. International Journal of Forecasting, **25**(1), 192–207.

BALZANELLA, ANTONIO, & IRPINO, ANTONIO. 2020. Spatial prediction and spatial dependence monitoring on georeferenced data streams. Statistical Methods & Applications, **29**(1), 101–128.

DE A.T. DE CARVALHO, FRANCISCO, BALZANELLA, ANTONIO, IRPINO, ANTONIO, & VERDE, ROSANNA. 2021. Co-clustering algorithms for distributional data with automated variable weighting. Information Sciences, **549**, 87–115.

DE CARVALHO, F. A. T., & LECHEVALLIER, Y. 2009. Partitional clustering algorithms for symbolic interval data based on single adaptive distances. Pattern Recognition, **42**(7), 1223–1236.

IRPINO, A., & VERDE, R. 2015. Basic statistics for distributional symbolic variables: a new metric-based approach. Advances in Data Analysis and Classification, **9**(2), 143–175.

Hidden Markov and regime switching copula models for state allocation in multiple time-series

Francesco Bartolucci[1], Fulvia Pennoni[2], and Federico P. Cortese[3]

[1] Department of Economics, University of Perugia
(e-mail: francesco.bartolucci@unipg.it)

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca
(e-mail: fulvia.pennoni@unimib.it)

[3] Department of Economics, Management and Statistics, University of Milano-Bicocca
(e-mail: f.cortese5@campus.unimib.it)

Abstract: We consider hidden Markov and regime-switching copula models as approaches for state allocation in multiple time-series, where state allocation means prediction of the latent state characterizing each time occasion based on the observed data. This dynamic clustering, performed under the two model specifications, takes the correlation structure of the time-series into account. Maximum likelihood estimation of the model parameters is carried out by the expectation-maximization algorithm. For illustration we use data on the market of cryptocurrencies characterized by periods of high turbulence in which interdependence among assets is marked.

Keywords: daily log-returns, expectation-maximization algorithm, forecast, latent variables, model-based clustering

1 Introduction

In the analysis of multiple time-series, state allocation, namely prediction of the state or regime underlying the observed data at a certain time occasion, is an important task, especially in finance and related fields. This type of clustering is dynamic because a different state may be predicted at every time occasion and may be based on models representing each time-specific state by a discrete latent variable assuming, typically, a few possible values. In this contribution, we compare two different model specifications of this type: multivariate hidden Markov (HM) models (Zucchini *et al.*, 2017) and regime-switching (RS) copulas (Rodriguez, 2007).

Among HM models we consider, in particular, those based on the assumption that the time-specific vector of observable variables follows a conditional Gaussian distribution with parameters depending on the latent state.

RS copulas are instead based on a copula function, which may be chosen among the Clayton, the Gumbel, the Gaussian, or the Student-*t*, with parameters governed by a hidden Markov process of first-order so as to flexibly account for the correlation patterns between each pair of series.

The expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) is used for maximum likelihood estimation of the parameters of both models. Model selection is performed to choose the most appropriate number of hidden states and evaluate the level of chain homogeneity over time (Bartolucci *et al.*, 2013). For the HM model, this selection is based on the Bayesian Information Criterion (BIC), and for RS copulas, it is also based on a goodness-of-fit procedure relying on the Cramér-von Mises statistic.

As an illustration we consider the problem of state allocation in analyzing time-series of the main cryptocurrencies daily log-returns over a three-year period.

2 Hidden Markov and Regime-Switching Copula Models

Let $\boldsymbol{y}_t$, $t = 1,2,\ldots$, be the vector where each element y_{tj}, $j = 1,\ldots,r$, corresponds to the value of time-series j at time occasion t, with r denoting the number of time-series under consideration. The main assumption of the multivariate HM model is that the random vectors $\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots$ are conditionally independent given a hidden process $u_1,u_2,\ldots$ that follows a first-order Markov chain with k states, labeled from 1 to k. This process is governed by the initial probabilities $\pi_u = p(u_1 = u)$, $u = 1,\ldots,k$, and the transition probabilities $\pi_{u|\bar{u}} = p(u_t = u|u_{t-1} = \bar{u})$, $t = 2,\ldots$, $\bar{u},u = 1,\ldots,k$. We assume a Gaussian distribution for the observations at every time occasion, that is, $\boldsymbol{y}_t \mid u_t = u \sim N_r(\boldsymbol{\mu}_u,\boldsymbol{\Sigma}_u)$, where $\boldsymbol{\mu}_u$ and $\boldsymbol{\Sigma}_u$ are the mean vector and variance-covariance matrix for latent state u. The above assumptions imply that the conditional distribution of the time-series $\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots$, given the sequence of hidden states, may be expressed as $f(\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots \mid u_1,u_2,\ldots) = \prod_t \phi(\boldsymbol{y}_t;\boldsymbol{\mu}_{u_t},\boldsymbol{\Sigma}_{u_t})$, where $\phi(\cdot;\cdot)$ denotes the density of the multivariate Gaussian distribution. The manifest distribution of the multiple time-series has the following density function:

$$f(\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots) = \sum_{u_1} \pi_{u_1}\phi(\boldsymbol{y}_1;\boldsymbol{\mu}_{u_1},\boldsymbol{\Sigma}_{u_1})\sum_{u_2}\pi_{u_2|u_1}\phi(\boldsymbol{y}_2;\boldsymbol{\mu}_{u_2},\boldsymbol{\Sigma}_{u_2})\cdots.$$

Concerning the copula model, we first consider only the bivariate case, so we define $\boldsymbol{y}_t = (y_{t1},y_{t2})$ as a vector with elements y_{tj}, $j = 1,2$, corresponding to the observation for time-series j at time $t = 1,2,\ldots$ and F_1 and F_2 as the marginal cdfs of each time-series. Sklar's theorem (Sklar, 1959) allows us to separate the fitting of the marginal cdfs from the fitting of the joint distribution, represented by a copula function. This approach consists in estimating the two marginal distributions, obtaining $\hat{F}_1$ and $\hat{F}_2$, and then computing the normalized ranks of the pseudo-observations $\tilde{\boldsymbol{e}}_t = (\tilde{e}_{t1}, \tilde{e}_{t2})$ as $\tilde{e}_{tj} = \text{rank}(\hat{z}_{tj})/(T+1)$, with $\hat{z}_{tj} = \hat{F}_j(y_{tj})$, and T being the number of observed time occasions. Finally, for the pseudo-observations $\tilde{\boldsymbol{e}}_t$, an RS copula model is assumed based on a hidden homogeneous Markov process denoted as $v_1, v_2, \ldots$, with k states. The copula density indicated with $c(\cdot;\cdot)$ may be chosen among the Clayton, the Gumbel, the Gaussian, or the Student-t copulas, with state-specific parameter β_v. The density of the pseudo-observations is given by

$$f(\tilde{\boldsymbol{e}}_1, \tilde{\boldsymbol{e}}_2, \ldots) = \sum_{v_1} \pi_{v_1} c(\tilde{\boldsymbol{e}}_1; \beta_{v_1}) \sum_{v_2} \pi_{v_2|v_1} c(\tilde{\boldsymbol{e}}_2; \beta_{v_2}) \cdots ,$$

and it is based on the initial and transition probabilities defined as above.

Given that the state sequence is not observable, a full maximum likelihood approach for estimating the parameters of both models is carried out through the EM algorithm. Following the current literature, model selection for the HM model is based on the BIC, and for the RS copula it is also performed through a goodness-of-fit procedure consisting in calculating a p-value referred to the Cramér-von Mises statistic for the hypothesis of correct model specification.

We compare the performance of HM models and RS copulas focusing on the crucial aspect of state allocation. The optimal state allocation is performed by finding the optimal joint sequence $\tilde{u}_1, \tilde{u}_2, \ldots$ (or $\tilde{v}_1, \tilde{v}_2, \ldots$) of unknown states given the corresponding observations. This clustering procedure, also known as global decoding, is achieved through the Viterbi algorithm (Viterbi, 1967), which is a dynamic programming algorithm.

We also aim at extending the RS copula approach to an arbitrary number of time-series r rather than to only 2. In this regard, we propose the composite likelihood approach (Varin *et al.*, 2011) for estimation, which is based on considering all possible ordered pairs of time-series among the available ones.

3 Application

As an illustration, for the HM model we consider the joint daily log-returns[*] of the five cryptocurrencies Bitcoin, Ethereum, Ripple, Litecoin, and Bitcoin

[*] provided by the Crypto Asset Lab: `https://cryptoassetlab.diseade.unimib.it/`.

Cash, for the period 2017-2020. For the RS copulas, allowing only for bivariate associations, we define four copulas where the bivariate vector of observations consists of the Bitcoin and each of the other four cryptocurrencies. Results for the HM model show that the minimum value of the BIC is reached considering a five-state heteroschedastic structure. According to these estimates, there are three negative regimes (in terms of estimated expected log-returns), with relatively high and positive correlations of Bitcoin with all the other cryptocurrencies, and two states with positive returns and lower correlations. Regarding the global decoding, these two states are the most likely in the first year of observation, and the other three states characterize the last two years.

Concerning the RS copulas, and considering as an example the couple of cryptocurrencies Bitcoin-Ethereum, we observe that a three-regime Clayton copula provides the best fit. Given that the Clayton copula allows for explicit computation of the lower tail correlation index, we estimate that two regimes provide zero or low values for the lower tail index, and the third regime provides high values for it. According to the optimal state sequence, we estimate that there is substantial interchangeability between the first two states in the whole period, whereas the third state is the most likely for the last year of observation.

References

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2013. *Latent Markov Models for Longitudinal Data*. Boca Raton, FL: Chapman & Hall/CRC.

DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

RODRIGUEZ, J. C. 2007. Measuring financial contagion: A copula approach. *Journal of Empirical Finance*, **14**, 401–423.

SKLAR, M. 1959. Fonctions de repartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, **8**, 229–231.

VARIN, C., REID, N., & FIRTH, D. 2011. An overview of composite likelihood methods. *Statistica Sinica*, **21**, 5–42.

VITERBI, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, **13**, 260–269.

ZUCCHINI, W., MACDONALD, I. L., & LANGROCK, R. 2017. *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton, FL: CRC.

BOOSTING MULTIDIMENSIONAL IRT MODELS

Michela Battauz [1] and Paolo Vidoni[1]

[1] Department of Economics and Statistics, University of Udine, (e-mail: michela.battauz@uniud.it, paolo.vidoni@uniud.it)

ABSTRACT: Multidimensional IRT models can be used to analyze the latent variables that underlay the responses given to a test or questionnaire. However, these models are not only difficult to estimate, but they also suffer of the rotational indeterminacy typical of factor analysis models. In this paper, we propose a boosting algorithm that, starting from a model that includes only the intercepts, sequentially updates a pair of coefficients in a component-wise approach. The solution provided by the algorithm tends to be sparse and to facilitate the interpretation without requiring a posterior rotation.

KEYWORDS: negative curvature direction, regularization, sparse solution.

1 Introduction

IRT models are commonly applied in educational assessment and they are also considered, with increasing frequency, in the field of health and psychological measurement studies. In these models, the probability of observing a categorical response is a function of a single latent trait (simple IRT models) or of multiple latent traits (multiple IRT models) and of some item parameters (see for example Reckase, 2009). Various methods have been proposed for model estimation. However, in the multidimensional setting, serious computational problems may occur if the number of items is large and many latent variables have to be considered. Moreover, in this context, the interpretability of the solution is very important.

In this paper, the new statistical boosting procedure introduced in Battauz & Vidoni (2021) is applied for estimating multiple IRT models. More precisely, we consider a suitable likelihood-based boosting algorithm which may escape from a region of local non-convexity of the objective function, improve the optimization procedure, provide a more interpretable sparse solution and regularize the estimates. We apply this new procedure to the multidimensional two-parameter logistic IRT model for dichotomously scored outcomes. An example concerning a sample from the 2017 Eurobarometer survey is presented.

2 Multidimensional IRT models: definition and inference

The response variable for the subject i on item j is a Bernoulli random variable Y_{ij}, $i = 1,\dots,n$, $j = 1,\dots,J$, with one denoting a positive response. The responses of subject i are collected in the vector $\mathbf{Y}_i = (Y_{i1},\dots,Y_{iJ})^\top$. Let $\theta_i = (\theta_{i1},\dots,\theta_{iD})^\top$, $i = 1,\dots,n$, be a latent random vector, composed of independent standard normal variables. Furthermore, it is assumed that $(\mathbf{Y}_i,\theta_i)$ are independent across subjects and that observations Y_{ij} are conditionally independent given θ_i. With particular attention to the multidimensional two-parameter logistic (2PL) IRT model, the conditional probability of giving a positive response to a specific item is defined as

$$P_{ij} = P(Y_{ij} = 1|\theta_i;\beta_j,\alpha_{1j},\dots,\alpha_{Dj}) = \frac{\exp(\beta_j + \alpha_{1j}\theta_{i1}+,\cdots+\alpha_{Dj}\theta_{iD})}{1+\exp(\beta_j + \alpha_{1j}\theta_{i1}+,\cdots+\alpha_{Dj}\theta_{iD})},$$

where β_j is the intercept and α_{dj}, $d = 1,\dots,D$, are the slope parameters. The vector of unknown model parameters is $\gamma = (\alpha_1^\top,\dots,\alpha_D^\top,\beta^\top)^\top$, with $\alpha_d = (\alpha_{d1},\dots,\alpha_{dJ})^\top$, $d = 1,\dots,D$, and $\beta = (\beta_1,\dots,\beta_J)^\top$; the vector γ has dimension $J + JD$, which, in some applications, can be very large.

Given the responses $\mathbf{y}$, realization of $\mathbf{Y} = (\mathbf{Y}_1^\top,\dots,\mathbf{Y}_n^\top)^\top$, the marginal likelihood for γ can be obtained by integrating out the unobserved θ values from the complete likelihood $L(\gamma;\mathbf{y}) = \prod_{i=1}^n f(\mathbf{y}_i|\theta_i;\gamma)\phi(\theta_i)$, where $f(\mathbf{y}_i|\theta_i;\gamma)$ is a Bernoulli-type probability function based on P_{ij} and $\phi(\cdot)$ denotes the density of a multivariate standard normal distribution with independent components. Thus, the marginal log-likelihood does not have a closed-form expression, since the D-dimensional integral does not have an analytic solution and requires numerical approximations. The most common methods for estimating the item parameters are based on the EM algorithm, approximating the integrals using Gaussian or adaptive quadrature procedures, or on suitable MCMC algorithms for handling with the high dimension of the integrals.

3 The boosting algorithm

We consider the boosting algorithm introduced in Battauz & Vidoni (2021), with the negative log-likelihood as objective function. Starting from a model that includes only the intercept terms, only two parameters are updated at each iteration of the algorithm, hence following a component-wise approach. The starting point of the algorithm poses a very challenging issue, since the gradient is null making any gradient descent method unable to move from it. A peculiar feature of the method is that it exploits any local non-convexity of the objective function, since the gradient vector and the Hessian matrix are used to define two alternative directions. These are the classical Newton-type direction and a negative curvature direction given by the eigenvector associated with the most negative eigenvalue (if any) of a 2×2 submatrix of the Hessian matrix. More specifically, at step k of the boosting algorithm, the Newton-type direction for each pair of parameters indexed $b, c = 1, \ldots, J(D+1)$, $b < c$, is given by:

$$\mathbf{s}_{bc}^{(k)} = -\widehat{\mathbf{H}}_{bc}^{(k-1)^{-1}} \widehat{\mathbf{g}}_{bc}^{(k-1)}, \tag{1}$$

while the negative curvature direction is:

$$\mathbf{d}_{bc}^{(k)} = -sign\left\{ \left(\widehat{\mathbf{g}}_{bc}^{(k-1)} \right)^{\top} \widehat{\mathbf{u}}_{bc}^{(k-1)} \right\} \widehat{\mathbf{u}}_{bc}^{(k-1)}, \tag{2}$$

where $\widehat{\mathbf{g}}_{bc}^{(k-1)}$ and $\widehat{\mathbf{H}}_{bc}^{(k-1)}$ are the gradient and the Hessian computed at step $k-1$, and $\widehat{\mathbf{u}}_{bc}^{(k-1)}$ is the eigenvector corresponding to the minimum negative eigenvalue of $\widehat{\mathbf{H}}_{bc}^{(k-1)}$. The algorithm computes the variation of a quadratic approximation of the objective function for all the pairs of parameters in both the directions, and selects the one leading to the largest decrease. The algorithm represents a particular application of the optimization method proposed by Gould et al. (2000), who proved the convergence to second-order critical points. Since the algorithm converges to the maximum likelihood estimates, a suitable stopping criterion is necessary to obtain regularized estimates.

4 A real-data example

The proposal was applied to the responses of 1027 Italian citizens to some items of the 2017 Eurobarometer survey regarding the area that people thinks that the decisions should be made at the European level. Table 1 reports the items and the estimated parameters. The number of iterations of the algorithm as well as the number of latent variables were selected by 5-fold cross-validation. The table also reports the maximum likelihood estimates (MLEs) obtained with the R package mirt and using the quartimax rotation, which was chosen for the higher similarity of the solution. It is possible to observe that the MLEs tend to assume more extreme values, while the boosting procedure provides regularized estimates. Both the methods identify a first dimension strongly related to all the items. The interpretation of the second dimension seems a bit more clear using the boosting algorithm, since it reveals a positive correlation between the areas of terrorism, immigration, democracy and peace (that present the highest estimated discrimination parameters). However, the areas of energy supply, environment, investment and job creation are also related to this dimension.

Table 1. *Items of the Eurobarometer survey included in the analysis and parameter estimates.*

		boosting			MLE		
QC7	Areas where more decision-making should take place at a European level	β_j	α_{1j}	α_{2j}	β_j	α_{1j}	α_{2j}
1	Fighting terrorism	3.07	4.02	1.61	6.10	-8.80	2.83
2	Dealing with health and social security issues	1.02	3.37	0.00	1.10	-3.55	-0.68
3	Promoting equal treatment of men and women	1.41	3.34	0.00	1.45	-3.37	-0.54
4	Promoting democracy and peace	1.95	2.99	0.92	2.08	-3.37	0.21
5	Securing energy supply	1.78	3.27	0.44	1.87	-3.49	-0.06
6	Dealing with migration issues from outside the EU	2.36	3.32	1.06	2.49	-3.73	0.33
7	Protecting the environment	2.41	4.74	0.58	2.46	-4.87	-0.38
8	Stimulating investment and job creation	1.80	4.41	0.74	2.02	-5.06	-0.24

References

GOULD, N. I. M., LUCIDI, S., ROMA, M., & TOINT, PH. L. 2000. Exploiting negative curvature directions in linesearch methods for unconstrained optimization. *Optimization Methods and Software*, **14**(1-2), 75–98.

MICHELA BATTAUZ, PAOLO VIDONI. 2021. A new likelihood-based boosting algorithm for factor analysis models with binary data. *Submitted.*

RECKASE, MARK D. 2009. *Multidimensional Item Response Theory Models.* New York, NY: Springer Verlag.

UNDERSTANDING AND ESTIMATING CONDITIONAL PARAMETRIC QUANTILE MODELS

Matteo Bottai[1]

[1] Karolinska Institute, (e-mail: matteo.bottai@ki.se)

ABSTRACT: The talk gives an overview of conditional parametric models. it outlines their features and potentials with focus on their interpretation, modeling possibilities, and real-data examples. It is intended for a broad audience, including methodological and applied statisticians, data analysts, practitioners, and anyone who may be interested in knowing more about these models.

KEYWORDS: integrated loss function, quantile regression, quantile regression coefficients models

Shapley Lorenz methods for eXplainable Artificial Intelligence

Niklas Bussmann[1], Roman Enzmann[2], Paolo Giudici[1] and Emanuela Raffinetti[1]

[1] Department of Economics and Management, University of Pavia (Italy), (e-mail: niklas.bussmann01@universitadipavia.it, paolo.giudici@unipv.it, emanuela.raffinetti@unipv.it)

[2] University of Bonn (Germany), (e-mail: ryenzmann@hotmail.com)

Abstract: A trustworthy application of Artificial Intelligence (AI) requires to measure in advance its possible risks. When applied to regulated industries, such as banking, finance and insurance, Artificial Intelligence methods lack explainability and, therefore, authorities aimed at monitoring risks may not validate them. To solve this issue, eXplainable Artificial Intelligence (XAI) methods have to be developed.

In this paper, we introduce an alternative XAI method, based on Lorenz Zonoids, that is statistically normalised and therefore more suitable to the risk management context. The application, focused on data involving more than 15,000 small and medium companies asking for credit, allows to further stress the benefits deriving from our proposal.

Keywords: Artificial Intelligence, Lorenz Zonoids, risk management.

1 Introduction

The key requirement for trustworthy Artificial Intelligence (AI) methods is their attitude to measure the risks deriving from their use. When applied to regulated fields, such finance and health, AI methods need to be validated by national regulators. It is worth noting that AI methods typically rely on the implementation of complex machine learning models which provide high predictive accuracy at the expense of explainability. This represents a problem for the regulated industries, where comprehensible results have to be made available in order to detect risks, especially in terms of the factors which can cause them. To avoid that wrong actions can be taken as a consequence of "automatic" choices, AI methods need to explain the reasons of their classifications and predictions.

In this paper, we propose a new explainable Artificial Intelligence method, based on the combination between the Shapley value approach (see, e.g. Shpaley, 1953) and the Lorenz Zonoid tool described in Giudici and Raffinetti

(2020). Shapley values belong to the class of local explanation methods, as they aim to interpret individual predictions in terms of which variables mostly affect them. Lorenz Zonoids instead are a global explanation method, as they aim to interpret all model predictions as a whole, in terms of which variables most determine them, for all observations.

We apply our methodology to a challenging problem: the prediction of a binary variable, representing the credit default, through a large set of balance sheet variables.

Next section describes the methodology, while Section 3 illustrates the empirical findings obtained applying our proposal to financial data.

2 Methodology

Following GIudici and Raffinetti (2021), we consider, for financial risk management purposes, a global explainable AI method, named Shapley-Lorenz decomposition, which combines the interpretability power of the local Shapley value game theoretic approach (see, e.g. Shapley, 1953) with a more robust global approach based on the Lorenz Zonoid model accuracy tool (see, e.g. Giudici and Raffinetti, 2020).

The Lorenz Zonoids, originally introduced by Koshevoy and Mosler (1996), were further developed by Giudici and Raffinetti (2020) as a generalisation of the ROC curve in a multidimensional setting and, therefore, the Shapley-Lorenz decomposition has the advantage of combining predictive accuracy and explainability performance into one single diagnostics. Furthermore, the Lorenz Zonoid is based on a measure of mutual variability that is more robust to the presence of outlying (anomalous) observations, with respect to the standard variability around the mean.

The Shapley-Lorenz decomposition expression is the result of a combination between the Shapley value-based formula and the Lorenz Zonoid tool. Formally, given K explanatory variables, the contribution of the additional variable X^k, expressed in terms of the differential contribution to the global predictive accuracy, equals to

$$LZ^{X_k}(\hat{\pi}) = \sum_{X' \subseteq \mathcal{C}(X) \setminus X_K} \frac{|X^{'}|!(K-|X^{'}|-1)!}{K!}[LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'})], \quad (1)$$

where: $\hat{\pi}$ is the estimated probability of default; the term $[LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'})]$ measures the marginal contribution provided by the inclusion of variable X_k; K is the number of available predictors; $\mathcal{C}(X) \setminus X_k$ is the set of all the possible model configurations which can be obtained with $K-1$ variables, excluding variable X_k; $|X^{'}|$ denotes the number of variables included in each possible model.

Note that he Lorenz Zonoids $LZ(\hat{\pi}_{X'\cup X_k})$ and $LZ(\hat{\pi}_{X'})$ in equation (1) can be computed by resorting to the covariance operators, i.e.,

$$LZ(\hat{\pi}_{X'\cup X_k}) = \frac{2}{\sum_{i=1}^{n}\hat{\pi}_{iX'\cup X_k}} Cov(\hat{\pi}_{X'\cup X_k}, r(\hat{\pi}_{X'\cup X_k})) \quad \text{and}$$

$$LZ(\hat{\pi}_{X'}) = \frac{2}{\sum_{i=1}^{n}\hat{\pi}_{iX'}} Cov(\hat{\pi}_{X'}, r(\hat{\pi}_{X'})),$$

where $r(\cdot)$ denotes the rank score.

The Shapley-Lorenz decomposition presents as an agnostic eXplainable Artificial Intelligence method which can be applied to the predictive output, regardless of which model and data generated it.

3 Application

We apply our proposed method to data supplied by a European External Credit Assessment Institution (ECAI) specialised in credit scoring for P2P platforms focused on SME commercial lending. In summary, the analysis relies on a dataset composed of official financial information, extracted from the balance-sheets of 15,045 SMEs, mostly based in Southern Europe, for the year 2015. The information about the status (0 = active, 1 = defaulted) of each company one year later (2016) is also provided. The observed proportion of defaulted companies is equal to 10.9%. In order to lead our analysis, we apply a logistic regression model after the data is split in a training set (80%) and a test set (20%). We then calculate, on the same split, the contribution of each of the nineteen explanatory variables to the estimate of the probability of default, using two explainable AI methods: the Shapley value approach and the Lorenz-Shapley approach that we propose. Table 1 displays the result of the comparison. From Table 1 note that the variable which most contributes to the prediction of default, according to the sum of the Shapley values, is Variable 8: (Profit or Loss before tax + Interest paid)/Total asset, followed at a considerable distance by Variables 13 and 14 (both related to EBITDA) and by Variable 3 (Total Assets/Total Liabilities). In terms of G^2 (deviance), instead, the differences between Variable 8 (the highest contributor) and Variables 14, 15 and 3 are lower. The role that Variable 13 has in terms of Shapley value is replaced by Variable 15. The first column of Table 1, giving the Shapley Lorenz values, indicate, instead, that Variable 8, with a value of 0.16, and Variable 3, with a value of 0.11, are one magnitude order higher than the others. This indicates a more clear cut choice, with only two variables being selected: a measure of leverage, and a measure of profitability. In the latter case, only the most contributing one, among the several that measure profitability, is chosen.

Table 1: Marginal contribution of each explanatory variable in terms of: Shapley-Lorenz Zonoids, G^2 and total Shapley values

Variable	Shapley-Lorenz	G^2	Shapley
Total assets/Equity	0.00	0.16	2.53
(Long term debt + Loans)/Shareholders Funds	0.00	0.54	-202.80
Total assets/Total Liabilties	0.11	1088.12	-1273.97
Current assets/Current Liabilties	0.05	553.68	-641.69
(Current assets - Current assets: stocks)/Current Liabilties	0.00	479.06	-93.51
(Shareholders Funds + Non current liabilities)/Fixed assets	0.00	13.16	4180.56
EBIT/interest paid	-0.01	411.10	1504.44
(Profit or Loss before tax + Interest paid)/Total assets	0.16	1633.51	-13115.53
Return on Equity	0.05	826.96	-1993.98
Operating revenues/Total assets	0.06	17.36	-289.46
Sales/Total assets	-0.02	10.96	252.59
Interest paid/(Profit before taxes + Interest paid)	0.01	103.26	379.73
EBITDA/interest paid	0.02	418.00	-1697.31
EBITDA/Operating revenues	0.03	1254.63	-1419.43
EBITDA/Sales	0.02	1122.05	-785.95
Trade Payables/Operating revenues	0.00	14.73	-193.60
Trade Receivables/Operating revenues	0.05	475.40	-585.58
Inventories/Operating revenues	0.01	126.78	1190.47
Turnover	0.02	85.26	1072.37

References

GIUDICI, P., & RAFFINETTI, E. 2020. Lorenz Model Selection. *Journal of Classification*, **37**, 754-768.

GIUDICI, P., & RAFFINETTI, E. 2021. Shapley-Lorenz eXplainable Artificial Intelligence. *Expert Systems With Appications*, **167**.

KOSHEVOY, G., & MOSLER, K. 1996. The Lorenz Zonoid of a Multivariate Distribution. *Journal of the American Statistical Association*, **91**, 873-882.

SHAPLEY, L.S. 1953. A value for *n*-person games. *Contributions to the Theory of Games*, 307-317.

ROBUST CLASSIFICATION OF SPECTROSCOPIC DATA IN AGRI-FOOD: FIRST ANALYSIS ON THE STABILITY OF RESULTS

Andrea Cappozzo[1], Ludovic Duponchel[2],
Francesca Greselin[3] and Brendan Murphy[4]

[1] Department of Mathematics, Politecnico di Milano, (andrea.cappozzo@polimi.it)

[2] LASIR Lab, University of Lille, (ludovic.duponchel@univ-lille.fr)

[3] Department of Statistics and Quantitative Methods, University of Milano Bicocca, (francesca.greselin@unimib.it)

[4] School of Mathematics and Statistics, University College Dublin, (brendan.murphy@ucd.ie)

ABSTRACT: We investigate here the stability of the obtained results of a variable selection method recently introduced in the literature, and embedded into a model-based classification framework. It is applied to chemometric data, with the purpose of selecting a few wavenumbers (of the order of tens) among the thousands measured ones, to build a (robust) decision rule for classification. The robust nature of the method safeguards it from potential label noise and outliers, which are particularly dangerous in the field of food-authenticity studies. As a by-product of the learning process, samples are grouped into similar classes, and anomalous samples are also singled out. Our first results show that there is some variability around a common pattern in the obtained selection.

KEYWORDS: Variable selection, Robust classification, Label noise, Outlier detection, Near infrared spectroscopy, Mid infrared spectroscopy, Agri-food.

1 Introduction

Nowadays, many challenging classification problems, arising from scientific domains such as chemometrics, computer vision, engineering, and genetics, among others, have to deal with hundreds or thousands of variables on each sample. Many contributions in the literature show that inferential methods benefit greatly from the identification of a subset of relevant variables. Dimension reduction techniques, like Principal Component Analysis (PCA), projection to latent structures (PLS-DA), single class modeling (SIMCA) and kernel methods (SVM) are generally adopted to this aim. In some fields of application, like in food-authentication, mislabeled and adulterated spectra may appear both in the calibration and/or validation sets. This contamination produces dramatic effects on the model estimation, and consequently on its prediction accuracy. To overcome this issue, a recent proposal in the literature introduces a variable selection step within the Robust Eigenvalue Decomposition Discriminant Analysis framework (Cappozzo *et al.*, 2019). Under the realistic assumption that only a portion of the spectral region is relevant for class discrimination, the procedure i) robustly identifies a subset of wavenumbers onto which building the decision rule, ii) protects it from potential label noise and outliers, and iii) simultaneously identifies anomalous samples.

We will recall here the main idea onto which the stepwise algorithm works, redirecting the interested reader to Cappozzo *et al.* (2021) for a more detailed presentation. The detection of p relevant features (out of the whole collection of $P \gg p$ available variables) on which to train the classifier has many advantages. Firstly, parameter estimation and interpretation is enhanced; secondly, loss on predictive power due to the inclusion of irrelevant and redundant information is avoided. Finally, cost reduction on future data collection and processing is obtained.

In model-based discriminant analysis, the features that directly depend on the class membership itself are called *relevant* variables. Conversely, *irrelevant* or noisy variables do not contain any discriminating power. Their distribution is completely independent on the group structure. Lastly, *redundant* variables essentially contain discriminant information that is already provided by the relevant ones: their distribution is conditionally independent of the grouping variable, given the relevant ones.

The algorithm starts from the empty set and, at each iteration, the inclusion of a *relevant* variable into the model is evaluated, based on its robustly assessed discriminating power. In a similar fashion, the removal of an existing variable from the model is also considered. The procedure iterates between variable addition and removal until two consecutive steps have been rejected.

2 Stability study

In this section, the results of a bootstrap-based analysis will be presented using data produced using non-parametric re-sampling of the actual data. The aim is to investigate the stability of the variable selection procedure.

The data we analyze come from the chemometric challenge organized during the “Chimiométrie 2005” conference (Fernández Pierna & Dardenne, 2007). The learning scenario encompasses $N = 215$ training and $M = 43$ test

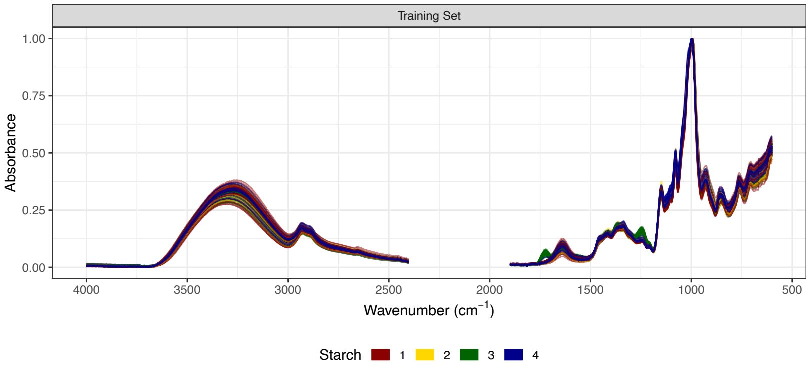


Figure 1. *Starches dataset: mid-infrared spectra of four starches classes.*

MIR spectra of starches of $G = 4$ different classes. For each sample, a total of $P = 2901$ absorbance measurements are recorded. A subset of training observations is displayed in Fig. 1. The aim of the competition was to discriminate the four different groups, defining a classification rule from the training set. In addition, outlier detection was advisable: four intentionally corrupted spectra were manually placed in the test set, as described in Fernández Pierna & Dardenne (2007).

For the first experiment 100 bootstrap datasets, of the same size as the actual dataset, were generated by sampling with replacement from the training set. A pattern in the selected variables arises from our results. For each bootstrapped sample, all models were fitted and the best-fit model was chosen using the BIC criterion, and the selected wavelengths were recorded. The chosen wavelengths show us which parts of the spectrum are of importance when classifying samples into different starches types. Results are shown in Fig. 2 through a raster plot. As we expect, there is some variability, due to the fact that the role of "relevant" and "irrelevant" variable is judged in terms of the set of already selected features. The wavelengths $997\,cm^{-1}$ and $995\,cm^{-1}$ correspond to spectral distributions of *amylose* and *amylopectin*, which are known to be present in different ratios across the starch classes. They have been selected with higher frequency, respectively 17 and 21 times in 67 runs.

3 Conclusions and further research

We developed a first stability analysis for a recent method for robust variable selection and classification, applied to spectrometric data. By a bootstrap simulation study on the learning set, although there has been variability in the

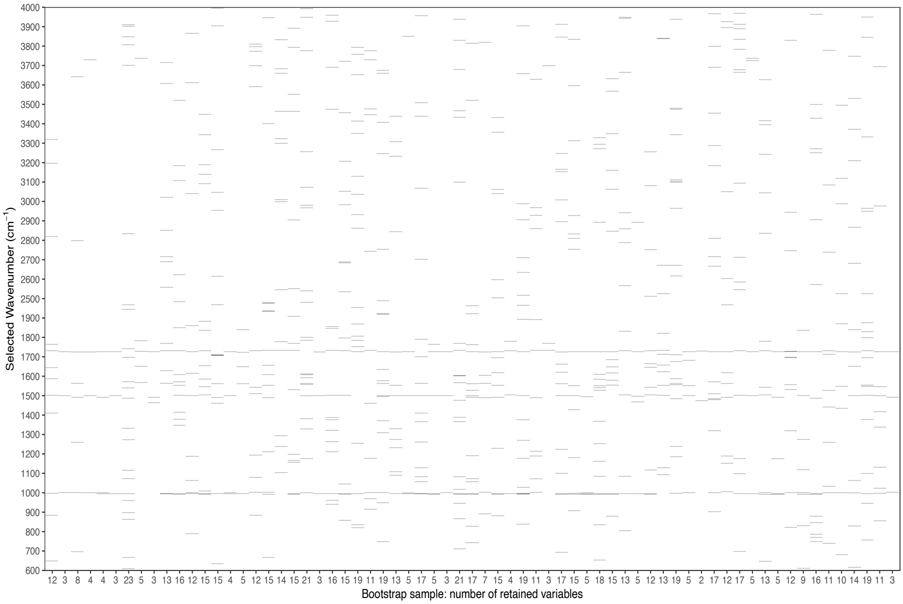


Figure 2. *Results of the stability analysis: for each of the 67 bootstrap samples, the selected wavenumbers are indicated in a raster plot.*

structure of the selected models, some stable pattern arises in results. Further research is still needed to cast more light on this topic. For instance, to investigate the sensitivity of the derived decision model, its accuracy on the test set is worth being analyzed, to establish the level of reliability in the resulting classification. This would mitigate the use of only a few real data examples and hence allows a more general discussion of the results.

References

Cappozzo, A., Greselin, F., & Murphy, T. B. 2019. A robust approach to model-based classification based on trimming and constraints. *Advances in Data Analysis and Classification*, 1–28.

Cappozzo, A., Duponchel, L., Greselin, F., & Murphy, T. B. 2021. Robust variable selection in the framework of classification with label noise and outliers: Applications to spectroscopic data in agri-food. *Analytica Chimica Acta*, **1153**, 338245.

Fernández Pierna, J. A., & Dardenne, P. 2007. Chemometric contest at “Chimiométrie 2005”: A discrimination study. *Chemometrics and Intelligent Laboratory Systems*, **86**(2), 219–223.

Issues in monitoring the EU trade of critical COVID-19 commodities

Andrea Cerasa[1], Enrico Checchi[1], Domenico Perrotta[1] and Francesca Torti[1]

[1] European Commission, Joint Research Centre, (e-mail: andrea.cerasa@ec.europa.eu, enrico.checchi@ec.europa.eu, domenico.perrotta@ec.europa.eu, francesca.torti@ec.europa.eu)

Abstract: The unexpected and constant increase of demand of commodities needed to manage the COVID-19 pandemic impacted the supply chain worldwide. Many countries, fearing shortages of those commodities, applied restrictions to the export of the national production.

The European Union, since the early stages of the pandemic, has monitored the procurement of these commodities by the EU Member States, to identify supply gaps, strong dependencies from extra-EU countries, as well as potential cases of frauds. Products like personal protective equipment, medicines, diagnostic kits, medical devices and (more recently) vaccines were scrutinized by a inter-service task force.

We illustrate some of the statistical issues encountered in analyzing these data from various perspectives, in particular the evolution in time of the traded prices and quantities of the most critical commodities. Robust statistical methods are still used to identify and rank spikes, level shifts and trends in hundreds of time series of Customs declarations.

Keywords: time series, international trade, COVID

SMOOTHED NON LINEAR PCA FOR MULTIVARIATE DATA

Marcello Chiodi [1]

[1] Department of Economics, Business and Statistics, University of Palermo, (e-mail: marcello.chiodi@unipa.it)

ABSTRACT: Principal Component Analysys is one of the wides known and used tool of linear explorative analysis with n observations on k numerical variables, in the most simple form. Besides the so called reduction of dimensionality achiveded by taking the first components as new reduced coordinates, the first principal axis can interpreted as the principal regression line, that is, the straight line which minimizes the sum of the orthogonal distances of the n points from the line, in a k dimensional space. Of course this interpretation of a principal lines relies on the assumption of linearity of relationships between variables, even if not conjointly normal. In this paper we propose an approach which searchs for a parametric curve $\mathbf{f}(t)$ in a k dimensional space, with some constrains for curvature or length. An extension is given to k components)

KEYWORDS: Smoothing splines, Principal regression, Multivariate curvature

1 Introduction

In a classical explorative phase of the analysis of a set $\mathbf{X}$ of data with n observations on k numerical variables, Principal Component Analysis (PCA) is often used to obtain reduction of dimensionality achieved by taking the first components as new reduced coordinates, but in general to have an insight in the multiple correlation structure among variables.

In so far, the first principal axis can interpreted as the principal regression line, that is, the straight line which minimizes the sum of the orthogonal distances of the n points from the line, in a k dimensional space. Of course this interpretation of a principal lines relies on the assumption of linearity of relationships between variables, even if not conjointly normal.

However, if the real interdependence structure between variables is not linear, the components could be not meaningful. Furthermore the distribution of the optimal distances from the first component could be not very regular.

Similar considerations could be made for components successive to the first, and this can be highlighted by appropriate residual analyses.

Indeed when looking for a dependence of a variable from another variable, we are not usually restricted to linear relationships, as estimated through linear regression, but we culd also use non parametric techniques, at least in an exploratory step of the analysis and with poor knowledge about the theoretical model which generated observed data. In this context usually some smooth functionis used which minimizes a compromise between fitting to data and smoothness.

If we are given a set of k variables, without a clear assignment of the roles of *dependent* and *explicative* variable, in some situation we would like to study multiple mutual interdipendence between variables, without the constraint of linearity.

Something similar is made in functional principal component analysis, when seeking for reduction of dimensionality with functional data.

In this paper we propose an exploratory tool, called smoothed PCA, which seek for function $\mathbf{f}(t)$ in a k dimensional space, close to observed points but sufficiently rough. An extension is given to k components.

2 Aim of the method

In our approach we searchs for a parametric curve $\mathbf{f}(t)$ in a k dimensional space, close to n observed points with some penalization or constrain $\mathbf{P}(\cdot)$ for curvature or length.

A first component is found solving a least squares penalized problem:

$$\min_{\mathbf{f}(t)} ||\mathbf{X} - \mathbf{f}(t)|| \; + \; \lambda \, \mathbf{P}(\mathbf{f}(t)) \tag{1}$$

As usual λ is a smoothing parametr which controls the amount of smoothing.

After a first component $\mathbf{f}_1(t)$ is found, a second component is similarly found, imposing a costrain of lack of correlation with the first component. Further components could be found in a similar way, even if this aspect is not uniquely solved till now.

The choice of λ could be made by means of cross validation techniques.

3 Explicit form of the approximant function.

The first problem we dealt with was the choice of the function $\mathbf{f}(t)$, which of course cannot be totally free. A natural choice was to seek for some family of parametric cubic splines. A possible choice, that we explored first, is to use

$$\mathbf{f}(t) = \sum_{l=1}^{m} \mathbf{c}_l B_l(t)$$

where the $\mathbf{c}_l$, $l = 1,2,\ldots,m$, are a set of $k-$dimensional vectors, and the $B_l(t)$ are a set of m Basis of splines (cubic) defined on some set of $m+4$ knots.

An alternative setting, which is the one we use in our presentation, is to define a function $\mathbf{f}(t)$ composed by $k-$components $f_j(t)$.

Each component $f_j(t)$ is a natural spline, with m knots z_l, with $l = 1,2,\ldots,m$, each interpolating m points for each of the j dimension. With this setting, the unknown quantities of the problems are the $m \times k$ coordinates of the m $k-$dimensional points $\mathbf{Q}_l$, with $l = 1,2,\ldots,m$.

This points could be maybe called *principal points*, but for now we simply use them as multivariate knots.

4 The roughness, or penalty, function.

The penalty function $\mathbf{P}(\cdot)$ is defined as a measure of the curvature in $\mathcal{R}_k$. Since here we have a curve in $\mathcal{R}_k$, the measure of curvature could be not so easy, but with the definition of $\mathbf{f}(t)$ as a set k natural splines, the curvature can be easily defined as the sum of the k curvatures of the single splines, based as usual on second derivatives, so that the simpler formulation is:

$$\mathbf{P}(\mathbf{f}(t)) = \sum_{j=1}^{k} \int [f_j''(t)]^2 dt$$

This formulation will allow to express the penalty $\mathbf{P}(\cdot)$ as a simple function of the higher coefficients of the piecewise polynomials which define the splines, and some tricks is used in order to manage with penalty term as it were a vector of residuals.

5 Numerical algorythms

The minimzation problem in (1), for a fixed value of λ is not an easy problem, since in the minimization problem the n optimal orthogonal projections $t_i, i = 1,2,...,n$ of the n observed points on the parametric curve $\mathbf{f}(t)$ must be found solving n optimization sub-problems, so that the problem cannot be splitted in k simpler penalized problems. For each possible curve $\hat{\mathbf{f}}(t)$, a set of optimal points should be recomputed.

At the present moment promising results are obtained with a double Levenberg-Marqadt type optimization, modified for the peculiarity of the problem. In our setting we tried to insert the penalization term in a least squares form.

A R package `smoothPCA` is under construction, which tries to use as much as possible existing optimized routines for the majority of steps.

The problem of the choice of the number of knots, m, is still open, even if it seems to be not so crucial as the choice of λ, for which some bounding values are proposed. Satisfactory solutions are obtained using as starting points a linear set of points $\mathbf{Q}_l$ computed along the line of the first principal component.

6 Exploratory analysis

The utility of the results of this techniques in exploratory analysis, relies in the possibility of giving a sort of multidimensional measure of conjoint non-linearity, together with the possibility of describing observed points in a reduced space obtained by non linear parametric transformations.

Some example will be presented on standard dataset

References

ALLEN, GENEVERA I., & WEYLANDT, MICHAEL. 2019. Sparse and Functional Principal Components Analysis. *2019 IEEE Data Science Workshop (DSW)*, Jun.

SILVERMAN, BERNARD W. 1996. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, **24**(1), 1 – 24.

ACCOUNTING FOR RESPONSE BEHAVIOR IN LONGITUDINAL RATING DATA

Roberto Colombi [1], Sabrina Giordano[2] and Maria Kateri[3]

[1] Department of Management, Information and Production Engineering, University of Bergamo, Italy (e-mail: roberto.colombi@unibg.it)

[2] Department of Economics, Statistics and Finance "Giovanni Anania", University of Calabria, Italy (e-mail: sabrina.giordano@unical.it)

[3] Institute for Statistics, RWTH Aachen University, Germany (e-mail: maria.kateri@rwth-aachen.de)

ABSTRACT: We present a hidden Markov model for repeated ordinal responses observed on some units at different time occasions. The responses reflect the levels of unobservable latent constructs and can be observed under two latent regimes according to whether the respondents are confident with their preference or take shelter in the extremes/middle points of the rating scale.

KEYWORDS: latent variables; response style; financial capability.

Hidden Markov models with two regimes

Consider one ordinal response observed on n units at T time occasions. So Y_{it} denotes the response of unit i, $i \in I = \{1,\dots,n\}$, at occasion t, $t \in \mathcal{T} = \{1,\dots,T\}$, with $Y_{it} \in \mathcal{C} = \{1,\dots,c\}$. The response is assumed to reflect the levels of unobservable latent constructs L_{it}, $i \in I$, $t \in \mathcal{T}$ and can be observed under two different latent regimes: *awareness* (AWR) and *middle or extreme categories response style* (EMRS) that are captured by binary latent variables U_{it}, $i \in I$, $t \in \mathcal{T}$. The presence of two regimes is based on the idea that when required to express their opinion on one item, respondents either identify their true preference into one category on the rating scale or, when in doubt or reluctant to disclose their opinion, take shelter by opting for the extreme or middle categories. These are the cases, for example, of patients asked to give a subjective assessment of their health or disability in daily living, or people required to evaluate their financial capability; all of them can feel confident or reluctant to answer. The proposal is a hidden Markov model (HMM) defined by two components that describe the distribution of the latent variables and the conditional distribution of the response given the latent variables. It generalizes the models by Bartolucci *et al.*, 2012 to a bivariate latent Markov process. Here, we describe the main features of the model proposed by Colombi *et al.*, 2021.

The latent Markov model. For every $i \in I$, $t \in \mathcal{T}$, the *latent construct* L_{it} (as: health status, financial capability) has a finite discrete state space $\mathcal{S}_L = \{1,\ldots,k\}$, while the *latent binary response style indicator* U_{it} has a state space $\mathcal{S}_U = \{1,2\}$, where 1 and 2 denote the EMRS and AWR states, respectively. The latent variables are independent across units and for every unit, $\{L_{it}, U_{it}\}_{t\in\mathcal{T}}$ is a first order bivariate Markov process with states (u,l), $u \in \mathcal{S}_U$, $l \in \mathcal{S}_L$. The initial probabilities ($t = 1$) of $\{L_{it}, U_{it}\}_{t\in\mathcal{T}}$ are $\pi_{i1}(u,l)$, and $\pi_{it}(u,l|\bar{u},\bar{l})$ are the transition probabilities. They are are simplified to $\pi_{it}(u,l|\bar{u},\bar{l}) = \pi_{it}^{U|L}(u|l,\bar{u})\pi_{it}^{L}(l|\bar{l}), t = 2,\ldots,T$, by assuming that L_{it}, given its past, does not depend on the past of U_{it} and the current U_{it} depends on its past and on the contemporaneous latent construct but not on the past of the latent construct. The row vectors $\mathbf{x}_i^{(m)}$ and $\mathbf{z}_{it}^{(m)}$, $m \in \{L,U\}$, stand for the covariates, not necessarily different, influencing the initial and transition probabilities, respectively, of the latent variables. Assuming independence between the latent variables at the first time, the latent model is specified by the following logit models: A) a baseline logit model for the initial probabilities of the latent construct $\log\frac{\pi_{i1}^{L}(l)}{\pi_{i1}^{L}(1)} = \alpha_{0l} + \alpha_{1l}'\mathbf{x}_i^{(L)}, l = 2,\ldots,k$; B) a logit model for the initial probabilities of the response style indicator $\log\frac{\pi_{i1}^{U}(1)}{\pi_{i1}^{U}(2)} = \bar{\alpha}_0 + \bar{\alpha}_1'\mathbf{x}_i^{(U)}$; C) baseline logit models for the marginal transition probabilities of the latent construct, with reference category the state $\bar{l}$ of the previous time point, i.e. for $\bar{l} \in \mathcal{S}_L$, $\log\frac{\pi_{it}^{L}(l|\bar{l})}{\pi_{it}^{L}(\bar{l}|\bar{l})} = \beta_{0l\bar{l}} + \beta_{1l\bar{l}}'\mathbf{z}_{it}^{(L)}, l \in \mathcal{S}_L, l \neq \bar{l}, t = 2,\ldots,T$; D) a logit model for the conditional transition probabilities of the response style indicator for each response style state $\bar{u}$ of the previous occasion and for each current state l of the latent construct $\log\frac{\pi_{it}^{U|L}(1|l,\bar{u})}{\pi_{it}^{U|L}(2|l,\bar{u})} = \bar{\beta}_{0l\bar{u}} + \bar{\beta}_{1l\bar{u}}'\mathbf{z}_{it}^{(U)}, l \in \mathcal{S}_L, \bar{u} \in \mathcal{S}_U, t = 2,\ldots,T$.

The observation model. Independence is assumed among units. The conditional probability functions of Y_{it}, given the EMRS $(1,l)$ and AWR $(2,l)$ latent states are both time and subject invariant, denoted by $f(y|l,u)$, $u \in \mathcal{S}_U$, $l \in \mathcal{S}_L$, $y \in \mathcal{C}$, for $t \in \mathcal{T}$, $i \in I$. Given the EMRS regime, $f(y|l,1)$, $l \in \mathcal{S}_L$, is parameterized by the logits $\log\frac{f(y|l,1)}{f(y-1|l,1)} = \phi_{0l} + \phi_{1l}s(y), y = 2,\ldots,c$, where the scores are known constants $s(y) = (\frac{c}{2} - y)/\sqrt{\sum_{y=1}^{c-1}(y-c/2)^2}$, $y \in \mathcal{C}$, ϕ_0 governs the skewness, ϕ_1 the U and bell shape. Given the AWR regime, $f(y|l,2)$, $l \in \mathcal{S}_L$, is parameterized by the logits $\log\frac{f(y|l,2)}{f(y-1|l,2)} = \varphi_{y_l}, y = 2,\ldots,c$.

Application to Bank of Italy data. We applied the model to the panel data from the Survey on Household Income and Wealth (Bank of Italy), collected every 2 years from 2006 to 2016 on 1109 Italian households. The ordinal re-

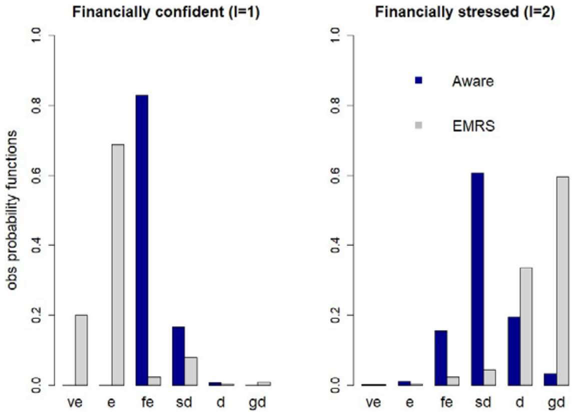


Figure 1. *Observation probability functions of AWR and EMRS respondents in the two latent states of the perceived financial condition.*

sponse of interest is the perception of the household's financial ability to make ends meet (ve = very easily, e = easily, fe = fairly easily, sd = with some difficulty, d = with difficulty, gd = with great difficulty), the covariates are: G (female, *male*), J (Jse: self-employee, Jhrs: housekeeper/retired/student, *employee*), CH (with children, *no children*), D (with debts, *no debts*), S (with savings, *no savings*), E (up to secondary school, *over high school*), R (no risk averse in managing financial investments, *risk averse*), with the reference categories being in italics. The minimum BIC corresponds to the model with $k = 2$ states, meaning that households can be grouped according to whether they feel financially confident ($l = 1$) or deal with financial stress ($l = 2$). Fig. 1 allows us to characterize the choices of the respondents in 4 latent states. Individuals, in the financially confident latent state, when in doubt about their perception, tend to choose with more chance the optimistic extreme points, AWR people instead are more incline to the intermediate rates. Reluctant households (EMRS) in the latent group that deals with financial stress have the highest probabilities of reporting great difficulties, AWR people in the same group are more likely to point out just some difficulties. The behavior in the 4 stata is well distinguished, and optimistic/pessimistic choices are mainly due to the EMRS tendency. By the sign of the estimates in Table 1 row 1, we deduce that at the first occasion women, employees, people without savings, with high education and risk averse are with higher probability in a worse financial sta-

Table 1. *Estimates (EM algorithm) of the parameters of logit models A, B, C, D.*

parameters	cst	G	Jse	Jhrs	CH	D	S	E	R
$(\alpha_{02}, \alpha_2)'$	2.8	0.44*	-1.38*	-0.75*	-0.15	0.02	-1.44*	-1.86*	-0.35*
$(\bar{\alpha}_0, \bar{\alpha}_1)'$	-0.06	-0.03	0.16	0.08	-0.04	0.32	0.63*	0.04	0.14
$(\beta_{021}, \beta_{121})'$	-0.86	1.32*	0.27	-0.49	-0.89*	0.48	-1.69*	-1.16*	-0.17
$(\beta_{012}, \beta_{112})'$	-11.93	0.18	-0.91	-0.21	-0.36	-0.23	8.44*	1.38*	-8.83*
$(\bar{\beta}_{011}, \bar{\beta}_{111})'$	1.10	0.45	-0.29	0.00	-0.20	0.13	-0.79*	-0.47*	-0.06
$(\bar{\beta}_{021}, \bar{\beta}_{121})'$	-3.36	-0.05	1.09*	-0.33	0.45	-0.37	1.97*	0.81*	-0.37
$(\bar{\beta}_{012}, \bar{\beta}_{112})'$	1.91	-0.07	-0.35	-0.23	0.00	-0.05	-0.19	-0.29	-0.39*
$(\bar{\beta}_{022}, \bar{\beta}_{122})'$	1.69	-0.50	-0.34	-0.08	0.10	-0.07	1.80*	-0.09	-0.37

cst: constant – $*$ 95% confidence interval does not contain zero

tus. Further, responders with savings show a major propensity to a response style at the beginning of the survey (row 2). From row 3, it seems that, in two consecutive moments, women move from a financially confident ($l = 1$) condition to a worse status ($l = 2$) with higher probability, while low-educated households with children and savings more likely tend to rest in the previous more comfortable financial status ($l = 1$). Individuals who have savings and a low education pass with greater probability from the financial stressed status ($l = 2$) to the better condition ($l = 1$), while financially stressed households tend to remain in the same worst status with greater probability when they are no risk averse (row 4). From rows 5-6, it is more likely to change from the EMRS status ($\bar{u} = 1$) to an AWR behavior ($u = 2$) for low educated persons with savings, who currently belong to the group of financially confident households, while self-employee and low educated respondents with savings show greater probability of remaining in the EMRS status if in the previous occasion were reluctant ($\bar{u} = 1$) and in the current time are financially stressed ($l = 2$). Who is no risk averse and in the current moment feels to be financially confident has higher probability of keeping the previous awareness in revealing the own financial capability. On the other hand, individuals with savings, being in the latent financially worrying status, tend with more propensity to give up on the previous AWR behavior and opt for a response style, rows 7-8.

References

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2012. *Latent Markov Models for Longitudinal Data.* CRC Press.

COLOMBI, R., GIORDANO, S., & KATERI, M. 2021. Hidden Markov models for longitudinal rating data with dynamic response styles: evidence on household financial capability. *Submitted.*

NETWORK-BASED SEMI-SUPERVISED CLUSTERING OF TIME SERIES DATA

Claudio Conversano[1], Giulia Contu[1], Luca Frigau[1] and Carmela Cappelli[2]

[1] Department of Economics and Business, University of Cagliari, (e-mail: conversa@unica.it, giulia.contu@inica.it, frigau@inica.it)

[2] Department of Humanities, University of Naples Federico II, (e-mail: carmela.cappelli@unina.it)

ABSTRACT: Semisupervised clustering extends standard clustering methods to the semisupervised setting, in some cases considering situations when clusters are associated with a given outcome variable that acts as a "noisy surrogate", that is a good proxy of the unknown clustering structure. A novel approach to semisupervised clustering associated with an outcome variable named network-based semisupervised clustering (NeSSC) has been recently introduced (Frigau *et al.*, 2021). It combines an initialization, a training and an agglomeration phase. In the initialization and training a matrix of pairwise affinity of the instances is estimated by a classifier. In the agglomeration phase the matrix of pairwise affinity is transformed into a complex network, in which a community detection algorithm searches the underlying community structure. Thus, a partition of the instances into clusters highly homogeneous in terms of the outcome is obtained. A particular specification of NeSSC, called Community Detection Trees (Co-De Tree), uses classification or regression trees as classifiers and the Louvain, Label propagation and Walktrap as possible community detection algorithm. NeSSC is based on an ad-hoc defined stopping criterion and a criterion for the choice of the optimal partition of the original data. In this presentation, we provide a new specification of the NeSSC algorithm that allows us to perform clustering of time series data. This specification is based on the integration between Co-De Tree and the Atheoretical Regression Tree (ART) approach introduced by (Cappelli *et al.*, 2013; Cappelli *et al.*, 2015). ART exploits the concept of contiguous partitions within the framework of Least Squares Regression Trees using as a single covariate an arbitrary sequence of completely ordered numbers $K = 1, 2, \dots, i, \dots, N$. Tree-regressing the response variable Y on this artificial covariate resorts to create and check at any node h all possible binary contiguous partitions of the $Y_i \in h$. These splits are the only ones that need to be checked to detect the binary partition that minimizes the sum of squares and, indeed, they are generated by using K as covariate. In other words, for the contiguity property the best split lays in K (or in its subintervals after the split of the root note has taken place) and the tree algorithm, based on the classical "reduction in impurity" splitting criterion is forced to identify it. In general, the use of K as covariate enables ART to generate G different groups having different means. The effectiveness of the proposed NeSSC-ART combined approach for time series clustering is demonstrated on simulated and real data

KEYWORDS: network-based semisupervised clustering, community detection trees, atheoretical regression tree.

References

CAPPELLI, C, D'URSO, P, & DI IORIO, F. 2013. Change point analysis of imprecise time series. *Fuzzi Sets and Systems*, **225**, 23–38.

CAPPELLI, C, D'URSO, P, & DI IORIO, F. 2015. Regime change analysis of interval-valued time series with an application to PM10. *Chemiometrics and Intelligent Laboratory System*, **146**, 337–346.

FRIGAU, L, CONTU, G, MOLA, F, & CONVERSANO, C. 2021. *NNetwork-based semisupervised clustering*. Vol. 37.

Characterising longitudinal trajectories of COVID-19 biomarkers within a latent class framework

Federica Cugnata [1], Chiara Brombin[1], Pietro E. Cippà [2], Alessandro Ceschi [3],
Paolo Ferrari [4] and last Clelia Di Serio [1]

[1] University Centre for Statistics in the Biomedical Sciences (CUSSB), Vita-Salute San Raffaele University, (e-mail: cugnata.federica@unisr.it, chiara.brombin@unisr.it, clelia.diserio@unisr.it)

[2] Department of Medicine, Division of Nephrology, Ente Ospedaliero Cantonale, Bellinzona and Faculty of Medicine, University of Zurich, (e-mail: Pietro.Cippa@eoc.ch

[3] Faculty of Medicine, University of Zurich, Biomedical Faculty, Università della Svizzera Italiana, Lugano, Institute of Pharmacology and Toxicology, Ente Ospedaliero Cantonale, Bellinzona, (e-mail: Alessandro.Ceschi@eoc.ch)

[4] Department of Medicine, Division of Nephrology, Ente Ospedaliero Cantonale, Bellinzona and Biomedical Faculty, Università della Svizzera Italiana, Lugano, (e-mail: Paolo.Ferrari@eoc.ch)

Abstract: In COVID-19 clinical research, identifying homogeneous subgroups of patients is essential for tailoring treatments. To address this issue from a statistical point of view, models accounting for unobservable heterogeneity in patients are needed. We propose latent class mixed models (LCMMs) to model trajectories of clinically relevant biomarkers for COVID-19 and we compared patients in the uncovered different classes with respect to their baseline clinical characteristics and COVID-19 outcomes.

Keywords: Latent class mixed model, C-Reactive Protein, serum creatinine

1 Introduction

One of the main goals in COVID-19 clinical research is to identify patients' characteristics associated with different degree of disease severity. Most of the published paper focus on patients' characteristics at hospital admission linking them to the final outcome either intensive care unit (ICU) admission or death. In this work we applied an alternative approach to evaluate the dynamics of commonly monitored biomarkers while uncovering subgroups of patients with specific longitudinal response pattern. In particular, here we focus on the trajectories of serum creatinine and C-Reactive Protein (CRP) from the hospital admission.

2 Sample description

A sample of 512 hospitalized patients, admitted by Ente Ospedaliero Cantonale COVID-19 dedicated hospital between March 1-May 1 2020, diagnosed with COVID-19 and with at least two determinations of Serum creatinine (3546 observations) or CRP (3592 observations) has been considered for the analysis. Diagnosis of COVID-19 was based on a positive nasopharyngeal swab specimen tested with real-time RT-PCR assay or high clinical suspicion. The study was approved by the Ethical Committee of the Canton of Ticino, Switzerland. Demographic and clinical characteristics along with the comorbidities and symptoms of COVID-19 were recorded at admission time. Clinical and laboratory parameters have been regularly monitored every 48h during hospitalization. The median patients' age was 72 years (IQR [60.75, 80.00]) ranging from 22 to 97 years; 317 (61.9%) were male. 379 patients (74%) were discharged, 95 patients (18.6%) died and 7.4% were still hospitalized. 116 patients (22.7%) in total were admitted to the ICU.

3 Statistical methods

To identify groups of patients with distinct biomarkers' trajectories over time, latent class linear mixed model (LCMM Proust-Lima *et al.*, 2017) were applied. LCMMs generalize traditional Linear Mixed Effects (LME) models, assuming that the population is heterogeneous and G unobserved sub-populations (latent classes), with their own mean profiles of trajectories, may be identified. Consistently with the literature on latent variable modelling the approach requires the specification of a structural latent model, i.e., a standard linear mixed model without measurement errors, along with a measurement model, linking the latent process to the outcome of interest. When heterogeneous population is assumed, for a subject i belonging to the class c_i equal to g ($g = 1,\dots,G$), a latent class-specific process can be defined as

$$\Lambda_i(t_{ij})|_{c_i=g} = X_{1i}(t_{ij})'\beta + X_{2i}(t_{ij})'\gamma_g + Z_i(t_{ij})'u_{ig} + w_i(t_{ij})$$

where t_{ij} denotes the time of measurement for subject i ($i = 1,\dots,N$) at occasion j ($j = 1,\dots,n_i$), $X_{1i}(t_{ij})$ and $X_{2i}(t_{ij})$ are vectors of time-dependent covariates respectively with common fixed effects β over classes and class-specific fixed effects γ_g, $Z_i(t_{ij})$ is a vector of time-dependent covariates associated with individual class-specific random effects u_{ig} and $w_i(t_{ij})$ represents an autocorrelated process. Then a measurement model is defined as $Y_{ij}|_{c_i=g} = H(\Lambda_i(t_{ij})|_{c_i=g} + \varepsilon_{ij};\eta)$ where H is a parametrized monotonic increasing link function (linear, splines, thresholds, etc. depending on the type of the longitudinal markers), ε_{ij} are independent normally distributed errors and represents a noisy latent process at time. Every subject is assigned to one latent class only. For each subject, the latent class membership is described by a latent variable c_i that equals g if i belongs to class g and probability of latent class membership is modeled using a multinomial logistic regression according to covariates X_{3i}:

$$\pi_{ig} = P(c_i = g|X_{3i}) = \frac{e^{\xi_{0g}+X'_{3i}\xi_{1g}}}{\sum_{l=1}^{G} e^{\xi_{0l}+X'_{3i}\xi_{1l}}}$$

where ξ_{0g} is the intercept for class g and ξ_{1g} is the vector of class-specific parameters related to the time-independent covariates X_{3i}.

Since we are specifically interest in identifying different dynamics over time for biomarkers only the measurement time from the hospital admission has been considered as covariate. Splines link functions (with 5 equidistant knots; Ramsay, 1988) were considered to account for nonlinearities in the longitudinal response. Several LCMMs were estimated assuming different number of latent classes and BIC criterion was used to select the optimal number of latent classes. In presence of more than two classes, Fisher exact test and Kruskal-Wallis test were used to compare patients clinical features in different latent classes.

4 Results

The best model for serum creatinine included two latent classes, with 453 subjects assigned to class 1 and 42 to class 2 (BIC = 974.92). At baseline, class 2 differs from class 1 (p-value<0.001) and for patients in class 1 creatine significantly declined over time (p-values<0.0001), while class 2 remains stable. The class-specific mean predicted trajectories are reported in Figure 1(a). Average posterior probabilities of falling into the class in which the subjects were classified are equal to 0.957 and 0.804. Examining differences among the two classes, it emerged that they are associated to diabetes (p-value<0.001), cardiovascular disease (p-value<0.001) and cough (p-value=0.044). Moreover considering the patients assigned to the class 1, 20.5% were admitted to the intensive care unit and 15% died whereas considering the patients assigned to the class 1, 52.4% were admitted to the intensive care unit and 61.9% died (both p-values<0.001).

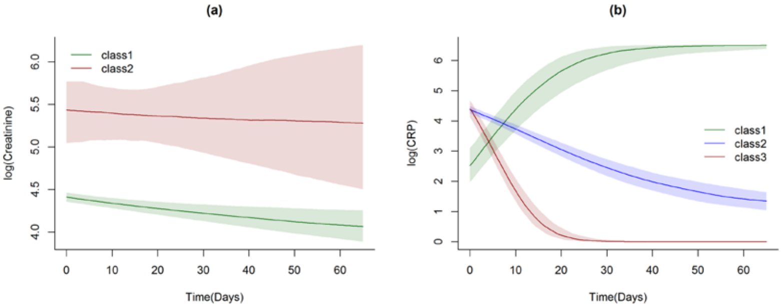


Figure 1. *Class-specific mean predicted trajectories for serum creatinine (a) and C-Reactive Protein (b)*

With reference to the model for CRP we found that the three latent classes model was the best in terms of BIC, with 30 subjects assigned to class 1, 411 to class 2 and 64 to class 3 (BIC=9716.96). At baseline, class 2 and class 3 differ from class 1 (both p-values<0.001). Moreover, for patients in class 1 CRP significantly increased over time (p-values<0.001) whereas for patients in class 2 and class 3 CRP significantly declined over time (both p-values<0.001) with larger decrease for class 3. The class-specific mean predicted trajectories are reported in Figure 1(b). Average posterior probabilities of falling into the class in which the subjects were classified are equal to 0.807, 0.873 and 0.773. These classes significantly differ on age (median age in class 1 is 76.50, in class 2 is 72.00 and in class 3 is 67.00, p-value=0.001) and they are associated to diabetes, cardiovascular disease and respiratory symptoms. Moreover the classes are associated to the outcome of the disease (p-value<0.001), the percentage of death is 66.7% in the class 1, 17.8% in the class 2 and 0% in the class 3.

In order to better understand the relationship between biomarkers evolution and COVID-19 outcome as a matter of future research the same latent class framework will be considered and in particular multi-process joint latent class mixed models will be applied.

References

PROUST-LIMA, CÉCILE, PHILIPPS, VIVIANE, & LIQUET, BENOIT. 2017. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software*, **78**(2), 1–56.

RAMSAY, JAMES O. 1988. Monotone regression splines in action. *Statistical science*, **3**(4), 425–441.

SENDER AND RECEIVER EFFECTS IN LATENT SPACE MODELS FOR MULTIPLEX DATA

Silvia D'Angelo [1]

[1] School of Mathematics and Statistics, University College Dublin, (e-mail: silvia.dangelo@ucd.ie)

ABSTRACT: Network and multidimensional network (multiplex) data often entail transitivity and heterogeneity of the nodes. This last aspect is particularly of interest in multiplex data, as nodes' tendencies to send or receive links is often network-dependent. Here, a class of latent space models is discussed. This class allows both to account for different levels of complexity in nodes' heterogeneity and for recurring symmetric relations between the nodes, via the inclusion of a shared latent space. The frameworks is quite general, as both weighted and binary networks are considered. Inference is carried out within a hierarchical Bayesian framework, while a Markov Chain Monte Carlo algorithm is used for estimation of model parameters.

KEYWORDS: latent space models, multiplex, Markov chain Monte Carlo

1 Introduction

Network data are relational data representing interactions among a set of actors, the nodes. Interactions among pairs of nodes are represented as links binding them, the edges. Depending on the type of relation represented in a network, such links can either be binary, indicating the presence or absence of a relation, or weighted, expressing the "strength" of the interaction between pairs of nodes. Moreover, when multiple relationships are observed among the same set of nodes, a particular type of network can be defined, that is a multidimensional network (or multiplex). Observed network data can display different characteristics, and these may have a direct impact on their structure. Two common features are transitivity ("a friend of my friend is my friend") and heterogeneity of the nodes. Building on a previous work (D'Angelo *et al.*, 2020), we propose to address the presence of the first feature by defining a shared, low dimensional, latent space (see Hoff *et al.*, 2002 and Gollini & Murphy, 2016) underlying the network or multidimensional network. Nodes are embedded in such latent space, with the main assumption that there proximity denotes similarity and hence a larger probability to interact in the observed network. Node-specific sender and receiver effects are then introduced to flexibly model heterogeneity in the data (see Hoff, 2005). Last, the possibility of different link functions can be considered (see Sewell & Chen, 2016), to adapt the framework to either binary or weighted networks.

2 The models

Given a set of n nodes, $i, j = 1, \dots, n$, we can define a multidimensional network as a collection of K adjacency matrices: $\mathbf{Y} = \left\{\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(k)}, \dots, \mathbf{Y}^{(K)}\right\}$. A single network can be viewed as a specific case of $\mathbf{Y}$, when $K = 1$. In the case of binary networks, the general entry $y_{ij}^{(k)}$ will be either 1, if nodes i and j are connected, or 0, if they are not. Instead, in weighted networks $y_{ij}^{(k)}$ will correspond to the weight associated to the interaction between nodes i and j in network k, that is the "strength of their interaction". Generally, we assume that:

$$f\left(\mathbf{E}\left[y_{ij}^{(k)}\right]\right) = \alpha^{(k)}\phi_{ij}^{(k)} - \beta^{(k)} d_{ij},$$

where $f(\cdot)$ is some link function, depending on the type of edges considered. Similar specifications to those employed in generalized linear mixed models can be used for $f(\cdot)$ (Sewell & Chen, 2016). $\alpha = \left\{\alpha^{(k)}\right\}_{k=1}^{K}$ and $\beta = \left\{\beta^{(k)}\right\}_{k=1}^{K}$ are intercept and scale network-specific parameters, adapting the shared latent space structure to the different networks. Distances between pairs of nodes in the latent space are indicated via d_{ij}, here taken to be the squared Euclidean distance between i and j in the latent space. Last, $\phi_{ij}^{(k)}$ represents the effect of the dyad-specific heterogeneity on the probability of an interaction in network k between nodes i and j. More specifically, it is assumed that: $\left[\phi_{ij}^{(k)}\right] = \left[g\left(\theta_i^{(k)}, \gamma_j^{(k)}\right)\right]$, where $\theta_i^{(k)}$ and $\gamma_j^{(k)}$ are, respectively, the sender effect of node i and the receiver effect of node j, in network k. To flexibly describe different levels of heterogeneity in multidimensional networks, we define three increasing complexity scenarios for either the sender and receiver parameters. A NULL scenario, where no effect is present ($\theta_i^{(k)} = 0$ and/or $\gamma_j^{(k)} = 0$), a CONSTANT scenario, where node-specific effects do not vary across networks ($\theta_i^{(k)} = \theta_i$ and/or $\gamma_j^{(k)} = \gamma_j$), and a VARIABLE scenario, where effects are present and network-specific ($\theta_i^{(k)}$ and/or $\gamma_j^{(k)}$). Depending on the presence or absence of the effects, $g(\cdot,\cdot)$ can be defined as:

$$g(\cdot,\cdot) = \begin{cases} 1 \quad \text{if both effects are NULL} \\ \theta_i^{(k)} \quad \text{if receiver effects are NULL} \\ \gamma_j^{(k)} \quad \text{if sender effects are NULL} \\ \frac{\theta_i^{(k)}+\gamma_j^{(k)}}{2} \quad \text{if neither the receiver and the sender effects are NULL} \end{cases}$$

Different combinations between $g(\cdot,\cdot)$ specifications and the three scenarios give rise to a set of 9 latent space models, incorporating varying degrees of heterogeneity.

Last, inference is carried out within a hierarchical Bayesian framework, and a Markov Chain Monte Carlo algorithm is employed for estimation of model parameters.

3 Conclusion

A class of latent space models for network and multidimensional networks is discussed. The models allow to flexibly account for transitivity and heterogeneity in network data, for both binary and weighted edges. Currently, only the class of models for binary networks is implemented in the *spaceNet* R package (`https://CRAN.R-project.org/package=spaceNet`), with the plan of including those for weighted networks in the near future.

References

D'Angelo, S., Alfò, M., & Murphy, T.B. 2020. Modeling node heterogeneity in latent space models for multidimensional networks. *Statistica Neerlandica.*, **74**, 324–341.

Gollini, I., & Murphy, T.B. 2016. Joint modeling of multiple network views. *Journal of Computational and Graphical Statistic.*, **25**, 246–265.

Hoff, P. 2005. Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association.*, **100**, 286–295.

Hoff, P., Raftery, A., & Handcock, M. 2002. Latent space approaches to social network analysis. *Journal of the American Statistical Association.*, **97**, 1090–1098.

Sewell, D., & Chen, Y. 2016. Latent space models for dynamic networks with weighted edges. *Social Networks.*, **44**, 105–116.

DTW-BASED ASSESSMENT OF THE PREDICTIVE POWER OF THE COPULA -DCC-GARCH-MST MODEL DEVELOPED FOR EUROPEAN INSURANCE INSTYTUTIONS

Anna Denkowska[1] and Stanisław Wanat[2]

[1]Department of Mathematics, Cracow University of Economics, Kraków, Poland.
(e-mail: anna.denkowska@uek.krakow.pl)
[2]Department of Mathematics, Cracow University of Economics, Kraków, Poland.
(e-mail: wanats@uek.krakow.pl)

ABSTRACT: We are investigating the possibilities of using the Dynamic Time Warping algorithm in two ways. A first way of using DTW is to assess the suitability of the Minimum Spanning Trees' topological indicators, which are constructed based on the tail dependence coefficients determined by the copula-DCC-GARCH model in order to establish the links between insurance companies in the context of potential shock contagion. A second way consists in using the DTW algorithm to group institutions by the similarity of their contribution to systemic risk, as expressed by DeltaCoVaR. The results obtained confirm the effectiveness of MST topological indicators for SR identification and evaluation of indirect links between insurance institutions.

KEYWORDS: time series analysis, Minimum Spanning Trees, topological indicators of the MST, Dynamic Times Warping, insurance sector, systemic risk Section Heading

1. Introduction

Our motivation is report of the European Insurance and Occupational Pensions Authority (EIOPA, 2017), encouraging to study the dynamics of interconnectedness between institutions. In the present article we use the Dynamic Time Warping (DTW - algorithm to determine the similarity between time series, which may be of different length and are distorted (stretched or shifted) in relation to the time axis) in two ways in the different market states: 1) to evaluate the suitability of Minimum Spanning Trees' topological indicators in the context of SR; 2) to construct the MST, to establish the similarity between the time series of the DeltaCoVaR. In the paper we analyze the dynamics of indirect connections between insurance companies that result from market price channels. We propose as in (Denkowska and Wanat, 2020) a hybrid approach to the analysis of interlinkages dynamics based on combining the copula-DCC-GARCH model and Minimum Spanning Trees (MST - connected and acyclic graph with the smallest sum of weights assigned to each edge; vertices are insurance institutions and the edges connect those lying at relatively small distances). The MST topology shows the links between institutions in the context of the possibility of propagating SR. We establish the similarity of time series' topological indicators of MST in periods of financial crises and outside of crises. Moreover, we examine the contribution of a single insurer to the systemic risk of the European insurance sector using the measure DeltaCoVaR (cf. Denkowska and Wanat, 2021).

SR in the financial sector was analyzed by: Bierth et al. (2015), Kanno (2016), Giglio et al. (2016), Kaserer (2018) and risk infection is studied by Hautsch et al. (2015). The paper (Petitjean et al. 2011) shows that the non-parametric DTW measure of similarity is better than other measures, such as the Pearson correlation coefficient.

2. Data and Methodology

We study the stock quotes of 38 European insurance institutions, most of them from the list of the top 50 insurance companies in Europe based on total assets. We analyze weekly logarithmic returns for the period from January 7th, 2005 to December 20th, 2019.

As in (Denkowska and Wanat, 2020) we carry out the analysis of the dynamics of interconnections between insurance companies using a new hybrid approach based on the combination of the copula-DCC-GARCH model and MST. For each period *t*, we determine the "distance" matrix between insurance companies using the metric: $d_t(i,j) = \sqrt{2(1-\lambda_t^L(i,j))}$ and the Kruskal algorithm (Mantegna and Stanley, 1999), we construct MST_t with 38 vertices and 37 edges.

Based on the trees thus obtained MST_t $(t = 1 \ldots T)$ we determine the time series of the following topological network indicators (Denkowska and Wanat, 2020): Average Path Length (APL - the average number of steps taken along all the shortest paths connecting all possible pairs of network nodes), Maximum Degree (Max.deg- highest number of edges arising from a vertex). Parameters „alpha" of the power law of the degree distribution, Network Diameter (length of the longest geodesic path between any two nodes), Rich Club Effect (RCE-well-connected vertices connect also one with another), Assortativity (graphical measure of the way vertices connect due to their degree).

Next we determine the DTW distance between the series in the following periods:

- the period of two subprime crises and excessive public debt; (February 8th, 2008- March 1st, 2013 - Subprime Mortgage Crisis (SMC)
- the period of crisis associated with the beginning of the migration crisis in Europe; (7th, 2015 to September 23rd, 2016) - Immigrant (I)
- the period of the crisis in France associated with strikes, and in Italy due to the ever-growing public debt; (April 21st, 2017 - May 11th, 2018 - France and Italy Crisis(FIC),
- the period- normal state – Normal (N), i.e. our periods: N_1 (January 7th, 2005 - February 1th, 2008), N_2 (March 8th, 2013 - July 31th, 2015), N_3 (September 30th, 2016 – April 14th, 2017), N_4 (May 18th, 2018 – December 20th, 2019).

DTW is one of the algorithms for measuring the similarity between two time series of different length that may differ in time (Raihan, 2017).

By examining the contribution to SR of all the analyzed insurance institutions, we establish a standard DeltaCoVaR measure for each of them described in the paper (Denkowska and Wanat, 2021).

2. Empirical results and discussion

MST's topological indicators constructed based on tail dependencies present different behaviors in the distinguished market states (Fig. 1). The analysis shows that during crises, MSTs shrink, as evidenced by the decreasing APL and Diameter and the growing Max.Deg. which is favorable to the potential spread of undesirable effects of the shocks on the insurance market. MSTs are scale-free in the studied period. The mean RCE for k = 4, where k is the degree of the vertex, is on a similar level. MSTs are non-assortative according to the previous definition, as the numbers are negative throughout the period considered.

The DTW results indicate a greater similarity of the APL time series fragments, separately in the periods of SMC, I, FIC, and in normal periods. The Diameter time series is noticeably divided into the group of SMC and FIC crises and a separate group of Normal states. The Max.Degree indicator remained at a similar level during the crises. During the entire period 2005-2019 MSTs are scale-free, as alpha has values in the range (2, 3). MSTs are not assortative in the entire analyzed period.

We present the average DeltaCoVaR for all analyzed institutions. We study the similarity of a fragment of this time series from the SMC period to other periods of crises or normal periods. SMC stands out in a separate group. Thus, not only the size of the SR contribution is observable on the basis of the time series itself, but also the dynamics of this contribution as assessed by the DTW is different. Also, the FIC or I crises are outside the group of similarities with most normal periods.

Now, using Kruskal's algorithm we construct MSTs based on the DTW$(\boldsymbol{i,j})$ distance matrix which show the similarity of DeltaCoVaR time series between pairs $(\boldsymbol{i,j})$ of insurers in states SMC, I, FIC and N. As a result of this analysis, we found that during the SMC crisis, the MST graph has the most compressed structure, as evidenced by the smallest APL, the largest MaxDegree and the smallest Assortativity (Tab. 1).

3. Conclusions

The presented analysis is the first work in the literature in which the possibilities of identifying SR in the insurance sector with the use of a hybrid model are determined by the copula-DCC-GARCH based MST with the DTW algorithm. Then, we use the DTW algorithm to analyze the similarity in different market regimes time series of the MST topological indicators. The results obtained confirm the possibility of identifying SR in the insurance sector using the presented model.

Funding: Cracow University of Economics, POTENTIAL Program, project number 26/EIM/2021/POT

Table 1. DTW (DeltaCoVaR)- based MST topological indicators.

	N	SMC	I	FIC
APL	7.18	6.50	8.67	8.42
max.deg	4.00	7.00	4.00	4.00
alpha	2.03	3.6	3.81	3.85
RCE (k=2)	0.43	0.44	0.30	0.14
diameter	0.01	0.02	0.02	0.01
Assort.	-0.34	-0.41	-0.29	-0.30

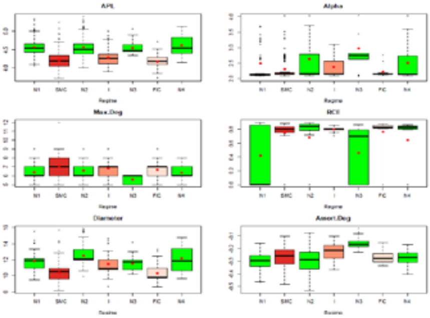


Figure 1. Topological indicators.

References

BIERTH, C., IRRESBERGER, F., & WEIß, G. N. 2015. Systemic risk of insurers around the globe. *Journal of Banking & Finance.*, **55**, 232-245.

DENKOWSKA, A., & WANAT, S. 2020. A Tail Dependence-Based MST and Their Topological Indicators in Modeling Systemic Risk in the European Insurance Sector. *Risks.*, **8(2)**, 1-39.

DENKOWSKA, A., & WANAT, S. 2021. A dynamic MST-deltaCoVaR model of systemic risk in the European insurance sector. *Statistics in Transition new series.*, **22(2)**, 173–188.

EIOPA. 2017. Systemic risk and macroprudential policy in insurance. Luxembourg: Publications Office of the EU.

GIGLIO, S., & KELLY, B., & PRUITT, S. 2016. Systemic risk and the macroeconomy: An empirical evaluation. *Journal of Financial Economics.*, **119(3)**, 457-471.

HAUTSCH, N., & SCHAUMBURG, J., & SCHIENLE, M. 2015. Financial Network Systemic Risk Contributions. *Review of Finance.*, **19(2)**, 685–738.

KANNO, M., 2016. The network structure and systemic risk in the global non-life insurance market. *Insurance: Mathematics and Economics.*, **67**, 38–53.

KASERER, C., & KLEIN, C. 2018. Supplementary Material to 'Systemic Risk in Financial Markets: How Systemically Important Are Insurers?' *Journal of Risk and Insurance.*, **86(3)**, 729-759.

PETITJEAN, F., & KETTERLIN, A., & GANCARSKI, P. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition.*, **44(3)**, 678–693.

RAIHAN, T. 2017. Predicting US recessions: A dynamic time warping exercise in economics. SSRN 3047649

TWO–STEP ESTIMATION OF MULTILEVEL LATENT CLASS MODELS WITH COVARIATES

Roberto Di Mari[1], Zsuzsa Bakk[2],Jennifer Oser[3]
and Jouni Kuha[4]

[1] Department of Economics and Business, University of Catania, (e-mail: roberto.dimari@unict.it)

[2] Department of Methodology and Statistics,Leiden University, The Netherlands

[3] Department of Politics and Government, Ben-Gurion University, Israel

[4] Department of Statistics, London School of Economics and Political Science, London, UK

ABSTRACT: In this article we present a two-step estimation approach applied to multilevel latent class analysis (LCA) with covariates. In the first step, the measurement model for the low-level and the high-level latent class variables is estimated. In the second step, covariates are added as predictors of latent class memberships, keeping the measurement model parameters fixed at their first step values. Separating the estimation of the structural from the measurement model generates a significant computational gain with respect to simultaneous estimation, greatly simplifying model building. Finite sample properties of the resulting estimator are investigated in a broad simulation study.

KEYWORDS: multilevel latent class analysis; covariates; two-step estimation; pseudo maximum likelihood

1 Introduction

Latent class (LC) analysis is an approach used to create a clustering of a set of observed variables, based on an underlying unknown classification. In the multilevel extension of the baseline LC model, the respondents are assumed to belong to higher level groups - e.g. students nested in schools, or households in countries. Multilevel LCA is becoming increasingly popular in various fields. In most applications the focus is on lower level clustering, and on the difference in the distribution of the lower level classes in higher level units.

In LCA creating a clustering is usually only the first step for applied researchers. The research interest often lies in including external variables as clustering predictors at a later stage of the analysis. While in single level LCA different approaches are available for relating LC membership to external variables, in multilevel settings only two classical approaches are used, both known to be suboptimal, namely the one-step and classical three-step approaches. Using the one-step approach the full LC model including covariates is estimated simultaneously (for example, Mutz & Daniel, 2013). Using the alternative three-step approach, after estimating the measurement model in step 1, respondents are assigned to latent classes in step 2, and this posterior assigned class membership is related to the predictors of interest through a multinomial logistic regression in the third step (for example Tomczyk *et al.*, 2015). However in the second step a classification error is introduced, that if not corrected for induces systematic bias in the step 3 model.

In the current paper we introduce a two-step approach, extending Bakk & Kuha (2018)'s work to the multilevel LC model as an alternative to the one-step and classical three-step approaches, since both are known to be sub-optimal in single level LC models.

2 The multilevel latent class model

Consider the vector of responses $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijK})$, where Y_{ijk} denotes the response of individual i in group j on the k-th categorical indicator variable, with $1 \leq k \leq K$ and $1 \leq j \leq J$, where K denotes the number of categorical indicators and J the number of level 2 units. In addition, we let n_j denote the number of level 1 units within the j-th level 2 unit, with $1 \leq j \leq J$. For simplicity of exposition, we focus on dichotomous indicators.

Adopting the nonparametric approach (Laird, 1978), multilevel LC analysis is an extension of the LC models (Goodman, 1974), assuming that level 1 units belong to one of the T categories belong to T categories ("latent classes") of an underlying categorical latent variable X, whereas level 2 units belong to one of the M categories of the group level latent class W. The model for $\mathbf{Y}_{ij}$ can then be specified as

$$P(\mathbf{Y}_{ij}) = \sum_{m=1}^{M} P(W_j = m) \sum_{t=1}^{T} P(X_{ij} = t | W_j = m) P(\mathbf{Y}_{ij} | X = t) \qquad (1)$$

where $P(W_j = m) = \pi_m$ is the probability of group j to belong to class m. $P(X_{ij} = t | W_j = m)$ is the probability that individual i in group j belongs to class t given group membership m. The term $P(\mathbf{Y}_{ij} | X = t)$ is the class-specific probability of observing a pattern of responses given that a person belongs to class t under the common assumption that item-conditional probabilities not to depend on the level 2 unit (Vermunt, 2003; Lukociene *et al.*, 2010). Furthermore we make the "local independence" assumption that the K indicator are independent within latent classes, leading to

$$P(\mathbf{Y}_{ij}) = \sum_{t=1}^{T} P(X_{ij} = t) \prod_{k=1}^{K} P(Y_{ijk} | X_{ij} = t). \qquad (2)$$

The multilevel LC model of Equation (1) can be parametrized by means of multinomial logistic regressions as follows

$$P(Y_{ijk}|X_{ij}=t) = \frac{\exp(\beta_t^k)}{1+\exp(\beta_t^k)}, \tag{3}$$

for the item-class probabilities,

$$P(W_j=m) = \frac{\exp(\delta_{0m})}{1+\sum_{l=2}^{M}\delta_{0l}}, \tag{4}$$

for the group-level membership probabilities, and

$$P(X_{ij}=t|W_j=m) = \frac{\exp(\gamma_{tm})}{1+\sum_{s=2}^{T}\exp(\gamma_{sm})} \tag{5}$$

for the individual latent class probabilities.

Under the parametrizations (3), (5) and (4), given a sample of J groups, the model parameters can be found by maximizing

$$\log L(\theta_1) = \sum_{j=1}^{J}\log P(\mathbf{Y}_{ij}), \tag{6}$$

with respect to $\theta_1 = (\delta_{02},\ldots,\delta_{0M},\beta^1_{2_1},\ldots,\beta^K_{T_J})'$.

Level 1 and level 2 covariates can be included to predict class membership. Denoting one level 2 covariate by Z_{1j} and a level 1 covariate by Z_{2ij} the multinomial logistic regression for X_{ij} with a random intercept can be written as:

$$P(X_{ij}=t|W_j,Z_{1j},Z_{2ij}) = \frac{\exp(\gamma_{0tm}+\gamma_{1t}Z_{1j}+\gamma_{2t}Z_{2ij})}{\sum_{s=1}^{T}\exp(\gamma_{0sm}+\gamma_{1s}Z_{1j}+\gamma_{2s}Z_{2ij})}. \tag{7}$$

A random slope for the level 1 covariate can be obtained by replacing γ_{2t} by γ_{2jt}. Level 2 covariates can be used also to predict group class membership, but for simplicity we present only a model with covariates on the level 1 LC variable.

Under the parametrization (7) that now includes covariates, the model for $\mathbf{Y}_{ij}|\mathbf{Z}_j$, where $\mathbf{Z}_j = (Z_{1j},Z_{2ij})'$, can be specified as

$$P(\mathbf{Y}_{ij}|\mathbf{Z}_j) = \sum_{m=1}^{M}P(W_j=m)\sum_{t=1}^{T}P(X_{ij}=t|W_j=m,Z_{1j},Z_{2ij})\prod_{k=1}^{K}P(Y_{ijk}|X_{ij}=t), \tag{8}$$

which depends on the vector of unknown parameters $\theta = (\theta_1,\theta_2)'$, where $\theta_2 = (\gamma_{12},\ldots,\gamma_{1T},\gamma_{22},\ldots,\gamma_{2T})'$. The one step approach finds $\hat{\theta}$ by maximizing

$$\log L(\theta) = \sum_{j=1}^{J}\log P(\mathbf{Y}_j|\mathbf{Z}_j), \tag{9}$$

with respect to θ.

3 A stepwise estimator for multilevel LC model with covariates

Step 1: the ML estimate $\widehat{\theta}_1$ of θ_1 is found as the maximizer of the log-likelihood of the simple multilevel LC model without covariates.

Step 2: covariates are added to the model. The log-likelihood (9) is maximized only with respect to θ_2, and θ_2 is kept fixed at its first step estimates.

Our 2-step estimator is an instance of pseudo maximum likelihood estimation (Gong & Samaniego, 1981). Such estimators are consistent under very general regularity conditions (see, for instance, Gourieroux & Monfort, 1995). We propose to compute the step-two standard errors to account for the uncertainty about the fixed parameters in the calculation applying the approach proposed by Bakk & Kuha (2018) for single level LC models to the multilevel setting.

We will setup a simulation study to assess the finite sample properties of the proposed estimator. To do so, we will generate data with varying sample sizes at both the lower and higher level, with different levels of class separation and association between the covariates and class membership. We expect that the proposed two-step estimator will be unbiased (similarly to the one-step approach) as opposed to the three step approach, and will be slightly less efficient than the one-step estimator.

References

BAKK, Z, & KUHA, J. 2018. Two-Step Estimation of Models Between Latent Classes and External Variables. *Psychometrika*, **83**, 871–892.

FINCH, W HOLMES, & FRENCH, BRIAN F. 2014. Multilevel latent class analysis: Parametric and nonparametric models. *The Journal of Experimental Education*, **82**(3), 307–333.

GONG, GAIL, & SAMANIEGO, FRANCISCO J. 1981. Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 861–869.

GOODMAN, LEO A. 1974. The Analysis of Systems of Qualitative Variables When Some of the Variables Are Unobservable. Part I: A Modified Latent Structure Approach. *American Journal of Sociology*, 79–259.

GOURIEROUX, CHRISTIAN, & MONFORT, ALAIN. 1995. *Statistics and Cconometric Models*. Vol. 1. Cambridge University Press.

LAIRD, NAN. 1978. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73**(364), 805–811.

LUKOCIENE, O., VARRIALE, R., & VERMUNT, J.K. 2010. The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociological Methodology*, **40**(1), 247–283.

MUTZ, R., & DANIEL, H.D. 2013. University and student segmentation: Multilevel latent-class analysis of students' attitudes towards research methods and statistics. *British Journal of Educational Psychology*, **83**(2), 280–304.

TOMCZYK, SAMUEL, HANEWINKEL, REINER, & ISENSEE, BARBARA. 2015. Multiple substance use patterns in adolescents: A multilevel latent class analysis. *Drug and alcohol dependence*, **155**, 208–214.

VERMUNT, JEROEN K. 2003. Multilevel Latent Class Models. *Sociological Methodology*, **33**(1), 213–239.

CLUSTERING DATA WITH NON-IGNORABLE MISSINGNESS USING SEMI-PARAMETRIC MIXTURE MODELS

Marie Du Roy de Chaumaray[1] and Matthieu Marbac[1]

[1] Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France, (e-mail: marie.du-roy-de-chaumaray@ensai.fr, matthieu.marbac-lourdelle@ensai.fr)

ABSTRACT: We are concerned in clustering continuous data sets subject to non-ignorable missingness. Clustering is achieved by a semi-parametric mixture that, for each subject, considers the joint distribution of the observed variables and the response-data indicator vector. Estimation is performed by maximizing the smoothed likelihood via a Majoration-Minimization algorithm.

KEYWORDS: Clustering, Mixture Model, Non-ignorable Missigness, Smoothed Likelihood.

1 Introduction

Mixture models permit to achieve the clustering purpose in a rigorous context but the case where data have missingness is generally neglected. Moreover, the missing not at random scenario (MNAR; Little & Rubin, 2019), where the missingness mechanism depends on the missing values even conditionally on the observed variables, generally requires the missingness mechanism to be considered to obtain consistent estimators. However, few statistical methods permit this scenario for clustering.

Two clustering approaches allow data subject to the MNAR scenario to be analyzed. Chi *et al.* , 2016 introduce the K-POD algorithm that extends the K-means algorithm to the case of missing data even if the missing mechanism if unknown. However, this approach suffers from the standard drawbacks of the K-means algorithm (*i.e.,* assumptions of spherical clusters and equals proportions of the clusters). Alternatively, using a *selection model* approach Miao *et al.* , 2016 proposed a specific Gaussian mixtures and t-mixtures to analyze data under MNAR scenario. For such approach, the missingness mechanism must be specified (probit and logit distributions are generally used). However, this approach produces strong bias if the parametric assumptions (made on the distribution of the variables or on the missingness mechanism) are violated.

In this paper, clustering is performed via a mixture model that uses a *pattern-mixture model* approach with non-parametric distributions. Thus, no assumptions are made on the data distribution or on the missingness mechanism except that the variables are independent within components. Note that this assumption is quite standard for semi-parametric mixtures (Levine *et al.* , 2011; Kasahara & Shimotsu, 2014). For each mixture component, we estimate, for each variable, its probability to be observed and its conditional distribution given the variables is observed. We emphasize that our concern is clustering and not imputation or density estimation. Indeed, without adding assumptions, the distribution of the variables within component cannot be estimated by our procedure.

2 Mixture for nonignorable missingness

2.1 The data

The observed sample is composed of n independent and identically distributed subjects arisen form K homogeneous subpopulations. Each subject is described by d continuous variables and some realizations of these variables may be unobserved. The probability, for a variable, to be not observed is allowed to depend on the values of the variable itself and the subpopulation membership.

Each subject i is described by a vector of three variables $(X_i^\top, R_i^\top, Z_i^\top)^\top$ where $X_i \in \mathbb{R}^d$ is set of continuous variables, $R_i = (R_{i1}, \dots, R_{id})^\top \in \{0,1\}^d$ indicates whether X_{ij} is observed ($R_{ij} = 1$) and $Z_i = (Z_{i1}, \dots, Z_{iK})^\top$ indicates the subpopulation of subject i ($Z_{ik} = 1$ if subject i belongs to subpopulation k and otherwise $Z_{ik} = 0$). Each subject belongs to one subpopulation such that $\sum_{k=1}^K Z_{ik} = 1$. The realizations of Z_i are unobserved and a part of the realizations of X_i can be unobserved too. Therefore, the observed variables for subject i are $(X_i^{\text{obs}\top}, R_i^\top)^\top$ where X_i^{obs} is composed of the elements of X_i such that $R_{ij} = 1$ and the unobserved variables for subject i are $(X_i^{\text{miss}\top}, Z_i^\top)^\top$ where X_i^{miss} is composed of the elements of X_i such that $R_{ij} = 0$.

2.2 General mixture model

We use mixture models in a purpose of clustering and not for density estimation. Clustering aims to estimate the subpopulation memberships given the observed variables (*i.e.,* the realization of Z_i given $(X_i^{\text{obs}\top}, R_i^\top)^\top$) without assumption on the missingness mechanism (*i.e.,* no assumption are made on the conditional distribution of $R_i \mid X_i, Z_i$). The probability distribution function (pdf) of $(X_i^\top, R_i^\top)^\top$ for subpopulation k (*i.e.,* $Z_{ik} = 1$) is denoted by $g_k(\cdot)$. Using the *pattern-mixture model*, the pdf $(X_i^\top, R_i^\top)^\top$ is defined by the pdf of the K-component mixture

$$g(x_i, r_i; \theta) = \sum_{k=1}^{K} \pi_k g_k(x_i, r_i; \tau_k) \text{ with } g_k(x_i, r_i; \tau_k) = g_k(r_i; \tau_k) g_k(x_i \mid r_i), \quad (1)$$

where $\pi_k > 0$, $\sum_{k=1}^{K} \pi_k = 1$ and $g_k(\cdot; \tau_k)$ is pdf of component k. The couples of variables $(X_{ij}, R_{ij})^\top$ are assumed to be conditionally independent given Z_i. Thus, the distribution of $R_i \mid Z_i$ is a product of Bernoulli distributions and the conditional density of $X_i \mid Z_i, R_i$ is defined as the product of univariate densities. Thus, from (1), the pdf of component k is also defined as

$$g_k(r_i; \tau_k) = \prod_{j=1}^{d} \tau_{kj}^{r_{ij}} (1 - \tau_{kj})^{1-r_{ij}} \text{ and } g_k(x_i \mid r_i) = \prod_{j=1}^{d} p_{kj}^{r_{ij}}(x_{ij}) q_{kj}^{1-r_{ij}}(x_{ij}),$$

where $\tau_k = (\tau_{k1}, \ldots, \tau_{kd})^\top$, τ_{kj} is the probability that X_{ij} is observed given that subject i belongs to subpopulation k, $p_{kj}(\cdot)$ is the conditional density of X_{ij} given $Z_{ik} = 1$ and $R_{ij} = 1$ and $q_{kj}(\cdot)$ is the conditional density of X_{ij} given $Z_{ik} = 1$ and $R_{ij} = 0$. Integrated out the unobserved variables X_i^{miss}, we have

$$g(x_i^{\text{obs}}, r_i; \theta) = \sum_{k=1}^{K} \pi_k g_k(x_i^{\text{obs}}, r_i; \tau_k), \text{ with } g_k(x_i^{\text{obs}}, r_i; \tau_k) = g_k(r_i; \tau_k) \prod_{j=1}^{d} p_{kj}^{r_{ij}}(x_{ij}),$$

where θ groups all the finite parameters (π_k and τ_k) and all the infinite parameters $p_{kj}(\cdot)$. For clustering, the *pattern-mixture model* should be preferred to *selection model* because it does not require to specify the missingness mechanism, allows this mechanism to be nonignorable and permits to easily obtain the conditional probabilities of the subpopulation membership given the distribution of the observed values defined by

$$\mathbb{P}(Z_{ik} = 1 \mid x_i^{\text{obs}}, r_i) = \frac{g_k(x_i^{\text{obs}}, r_i; \tau_k)}{\sum_{\ell=1}^{K} \pi_\ell g_\ell(x_i^{\text{obs}}, r_i; \tau_\ell)}.$$

Note that we do not need to estimate $q_{kj}(\cdot)$ for the clustering purpose but that this implies that we are not able to estimate the distribution of $X_i \mid Z_i$. Thus, this approach does not permit to estimate the marginal distribution of

$X_i \mid Z_i$ without adding assumptions on the missing mechanism. This implies that the proposed approach can be used for clustering but not for density estimation. Model identifiability is obtained by extending Theorem 8 in Allman *et al.* , 2009. Parameter estimation is performed by maximizing the smoothed likelihood over θ via a MM algorithm like in Levine *et al.* , 2011. More details are given in Du Roy de Chaumaray & Marbac, 2020.

3 Conclusion

The proposed method allows continuous data set with non-ignorable missingness to be clustered with no more assumption than the independence within components. Selecting the number of components is a difficult task that could be achieved by extending the approach of Kasahara & Shimotsu, 2014 to the mixed-type data. A procedure of bandwidth selection should be investigated.

References

ALLMAN, ELIZABETH S, MATIAS, CATHERINE, RHODES, JOHN A, *et al.* . 2009. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, **37**(6A), 3099–3132.

CHI, JOCELYN T, CHI, ERIC C, & BARANIUK, RICHARD G. 2016. k-pod: A method for k-means clustering of missing data. *The American Statistician*, **70**(1), 91–99.

DU ROY DE CHAUMARAY, MARIE, & MARBAC, MATTHIEU. 2020. Clustering Data with Nonignorable Missingness using Semi-Parametric Mixture Models. *arXiv preprint arXiv:2009.07662*.

KASAHARA, HIROYUKI, & SHIMOTSU, KATSUMI. 2014. Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(1), 97–111.

LEVINE, MICHAEL, HUNTER, DAVID R, & CHAUVEAU, DIDIER. 2011. Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, 403–416.

LITTLE, RODERICK JA, & RUBIN, DONALD B. 2019. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

MIAO, WANG, DING, PENG, & GENG, ZHI. 2016. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, **111**(516), 1673–1683.

SPATIAL-TEMPORAL CLUSTERING BASED ON B-SPLINES: ROBUST MODELS WITH APPLICATIONS TO COVID-19 PANDEMIC

Pierpaolo D'Urso [1], Livia De Giovanni[2] and Vincenzina Vitale[1]

[1] Department of Social and Economic Sciences, Sapienza University of Rome, P.za Aldo Moro, 5 - 00185 Rome, Italy, (e-mail: pierpaolo.durso@uniroma1.it, vincenzina.vitale@uniroma1.it)

[2] Department of Political Sciences, LUISS university, Viale Romania, 32 - 00197 Rome, Italy, (e-mail: ldegiovanni@luiss.it)

ABSTRACT: Robust fuzzy C-Medoids clustering models based on B-splines with spatial penalty term have been proposed to cluster Italian regions according to the daily time-series of the cumulative COVID-19 cases over population (per 10000 inhabitants) and of the cumulative COVID-19 deaths over population (per 10000 inhabitants), spanning from 2020-02-24 to 2021-02-08. Both spatial and time components have been efficiently embedded in the model. Furthermore the use of B-splines coefficients allows to reduce consistently the computational burdern.

KEYWORDS: B-splines, robust distance, PAM algorithm, COVID-19 data, contiguity matrix

1 Introduction

The new 2019 coronavirus that has originated the COVID-19 disease spread out quickly from the Chinese city of Wuhan worldwide giving rise to a pandemic whose huge effects on national health systems are still evident. It has been well known that Italy, its Northern regions in particular, was the first country facing the outbreak in February 2020 to such an extent that the Italian government needed to impose a nationwide lockdown (on 9 March 2020) to dastrically reduce the incidence rate and the overfloading of the intensive care units. Italy faced other two outbreak waves, in October 2020 and then in March 2021, involving all territories, the Southern ones too. Three and then five risk profiles have been identified by the Scientific committee engaged by the national authorities to monitor pandemic's dynamic in order to differentiate the restrictive measures in the territories. At the beginning of 2021, the COVID-19 vaccination campaign has been started and is ongoing to this day. The Italian Civil Protection Department provides, daily, all information related to the COVID-19 outbreak in Italy at the regional level and, for some variables, also at the provincial level even if data reliability is low due to misreporting and lack of uniformity in the number of swabbed people per region. In this study we focused on the daily time-series of the cumulative cases over population (per 10000 inhabitants) and on the cumulative deaths over population (per 10000 inhabitants), spanning from 2020-02-24 to 2021-02-08, at a regional level. The aim of the work is to cluster Italian regions based on the aforementioned rates, separately. Being spatial time series, the proposed clustering appoaches embedded both the spatial and time components by including a spatial penalization term in the objective function, as proposed by D'Urso *et al.*, 2019, and a suitable transformation of the time series onto (finite dimensional) vectors of cubic B-splines basis coefficients. To deal with noisy data and outliers, three robust approaches have been proposed, one based on a exponential transformation of the distance, the other two based on the trimming and noise approach, respectively. The paper is structured as follows. Section 2 focuses on the proposed clustering models while Section 3 on the application to COVID-19 data.

2 The Spatial-temporal clustering based on B-splines: robust methods

In a formal way, a spatial-time data matrix can be algebraically defined as (D'Urso, 2000):

$$\mathbf{X} \equiv \{x_i(t) : i = 1,\ldots,I; t = 1,\ldots,T\} \tag{1}$$

where i indicates the generic spatial unit and t the generic time. The time series $\{(t, x_i(t))\}$ could be seen as the result of collecting a variable X on unit i at the T times $\{t = 1,\ldots,T\}$. We can model each time series by a simple linear least-squares fit as:

$$x_i(t) = \sum_{s=1}^{p} b_i^s B_s(t) + \varepsilon_i,\ t = 1,\ldots T$$

where $\{B_s(\cdot)\}_{s=1}^{p}$ are p-dimensional functional basis. For the I time series $\mathbf{x}_i$, $i = 1,\ldots,I$, we will have I vectors of fitted coefficients $\mathbf{b}_i = (b_i^1,\ldots,b_i^s,\ldots,b_i^p)'$, $i = 1,\cdots,I$. For sake of simplicity, we show the results with reference to the Spatial-Temporal based on Exponential distance Fuzzy C-Medoids clustering model (ST-BS-Exp-FCMd), even if the same problem can be addressed by using the Spatial-Temporal Fuzzy Trimmed C-Medoids clustering model (ST-BS-Tr-FCMd) and the Spatial-Temporal Fuzzy C-Medoids clustering model with Noise Cluster (ST-BS-Noise-FCMd). The ST-BS-Exp-FCMd model is defined as follows:

$$\begin{aligned} &\min: \sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}[1-\exp(-\beta\|\mathbf{b}_i-\widetilde{\mathbf{b}}_c\|^2)]+\frac{\gamma}{2}\sum_{i=1}^{I}\sum_{c=1}^{C} u_{ic}^{m}\sum_{i'=1}^{I}\sum_{c'\in C_c} p_{ii'}u_{i'c'}^{m} \\ &\quad s.t. \sum_{c=1}^{C} u_{ic}=1,\; u_{ic}\geq 0 \end{aligned} \tag{2}$$

where $\mathbf{b}_i$ and $\widetilde{\mathbf{b}}_c$ are the vectors of coefficients of the B-spline representation of the i-th spatial time series and of the c-th spatial medoid (c=1,...,C) respectively, while $m > 1$ is well-known fuzziness parameter. The β parameter is set as the inverse of the variability of the data and appropriately tunes the distance according to the variability of the data.
As far as the spatial penalty term is concerned, γ is the tuning parameter of spatial information. The spatial proximity among the I objects, has been taken into account by means of the contiguity matrix $\mathbf{P}_{I\times I}$ where the generic element $p_{ii'}$=1 if the object i is contiguous to the object i', 0 otherwise. The u_{ic} is the membership degree of the unit i belonging to the cluster c:

$$u_{ic}=\frac{\left[[1-\exp(-\beta\|\mathbf{b}_i-\widetilde{\mathbf{b}}_c\|^2)]+\gamma\sum_{i'=1}^{I}\sum_{c'\in C_c} p_{ii'}u_{i'c'}^{m}\right]^{-\frac{1}{m-1}}}{\sum_{c'=1}^{C}\left[[1-\exp(-\beta\|\mathbf{b}_i-\widetilde{\mathbf{b}}'_c\|^2)]+\gamma\sum_{i'=1}^{I}\sum_{c''\in C_{c'}} p_{ii'}u_{i'c''}^{m}\right]^{-\frac{1}{m-1}}} \tag{3}$$

For $\gamma = 0$, the ST-BS-Exp-FCMd model reduces to its no-spatial version, the BS-Exp-FCMd clustering model, while the ST-BS-Tr-FCMd and ST-BS-Noise-FCMd models to their no-spatial versions, the BS-Tr-FCMd and BS-Noise-FCMd models, respectively.

3 Clustering of Italian regions - COVID-19 data

In this study, we show the results with reference to the ST-BS-Exp-FCMd model applied to cluster the I=20 Italian regions during T=351 times represented by the days from 2020-02-24 to 2021-02-08. The optimal number of clusters has been identified running the model, with $\gamma = 0$ and $m = 1.5$, and choosing the number of groups that maximizes the Fuzzy Silhouette index. Then fixed C, the optimal value of γ has been chosen according a heuristic procedure based on maximization of the spatial autocorrelation measure introduced in Coppi *et al.*, 2010. To assign each region to a specific cluster we have set the cut-off value $u_{ic} \geq 0.6$ (Maharaj & D'Urso, 2011). The clustering results are reported in Table 1 for both models[*]. For each one, three clusters have been selected, whose medoids are denoted in each column header. For the *Total cases over population*, two fuzzy units has been identified, Basilicata and Sicily, the latter characterized by an anomalous increase of infections during the second wave. For the *The total deaths over population* due to COVID-19 desease, two fuzzy units have been identified, Aosta Valley and Sicily; the former is a global outlier, the latter a local one. We argue that the three clusters matched with three risk levels, from the highest to the lowest one. The main advantages of this methodology consist in data reduction obtained by using B-splines coefficients, in robustness by using the exponential, trimming and noise approaches while spatial information is taken into account adding a penalty term in the objective function.

Table 1. *Total cases (columns 1-3) and Total deaths (columns 4-6) over population (per 10000 inhabitants) - 3 clusters memberships*

		Model with no spatial penalty ($\gamma = \gamma_{opt}$) for Total cases over pop.			**Model with spatial penalty ($\gamma = \gamma_{opt}$) for Total deaths over pop.**		
	Region	**Piedmont**	**Lazio**	**Calabria**	**Trentino-South Tyrol**	**Lazio**	**Calabria**
1	Piedmont	**1.000**	0.000	0.000	0.956	0.023	0.021
2	Aosta Valley	0.697	0.154	0.150	*0.563*	0.218	0.218
3	Lombardy	0.992	0.004	0.003	0.812	0.094	0.094
4	Trentino-South Tyrol	0.936	0.035	0.029	**1.000**	0.000	0.000
5	Veneto	0.849	0.083	0.068	0.975	0.014	0.011
6	Friuli-Venezia Giulia	0.688	0.222	0.089	0.850	0.095	0.055
7	Liguria	0.881	0.087	0.032	0.834	0.097	0.070
8	Emilia-Romagna	0.784	0.16	0.056	0.846	0.093	0.061
9	Tuscany	0.095	0.860	0.045	0.194	0.708	0.097
10	Umbria	0.031	0.953	0.016	0.004	0.99	0.007
11	Marche	0.019	0.964	0.017	0.173	0.748	0.079
12	Lazio	0.000	**1.000**	0.000	0.000	**1.000**	0.000
13	Abruzzo	0.000	0.999	0.000	0.011	0.978	0.011
14	Molise	0.013	0.943	0.044	0.000	1.000	0.000
15	Campania	0.020	0.964	0.015	0.001	0.997	0.002
16	Apulia	0.018	0.910	0.072	0.001	0.998	0.001
17	Basilicata	0.044	0.433	*0.523*	0.042	0.744	0.215
18	Calabria	0.000	0.000	**1.000**	0.000	0.000	**1.000**
19	Sicily	*0.376*	0.279	*0.345*	*0.563*	0.298	0.139
20	Sardinia	0.027	0.036	0.937	0.000	0.001	0.999

References

COPPI, R., D'URSO, P., & GIORDANI, P. 2010. A fuzzy clustering model for multivariate spatial time series. *J. Classification*, **27**(1), 54–88.

D'URSO, P. 2000. Dissimilarity measures for time trajectories. *Stat. Methods Appl.*, **9**(1-3), 53–83.

D'URSO, P., DE GIOVANNI, L., DISEGNA, M., & MASSARI, R. 2019. Fuzzy clustering with spatial–temporal information. *Spatial Statistics*, **30**, 71–102.

MAHARAJ, E A, & D'URSO, P. 2011. Fuzzy clustering of time series in the frequency domain. *Information Sciences*, **181**(7), 1187–1211.

[*] One notices that, in the contiguity matrix, Calabria and Sicily have been considered contiguous since they have very frequent ferry connections.

PIVMET: PIVOTAL METHODS FOR BAYESIAN RELABELLING IN FINITE MIXTURE MODELS

Leonardo Egidi [1], Roberta Pappadà [1], Francesco Pauli [1] and Nicola Torelli [1]

[1] Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'Bruno de Finetti', Università degli Studi di Trieste (e-mail: legidi@units.it, rpappada@units.it, francesco.pauli@deams.units.it, nicola.torelli@deams.units.it)

ABSTRACT: The identification of groups' prototypes, i.e. elements of a dataset that are representative of the group they belong to, is relevant to the tasks of clustering, classification and mixture modeling. The R package pivmet includes different methods for extracting pivotal units from a dataset, to be exploited for a Markov Chain Monte Carlo (MCMC) relabelling technique for dealing with label switching in Bayesian estimation of mixture models. Moreover, consensus clustering based on pivotal units may improve classical algorithms (e.g. k-means) by means of a careful seeding.

KEYWORDS: pivotal unit, mixture model, relabelling, consensus clustering.

1 Introduction

The identification of some units which may be representative of the group they belong to is often a matter of statistical importance and can help avoiding an extra amount of work when processing the data. The advantage of such pivotal units (hereafter called pivots) is that they are somehow chosen to be as far as possible from units in the other groups and as similar as possible to the units in the same group, and may be beneficial in many statistical frameworks, such as clustering, classification, and mixture modeling.

The pivmet R package (Egidi *et al.*, 2021) implements various pivotal selection criteria, graphical tools and the relabelling method (Papastamoulis, 2016) described in Egidi *et al.*, 2018 to deal with 'label switching' (Redner & Walker, 1984), a well-known phenomenon causing nonidentifiability of the mixture parameters during the MCMC sampling (Frühwirth-Schnatter, 2001). Compared to other packages, it allows the user to fit their own mixture model using data augmentation with component memberships either via the JAGS (Plummer, 2018) or the Stan (Carpenter *et al.*, 2017) software, by specifying suitable prior distributions. Pivotal units are detected via the similarity matrix derived from the MCMC sample—whose elements are the estimated probabilities that any two units in the observed sample are drawn from the same component—and used to relabel the chains. Such units may be fruitfully used in Dirichlet process mixture models (DPMM) (Ferguson, 1973, Neal, 2000), a class of models that naturally sorts data into clusters, and in data clustering to guarantee a better final clustering solution starting from a careful seeding based on well-separated statistical units.

The aim of the paper is to provide a quick overview of the computational capabilities of our package in the field of Bayesian mixture models.

2 Finite mixtures of Gaussian distributions

Consider a multivariate mixture of Gaussian distributions, let $\boldsymbol{y}_i \in \mathbb{R}^d$ and assume that

$$\boldsymbol{y}_i \sim \sum_{j=1}^{k} \eta_j \mathcal{N}_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \;\; i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\mu}_j \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_j$ is a $d \times d$ positive definite covariance matrix. We assume the following prior specification for the parameters in (1):

$$\begin{aligned} \boldsymbol{\mu}_j \sim & \mathcal{N}_2(\boldsymbol{\mu}_0, S_2), \;\; \boldsymbol{\Sigma}_j^{-1} \sim \text{Wishart}(S_3, d+1) \\ \eta \sim & \text{Dirichlet}(\boldsymbol{\alpha}), \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha}$ is a k-dimensional vector and S_2 and S_3 are positive definite matrices. We fix $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\alpha} = (1, \ldots, 1)$ and assume S_2 and S_3 are diagonal matrices, with diagonal elements equal to 10^5. We simulate a sample of size $n = 150$ from a bivariate Gaussian distribution with the function `piv_sim` and we fit the model using the JAGS option. From the bivariate traceplot chains for each mean component $\mu_{j,1}$ and $\mu_{j,2}$ in Figure 1 we clearly note that label switching has occurred and the relabelling algorithm fixed it, by isolating the four bivariate high-density regions.

3 Dirichlet Process Mixture Models

DPMM are useful tools for non-parametric density estimation, and, more generally, the choice of a Dirichlet process prior avoids the specification of an inappropriate parametric form. The DPMM has the following form:

$$\begin{aligned} y_i &\sim K(y_i|\theta_i), \; i = 1, \ldots, n, \\ \theta_i &\sim F, \;\; F \sim \text{ DP}(\alpha, G), \end{aligned} \tag{3}$$

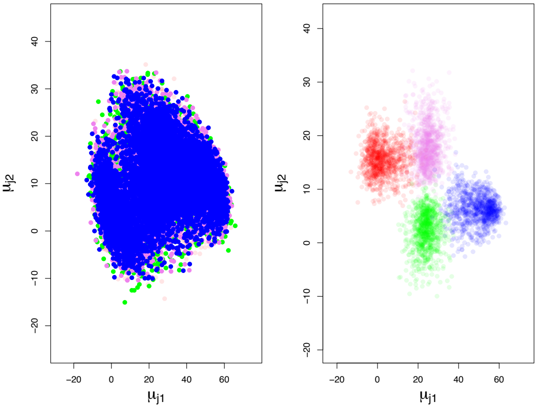


Figure 1. *Bivariate mixture data: scatterplot for the mean parameters obtained via JAGS sampling (left plot) and relabelled estimates (right plot) via the* `maxsumdiff` *pivotal criterion.*

where $K(\cdot)$ is a parametric kernel function which is usually continuous, F is an unknown probability distribution, DP is the nonparametric Dirichlet process prior with concentration parameter α and *base measure* G, which encapsulates any prior knowledge about F. A common choice for $K(\cdot)$ is a Gaussian mixture model, so that $K(y_i|\theta_i) = \mathcal{N}(\mu_i, \sigma_i^2)$. The DPMM sorts the data into clusters, corresponding to the mixture components. Thus, it may be seen as an infinite dimensional mixture model which generalizes finite mixture models. Thus, pivotal units detection may be quite relevant for this class of models in order to identify distinct groups characteristics. We generate $n = 200$ data from a student$-t$ distribution with 3 degrees of freedom and we draw posterior samples for $\mu_1, \mu_2, \dots, \mu_k$ via the `dirichletprocess` package. Figure 2 represents posterior density estimation for the simulated dataset along with nine pivotal units (blue points). detected by `pivmet` via the `maxsumdiff` pivotal criterion.

References

Carpenter, Bob, Gelman, Andrew, Hoffman, Matthew D, Lee, Daniel, Goodrich, Ben, Betancourt, Michael, Brubaker, Marcus A, Guo, Jiqiang, Li, Peter, & Riddell, Allen. 2017. Stan: a probabilistic programming language. *Journal of Statistical Software*, **76**(1), 1–32.

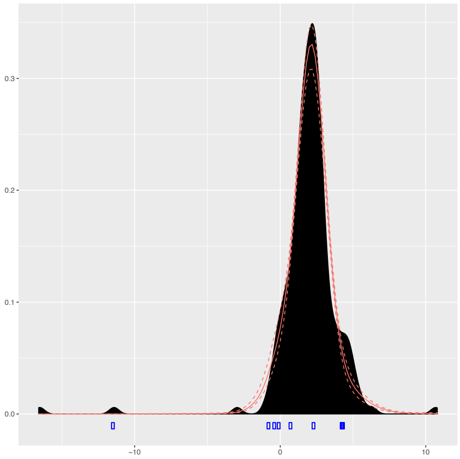


Figure 2. *Posterior density estimation (red line) for a sample of $n = 200$ data points from a student$-t$ distribution. Blue points below the x-axis denote the pivotal units.*

Egidi, Leonardo, Pappadà, Roberta, Pauli, Francesco, & Torelli, Nicola. 2018. Relabelling in Bayesian mixture models by pivotal units. *Statistics and Computing*, **28**(4), 957–969.

Egidi, Leonardo, Pappadà, Roberta, Pauli, Francesco, & Torelli, Nicola. 2021. pivmet: Pivotal methods for Bayesian relabelling and k-means clustering. *arXiv preprint arXiv:2103.16948*.

Ferguson, Thomas S. 1973. A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

Frühwirth-Schnatter, Sylvia. 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**(453), 194–209.

Neal, Radford M. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, **9**(2), 249–265.

Papastamoulis, Panagiotis. 2016. label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software, Code Snippets*, **69**(1), 1–24.

Plummer, Martyn. 2018. *rjags: Bayesian graphical models using MCMC*. R package version 4-8.

Redner, Richard A., & Walker, Homer F. 1984. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, **26**(2), 195–239.

CLUSTER VALIDITY BY RANDOM FORESTS

Tahir Ekin[1] and Claudio Conversano[2]

[1] McCoy College of Business, Texas State University,
(e-mail: tahirekin@txstate.edu)

[2] Department of Economics and Business Sciences, University of Cagliari,
(e-mail: conversa@unica.it)

ABSTRACT: Clustering is a widely used unsupervised method, which is characterized by the lack of an outcome variable that supervises the analysis. In literature, several indices have been developed to assess the goodness of cluster partition. Nonetheless, they usually suffer from computational limitations and therefore may not be appropriate in big data circumstances. We propose a method that validates the outputs of multiple clustering algorithms and is scalable for large number of observations. It utilizes machine learning classifiers to automatically rank the clustering outputs accounting for the coherence of the partitions with the data patterns. We illustrate the performance of the proposed method by applying it to simulated clustering datasets, as well as to big data situations in health care fraud detection.

KEYWORDS: cluster validity, classifiers, big data

Robust estimation of parsimonious finite mixture of Gaussian models

Luis Angel García-Escudero[1], Agustín Mayo-Iscar[1] and Marco Riani[2]

[1] Universidad de Valladolid, (e-mail: `lagarcia@uva.es`, `agustin.mayo.iscar@uva.es`)

[2] Università degli Studi di Parma, (e-mail: `marco.riani@unipr.it`)

Abstract: Maximum likelihood estimators are typically robustified by using impartial trimming. For robustifying mixture models' estimators is necessary to apply additionally constraints, for avoiding spurious solutions. We propose robust estimators, based on the joint application of trimming and constraints, for the classical collection of 14 parsimonious models of Celeux and Govaert. They include different versions of constraints, for being jointly applied, in order to get a more flexible methodology. Feasible algorithms for these estimators, EM and ECM type, have been developed. Empirical evidences about the performance of these estimators, when applied, both, to artificial and to real data will be provided.

Keywords: trimming, constrained estimation, mixture models

A RISK INDICATOR FOR CATEGORICAL DATA

Silvia Facchinetti[1] and Silvia Angela Osmetti[1]

[1] Department of Statistical sciences, Università Cattolica del Sacro Cuore – Milano, (e-mail: silvia.facchinetti@unicatt.it, silvia.osmetti@unicatt.it)

ABSTRACT: In this paper we present a suitable measure of risk for data expressed on an ordinal scale. The proposed indicator is based on the cumulative probabilities of the ordinal variable that represents the level of severity for different risk events. The method relies on the construction of a Criticality index which may be used as an initial view of the level of risk, for comparisons among environments, to indicate how risk changes over time, and to identify appropriate interventions. Along with the description of the methodology, we present two examples of application in statistical quality control field and in cyber risk evaluation.

KEYWORDS: categorical variables, risk measure, Criticality index, ordinal data.

1 Methodological proposal

The most common approach of risk modeling is a quantitative approach. When data are available only on an ordinal scale, companies often use approaches based on categorical data improperly treating the data as quantitative.

In this paper we propose a risk indicator which can exploit ordinal data to rank risks by their "criticality", so to prioritise preventive actions aimed at mitigating and reducing their impact ex-ante rather than ex-post.

Let $X \sim \{x_k, p_k; k = 1, 2, \ldots, K\}$ be a categorical random variable (r.v.) with ordered categories x_k and probabilities p_k that represents a severity variable. In the loss data framework, the severity is a continuous r.v., while in the context of ordinal risk data, the severity is generally expressed on an ordinal scale, characterised by K distinct levels, ordered according to the corresponding magnitude. We define the *Criticality Index* as follows:

$$I = \frac{1}{K-1} \sum_{k=1}^{K-1} (K-k) p_k \tag{1}$$

It is a normalized index with values in $[0, 1]$, that provides a risk measure easy to interpret, with extreme values univocally defined, and intermediate values expressed as a percentage.

This index can be estimated by its empirical counterpart, replacing the probabilities p_k with their estimators $\hat{p}_k = r_k/n$, where r_k is the number of observations in the sample equal to the category x_k ($r_k \in \mathbf{N}$ and $\sum_{k=1}^{K} r_k = n$). Henceforth, we use I to indicate the index defined in Equation (1), and $\hat{I} = \frac{1}{K-1}\sum_{k=1}^{K-1}(K-k)\hat{p}_k$ to denote its estimator.

The *Criticality Index* estimator is an unbiased and consistent estimator for I (Facchinetti & Osmetti, 2018). In fact, since every r_k $(k = 1,2,\ldots,K)$ follows a binomial distribution with parameters (n, p_k), the expression $\sum_{k=1}^{K-1}(K-k)\frac{r_k}{n}$ is a mixture of binomial r.v.s $(k = 1,2,\ldots,K-1)$. Therefore, since $E(r_k) = np_k$ and $Var(r_k) = np_k(1-p_k)$, the mean and the variance of $\hat{I}$ are respectively:

$$E(\hat{I}) = \frac{1}{K-1}\sum_{k=1}^{K-1}(K-k)\frac{E(r_k)}{n} = \frac{1}{K-1}\sum_{k=1}^{K-1}(K-k)p_k = I \tag{2}$$

$$Var(\hat{I}) = \frac{1}{n(K-1)^2}\left[\sum_{k=1}^{K-1}(K-k)^2 p_k(1-p_k) - 2\sum_{k=1}^{K-1}(K-k)p_k\sum_{l=1}^{k-1}(K-l)p_l\right] \tag{3}$$

Finally, from (3), $\lim_{n\to\infty} Var(\hat{I}) = 0$.

Moreover, a Kolmogorov–Smirnov test for discrete r.v.s (Facchinetti & Osmetti, 2013) ensure that the *Criticality Index* estimator is asymptotically normally distributed, with the mean and variance given in (2) and (3).

2 Applications

In this section we present two examples of application of the *Criticality Index*. First, in statistical quality control field, the proposed index appears naturally suitable for measuring the risk of failure of a product in the testing and recall phases of products, or in similar situations where quality is expressed on an ordinal scale. Second, in cyber risk evaluation, since the data are very sensitive and it is unlikely that a private institution is willing to disclose them, we consider a classification of cyber risk loss data into severity levels and we apply the proposed methodology to measure cyber risks, using ordinal data.

2.1 Statistical quality control

We apply the *Criticality Index* on real data concerning severity, detection, and the occurrence of defects in the component of hose assemblies (stripes, guard, fitting, and hose) produced by a sales company of multinational manufacturer.

Severity is a measure of the gravity of a particular type of defect on a 3-point scale (serious, medium, minor defect); detection is a measure of the ease of identifying a failure mode on a 3-point scale (low, medium, high detection); occurrence is the frequency of a particular type of defect in a product.

These information are typically available in companies that apply Failure Mode and Effects Analysis (FMEA) to identify potential failures that could affect the customer's expectations of product quality or process performance (Sellappan & Palanikumar, 2013).

To obtain a global measure of risk, we summarize the *Criticality Indices* related to the severity and detection in a Criticality Impact Chart (Figure 1).

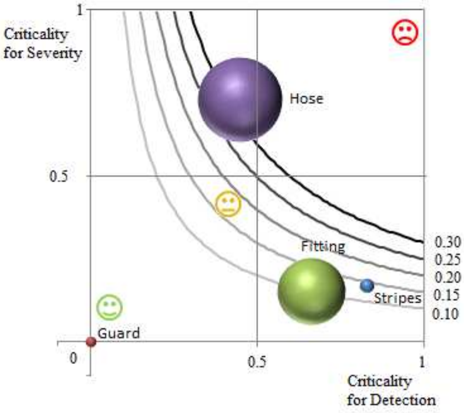


Figure 1. *Criticality Impact Chart for severity and detection.*

For each component of the product, we plot a ball with coordinates given by the levels of risk ($\hat{I}$) for severity and for detection. The dimension of the balls is related to the occurrence of each component. The lines represent the locus of points with equal joint level of risk. We observe that hose is the component with the highest joint level of risk. For guard, a situation of minimum heterogeneity occurs and, thus, the indices assume their minimum value.

This graph may be very useful for companies wanting to prioritize interventions on the production line of a finite product, as well as those wanting to improve related process controls.

2.2 Cyber risk

We apply our proposal to real data on serious cyber attacks occurs worldwide in 2017, described by Clusit (Italian Association for Information Security) in its Report on ICT Security in Italy (Antonielli *et al.*, 2018).

We consider for each type of attack the ordinal variable severity on a 3-point scale (critical, high, medium severity). In Table 1, we report the *Criticality Index* estimates, the standard errors and the associated 90% asymptotic confidence intervals (Facchinetti & Osmetti, 2018).

Table 1. *Criticality index estimates, standard errors and* 90% *CIs for type of attack.*

Type of attack	$\hat{I}$	SE	CI (90%)
Cybercrime	0.239	0.015	0.214-0.264
Hacktivism	0.342	0.045	0.268-0.416
Espionage/Sabotage	0.973	0.014	0.950-0.996
Information Warfare	0.952	0.027	0.908-0.995

From Table 1 we obtain that Espionage and Information Warfare dominate Hacktivism in terms of severity, followed by Cybercrime. Our proposed measure can thus be employed as a simple and effective measurement to prioritise cyber risk.

References

ANTONIELLI, A., BECHELLI, L., BOSCO, F., & BUTTI, G., ET AL. 2018. *Rapporto Clusit 2019 sulla Sicurezza ICT in Italia*. Clusit.

FACCHINETTI, S., & OSMETTI, S.A. 2013. A goodness-of-fit test for maximum order statistics from discrete distributions. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation*, **4**, 9–21.

FACCHINETTI, S., & OSMETTI, S.A. 2018. A risk index for ordinal variables and its statistical properties: A priority of intervention indicator in quality control framework. *Quality and Reliability Engingeering International*, **34**, 265–275.

FACCHINETTI, S., GIUDICI, P., & OSMETTI, S.A. 2020. Cyber risk measurement with ordinal data. *Statistical Methods and Applications*, **29**, 173–185.

SELLAPPAN, N., & PALANIKUMAR, K. 2013. Modified Prioritization Methodology for Risk Priority Number in Failure Mode and Effects Analysis. *International Journal of Applied Science and Technology*, **3**, 27–36.

ADDITIVE QUANTILE REGRESSION VIA THE QGAM R PACKAGE

Matteo Fasiolo[1]

[1] University of Bristol, (e-mail: matteo.fasiolo@bristol.ac.uk)

ABSTRACT: Generalized additive models (GAMs) are flexible non-linear regression models, which can be fitted efficiently using the approximate Bayesian methods provided by the mgcv R package. While the GAM methods provided by mgcv are based on the assumption that the response distribution is modelled parametrically, in this talk I will discuss more flexible methods that do not entail any parametric assumption. In particular, I will introduce the qgam package, which is an extension of mgcv providing fast calibrated Bayesian methods for fitting quantile GAMs (QGAMs) in R. QGAMs are based on a smooth version of the pinball loss of Koenker (2005), rather than on a likelihood function, hence jointly achieving satisfactory accuracy of the quantile point estimates and coverage of the corresponding credible intervals requires adopting the specialized Bayesian fitting framework of Fasiolo et al. (2020), which is implemented by the qgam package.

KEYWORDS: Bayesian quantile regression, generalized additive models, regression splines, calibrated Bayes, fast Bayesian inference.

GAUSSIAN MIXTURE MODELS FOR HIGH DIMENSIONAL DATA USING COMPOSITE LIKELIHOOD

Michael Fop [1], Dimitris Karlis[2], Ioannis Kosmidis[3], Adrian O'Hagan[1],
Caitriona Ryan[4] and Isobel Claire Gormley[1]

[1] School of Mathematics and Statistics, University College Dublin, Ireland. (e-mail: michael.fop@ucd.ie, claire.gormley@ucd.ie)

[2] Department of Statistics, Athens University of Economics and Business, Greece.

[3] Department of Statistics, University of Warwick, UK.

[4] Hamilton Institute, Maynooth University, Ireland.

ABSTRACT: The use of finite Gaussian mixture models (GMMs) is a well established approach to performing model-based clustering. Despite its popularity, its widespread use is hindered by its inability to transfer to high-dimensional data applications. This is often due to the difficulties related to dealing with high-dimensional covariance matrices and joint densities. Here we propose a composite likelihood framework to enable the use of GMMs for clustering high-dimensional data. The framework is specified by approximating the likelihood of a GMM by means of a block-pairwise composite likelihood, which allows the decomposition of the potentially high-dimensional density into terms of smaller dimensions. A computationally efficient expectation maximization algorithm is developed to facilitate estimation. Performance of the approach is demonstrated through simulated and real data examples.

KEYWORDS: composite likelihood, high-dimensional data, model-based clustering.

1 Introduction

Model-based clustering of continuous data is routinely implemented by means of Gaussian mixture models (GMMs). Despite the popularity of GMMs, their widespread use is curtailed by their inability to transfer to settings where the number of variables p is large compared to the sample size N. Difficulties with storage and manipulation of the multivariate Gaussian distribution's covariance matrix in such settings lead to increased computational cost, and it often makes the approach impractical. Further, in settings where the number of variables $p > N$, fitting a GMM with an unconstrained covariance matrix is infeasible. To overcome these issues, parsimonious models based on covariance matrix factorization and/or strict independence restrictions have been proposed (Bouveyron & Brunet-Saumard, 2014). Despite their advantages, these methods still struggle in high-dimensional data settings ($p \gg N$) and have several limitations in the case of highly correlated variables.

Recently, Ranalli & Rocci, 2016; Ranalli & Rocci, 2017 proposed the use of composite likelihood (CL) to estimate the parameters of a finite mixture model for ordinal and mixed mode data. The CL approach (see Varin *et al.*, 2011) uses smaller dimensional marginal and/or conditional pseudo-likelihoods to estimate the parameters of the model. The use of CL avoids the need to fully specify the underlying joint distribution and estimates parameters from a product of lower dimensional likelihoods. Such an approximation is very helpful when the full model is difficult to specify or manipulate. The CL framework assists in avoiding the computational problems often arising from the need to deal with a multi-dimensional joint distribution. In addition, the specification of appropriate conditional likelihoods allows the modelling of the dependence structure by means of lower dimensional terms.

Here the CL approach is exploited to enable clustering of high-dimensional data using GMMs. Lower dimensional terms corresponding to Gaussian multivariate marginal distributions are involved in the construction of the pseudo-likelihood, thus avoiding the use of high dimensional covariance matrices (and their inversion), which is advantageous in $p \gg N$ scenarios. We embed GMMs in the CL framework to serve a dual purpose: to facilitate the use of GMMs in high-dimensional scenarios, while capturing at the same time the complex dependence structures which are often present in such settings.

2 Block-pairwise composite likelihood for GMMs

To deal with the complexities arising in high-dimensional settings $(p \gg n)$, we decompose the likelihood of a GMM into terms of tractable size, corresponding to lower dimensional Gaussian distributions in which n is larger than the number of variables involved in each term. To do so, we define a general composite likelihood based on pairs of blocks of variables.

Suppose the vector $\mathbf{X}$ of p variables is partitioned into a set of K non-overlapping blocks $\mathcal{B} = \{B_1, \dots B_j, \dots B_K\}$. For ease of exposition, we take blocks having the same size b. Let $\mathcal{S}$ denote the set of all possible $\binom{K}{2}$ pairs constructed using the blocks in $\mathcal{B}$. The generic element $S_l \in \mathcal{S}$ is given by a pair of blocks, such that $S_l = B_j \cup B_k$, with $j \neq k$, $\forall\ j,k = 1, \dots, K$. We then define the following *block-pairwise composite likelihood* (BCL)

$$\begin{aligned}\mathrm{BCL}(\boldsymbol{\Theta}) &= \prod_{S_l \in \mathcal{S}} \left\{ \prod_{i=1}^{N} \sum_{g=1}^{G} \tau_g \phi(\mathbf{x}_i^l; \boldsymbol{\mu}_g^l, \boldsymbol{\Sigma}_g^l) \right\} \\ &= \prod_{j=1}^{K-1} \prod_{k>j} \left\{ \prod_{i=1}^{N} \sum_{g=1}^{G} \tau_g \phi\left(\{\mathbf{x}_i^j, \mathbf{x}_i^k\}; \boldsymbol{\mu}_g^{(j,k)}, \boldsymbol{\Sigma}_g^{(j,k)} \right) \right\},\end{aligned} \tag{1}$$

where $\mathbf{x}_i^l = \{\mathbf{x}_i^j, \mathbf{x}_i^k\}$ is the observation i measured on the variables included in the pair of blocks $S_l = B_j \cup B_k$, $\phi(\cdot)$ is the density of a multivariate Gaussian of dimension $2b$ and $\boldsymbol{\mu}_g^l$ and $\boldsymbol{\Sigma}_g^l$ are the component parameters that relate to the variables in the pair S_l. The second expression in (1) makes explicit that $\boldsymbol{\mu}_g^{(j,k)}$ and $\boldsymbol{\Sigma}_g^{(j,k)}$ are the parameters of the joint Gaussian distribution of $\{\mathbf{x}_i^j, \mathbf{x}_i^k\}$.

In (1), each product term in the curly brackets is the likelihood of a GMM over the variables in the pair S_l. Hence, by setting $b \ll n/2$, the potentially high-dimensional GMM likelihood is decomposed into a number of terms involving lower dimensional Gaussian distributions, enabling computationally efficient inference. As the BCL approach works with low dimensional Gaussian distributions, estimation and inversion of large covariance matrices are avoided, facilitating the use of GMMs in high-dimensional scenarios.

As long as $1 < b \ll n/2$, there is computational advantage in the BCL approach. However, for certain situations, $\binom{K}{2}$ can mean an intractable number of terms in the log-likelihood. To further reduce the complexity, instead of looking at all possible $2b$-dimensional marginal distributions, we define a restricted subset $\mathcal{S}^* \subset \mathcal{S}$ of pairs of blocks. In particular, we take this set as a sequential enumeration of pairs of blocks: $\mathcal{S}^* = \bigcup_{j=1}^{K-1} B_j \cup B_{j+1}$. We then define the following *restricted block-pairwise composite likelihood* (RBCL):

$$\begin{aligned}\mathrm{RBCL}(\boldsymbol{\Theta}) &= \prod_{i=1}^{N} \left\{ \prod_{S_l \in \mathcal{S}^*} \sum_{g=1}^{G} \tau_g \phi(\mathbf{x}_i^l; \boldsymbol{\mu}_g^l, \boldsymbol{\Sigma}_g^l) \right\} \\ &= \prod_{i=1}^{N} \left\{ \prod_{\substack{j=1 \\ k=j+1}}^{K-1} \sum_{g=1}^{G} \tau_g \phi(\{\mathbf{x}_i^j, \mathbf{x}_i^k\}; \boldsymbol{\mu}_g^{(j,k)}, \boldsymbol{\Sigma}_g^{(j,k)}) \right\}.\end{aligned} \tag{2}$$

Importantly, compared to the $\binom{K}{2}$ pairs of BCL, the number of pairs for RBCL is linear in K. The formulation of the GMM in terms of BCL and RBCL significantly reduces the complexity in computing and dealing with high-dimensional likelihood terms and covariance matrices.

3 Estimation

The complete-data composite log-likelihood decomposes into the sum of a number of standard GMM complete data log-likelihood terms, each related to the marginal joint Gaussian distribution of the block pair $\mathbf{x}_i^l = \{\mathbf{x}_i^j, \mathbf{x}_i^k\}$. However, these terms are not independent of each other due to the coupling of the parameters in the factorization. Therefore, maximization of the BCL and RBCL is carried out by means of an Expectation-Conditional-Maximization algorithm (Meng & Rubin, 1993). In particular, the maximization step involves a series of conditional maximization passes, where the pairs are scanned in a sequential manner, such that the joint distribution of each pair is rewritten as the product of a marginal distribution estimated at the previous step and the conditional distribution of a block given the block of the previous step. The optimization is based on the conditional estimation procedure outlined in Fop *et al.*, 2021. While the main purpose of employing the CL framework when clustering high-dimensional data is computational efficiency, further work will explore the statistical properties (e.g. consistency) of the resulting parameter estimates. Performance of the CL approach and estimation procedure are demonstrated through simulated and real data examples.

References

Bouveyron, Charles, & Brunet-Saumard, Camille. 2014. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, **71**, 52–78.

Fop, M., Mattei, P.A., Bouveyron, C., & Murphy, T.B. 2021. Unobserved classes and extra variables in high-dimensional discriminant analysis. *arXiv:2102.01982*.

Meng, Xiao-Li, & Rubin, Donald B. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**(2), 267–278.

Ranalli, Monia, & Rocci, Roberto. 2016. Mixture models for ordinal data: a pairwise likelihood approach. *Statistics and Computing*, **26**(1-2), 529–547.

Ranalli, Monia, & Rocci, Roberto. 2017. Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis*, **110**, 87–102.

Varin, Cristiano, Reid, Nancy, & Firth, David. 2011. An overview of composite likelihood methods. *Statistica Sinica*, 5–42.

ON MODEL-BASED CLUSTERING USING QUANTILE REGRESSION

Carlo Gaetan [1], Paolo Girardi[2] and Victor Muthama Musau [3]

[1] DAIS, Ca' Foscari of Venice (Italy) , (e-mail: gaetan@unive.it)

[2] Department of Developmental and Social Psychology, University of Padova (Italy), (e-mail: paolo.girardi@unipd.it)

[3] Department of Pure and Applied Sciences, Kirinyaga University (Kenya), (e-mail: vmusau@kyu.ac.ke)

ABSTRACT: Clustering general regression functions or curves can suffer of lack of robustness when we consider the usual Gaussian assumption. In this note we introduce a new model-based clustering method that tries to overcome this limitation.

KEYWORDS: Functional data, hierarchical Bayesian model, MCMC algorithm

1 Introduction

Unlike the classical clustering approaches such as agglomerative hierarchical clustering and K-means clustering, which are largely heuristic and not based on formal statistical models, model-based clustering takes a likelihood based approach thus permitting inference to be drawn on the clusters. These techniques are based on the finite mixture model theory (Fraley & Raftery, 2002), where each mixture component corresponds to a cluster. However, fundamental concerns remain about robustness and in particular the choice of distribution representing the within cluster density. The Gaussian mixture models are historically the most popular tool for model-based clustering. However, if the distribution of the observed variable is characterized by asymmetry and presence of outliers, a Gaussian distribution may not be an appropriate within cluster density. The direct link that exists between univariate quantile regression approach and the Asymmetric Laplace Distribution (ALD) forms our basis of introducing a clustering model based on finite mixture of ALDs to group individuals subject to heterogeneity due to regressor variables.

2 Methodology

We start by considering a vector, $\mathbf{y} = (y_1, \ldots, y_T)'$ of responses y_t and the associated design matrix $\mathbf{X} = (x_1, \ldots, x_T)'$ that collects the vectors x_t of L covariates. Further, let $Q_p(y_t|x_t)$, for $0 < p < 1$, be the pth quantile regression function of y_t given x_t which can be modelled as $Q_p(y_t|x_t) = x_t'\beta$, where β is a vector of unknown parameters to be estimated. The regression coefficient estimate is obtained by minimizing (Koenker & Bassett, 1978)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{t=1}^{T} \rho_p(y_t - x_t'\beta) \tag{1}$$

where $\rho_p(\cdot)$ is the check loss function defined by $\rho_p(x) = x(p - I(x < 0))$ and $I(\cdot)$ denotes the usual indicator function. Koenker and Machado (1999) showed that there is a direct relationship between minimizing (1) and the maximum likelihood theory using independently distributed asymmetric Laplace variable with density

$$\text{ald}(y_t|\beta, \sigma, p) = \frac{p(1-p)}{\sigma} \exp\left\{ -\rho_p \left(\frac{y_t - x_t'\beta}{\sigma} \right) \right\} \tag{2}$$

where $\sigma > 0$ and $0 < p < 1$ represents the skewness parameter that can be used directly to model any quantile of interest.

According to the finite mixture framework theory we define the likelihood of our mixture model for a single vector $\mathbf{y}$ as

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}, p|\mathbf{y}) = \sum_{k=1}^{K} \alpha_k \prod_{t=1}^{T} \text{ald}_1(y_t|\beta_k, \sigma_k, p) = \sum_{k=1}^{K} \alpha_k \text{ALD}(\mathbf{y}|\beta_k, \sigma_k, p) \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1', ..., \beta_K')'$, $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_K)'$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)'$ is the vector of the mixing proportions for the K clusters which satisfy the conditions $0 < \alpha_k < 1$ and $\sum_{k=1}^{K} \alpha_k = 1$.

We now consider a set $\mathcal{Y} = \{\mathbf{y}_i, i = 1, ..., n\}$ of n vectors $\mathbf{y}_i = (y_{i1}, ..., y_{iT})'$ of independent observations. and we want to split the data set $\mathcal{Y}$ into K clusters. According the mixture model (3) the cluster membership $c_i \in \{1, ..., K\}$, where $c_i = k$ indicates that the ith vector $\mathbf{y}_i$ belongs to cluster k is a multinomial random variable with parameter $\boldsymbol{\alpha}$.

We adopt a Bayesian approach to make inference on the model parameters $\boldsymbol{\psi} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\sigma}')'$. Moreover it is possible to get the posterior probability of membership of a single vector, $\Pr(c_i = \cdot|\mathcal{Y})$. In doing this we first note that

Kozumi and Kobayashi (2011) represent the density (2) as a location scale mixture of Gaussian distributions i.e.

$$y_t = x_t'\beta + \theta w_t + \omega\sqrt{\sigma w_t}\nu_t \tag{4}$$

where $\nu_t \sim N(0,1)$, and w is an exponential random variable with $E(w) = \sigma$. Here ν and w are mutually independent and $\theta = (1-2p)/\{p(1-p)\}$ and $\omega^2 = 2/\{p(1-p)\}$.

Equation (4) constitutes the first stage of a hierarchical Bayesian model where the prior distribution on the cluster specific parameters and as well as the mixing proportions are specified as conjugate priors to having closed form conditional posterior densities which are easy to sample from in a MCMC algorithm.

A conjugate prior for the mixing proportions $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_K)'$ is the Dirichlet distribution, $\boldsymbol{\alpha} \sim D(\zeta_1,...,\zeta_K)$. A straightforward prior for β_k is the multivariate Gaussian distribution, $\mathcal{N}(b_0,\Sigma_0)$ where by setting $b_0 = 0$ and $\Sigma_0 = aI$, for $a \gg 0$, leads to an improper prior. Finally we propose the inverse gamma distribution, $\text{IG}(s_0,d_0)$, as the prior for σ_k where the shape and scale parameters, s_0 and d_0 respectively, are known.

Musau (2021) gives a complete account on how we can devise an MCMC algortihm for sampling from the posterior distribution of $\boldsymbol{\psi}$.

3 Numerical results

We exemplify our proposal with a clustering problem for functional data. We consider the well-known Canadian temperature dataset available in the R package `fda`. The dataset consists of the daily measured temperatures from 35 Canadian weather stations across the country.

Under functional data framework (Ramsay & Silverman, 2005), daily temperature data, y_t, can be described by a linear combination of $L = 65$ cubic spline basis functions, $y_t \simeq \sum_{j=1}^{L} \beta_j B_j(t) = x_t'\beta$, with knots which are equally distributed over the range of time.

The funHDDC clustering algorithm (Bouveyron & Jacques, 2011) on this data selects $K = 4$ as the optimal number of clusters. Figure 1 (left panel) summarize the resulted clusters.

For each of the 35 stations we randomly introduce outliers (y_t=0) at 10% of the total observation points. This distorts the general trend of the data, as shown in right panel of Figure 1, making reconstruction of the clusters difficult.

We apply our mixture model setting $p = 0.5$, i.e. we consider a robust median regression and we compare its performance in reconstructing the 4

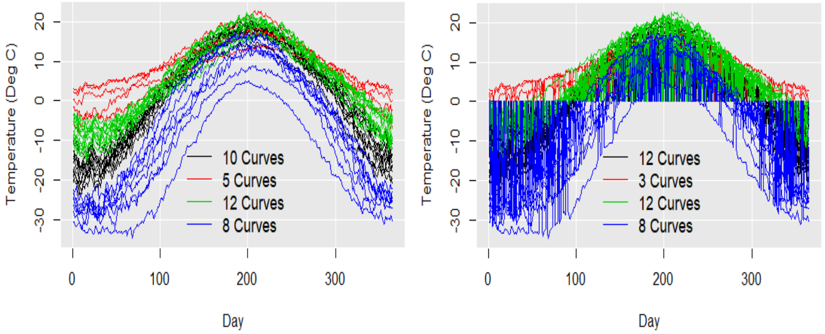


Figure 1. *Clustering of the 35 temperature curves as obtained by funHDDC algorithm (left panel) and results with curves contaminated by outliers (right panel).*

clusters with the previous algorithm, leading to a perfect agreement. These results generally indicate a good performance of our proposed algorithm when clustering data characterized by outlying observations.

References

BOUVEYRON, C., & JACQUES, J. 2011. Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, **5**, 281–300.

FRALEY, C., & RAFTERY, A.E. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

KOENKER, R., & BASSETT, G. 1978. Regression quantiles. *Econometrica*, **46**, 33–50.

KOENKER, R., & MACHADO, J. A.F. 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**, 1296–1310.

KOZUMI, H., & KOBAYASHI, G. 2011. Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, **81**, 1565–1578.

MUSAU, V.M. 2021. *Model-based Clustering Using Quantile Regression*. Ph.D. thesis, University of Padua, Italy.

RAMSAY, J.O., & SILVERMAN, B.W. 2005. *Functional Data Analysis*. Springer, New York.

SOCIOECONOMIC INEQUALITIES AND CANCER RISK: MYTH OR REALITY?

Carlotta Galeone [1]

[1] University of Milan, (e-mail: carlotta.galeone@statinfo.org)

ABSTRACT: Accurate quantification of the impact of low socioeconomic position (SEP) on selected no communicable diseases, including diabetes several cancers, is needed. There is increasing evidence that low SEP is a strong determinant of morbidity and premature mortality concerning cancer risk. This is mainly caused by a delay in screening uptake and consequent timeliness of symptomatic presentation and lifestyle (including diet, smoking habit and physical activity). The accurate quantification of the relation between SEP and cancer risk is crucial to plan public health interventions for cancer incidence and socioeconomic disparities reduction. The recent advent of collaborative and interdisciplinary research by pooling a large amount of worldwide epidemiological data in multi-institutional data consortia is the answer to this gap in knowledge. In fact, data analyses of epidemiological consortia will allow to define and quantify the associations of interest with a higher degree of accuracy, explore subgroups of the population, and investigate the interactions between environmental, genetic, and socioeconomic factors. The Stomach Cancer Pooling (StoP) Project and the International Head and Neck Cancer Epidemiology (INHANCE) are two example of large data consortia, in which the University of Milan is proactively involved. Their large sample size allowed investigators to address the effects of education and household income on the onset and evolution of the disease. INHANCE findings suggested that low education and low income are risk factors for head and neck cancer, independent of tobacco smoking and alcohol consumption. The collaborative pooled analysis within the StoP consortium showed a strong inverse relation between SEP indicators and gastric cancer risk, with a 40% decreased risk among individuals with intermediate/high education status than less educated study subjects. In conclusion, social epidemiology is crucial to understand the sociostructural factors related to health and disease. In an era of fast inter-diffuse communication and data-sharing, large data consortia are among the most effective strategies to create new social epidemiological useful evidences. In these example of data consortia, SEP is strongly related to a number of cancers. Health education campaigns targeting socioeconomically disadvantaged in vulnearal populations are probabily the most efficacious stategy to redurce the cancer burden in the world.

KEYWORDS: socioeconomic inequalities, cancer risk, INHANCE, StoP.

PARAMETER-WISE CO-CLUSTERING FOR HIGH DIMENSIONAL DATA

Michael Gallaugher[1], Christophe Biernacki[2] and Paul McNicholas[3]

[1] Baylor University, (e-mail: Michael_Gallaugher@baylor.edu)

[2] University Lille 1, (e-mail: christophe.biernacki@inria.fr)

[3] McMaster University, (e-mail: paulmc@mcmaster.ca)

ABSTRACT: Due to the "big data" phenomenon, data is becoming increasingly higher dimensional. As such, new techniques need to be developed to handle higher dimensional data, and this is especially true in clustering. One such clustering method for high dimensional data is co-clustering where the aim is to cluster both rows and columns resulting in data blocks, or co-clusters, where observations within each block are independent and identically distributed. Although highly parsimonious, co-clustering can be quite inflexible. In this talk, a method that clusters columns according to both means and variances, while assuming normality, will be presented. The proposed model increases flexibility while maintaining a high degree of parsimony. Both simulated and real data will be used for illustration.

KEYWORDS: model-based clustering, mixture model, co-clustering, high-dimensional data

QUANTIFYING THE IMPACT OF COVARIATES ON THE GENDER GAP MEASUREMENT: AN ANALYSIS BASED ON EU-SILC DATA FROM POLAND AND ITALY

Francesca Greselin[1] and Alina Jędrzejczak[2]

[1] Department of Statistics and Quantitative Methods, University of Milano Bicocca, Italy (francesca.greselin@unimib.it)

[2] Department of Statistical Methods, University of Łódź, Poland (alina.jedrzejczak@uni.lodz.pl)

ABSTRACT: High income inequality, accompanied by substantial regional differentiation, is still a great challenge for social policy makers in many European countries. One of the important elements of this phenomenon is inequality between income distributions of men and women. Using data coming from EU-SILC 2018, we compare the distribution of income for Italy and Poland, and analyze gender gap in these countries. We are interested here to uncover the socioeconomic factors that could contribute to explain the differences observed in the income distribution for men and women.

KEYWORDS: Income inequality, Gender gap, Gini index, new Zenga index, relative distribution method, Dagum, Italy, Poland.

1 Introduction

Substantial regional disparities and income inequality is a great challenge for policymakers in many European countries, nowadays. One of the critical elements of this phenomenon is the inequality between income distributions of men and women. The gender pay gap can be a problem from a public policy perspective because it reduces economic output and means that women are more likely to be dependent upon welfare payments, especially in old age.

Many studies analyze income inequality across the European Union (EU) countries and regions for social and economic policies. The focus of the present paper is on income distributions across Poland and Italy, to compare countries with different economic backgrounds. Poland still suffers the transition from a centrally-planned economy to a market-based economy, and Italy is a former well-established market economy. Moreover, according to the Tárki

European Social Report (TÁRKI, 2008), a study on intolerance to income inequality across countries confirmed a markedly lower level of acceptance of inequality in the post-socialist bloc than in the other European countries. The calculations were based on microdata from the European Union Statistics on Income and Living Condition (EU-SILC) (Eurostat, 2018).

Several methods can be applied to the measurement of the income discrepancy between men and women. Among them, summary measures remain an important tool for the comparison of distributional changes. However, to uncover the factors contributing to the gender discrepancy, it is useful to move beyond the typical focus on average or median earnings differences, towards a view on how the entire distribution of women's earnings relative to men's compares. Indeed, inequality is a property of a distribution. A prominent feature of these methods is the use of the "relative distribution", a transformation of the data from two distributions into a single distribution that contains all the information needed for scale-invariant comparison (Handcock & Morris, 2006). In a previous paper (Greselin & Jędrzejczak, 2020) the authors highlighted remarkable differences between Poland and Italy, especially related to the discrepancy across regions between men and women. The next natural step is hence to search for the socioeconomic factors that could explain the differences observed in the income distribution for men and women.

2 Quantifying the covariates effects

Often there are covariates which vary systematically by the compared populations, and the impact of these covariates is of interest. We will follow the approach introduced by Handcock & Morris (2006). The overall relative distribution is decomposed into a first component representing the effect of changes in the marginal distribution of a covariate, and a second component defining the residual changes. The first term is the composition effect, which measures the shift in the covariates from one population to the other. The second term is obtained by adjusting the reference (men) population to have the same marginal covariate composition as the comparison (women) population. By holding the population composition constant across the gender groups, differences in the covariate-response relationships can be correctly identified.

Let (Y_0, Z_0) and (Y, Z) denote random vectors describing the reference and comparison populations, where Y_o and Y are the response variable, while Z_0 and Z are the categorical covariates, with support $1, 2, ...K$. Let $\pi_{k}{}_{k=1}^{K}$ and $\pi^{0}_{k}{}^{K}_{k=1}$ be the probability mass function of Z and Z_0, respectively. These probability mass functions represent the population composition with respect to the covariate. The marginal density of Y can be written as $f(y) := \sum_{k=1}^{K} \pi_k f_{Y|Z}(y|k)$, where $f_{Y|Z}(y|k)$ denotes the conditional densities of Y given that $Z = k$, for $k = 1,\ldots,K$. An analogous definition holds for $f_0(y) := \sum_{k=1}^{K} \pi_k^0 f_{Y_0|Z_0}(y|k)$. Now any differences between $f(y)$ and $f_0(y)$ are a result of the differences in the conditional densities $f_{Y_0|Z_0}(y|k)$ and $f_{Y|Z}(y|k)$, for $k = 1,\ldots,K$. These represent differences in the covariate-response relationship between the two populations.

We can construct a counter-factual distribution for the compositional difference using these ideas. We define the distribution of Y_0 *composition-adjusted* to Y to be Y_{0C} with density: $f_{0C}(y) := \sum_{k=1}^{K} \pi_k f_{Y_0|Z_0}(y|k)$. It corresponds to a counter-factual population with the covariate composition of the comparison population and the covariate-response relationship of the reference population. Comparisons of $f_{0C}(y)$ to $f(y)$ hold the population composition constant, and therefore isolate differences in the covariate-response relationship. By contrast, $f_0(y)$ and $f_{0C}(y)$ have the same covariate-response relationship and comparisons between them isolate the impact of the compositional shifts. Using the composition-adjusted response distribution, we can decompose the overall relative distribution into a component that represents the effect of changes in the marginal distribution of the covariate (the composition effect), and a component that represents the residual changes. In terms of density ratios, we have:

$$\frac{f(y_r)}{f_0(y_r)} = \frac{f_{0C}(y_r)}{f_0(y_r)} \times \frac{f(y_r)}{f_{0C}(y_r)} \tag{1}$$

Figure 1 graphically shows the decomposition of the relative income distribution of women in relation to men, assuming the position (managerial or not, variable PL150: Managerial position) as the explanatory variable. The first panel from the left shows the (uncorrected) relative density of income differences between men and women, the middle panel represents the effects of differences in the distributions of the explanatory variable, and the right panel represents the counterfactual distribution - i.e. the expected relative density for men's and women's income distributions with assuming the same profiles of positions held in both groups. The comparison of the three relative densities provides a useful tool for assessing the relative magnitude and nature of the impact of individual components. It can be noticed that the distribution presented in the middle panel is U-shaped and in the central part close to the uniform distribution. This means that the difference in the structure of management positions observed between the two cohorts of the distribution in central deciles has little effect on the observed income gap. On the other hand, greater differences were observed in extreme decile groups, which suggests a

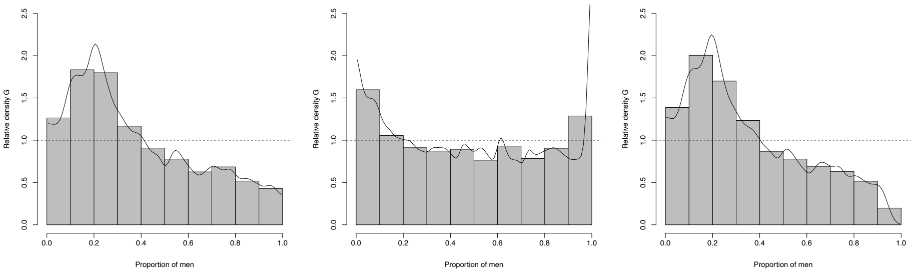


Figure 1. *The three plots for Polish data, to assess the effect of managerial position*

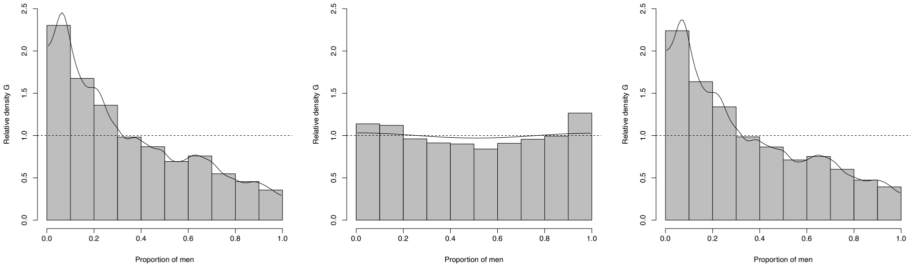


Figure 2. *The three plots for Italian data, to assess the effect of managerial position*

certain polarization of income of these groups in relation to the position held. Women from the last decile occupy higher positions, which, however, does not translate into their earnings. As a result, the income gap in these groups, adjusted by the type of position held in the counterfactual distribution, widens (right panel). On the other hand, results on Italian data are somehow different.

3 Conclusions and further research

We developed a first analysis on the covariate effects for studying gender gap. Besides the univariate case, also the adjustment for multivariate covariate is worth to be considered and is the object of ongoing work.

References

GRESELIN, F., & JĘDRZEJCZAK, A. 2020. Analyzing the Gender Gap in Poland and Italy, and by Regions. *International Advances in Economic Research*, **26**(4), 433–447.

HANDCOCK, M. S., & MORRIS, M. 2006. *Relative distribution methods in the social sciences*. Springer Science & Business Media.

TÁRKI. 2008. TÁRKI European social report. *Czech Sociological Review*.

A TRANSDIMENSIONAL MCMC SAMPLER FOR SPATIALLY DEPENDENT MIXTURE MODELS

Alessandra Guglielmi [1], Mario Beraha[1], Matteo Gianella[1], Matteo Pegoraro[2] and Riccardo Peli[2]

[1] Department of Mathematics, Politecnico di Milano, (e-mail: alessandra.guglielmi@polimi.it)

[2] MOX, Department of Mathematics, Politecnico di Milano

ABSTRACT: We consider the problem of spatially dependent areal data, where for each area independent observations are available, and propose to model the density of each area through a finite mixture of Gaussian distributions. The spatial dependence is introduced via a novel joint distribution for a collection of vectors in the simplex, that we term logisticMCAR. We also discuss a generalization of the mixture model with a random number of components, introducing a reversible jump algorithm to sample from the full posterior. Through simulated data examples we check the performance of our algorithm. Moreover, we present an application on a real dataset of Airbnb listings in the city of Amsterdam, also showing how to easily incorporate for additional covariate information in the model.

KEYWORDS: finite mixture models, spatial density estimation, logistic normal, multivariate CAR models, reversible jump.

1 Introduction

Mixture models (Frühwirth-Schnatter *et al.*, 2019) provide a natural framework for density estimation. Though mixtures are often used under the assumption of exchangeable samples from a unique unknown distribution, such models may be adopted to model data that show spatial dependence. In this work we focus on areal data, considering the problem of modelling data from I different groups, where each group corresponds to a specific areal location. More in detail, we assume that the spatial domain Ω is divided into I areas and, for each area, there is a vector of observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iN_i})$ from the same variable, each value y_{ij} corresponding to a different subject j in area i. We further assume that data, within each areal unit i, are independent and identically distributed (i.i.d.) from an area-specific density f_i; the problem we address is the joint estimation of spatially dependent densities $f_1, \ldots, f_I$. We take the

Bayesian viewpoint and we specify a prior for dependent densities $(f_1,\dots,f_I)$ encouraging distributions associated to areas that are spatially close to be more similar than those associated to areas that are far away.

In this paper, we consider first the same framework as in Beraha *et al.*, 2020, where we assume a finite mixture with a fixed number of components H in each area I and introduce spatial dependence via a suitable prior on the weights of the mixtures, i.e., the *logistic multivariate CAR prior*. We will show how specific features of the proposed model include (i) a sparse mixture specification as meant in Malsiner-Walli *et al.*, 2016 and (ii) densities corresponding to areal units which belong to two different connected components in the proximity graph (=matrix) may behave differently.

As it happens with finite mixture models, the choice of the appropriate number H of components is crucial. Under the Bayesian approach, it is straightforward to frame H random and compute the posterior distribution for all parameters, including H. In this case, Markov Chain Monte Carlo (MCMC) algorithms for posterior inference are called *transdimensional*, and they are not easy to design. Examples of such transdimensional MCMC algorithms include the reversible jump MCMC sampler in Green, 1995 and the MCMC algorithm based on birth-and-death processes in Stephens, 2000. Hence, we extend the model above (see Beraha *et al.*, 2020 for details) by assuming a prior on the number H of components and we propose a transdimensional sampler via a reversible jump MCMC algorithm. The approach we follow to design a reversible jump move is based on Norets, 2021.

2 Details on the model and the reversible jump algorithm

As an extension of the model in Beraha *et al.*, 2020 to a random number of components, we assume the following:

$$
\begin{aligned}
y_{ij} \mid \boldsymbol{w}_i, \boldsymbol{\tau}, H &\overset{\text{iid}}{\sim} \sum_{h=1}^{H} w_{ih}\, \mathcal{N}(\boldsymbol{\tau}_h) \qquad j = 1,\dots,N_i && (1)\\
\boldsymbol{\tau}_h \mid H &\overset{\text{iid}}{\sim} P_0 \qquad\qquad\qquad h = 1,\dots,H\\
(\boldsymbol{w}_1,\dots,\boldsymbol{w}_I) \mid \rho, \sigma^2, H &\sim \text{logisticMCAR}(\mathbf{0}, \rho, \sigma^2 \mathbf{I}; G) && (2)\\
\sigma^2 \sim \text{inv-gamma}(\alpha,\beta), \quad \rho &\sim beta(a,b), \quad H \sim \pi(H)
\end{aligned}
$$

Here $\boldsymbol{\tau}_h$ represents mean and variance of the Gaussian component in the mixture (1), $\mathbf{I}$ is the $(H-1)\times(H-1)$ identity matrix, while $\boldsymbol{w}_i = (w_{i1},\dots,w_{iH})^T$. The distribution P_0 is the normal–inv-gamma density that is conjugate to the

Gaussian distribution $\mathcal{N}(\boldsymbol{\tau}_h)$ in (1). The spatial prior logisticMCAR is defined through a logistic transformation of Gaussian multivariate CAR models for auxiliary parameters $\widetilde{\boldsymbol{w}}_i$s. Parameters in (2) include the proximity matrix G, in this paper fixed as $g_{ij} = 1$ if areas i and j are neighbours and $g_{ij} = 0$ otherwise, a positive parameter ρ of the multivariate CAR specification – $\rho = 0$ corresponding to the transformed weights being independent – and a positive parameter σ^2 representing the conditional variance of the multivariate CAR model. See Beraha *et al.*, 2020 for the definition of such prior.

As mentioned above, when H has the prior distribution $\pi(H)$ with support $\{1,2,\dots\}$, such a model requires a transdimensional sampling scheme for posterior inference. Reversible Jump MCMC samplers (Green, 1995) provide a general framework for transdimensional simulation schemes. Given the current state of the chain $\boldsymbol{\theta} = (H, \boldsymbol{\theta}_H)$, with $\boldsymbol{\theta}_H = (\widetilde{\boldsymbol{w}_1}, \dots, \widetilde{\boldsymbol{w}_I}, \boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_H)$, with $\widetilde{\boldsymbol{w}}_i \in \mathbb{R}^H$, the next state $\boldsymbol{\theta}' = (H', \boldsymbol{\theta}_{H'})$ is (i) sampled from a proposal distribution $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$, and (ii) accepted with probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}')$. Usually, the proposal distribution is defined in two steps. If $\boldsymbol{\theta}_H \in \mathbb{R}^{n_H}$ and $\boldsymbol{\theta}' \in \mathbb{R}^{n_{H'}}$, with $n_{H'} > n_H$ and $d = n_{H'} - n_H$, first a random vector $\boldsymbol{u} \in \mathbb{R}^d$ is sampled from a distribution $q_d(\boldsymbol{u})$ and then $\boldsymbol{\theta}_{H'}$ is defined as $g_{H\to H'}(\boldsymbol{\theta}_H, \boldsymbol{u})$ for a suitable mapping function $g_{H\to H'}$. Since both q_d and $g_{H\to H'}$ are arbitrary, the definition of a suitable reversible jump move is usually a difficult task.

The approach we follow to design a reversible jump move is based on Norets, 2021, who introduces auxiliary priors and proposals for generic nested models indexed by H in $\{1,2,\dots\}$ and a prior for $(H, \boldsymbol{\theta}_H)$ the form $\pi(\boldsymbol{\theta}_H \mid H)\pi(H)$. Let $\boldsymbol{\theta}_\infty$ denote the infinite vector of all parameters for the *largest* model, i.e., the mixture model with infinite components, and let $[\boldsymbol{\theta}_\infty]_{H'}$ be the H'-th entry of $\boldsymbol{\theta}_\infty$, with $H' > H$. Since models are nested, the unknown parameters are nested as well, i.e., if $H' = H+1$, $[\boldsymbol{\theta}_\infty]_{H+1} = (\widetilde{w}_{1H+1}, \dots, \widetilde{w}_{IH+1}, \boldsymbol{\tau}_{H+1})$. The key point is the approximation of the conditional posterior distribution of $[\boldsymbol{\theta}_\infty]_{H+1}$ with a multivariate Gaussian distribution centred at the mode of the conditional posterior of $[\boldsymbol{\theta}_\infty]_{H+1}$ given $\boldsymbol{y}, H+1, \boldsymbol{\theta}_H$. In this way, we sidestep the artificial construction of proposal distributions and mapping functions whilst ensuring quasi-optimal properties of the resulting sampler in terms of chain mixing and sampler efficiency.

To illustrate our algorithm, we consider the case of $I = 9$ areas in a square unit area domain and we simulate data for each area i from

$$y_{ij} \overset{\text{iid}}{\sim} w_{i1}\mathcal{N}(-5,1) + w_{i2}\mathcal{N}(0,1) + w_{i3}\mathcal{N}(5,1) \quad j = 1,\dots,25. \tag{3}$$

Note that the number of samples in each location is small, so that the sharing of information between neighbouring mixtures is a key point. The *true* weights

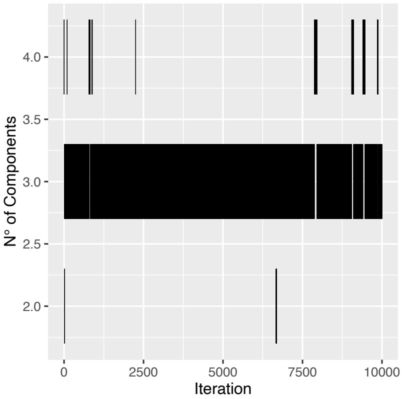


Figure 1: Posterior distribution (traceplot) of H.

$\boldsymbol{w}_i$, $i = 1,\ldots,I$, are set to the inverse of the logistic transformation of $\widetilde{\boldsymbol{w}}_i$ by definition, while the transformed weights $\widetilde{\boldsymbol{w}}_i$ are fixed as $\widetilde{w}_{i1} = 3(s_i - \bar{s}) + 3(t_i - \bar{t})$, $\widetilde{w}_{i2} = -3(s_i - \bar{s}) - 3(t_i - \bar{t})$, where (s_i, t_i) are the coordinates of the center of area i and $(\bar{s}, \bar{t})$ the coordinates of the grid center.

From Figure 1, which displays the posterior distribution of H (no burn-in and no thinning), it is clear that the true value is recovered by our reversible jump sampler.

References

BERAHA, M., PEGORARO, M., PELI, R., & GUGLIELMI, A. 2020. Spatially dependent mixture models via the Logistic Multivariate CAR prior. *arXiv:2007.14961*.

FRÜHWIRTH-SCHNATTER, S., CELEUX, G., & ROBERT, C. P. 2019. *Handbook of Mixture Analysis*. Boca Raton: CRC Press.

GREEN, PETER J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

MALSINER-WALLI, GERTRAUD, FRÜHWIRTH-SCHNATTER, SYLVIA, & GRÜN, BETTINA. 2016. Model-based clustering based on sparse finite Gaussian mixtures. *Statistics and Computing*, **26**, 303–324.

NORETS, A. 2021. Optimal auxiliary priors and reversible jump proposals for a class of variable dimension models. *Econometric Theory*, **37**, 49–81.

STEPHENS, M. 2000. Bayesian analysis of mixture models with an unknown number of components–an alternative to reversible jump methods. *Annals of Statistics*, **28**, 40–74.

NON-PARAMETRIC CONSISTENCY FOR THE GAUSSIAN MIXTURE MAXIMUM LIKELIHOOD ESTIMATOR

Christian Hennig [1], Pietro Coretto[2]

[1] Dipartimento di Scienze Statistiche "Paolo Fortunati", University of Bologna (e-mail: christian.hennig@unibo.it)

[2] Department of Economics and Statistics, University of Salerno, (e-mail: pcoretto@unisa.it)

ABSTRACT: Fitting Gaussian mixtures by maximum likelihood is a major model-based approach to clustering. Under certain constraints, for a fixed and known number of mixture components, it is known to be consistent assuming that the data were indeed generated by a Gaussian mixture. Here we state a nonparametric consistency result, showing that under general conditions that allow for distributions that are not Gaussian mixtures the suitably constrained maximum likelihood estimator for Gaussian mixtures is consistent for the value of its own canonical functional (population version).

KEYWORDS: model-based clustering, consistency, separation, k-means

1 Introduction

The Gaussian mixture model is probably the most popular approach to model-based cluster analysis, see, e.g., McLachlan & Peel, 2000. Given the number of mixture components, under suitable conditions (which obviously include that the model holds), the maximum likelihood (ML) estimator is consistent for estimating the parameters of the Gaussian mixture model, see Redner & Walker, 1984. Occasionally it is claimed that ML in Gaussian mixtures requires the mixture model to hold, whereas some other clustering methods are more universally applicable, because they are "nonparametric" and do not rely on model assumptions. Sometimes k-means is referred to as nonparametric (despite the fact that it can be derived as ML-estimator for a fixed partition model with spherical Gaussian clusters, see Bock, 1996), based on the nonparametric consistency theorem proved by Pollard, 1981, which shows that without assuming any parametric model, under fairly general conditions, k-means converges to its own canonical functional (population version).

Here we state that such a result can also be proved for the ML estimator for Gaussian mixtures, without requiring that the data are in fact generated from a Gaussian mixture. It is also of interest (and discussed in the conference presentation) to what extent it can be made sure that under certain (not necessarily Gaussian mixture) distributions with a clear clustering the value of the Gaussian mixture ML canonical functional can be interpreted appropriately as corresponding to the clusters in the population.

2 ML-estimation of Gaussian mixtures

The Gaussian mixture model is probably the most popular approach to model-based cluster analysis, see, e.g., McLachlan & Peel, 2000. Data are modelled as $p \geq 1$-dimensional Euclidean random variables $X_1, \ldots, X_n$ i.i.d., where the distribution of X_1 has density

$$\psi(x;\theta) = \sum_{j=1}^{G} \pi_j \phi(x;\mu_j,\Sigma_j), \tag{1}$$

where G is the number of mixture components (considered fixed here), $\phi(\cdot;\mu,\Sigma)$ is the p-variate Gaussian density with mean μ and covariance matrix Σ, $\pi_j \in [0,1]$ for $j = 1,2,\ldots,G$, $\sum_{j=1}^{G} \pi_j = 1$. The parameter vector θ contains all Gaussian parameters plus all proportion parameters.

The standard way of estimating θ is by maximum likelihood (ML). For $\tilde{X}_n = (X_1,\ldots,X_n)$, the log-likelihood is

$$l_n(\tilde{X}_n;\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \psi(X_i;\theta). \tag{2}$$

The ML-estimator is then

$$\theta_n(\tilde{X}_n) = \arg\max_{\theta \in \Theta_G} l_n(\tilde{X}_n;\theta). \tag{3}$$

The theory presented here will concern the global optimum $\theta_n(\tilde{X}_n)$, whereas algorithms used in practice such as the EM (McLachlan & Peel, 2000) cannot guarantee that this is indeed found.

The parameter space Θ_G cannot simply be the space of all parameter vectors that are in principle possible in (1), because l_n can degenerate if an eigenvalue of a component's covariance matrix converges to zero.

This can be dealt with constraining the ratio of any two of the eigenvalues of the within-component covariance matrices. See García-Escudero *et al.*, 2018 for a discussion of eigenvalue constraints in Gaussian mixture modelling. Let $\lambda_{j,k}$ be the kth eigenvalue of Σ_j, define

$$\Lambda(\theta) = \left\{\lambda_{j,k} :\ j = 1,2,\dots,G;\ k = 1,2,\dots,p\right\},$$
$$\lambda_{\min}(\theta) = \min_{j,k}\{\Lambda(\theta)\}, \lambda_{\max}(\theta) = \max_{j,k}\{\Lambda(\theta)\}.$$

Then, for given $\gamma < \infty$,

$$\Theta_G = \left\{\theta :\ \pi_j \geq 0\, \forall j \geq 1,\ \sum_{j=1}^{G} \pi_j = 1;\ \frac{\lambda_{\max}(\theta)}{\lambda_{\min}(\theta)} \leq \gamma\right\}. \tag{4}$$

3 Consistency and the canonical functional

The canonical functional (population version) of an estimator is a functional on the space of distributions that extends the estimator in a canonical manner so that it reproduces the estimator when applied to the empirical distribution of the dataset. Define

$$L(P;\theta) = \mathrm{E}_P \log(\psi(X;\theta)),\ L_G(P) = \sup_{\theta\in\Theta_G} L(\theta)$$

(population version of the log-likelihood function and its supremum; E_p denotes the expected value assuming $X \sim P$). Then, the canonical functional corresponding to θ_n is defined as

$$\theta^\star(P) = \arg\max_{\theta\in\Theta_G} L(P;\theta). \tag{5}$$

This definition (as well as (3)) implies existence and uniqueness of the argmax. These are not trivial. Uniqueness is in fact violated, because for mixture models the order of the mixture components is not identified, and for $G > 1$ (in case of existence) there are several maximisers of (3)) and (5). In this case, define $\theta_n(\tilde{X}_n)$ and $\theta^\star(P)$ as any maximiser, $S(P;\theta^\star(P))$ as the set of all maximisers θ with $L(P;\theta) = L(P;\theta^\star(P))$, and

$$\mathcal{K}(P,\varepsilon) = \left\{\theta \in \Theta_G :\ \inf_{\dot\theta\in S(\theta^\star(P))} \|\theta - \dot\theta\| < \varepsilon\right\} \quad \text{for any} \quad \varepsilon > 0.$$

The following assumptions are required:

A1 For every $x_1, \dots, x_G \in \mathbb{R}^p : P\{x_1, \dots, x_G\} < 1$.
A2 $L_G(P) > L_{G-1}(P)$ (implying $L_G(P) > -\infty$, which follows from existence of second moments).

A1 stops all covariance matrices from degenerating simultaneously. A2 guarantees the existence of the involved covariance matrices, and prevents a proportion parameter from being set to zero so that the corresponding mean and covariance matrix could take any value without changing the likelihood. From these it follows that

- $\theta_n(\tilde{X}_n)$ exists with probability arbitrarily close to 1 for large enough n,
- $\theta^\star(P)$ exists,

and ultimately the nonparametric consistency result:

Theorem 1. *Assume A1 and A2. Then for every* $\varepsilon > 0$ *and every sequence of maximisers* $\theta_n(\tilde{X}_n)$ *of* l_n*:*

$$\lim_{n \to \infty} P\left\{\theta_n(\tilde{X}_n) \in \mathcal{K}(P, \varepsilon)\right\} = 1.$$

This can be proved adapting results in Coretto & Hennig, 2017 (where corresponding statements are showed for a version including an additional “noise component”) to the Gaussian mixture case.

References

BOCK, H. H. 1996. Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis*, **23**, 5–28.

CORETTO, P., & HENNIG, C. 2017. Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. *Journal of Machine Learning Research*, **18**, 1–39.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., GRESELIN, F., INGRASSIA, S., & MAYO-ISCAR, A. 2018. Eigenvalues and constraints in mixture modeling: geometric and computational issues. *Advances in Data Analysis and Classification*, **12**, 203–233.

MCLACHLAN, G. J., & PEEL, D. 2000. *Finite Mixture Models*. New York: Wiley.

POLLARD, D. 1981. Strong Consistency of *K*-Means Clustering. *Annals of Statistics*, **9**, 135–140.

REDNER, R. A., & WALKER, H. F. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**, 195–239.

Improving the Reliability of a Nonprobability Web Survey

Yinxuan Huang[1] and Natalie Shlomo[1]

[1] Social Statistics, School of Social Sciences, University of Manchester,
(e-mail: Natalie.shlomo@manchester.ac.uk)

Abstract: In this paper we present robust weighting adjustments and imputation methods to compensate for selection bias in a nonprobability online web-survey taken from the WageIndicator (WI) programme (www.wageindicator.org). For the substantive study, we estimate the gender pay gap (GPG) using the 2016 WI survey data from the Netherlands. To calculate the adjustment weights, we use the 2016 EU-SILC data as a reference sample. Based on the study of GPG, we show that the combination of predictive mean matching and robust weighting adjustment techniques are able to compensate for the selection bias in the nonprobability web survey and ameliorate outcomes of the Blinder-Oaxaca decomposition model in terms of the degree of similarity relative to patterns found in representative probability samples in the Netherlands.

Keywords: Gender Pay Gap, Binder-Oaxaca decomposition, propensity score, predictive mean matching, sample matching

1. Introduction

One nonprobability web survey supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 730998 (InGRID-2 Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy) is the WageIndicator (WI) programme (www.wageindicator.org). It was initiated in The Netherlands in 2001 as a platform for employees and employers looking for information about income. The respondents of these multilingual web-survey are volunteers recruited through national WI websites and a wide range of websites of WI partners. Apart from questions on real wage data, working conditions, and demographic characteristics, WI web surveys also cover a wide range of topics related to job and life satisfaction, work-life balance and health. However, one of the key issues is *self-selection*.

In this paper we present an application using the 2016 Netherlands WI data to measure the gender pay gap (GPG) using log hourly wage. We selected from this dataset those that are employed or self-employed. The minimum age in the data was 17. We also deleted outliers with very small or very large log hourly wage which did not seem feasible since it is important to note that there is no interviewer screening of responses or edit checks to the web survey that are typically carried out as expected in a probability-based sample. The final sample size was 22,643. To adjust for the selection bias, it is necessary to identify a probability reference sample and for this purpose we used the 2016 Netherlands EU-SILC dataset. We selected only the employed and self-employed with a minimum of age of 17 to be consistent with the WI data. The EU-SILC sample size was 12,939.

2. Adjustment Weights and Imputation in the WageIndicator Web survey

We use quasi-randomisation approaches to account for the selection bias in the 2016 Netherlands WI dataset where the two main techniques are sample matching and post-hoc adjustments using propensity scores.

Sample Matching: we calculate a propensity score to estimate the probability of participation for the nonprobability WI dataset. The WI dataset is stacked to the EU-SILC dataset and we define $R_i = 1$ if *i* is in the WI dataset, otherwise $R_i = 0$. Using a logistic regression model, we estimate a propensity score of participation: $p_i = P(R_i = 1|\mathrm{x_i}) = \exp(\mathrm{x}_i\beta)/(1 + \exp(\mathrm{x}_i\beta))$ where $\mathrm{x_i}$ is a vector of covariates that are common in both datasets. The covariates are: age group (17-25, 26-35, 36-45, 46-55, 56-65, 66+), sex (Males, Females), employment (Employed, Self-employed), education (Elementary, Secondary, Tertiary, Missing), occupation (Manager, Professional, Technician, Clerical, Service sales, Agricultural, Craft/trade, Operators, Elementary, Missing). Then, within strata defined by sex and age group, we identify the record in the WI dataset and the record in the EU-SILC data having the closest propensity score and copied the WI log hourly wage to the EU-SILC record. We excluded those cases where WI log hourly wage was missing and allowed for up to 10 multiple donors from the WI dataset. For the substantive analysis on GPG, sample weights and all covariates used were those of the EU-SILC data, but the response variable of log hourly wage is from the WI dataset.

Propensity Score Adjustment: To calculate the propensity score, we use the method proposed in Chen, et al (2019) which utilizes the weights of the EU-SILC reference sample. The initial weight of the WI data d_i is the inverse propensity score. The final weight of the WI data is obtained by benchmarking to the EU-SILC weighted data using post-stratification and raking on the 5 covariates mentioned above: sex*age group*education and employment*occupation.

Missing Data: The calculation of adjustment weights for the WI dataset included item missing data and they were defined as separate categories for the variables education (16%) and occupation (21%). Besides these variables, there is missing data in log hourly wage (45%). Therefore, we carried out an imputation method whilst accounting for the adjustment weights to ensure that the imputation was applied on representative data. For this purpose we ran the MICE procedure with predicted mean matching (Van Buuren, et al. 2011) in R (package: mice.impute.pmm). Other variables in the imputation model with no missing data were sex, age group and urbanicity (Large cities, Small cities, Rural areas) and we also included the adjustment weight to account for the selection bias. We denote this approach by Weight/PMM. In addition, we carried out a different approach assuming a single imputation approach. We first imputed the WI dataset using a single iteration of the predictive mean matching and then calculated the adjustment weights with no missing data categories (any missing data in the EU-SILC were deleted). We denote this approach by PMM/Weight. A simulation study not shown here showed that both approaches provide similar point estimates of correlations and regression coefficients however the PMM/Weight approach had less variation compared to the Weight/PMM approach as is expected from single imputation.

3. Application Measuring the Gender Pay Gap

The advantage of using the WI data to measure the GPG is that is has the variable log hourly wage. In contrast, the EU-SILC data has only annual income from wages and therefore is dependent on confounders such as part-time work. To measure the GPG, we use the Blinder-Oaxaca decomposition (Oxaca, 1973, Blinder, 1973) which is available in the STATA package (Jann, 2008). The method explains the difference in the means of the log hourly wage by decomposing the gender gap into that part that is due to differences in the mean values of the independent variables in the model, and group differences in the effects (parameters) of the independent variables. The method calculates the size and significance of the overall pay gap between men and women, and also divides the gap into a part that is explained by differences in determinants of wages and a part that cannot be explained by such group differences. Moreover, since our analysis include employees and self-employed as reported by the respondents to the WI web survey, the Blinder-Oaxaca decomposition model is integrated with the Heckman's selection model to correct for self-choice in the labour market. All methods in the analyses used weights described in Section 2. As a benchmark for the analysis, the 2016 GPG in the Netherlands was around 15.6% based on the Structure of Earnings survey.

Table 1 shows the results of the Blinder-Oaxaca decomposition of the difference between log hourly earnings of men and women. The upper section exhibits the overall pay gaps between men and women under the different approaches: original WI, Weight\PMM, PMM\Weight and sample matching. In addition, the overall explained part and the unexplained part are also expressed as a percentage of the difference between log hourly earnings of men and women. The subcomponents of the explained part are displayed in the lower section of Table 1. The explanatory variables included in the analysis are age, education, occupation, and urbanicity.

Table 1: Oaxaca-Blinder decomposition of GPG with adjusted selection bias for men and women

	Original WI (unweighted no missing data)	**Weight\ PMM**	**PMM\ Weight**	**Sample Matching (EU-SILC weights**
		Overall		
Men	2.67	3.16	3.12	2.72
Women	2.43	2.70	2.62	2.61
Difference	0.24*	0.46***	0.50***	0.11
Total gap in logged hourly wage	9%	18%	16%	4%
Explained%	7%	27%	34%	3%
Unexplained%	93%	73%	66%	97%
		Detailed composition (%) of the explained gap		
Age Group	1%	-2%	-2%	18%
Education	33%	42%	41%	36%
Occupation	62%	64%	65%	40%
Urbanicity	4%	-4%	-4%	7%
n	10851	22,643	22,643	12,096

All approaches in Table 1 suggest a pay gap between men and women in favour of men. With regard to the size of the GPG (the difference between log hourly wage of men and women), the GPG detected in the original WI data and the sample matching approaches appear to be smaller than those detected in Weight/PMM and PMM/Weight approaches. The GPG is 9% in the original WI dataset and even less in the sample matching of 4% (where the difference was found to be not significant). The Weight/PMM and the PMM/Weight approaches, with the use of adjustment weights and imputation as explained in Section 2, have a GPG of 18% and 16%

respectively, and highly significant, which is approximately the expected level. We note that the results of this model are dependent on the explanatory variables that we have available.

4. Conclusions

In this substantive study of estimating the 2016 GPG for the Netherlands based on the 2016 WI nonprobability web-survey, we provide important lessons for others working with this type of data on how to improve the reliability of nonprobability online data collection for carrying out general inference. We demonstrate that choosing a probability-based reference sample and applying the robust estimation for propensity score calculations according to Chen et al. (2019) with benchmarking on the inverse propensity scores to produce final weight adjustments, and using predictive mean matching to impute missing data, can be used to overcome potential biases in a nonprobability sample. We also demonstrated that sample matching did not produce credible results for this application. We also show two approaches for carrying out imputations of item missing data: impute after the weighting adjustments and include the weight variable as a covariate in the imputation model; impute missing data within the nonprobability sample to obtain a complete dataset and then carry out the weighting adjustments. The approaches provide similar results albeit there is smaller variation in the impute/weight approach as it is typically based on a single imputation.

We note that none of the other studies using the online WI web-survey datasets attempt to adjust for the selection bias using a probability-based reference sample as we have shown here with the EU-SILC for the study of the GPG in the Netherlands. We provide evidence that we must undertake robust methods to improve the reliability of a web survey before carrying out statistical analyses, otherwise we can obtain severely biased results.

Acknowledgement: The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730998 (InGRID-2 Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy).

References

BLINDER, A. S. *1973. Wage Discrimination: Reduced Form and Structural Estimates. Journal of Human Resources, **8** (4), 436–455.*

CHEN, Y., LI, P. & WU, C. 2019. Doubly Robust Inference with Non-probability Survey Samples. *Journal of the American Statistical Association*, **115(532),** 2011-2021.

JANN, B. 2008. The Blinder-Oaxaca Decomposition for Linear Regression Models. *The Stata Journal,* **8(4),** 453-479.

OAXACA, R. *1973. Male-Female Wage Differentials in Urban Labour Markets. International Economic Review, **14** (3), 693–709.*

VAN BUUREN, S. & GROOTHUIS-OUDSHOORN, K. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software,* **45(3)**, 1355–1390.

A SEMI-BAYESIAN APPROACH FOR THE ANALYSIS OF SCALE EFFECTS IN ORDINAL REGRESSION MODELS

Maria Iannario[1] and Claudia Tarantola[2]

[1] Department of Political Sciences, University of Naples Federico II, (e-mail: maria.iannario@unina.it)

[2] Department of Economics and Management, University of Pavia, (e-mail: claudia.tarantola@unipv.it)

ABSTRACT: In this paper we propose a semi-Bayesian approach for the analysis of categorical data with an ordered outcome when a scaling component is considered. A recursive partitioning method yielding two trees –one for the location and one for the scaling– is used for selecting covariates, then a Bayesian approach for model estimation is implemented and an MCMC sampler is used to obtain posterior estimates. An analysis on risk perception concerning Covid-19 pandemic is carried out to assess the performance of the method.

KEYWORDS: Heterogeneity of variances, ordinal responses, scale effects, tree structure, MCMC.

1 Background and preliminaries

Ordinal regression models based on a rating procedure are common in different disciplines such as Economics, Marketing, Medicine and Psychology. If unobserved heterogeneity of variances is present, scale effects in regression structures with ordinal responses are needed. The modeling of scale effects in ordinal regression was already considered by McCullagh, 1980, who introduced the so-called location-scale model, extended in the Bayesian framework because of the flexibility in specifying models and richness and accuracy in providing parameter estimates (see Bürkner, 2017; Liddell & Kruschke, 2018). Variable selection in this framework represents a challenge since typically it is not known which variables contribute to the location and to the scaling component. Tree-based methods offer a nonparametric solution to investigate the interaction structure and automatically select variables (see Tutz & Berger, 2021). In our proposal we take into account covariates obtained for the two components by separate trees and implement an ordinal logit model with parameters estimated through a Bayesian approach.

2 Model description

Let $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)'$ be a random sample generated by an ordinal random variable $Y \sim G(y)$ on the support $\{1, \dots, k\}$, where k is a known integer. We interpret Y_i as the rating expressed by the i-th subject about a definite item. For each subject, we collect information $I_i = (y_i, \boldsymbol{x}_i)$, for $i = 1, 2, \dots, n$, where y_i is the observed value of the rating and $\boldsymbol{x}_i$ is a row vector of the matrix $\boldsymbol{X}$ which includes all the appropriate covariates. We indicate with Y_i^* the underlying (continuous) latent variable such that,

$$\alpha_{j-1} < Y_i^* \le \alpha_j \qquad \Longleftrightarrow \qquad Y_i = j, \qquad j = 1, 2, \dots, k,$$

where $-\infty = \alpha_0 < \alpha_1 < \dots < \alpha_k = +\infty$ are the thresholds of Y^*.

Assume that $p \ge 1$ covariates are relevant for explaining Y^* by the latent regression model

$$Y_i^* = \boldsymbol{x}_i \boldsymbol{\beta} + \sigma \varepsilon_i, \qquad i = 1, 2, \dots, n,$$

where σ is the standard deviation of the noise variable $\varepsilon \sim F_\varepsilon(.)$. Then, the probability mass function of Y_i, for $j = 1, 2, \dots, k$, is:

$$Pr(Y_i = j \mid \boldsymbol{\theta}, \boldsymbol{x}) = Pr(\alpha_{j-1} < Y_i^* \le \alpha_j) = F_\varepsilon\Big(\frac{\alpha_j - \boldsymbol{x}_i \boldsymbol{\beta}}{\sigma}\Big) - F_\varepsilon\Big(\frac{\alpha_{j-1} - \boldsymbol{x}_i \boldsymbol{\beta}}{\sigma}\Big).$$

Common choices for $F_\varepsilon(.)$ are the Gaussian, the logistic, and the (complementary) log-log distribution, whose related models are named probit, logit, and extreme value model, respectively. Here we focus on the logit link function. The parameter vector $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \sigma)'$ is split into the intercept values $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{k-1})'$, the covariates coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ and the scale parameter σ. The latter may depend on covariates yielding $\sigma_i = \boldsymbol{z}_i \boldsymbol{\gamma}$. Here $\boldsymbol{z}_i$ is a row vector of the matrix $\boldsymbol{Z}$ which includes all the $q \ge 1$ relevant covariates and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_q)'$ the covariates coefficients. Since we do not have relevant prior information, we use non informative priors on all parameters of interest, letting the data guide the behaviour of the posterior distributions. We rely on MCMC methods to obtain posterior samples.

3 Applicative section

We examine a set of data collected via a survey conducted in Italy during the 2020 COVID-19 lockdown (March 18 until May 3, 2020). The dataset consits of 2224 observations on 21 variables. Respondents were asked to express on

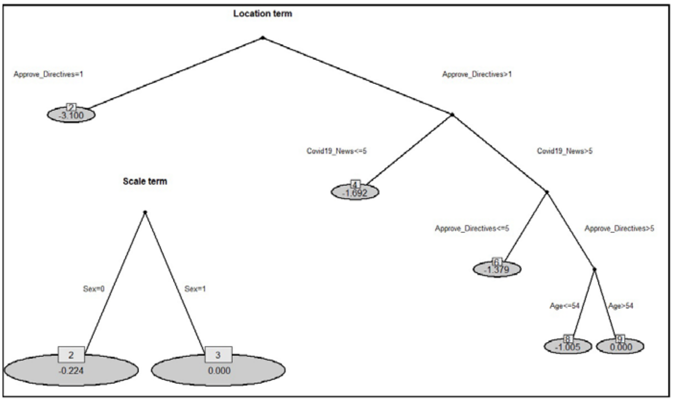


Figure 1. *Tree structures for the location and scale term of the Covid-19 data set. The parameter estimates are given in the terminal nodes.*

a five-points scale how risky they evaluate Covid-19 infection for the society (*Risk*).

The relevant covariates to include in the model are the ones reported in the tree structures of Figure 1, obtained following the approach of Tutz & Berger, 2021. In particular we examine the following covariates: *Sex*, gender of the respondent (1=female, 0=male); *Approve_Directives*, the respondents were asked to evaluate their agreement with the government directive on a scale from 1 (completely disagree) to 7 (completely agree); *Covid_News*, the respondents were asked to evaluate frequency of Covid news access and consumption on a scale from 1 (seldom) to 7 (often); *Age*, a dichotomous variable (0 if $Age \leq 54$, 1 otherwise).

The Bayesian estimates of the location and scale parameters are reported in Table 1 (posterior mean, MCMC Standard Error and 95% credible intervals). These results are obtained via the R package `brms` (Bayesian regression model using "Stan"); see Bürkner, 2017. The estimated thresholds are $\hat{\alpha}_1 = -1.11$*(0.30)*, $\hat{\alpha}_2 = 1.00$*(0.27)*, $\hat{\alpha}_3 = 1.82$ *(0.28)*, and $\hat{\alpha}_4 = 3.39$ *(0.31)*. We run in parrel 4 chains of 2000 iteration with a burnin period of 1000 iteration each; as previously mentioned default non informative priors have been used. Standard convergence diagnostics has been considered. The Bayesian estimates of the latent variables standard deviations are obtained from the posterior samples of log-disc (log-discrimination) with disc corresponding to the inverse of the standard deviation.

In Figure 2 we provide a visual representation of the estimated relation-

Table 1. *Bayesian estimates for the location-scale model*

	Estimate	SE	L-95% CI	U-95% CI
Approve_Directives	0.28	0.03	0.22	0.35
Covid19_News	0.25	0.04	0.17	0.33
Age	0.60	0.24	0.14	1.08
log_disc_Sex	-0.21	0.06	-0.33	-0.09
sd_disc_Sex	1.24	0.08	1.09	1.39

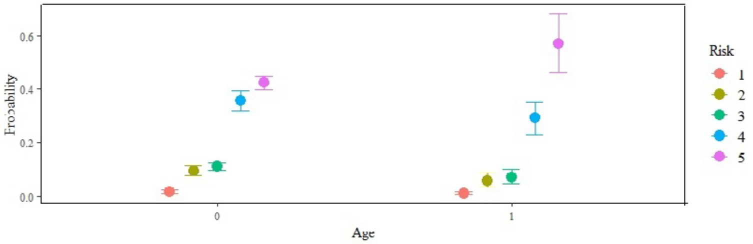


Figure 2. *Marginal effects of Age on Risk evaluation. Points indicate the posterior mean estimates and error bars corresponds to the 95% Credible Intervals.*

ship between *Age* and *Risk*. This figure displays the estimated probabilities of the five response categories for the two age groups. We notice that older people present a higher risk perception. The latter is also stated by respondents who approve the directives expressed by Italian Government and usually read and discuss Covid-19 news. *Sex* instead affects the scale component; higher variability in expressing risk perception is reported for females.

References

Bürkner, P. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, **80**(1), 1–28.

Liddell, T. M., & Kruschke, J.K. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, **79**, 328–348.

McCullagh, P. 1980. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B*, **42**, 109–142.

Tutz, G, & Berger, M. 2021. Tree-structured scale effects in binary and ordinal regression. *Statistics and Computing*, **17**, 31–17.

BEST APPROACH DIRECTION FOR SPHERICAL RANDOM VARIABLES

Jayant Jha[1]

[1] Institut de Neurosciences des Systèmes, Aix-Marseille University, Marseille, France, (e-mail: jayantjha@gmail.com)

ABSTRACT: The quantiles of projections are discussed for spherical random variables. The concept of best approach direction is defined for any quantile based on the ordering of projections in different directions. The usefulness of the concept is discussed when the preferred direction for a specified proportion of observations is of interest. The variation of best approach directions with quantiles is studied for different families of distributions on the sphere which helps in gaining insights into the symmetry, uniformity, and multimodality of the distributions. Exact polynomial-time algorithms are provided for the computation of its estimate on circle and spheres. The connected highest sample density regions for spherical observations can be directly derived from these estimates. Inferential properties of the estimator are studied. Simulations and real data analyses are performed to illustrate the results.

KEYWORDS: depth, directional data, quantiles, von Mises Fisher distribution

SIMPLE EFFECT MEASURES FOR INTERPRETING GENERALIZED BINARY REGRESSION MODELS

Maria Kateri [1]

[1] Institute for Statistics, RWTH Aachen University, Germany
(e-mail: maria.kateri@rwth-aachen.de)

ABSTRACT: In a statistical information theoretical setup, the logistic regression model has been extended to a family of binary regression models that are scaled through the ϕ-divergence. This generalized model provides a great flexibility and enables a precise fit but at the cost of not easily interpretable parameters. Here, we propose some simple measures that facilitate a straightforward and sound interpretation for the effects of quantitative and qualitative explanatory variables on a binary response.

KEYWORDS: logistic regression, ϕ-divergence, ordinal data, odds ratio.

1 Binary response models based on ϕ-divergence

In the context of regression modeling of the effects of p explanatory variables, $x_1, \ldots, x_p$, on a binary response Y, we consider a sample of size n with Y_i being the response of the i-th observation, $\boldsymbol{x}_i = (x_{i1}, \ldots x_{ip})$ the associated values of the explanatory variables, and we assume that $Y_1, Y_2, \ldots, Y_n$ are independent. The most well-known model for modeling $p_i = \mathrm{P}(Y_i = 1)$ in terms of the explanatory variables is the logistic regression model

$$p_i = Pr(Y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}, \quad i = 1, \ldots, n. \tag{1}$$

Kateri & Agresti, 2010 proved that in the above specified framework and in the class of models with explanatory variables that have fixed value $s_j = \sum_{i=1}^{n} y_i x_{ij}$ for $\sum_{i=1}^{n} p_i x_{ij}$, $j = 1, \ldots, p$, the logistic regression model (1) is the closest to the model of constant success probability $\mathrm{P}(Y_i = 1|\boldsymbol{x}_i) = \exp(\beta_0)/[1 + \exp(\beta_0)] = p^{(0)}$, in terms of the Kullback-Leibler (KL) divergence.

Based on this property and considering the general family of ϕ-divergences, which contains the KL as special case, Kateri & Agresti, 2010 introduced the generalized binary regression model

$$F\left(\frac{p_i}{p^{(0)}}\right) - F\left(\frac{1-p_i}{1-p^{(0)}}\right) = \sum_{j=1}^{p} \beta_j x_{ij}\ , \quad i = 1, \ldots, n, \tag{2}$$

with $p_i = p(\boldsymbol{x}_i)$ and $F = \phi'$, where ϕ is a twice differentiable, strictly convex real–valued function on $[0, +\infty)$, satisfying $\phi(1) = \phi'(1) = 0$, $0\phi(0/0) = 0$ and $0\phi(x/0) = x\phi_\infty$ with $\phi_\infty = \lim_{x\to\infty}[\phi(x)/x]$. In the class of models with fixed values $s_j = \sum_{i=1}^n y_i x_{ij}$ for $\sum_{i=1}^n p_i x_{ij}$, $j = 1, \ldots, p$, model (2) under the constraints $0 < p_i < 1$, is the closest to the model of constant success probability, $p_i = p^{(0)}$ for all i, in terms of the ϕ–divergence.

For $\phi(x) = x\log(x) - x + 1, x > 0$, the ϕ–divergence simplifies to the KL divergence and model (2) reduces to (1). The Pearsonian divergence corresponds to $\phi(x) = \frac{1}{2}(x-1)^2$, for which (2) simplifies to the linear probability model $p_i = p^{(0)}\big[1 + (1-p^{(0)})\sum_{j=1}^p \beta_j x_{ij}\big]$, with $-1/(1-p^{(0)}) < \sum_{j=1}^k \beta_j x_{ij} < 1/p^{(0)}$, for all i. For $\phi_\lambda(x) = \frac{1}{\lambda(\lambda+1)}[x^{\lambda+1} - x - \lambda(x-1)]$, $x > 0$, where λ is a real–valued parameter, the ϕ–divergence becomes the power divergence of Cressie and Read and (2) leads to

$$p_i = p^{(0)}\big[1 + \lambda(\beta_{0i} + \sum_{j=1}^p \beta_j x_{ij})\big]^{1/\lambda}, \ \ i = 1, \ldots, n, \tag{3}$$

with parameters β_{0i} satisfying suitable constraints to ensure that $p_i \in (0,1)$. When $\lambda = 0$, $\phi_0(x) = \lim_{\lambda\to 0}[\phi_\lambda(x)]$ and model (3) becomes (1). It reduces to the linear probability model for $\lambda = 1$. Model (3) can be expressed by the simpler equivalent form

$$p_i = \big[\tilde\beta_0 + \sum_{j=1}^p \tilde\beta_j x_{ij})\big]^{1/\lambda}, \ \ i = 1, \ldots, n. \tag{4}$$

2 Parameter interpretation and induced effect measures

The effect of any explanatory variable x_k, quantitative or qualitative, is interpreted conditional on the value of all other covariates in terms of the corresponding parameter β_k in the model, as usual in regression models. In particular, for quantitative x_k, the F-scaled odds ratio (OR)

$$\left[F\Big(\frac{p(\boldsymbol{x}_i)}{p^{(0)}}\Big) - F\Big(\frac{1-p(\boldsymbol{x}_i)}{1-p^{(0)}}\Big)\right] - \left[F\Big(\frac{p(\boldsymbol{x}_{i'})}{p^{(0)}}\Big) - F\Big(\frac{1-p(\boldsymbol{x}_{i'})}{1-p^{(0)}}\Big)\right] \tag{5}$$

opposing the F-scaled odds for any two covariate vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i'}$ differing only on their x_k component, equals $\beta_k(x_{ik} - x_{i'k})$, where β_k is the parameter in model (2). Is x_k categorical with c levels, then the associated parameters β_{kj}, $j = 2, \ldots, c$, equal the F-scaled ORs comparing level j to the reference level 1.

The necessity and practical importance of effect measures that are easy to calculate and straightforward to interpret has been underlined among others by Agresti & Kateri, 2017 and Agresti *et al.*, 2021, for ordinal and binary responses, respectively. These sources discuss existing effect measures, reviewing the related literature and propose new ones. The need for simple effect measures is even more important for the case of generalized models of type (2), for which the F-scaled ORs are by far more unattractive to deal with and interpret. Here, we extend measures proposed in Agresti & Kateri, 2017 and Agresti *et al.*, 2021 for the parametric family of models (3). The adaption of these measures for any other member of the ϕ–divergence based binary regression models family is straightforward.

For a quantitative covariate x_k, a common choice of simple effect measure for the logistic regression model is the rate of change of the response probability $p = \mathrm{P}(Y = 1|\boldsymbol{x} = \boldsymbol{x}^*)$ in x_k when all other covariates in $\boldsymbol{x}$ are kept fixed at value $\boldsymbol{x}^*$, which is $\partial p/\partial\beta_k = \beta_k p(1-p)$ and is known as partial effect. This rate depends on $\boldsymbol{x}$ and for given $\boldsymbol{x} = \boldsymbol{x}^*$ achieves its maximum $\beta_k/4$ at $p = 0.5$. The average partial effect over all $\boldsymbol{x}_i$ in a sample or the partial effect at the mean $\bar{\boldsymbol{x}}$ have been proposed as simple effect measures (s. Agresti *et al.*, 2021). These measures can also be defined for the generalized binary regression models presented above. For the linear probability model this rate equals $\partial p/\partial\beta_k = \beta_k p^{(0)}(1-p^{(0)})$, independent of $\boldsymbol{x}$, while for model (3) and using the alternative definition (4) for p, we have $\partial p/\partial\tilde{\beta}_k = \tilde{\beta}_k p^{1-\lambda}/\lambda$. This rate is increasing in β_k for $\lambda \in (0,1)$ and decreasing for $\lambda < 0$ or $\lambda > 1$. Similar measures can be defined for a binary covariate x_k, by replacing the rate of change by the difference between p's for $x_k = 0$ and $x_k = 1$, estimating these differences over all $\boldsymbol{x}_i$'s and averaging them. In case of a categorical covariate, this process can be followed for the differences between any pair of its levels.

3 Comparison of two ordinal responses

For the problem of comparing two independent groups of items based on their response on an ordinal scale of c levels, the data form a $2 \times c$ contingency table. Such cases can equivalently be analyzed by models treating the binary variable as response. The data in Table 1 are from an experiment on the use of drugs (sulfones and streptomycin) in the treatment of leprosy. The rows group the patients according to the degree of infiltration (a measure of a certain type of skin damage) present at the beginning of the experiment. The columns indicate the change in the overall clinical condition of the patient after 48 weeks of treatment. This data set has been analyzed by generalized binary regression models by Kateri & Agresti, 2010, considering equidistant scores for the response on clinical change. The corresponding logistic model fits well ($G^2 = 0.63$), as does also the linear probability model ($G^2 = 0.26$), both with $df = 3$. Fitting the power divergence model with λ as a parameter, gives the best fit for $\hat{\lambda} = 1.673$ ($G^2 = 0.05$, $df = 2$). Kateri & Agresti, 2010 commented that generalized models of this type, though very flexible, have a restricted scope for applications due to the lack of a simple interpretation.

Table 1. *Change in Clinical Condition (C1: Worse, C2: Stationary, C3: Slight Improvement, C4: Moderate Improvement, C5: Marked Improvement) by Degree of Infiltration in a study comparing two drugs against leprosy (Source: Cochran, 1954).*

Degree of	Clinical Change				
Infiltration	C1	C2	C3	C4	C5
High	1	13	16	15	7
Low	11	53	42	27	11

This drawback can be overcome by adopting for these models the measures for ordinal models introduced by Agresti & Kateri, 2017 (Section 5). These are the ordinal superiority measures Δ and γ, which in our case are $\Delta = \sum_{j>k} \pi_{1j}\pi_{2k} - \sum_{k>j} \pi_{1j}\pi_{2k}$ and $\gamma = \sum_{j>k} \pi_{1j}\pi_{2k} + \sum_{j} \pi_{1j}\pi_{2j}/2$, ranging in $[-1,1]$ and [0, 1], respectively, where π_{ij} is the (i,j) cell probability. For the logistic, linear, and power divergence models fitted on Table 1, Δ is estimated as $\hat{\Delta}_0 = 0.229$, $\hat{\Delta}_1 = 0.241$ and $\hat{\Delta}_{\hat{\lambda}} = 0.231$, respectively, while $\hat{\gamma}_0 = 0.614$, $\hat{\gamma}_1 = 0.620$, and $\hat{\gamma}_{\hat{\lambda}} = 0.616$. Thus under all three models it is estimated that there is about 62% change for a better clinical change at the high than the low group.

The models discussed so far are based on local F-scaled ORs. Treating the ordinal variable as response, we consider models and measures of ordinal superiority that are based on cumulative F-scaled ORs, and compare them.

References

AGRESTI, A., & KATERI, M. 2017. Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics*, **73**, 214–219.

AGRESTI, A., TARANTOLA, C., & VARRIALE, R. 2021. Simple ways to interpret effects in modeling binary data. *In:* KATERI, M., & MOUSTAKI, I. (eds), *Trends and Challenges in Categorical Data Analysis*. Springer (to appear).

KATERI, M., & AGRESTI, A. 2010. A generalized regression model for a binary response. *Statistics & Probability Letters*, **80**, 89–95.

Mixtures of Kato–Jones distributions on the circle, with an application to traffic count data

Shogo Kato[1], Kota Nagasaki[2] and Wataru Nakanishi[2]

[1] Institute of Statistical Mathematics,
(e-mail: skato@ism.ac.jp)

[2] Department of Civil and Environmental Engineering, Tokyo Institute of Technology,
(e-mail: k.nagasaki@plan.cv.titech.ac.jp,
nakanishi@plan.cv.titech.ac.jp)

Abstract: Kato–Jones distribution is a probability distribution on the circle that is unimodal and affords a wide range of skewness and kurtosis. Motivated by a multimodal skewed data set which appears in traffic engineering, we discuss some properties of mixtures of Kato–Jones distributions. A key reparametrization is done to achieve the identifiability of the proposed mixtures. With this reparameterazation, we consider two methods for parameter estimation, namely, a modified method of moments and the maximum likelihood method. These methods are seen to be useful for fitting the proposed mixtures to the traffic counter data set of interest.

Keywords: directional statistics, EM algorithm, maximum likelihood estimation, method of moments, road network analysis.

1 Introduction

Circular data are a set of observations which can be expressed as angles between $[0, 2\pi)$. For the modelling of circular data, a considerable number of probability distributions have been proposed in the literature. Among them, a flexible four-parameter family of distributions has been proposed by Kato & Jones, 2015. It is given by the density

$$g_{\mathrm{KJ}}(\theta; \mu, \gamma, \lambda, \rho) = \frac{1}{2\pi} \left\{ 1 + 2\gamma \frac{\cos(\theta - \mu) - \rho\cos\lambda}{1 + \rho^2 - 2\rho\cos(\theta - \mu - \lambda)} \right\}, \quad 0 \leq \theta < 2\pi,$$

where $0 \leq \mu < 2\pi$, $0 \leq \gamma < 1$, and $0 \leq \rho < 1$ and $0 \leq \lambda < 2\pi$ satisfy $(\rho\cos\lambda - \gamma)^2 + (\rho\sin\lambda)^2 \leq (1-\gamma)^2$. This distribution, which will be called Kato–Jones distribution, is unimodal, affords a very wide range of skewness and kurtosis, has clear interpretation of the parameters, and allows straightforward parameter estimation by both method of moments and maximum likelihood.

Motivated by a multimodal skewed data set which appears in traffic engineering, we consider the following mixtures of Kato–Jones distributions with density

$$\begin{aligned} f(\theta) &= \sum_{k=1}^{m} \pi_k g_{\mathrm{KJ}}(\theta; \mu_k, \gamma_k, \lambda_k, \rho_k) \\ &= \frac{1}{2\pi} \sum_{k=1}^{m} \pi_k \left\{ 1 + 2\gamma_k \frac{\cos(\theta - \mu_k) - \rho_k \cos\lambda_k}{1 + \rho_k^2 - 2\rho_k \cos(\theta - \mu_k - \lambda_k)} \right\}, \quad 0 \leq \theta < 2\pi, \end{aligned} \tag{1}$$

where $m \in \mathbb{N}$ is the number of the components of the mixture and $0 < \pi_1, \ldots, \pi_m < 1$ are the mixing proportions satisfying $\sum_{k=1}^{m} \pi_k = 1$.

Apart from our proposal (1), some mixtures of circular distributions have been proposed in the literature. The most attention have been paid to mixtures of the von Mises distributions (e.g., Wallace & Dowe, 2000; Mooney *et al.*, 2003; Banerjee *et al.*, 2005; Mulder *et al.*, 2020). The components of the mixtures, the von Mises distributions, are symmetric distributions with two parameters controlling location and mean resultant length. Recently, mixtures of the sine-skewed distributions have been discussed by Miyata *et al.*, 2020. The sine-skewed distribution is an extension of a circular distribution which can adopt a mildly asymmetric shape. However these existing models do not seem to be appropriate for our traffic data because one of the clusters of our data is strongly skewed.

In this short paper, we discuss two methods for parameter estimation for the mixture (1), namely, a modified method of moments and the maximum likelihood method. Then, using the proposed methods, we apply the proposed mixture (1) to the traffic data which show bimodality and asymmetry.

2 Parameter estimation

Let $\Theta_1, \ldots, \Theta_n$ be independent and identically distributed from the mixture (1). Note that, as it stands, the parameters, π_k and γ_k, of the mixture (1) can not be uniquely determined in parameter estimation and therefore the mixture (1) is not identifiable. In order to circumvent this problem, we reparametrize the parameters of the mixture (1). With this reparametrization, we discuss two methods for parameter estimation.

The first method is a modified version of the method of moments based on trigonometric moments. Kato & Jones, 2015, proposed a method of moments based on trigonometric moments for Kato–Jones distribution or, equivalently, the mixture (1) with $m = 1$. However their method can not be directly applied to our mixture (1) with general m because the resulting estimates are not always within the range of λ_k and ρ_k. In order to circumvent this problem, we propose a function to evaluate the error between the empirical and theoretic trigonometric moments. Then the estimates are obtained as the minimizer of the proposed function. An advantage of this method is that the estimates always belong to the parameter space and therefore are well-defined. In particular, for a single component mixture $m = 1$, this estimator converges to the method of moments estimator of Kato & Jones, 2015, under certain conditions. Some asymptotic properties such as the consistency and asymptotic normality also hold for the proposed estimator.

Second we consider the maximum likelihood estimation. As is the case for $m = 1$, there do not seem to be a closed-form expression for the maximum likelihood estimator for general m as well. Therefore we consider a numerical algorithm to estimate the maximum likelihood estimate of the mixture (1). We apply the EM algorithm to estimate the parameters of the mixture (1). This algorithm enables us to express the reparametrized mixing proportions of the mixture (1) in closed form in each step. The other parameters of the mixture need to be estimated numerically. However the estimation of these parameters is equivalent to weighted maximum likelihood estimation for a single Kato–Jones distribution and can be done in a similar manner as in Kato & Jones, 2015.

Our experiments suggest the following: The modified method of moments estimation is faster than the maximum likelihood estimation. There is no great difficulty in implementing the maximum likelihood estimation using the EM algorithm. The modified method of moments estimate provides a useful initial value of the EM algorithm for maximum likelihood estimation.

3 Application to traffic count data

Using the two proposed methods for parameter estimation, we fit the proposed mixture (1) to a traffic data set. The data of interest are the timestamps of all vehicles' passing recorded by a traffic counter at 20.4 kilopost of Kobe route, Hanshin Expressway, Japan. Kobe route is located in Osaka metropolitan area and connects two large cities of Japan, Osaka and Kobe. The data of the timestamps are converted from 24 hours to angles in $[0, 2\pi)$; for clarity, 0 corresponds to midnight, π to midday, etc. Our data show bimodality and one of the clusters of the data is strongly skewed.

In parameter estimation, we first estimate the parameters based on the modified method of moments. Then the maximum likelihood estimation is carried out by using the modified method of moments estimates as the initial values of the EM algorithm. The model estimated by the maximum likelihood method is a two-component $(m = 2)$ mixture of Kato–Jones distributions. The estimated model provides a reasonable fit to the data including the strongly skewed cluster of data. Details of the data analysis will be given in the talk.

References

BANERJEE, A., DHILLON, I.S., GHOSH, J., & SRA, S. 2005. Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, **6**, 1345–1382.

KATO, S., & JONES, M.C. 2015. A tractable and interpretable four-parameter family of unimodal distributions on the circle. *Biometrika*, **102**, 181–190.

MIYATA, Y., SHIOHAMA, T., & ABE, T. 2020. Estimation of finite mixture models of skew-symmetric circular distributions. *Metrika*, **83**, 895–922.

MOONEY, J.A., HELMS, P.J., & JOLLIFFE, I.T. 2003. Fitting mixtures of von Mises distributions: a case study involving sudden infant death syndrome. *Computational Statistics & Data Analysis*, **41**, 505–513.

MULDER, K., JONGSMA, P., & KLUGKIST, I. 2020. Bayesian inference for mixtures of von Mises distributions using reversible jump MCMC sampler. *Journal of Statistical Computation and Simulation*, **90**, 1539–1556.

WALLACE, C.S., & DOWE, D.L. 2000. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, **10**, 73–83.

How to design a directional distribution

John T. Kent [1]

[1] School of Mathematics, University of Leeds (e-mail: j.t.kent@leeds.ac.uk)

Abstract: One way to specify a model in directional statistics is to look for an exponential family which mimics the multivariate normal distribution under high concentration. However, in some important examples this strategy leads to an over-specified model, with a spare parameter. This paper revisits two standard distributions, the Fisher-Bingham distribution on the sphere and the bivariate von Mises distribution on the torus, and takes a fresh look at guidelines to specify this parameter.

Keywords: Fisher-Bingham distribution, bivariate von Mises distribution, directional statistics

1 Introduction

Directional statistics is concerned with data on circles, spheres and related manifolds. For this paper, we focus on two particular cases: the unit sphere S_{p-1} in $\mathbb{R}^p$, especially the case $p = 3$, and the torus $(S_1)^d$, especially $d = 2$. In each case it is possible to construct an exponential family which mimics the multivariate normal distribution under high concentration. But there is a problem. The models include one more parameter than necessary. Over-parameterized models can lead to problems of interpretation and fitting. Hence, if possible, it is usually better to choose a parameterization with the same number of parameters as the corresponding asymptotic multivariate normal distribution. Various suggestions have been made in the literature to fix the spare parameter, but these suggestions sometimes have severe limitations.

2 The Fisher-Bingham distribution

The 6-parameter Fisher-Bingham distribution (FB6) on the unit sphere S_2 in $\mathbb{R}^3$, after rotation to a standardized coordinate system, has the density

$$f(\boldsymbol{x}) \propto \exp\{\kappa x_3 + \beta_1 x_1^2 + \beta_2 x_2^2\}, \quad \boldsymbol{x} \in S_2 \tag{1}$$

with respect to the uniform distribution on S_2, where $\kappa > 0,\ -\infty < \beta_2 \leq \beta_1$. Here $\boldsymbol{x} = [x_1,\ x_2,\ x_3]^T$ is a unit vector, $\boldsymbol{x}^T\boldsymbol{x} = 1$, and the third coordinate axis can be viewed as the "north pole". The parameters of the model are κ, β_1, β_2 plus two parameters for location and one parameter for orientation about the north pole, making 6 parameters in all. Provided $\kappa > 0$ and $2\beta_1 \leq \kappa$, $2\beta_2 \leq \kappa$, the density is unimodal at the north pole. Under high concentration the distribution is asymptotically bivariate normal, a 5-parameter family.

How should the parameters in FB6 be constrained to yield a 5-parameter family? Two choices are:

(a) the *balanced FB5 distribution (FB5b)*, with $\beta_2 = -\beta_1 = \beta$ say, $0 \leq \beta \leq \kappa/2$. It was introduced in Kent, 1982 (without the adjective "balanced") and is sometimes known as the Kent distribution.

(b) The *extreme FB5 distribution (FB5e)*. Set $\beta_1 = 0$, $\beta_2 = -\delta$, say, where $\delta \geq 0$. It was introduced in Kent *et al.* , 2016.

The parameters β_1 and β_2 determine the eccentricity of the the distribution; in the limiting bivariate normal case, the eccentricity describes the ratio of the eigenvalues of the covariance matrix. Although both the balanced and extreme FB5 distributions can accommodate high eccentricity under high concentration, the balanced distribution is much less able to describe high eccentricity under low and moderate concentration. See Fig. 1 panels (a) and (b) for an example with moderate concentration, where the mode has been moved to the equator. Panel (a) gives the most eccentric choice possible with the FB5b distribution, and Panel (b) shows how FB5e can be much more eccentric.

3 Bivariate von Mises distribution

Represent points on the torus $S_1 \times S_1$ as a pair of angles θ_1 and θ_2. The bivariate von Mises distribution, after a suitable rotation of each circle, has density

$$f(\theta_1, \theta_2) \propto \exp\{\kappa_1 c_1 + \kappa_2 c_2 + \boldsymbol{v}_1^T B \boldsymbol{v}_2\}, \tag{2}$$

where B is a 2×2 parameter matrix (not necessarily symmetric) Mardia & Jupp, 1999; Mardia *et al.* , 2008; Kent *et al.* , 2008; Mardia *et al.* , 2012. Here the shorthand notation $c_j = \cos\theta_j, s_j = \sin\theta_j$ and $\boldsymbol{v}_j = [c_j,\ s_j]^T,\ j = 1,2$ for the first order trigonometric functions has been used. The density is unimodal with a mode at $\theta_1, \theta_2 = 0$, provided

$$\kappa_1 + b_{11} > 0,\ \kappa_2 + b_{11} > 0,\ b_{12} = b_{21} = 0,\ b_{22}^2 \leq (\kappa_1 + b_{11})(\kappa_2 + b_{11}). \tag{3}$$

After adding 2 parameters for location, the bivariate von Mises distribution, subject to the constraints in (3) forms a 6-parameter family (BVM6,

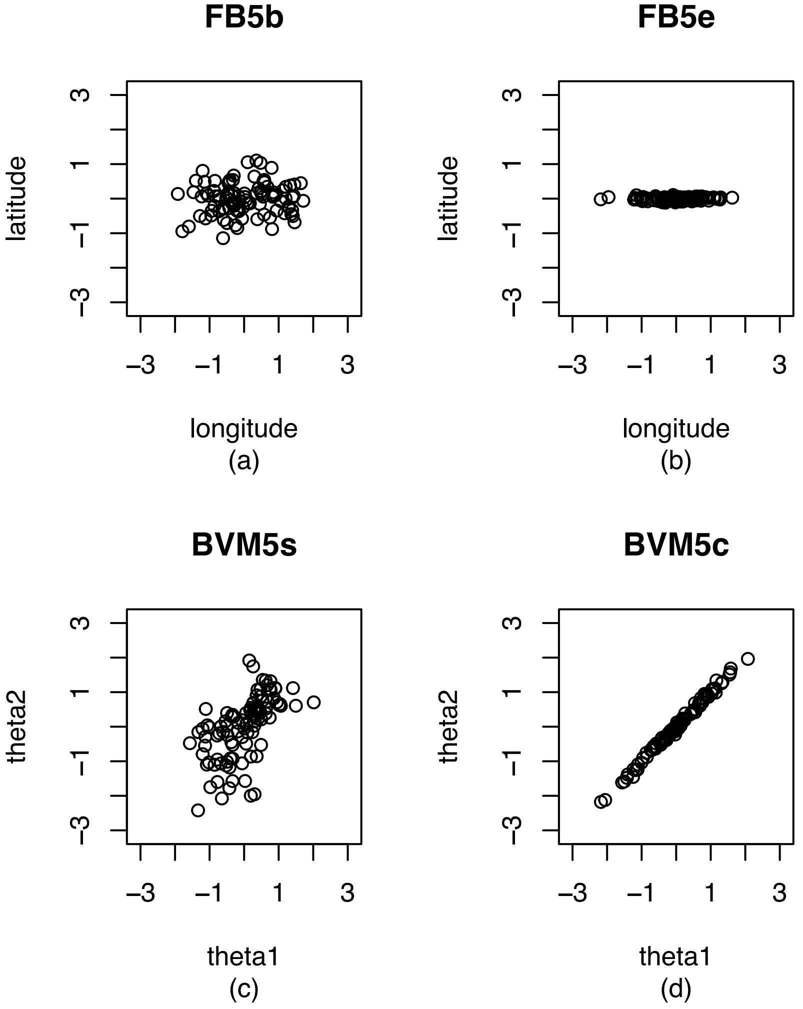


Figure 1. *Illustration of directional simulations. Panels (a) and (b) illustrate the FB5b and FB5e distributions. Note the spread in longitude is similar for both distributions, but FB5e has a much smaller spread in latitude than can be modelled by FB5b. Panels (c) and (d) illustrate the BVMs and BVMc distributions. Note the marginal spreads in* θ_1 *and* θ_2 *are similar to one another and similar for the two distributions. However, BVMc shows a much higher correlation between the two angles than can be modelled by BVMs.*

say). Under high concentration BVM6 is asymptotically bivariate normal, a 5-parameter family. Just as in the last section, BVM6 is over-parameterized.

In this case there are two well-established ways to constrain the spare degree of freedom:

(a) The *bivariate von Mises sine model (BVM5s)*, by setting $b_{11} = 0$.
(b) The *bivariate von Mises cosine model (BVM5c)*, by setting $b_{22} = |b_{11}|$.

Under high concentration both the sine and cosine model can accommodate high correlation between θ_1 and θ_2. However, under low and moderate concentration the cosine model accommodates high correlation more effectively. See Fig 1, panels (c) and (d) for an example with moderate concentration. Panel (c) gives the most highly correlated choice possible with the BVMs distribution, and Panel (d) shows how BVMc can exhibit much higher correlation.

References

KENT, J. T. 1982. The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society, Series B*, **44**, 71–80.

KENT, J. T., MARDIA, K. V., & TAYLOR, C. C. 2008. Modelling strategies for bivariate circular data. *Pages 70–73 of:* BARBER, S., BAXTER, P. D., GUSNANTO A., & MARDIA, K. V. (eds), *The Art and Science of Statistical Bioinformatics*. Leeds University Press.

KENT, J. T., HUSSEIN, I., & JAH, M. K. 2016. Directional Distributions in Tracking of Space Debris. *Pages 2081–2086 of: Proceedings of the 19th International Conference on Information Fusion (FUSION), Heidelberg, Germany*. IEEE.

MARDIA, K. V., & JUPP, P. E. 1999. *Directional Statistics*. New York: Wiley.

MARDIA, K. V., HUGHES, G., TAYLOR, C. C., & SINGH, H. 2008. A multivariate von Mises distribution with applications to bioinformatics. *Canadian Journal of Statistics*, **36**, 99–109.

MARDIA, K. V., KENT, J. T., ZHANG, Z., TAYLOR, C. C., & HAMELRYCK, T. 2012. Mixtures of concentrated multivariate sine distributions with applications to bioinformatics. *Journal of Applied Statistics*, **39**, 2475–2492.

IDENTIFYING MORTALITY PATTERNS OF MAIN CAUSES OF DEATH AMONG YOUNG EU POPULATION USING SDA APPROACHES

Simona Korenjak-Černe[1] and Nataša Kejžar[2]

[1] University of Ljubljana, School of Economics and Business, and Institute of Mathematics, Physics and Mechanics (e-mail: simona.cerne@ef.uni-lj.si)

[2] University of Ljubljana, Faculty of Medicine, Institute for Biostatistics and Medical Informatics (e-mail: natasa.kejzar@mf.uni-lj.si)

ABSTRACT: Young population is generally considered to be very healthy, so the most common causes of death in this population are often associated with risky behaviours. In fact, in the population aged 20-39, external causes of death account for more than half of the causes of death in EU countries (also in the US), while by far the most common causes of death in the general population are circulatory diseases and various cancers. The next most common causes of death in the 20-39 age group in the US are suicides and homicides, both of which are strongly associated with stress, therefore we examine them also for EU countries. Our application is based on the 2016 data, which at this point is the most recent complete data available, however the area is even more relevant nowadays in the pandemic and post-pandemic period with many extraordinary stressful situations.

In order to include as much information as possible from these data into our cluster analysis, we use symbolic data methods. By considering for each age-sex group not only the number of deaths but also their distribution among the main causes of death, we can include internal variability (in our case, variability by cause of death) in the analysis.

The main objective of the study is twofold: first, to identify groups of EU countries with similar mortality patterns, taking into account two-level information for each age and sex group, i.e. number of deaths and their distribution among the main causes of death; and second, to describe clusters of mortality patterns and to investigate possible links between the mortality patterns in the obtained clusters and some other socio-demographic indicators. In our study, we use a symbolic table for more informative data description and adaptations of compatible hierarchical and non-hierarchical clustering methods for group identification that allows us to consider this two-level information. To this end, we have extended our R package `clamix`.

KEYWORDS: symbolic data analysis, main death causes, young population.

ROBUST SUPERVISED CLUSTERING: SOME PRACTICAL ISSUES

Fabrizio Laurini [1] and Gianluca Morelli [1]

[1] Department of Economics and Management and Ro.S.A., University of Parma, (e-mail: fabrizio.laurini@unipr.it, gianluca.morelli@inicas.it)

ABSTRACT: A semi-automatic procedure for regression models, which leads to identify the optimal number of clusters, in a large and complex data set, is discussed. Robust methods usually suffer from high-computational load and we give practical clues when using the TCLUST-REG with the FSDA toolbox in Matlab

KEYWORDS: FSDA, Outliers, TCLUST-REG.

1 Introduction and motivation

The purpose of this paper is to provide the user with a set of semi-automatic tools in the context of regression clustering which can help to select the optimal number of groups (or more generally to find a set of relevant solutions), give insights about the optimal restriction factors among the variances of the estimated residual variances and finally enable to estimate the optimal trimming level keeping into account that it can depend on the chosen solution.

We made use of our Flexible Statistics for Data Analysis software package, the FSDA toolbox for MATLAB, which is available as "Add-On" inside MATLAB or on github.

2 Technical machinery

Let the multivariate covariates X and the response variable Y be defined on Ω with values in $\mathcal{X} \times \mathcal{Y} \subseteq R^{p-1} \times R$. Then, $\{x_i, y_i\}$, $i = 1, 2, \ldots, n$, represents a i.i.d. random sample of size n drawn from (X, Y). Assume that Ω can be partitioned into k groups, say Ω_1, Ω_2, …, Ω_k. Then, the general formulation of the regression clustering mixture model has a density which can be written as

$$p(x, y, \theta) = \sum_{g=1}^{k} p(y|x, \theta_{y,g}) p(x, \theta_{x,g}) \pi_g,$$

where $p(y|x,\theta_{y,g})$ is the conditional density of Y given x in Ω_g which depends on the vector of parameters $\theta_{y,g}$, $p(x,\theta_{x,g})$ is the marginal density of X in Ω_g which depends on the vector of parameters $\theta_{x,g}$, and π_g reflects the importance of Ω_g in the mixture with the usual constraints $\pi_g > 0$ and $\sum_{g=1}^{k}\pi_g = 1$. Vector θ denotes the full set of parameters $\theta = (\theta_{y,g}^T\ \theta_{x,g}^T)^T$. It is customary to assume that in each group g the conditional relationship between Y and x, $p(y|x,\theta_{y,g})$, has form $Y = \beta_{0,g} + x^T\beta_g + \varepsilon_g$, with proper parameters for all g components. Assuming normality and linearity implies the Gaussian Cluster Weighted Model (CWM) of Gershenfeld *et al.* , 1999, and can be written as

$$p(x,y,\theta) = \sum_{g=1}^{k} \phi(y;\beta_{0,g} + \beta_g^T x, \sigma_g^2)\phi_{p-1}(x,\mu_g,\Sigma_g)\pi_g.$$

This is linked to the clustering around regression that ignores the distribution of X. To accommodate for such an unrealistic assumption, in the so-called classification framework of model based clustering, the classification log-likelihood

$$L_{\text{Cla}}(\theta) = \sum_{i=1}^{n}\sum_{g=1}^{k} z_{ig}(\theta)\log\phi(y_i|b_{0g},x_i^T b_g,s_g^2)\phi_{p-1}(x_i,m_g,S_g)p_g. \qquad (1)$$

The target function (1) is unbounded when no constraints are imposed on the scatter parameters. It is necessary therefore to impose constraints on the maximization on the set of eigenvalues of the scatter matrices.

In the literature of robust regression it is widely known the effect of both vertical outliers in Y and outliers in X. Robustness can be achieved by discarding in each step of the maximization procedure a proportion of units equal to α, associated with the smallest contributions to the target likelihood. More precisely, for example in the mixture modeling context, the Trimmed CWM parameter estimates are based on the maximization of the following trimmed likelihood function $L_{\text{Mixt}}(\theta|\alpha,c_y,c_X)$ García-Escudero *et al.* , 2017

$$L_{\text{Mixt}}(\theta|\alpha,c_y,c_X) = \sum_{i=1}^{n} z^*(x_i,y_i)\log\left[\sum_{g=1}^{k}\phi(y_i|b_{0,g},b_g^T x,s_g^2)\phi_{p-1}(x_i,m_g,S_g)p_g\right], \qquad (2)$$

where $z^*(\cdot,\cdot)$ is a 0-1 trimming indicator function. A fixed fraction α of observations can be unassigned by setting $\sum_{i=1}^{n} z(x_i\, y_i) = [n(1-\alpha)]$. The TCLUST-REG García-Escudero *et al.* , 2010 can be considered as a particular case of TCWRM in which the contribution to the likelihood of $\phi_{p-1}(x_i,m_g,S_g)$ is set equal to 1.

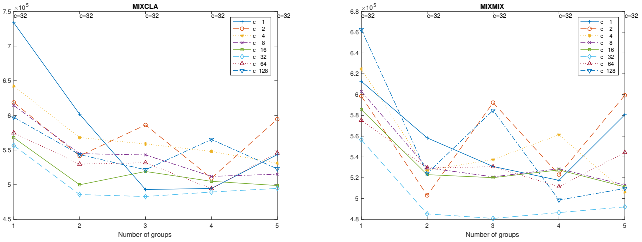


Figure 1. *Maximised likelihood for models Cla and Mixt (from left to right respectively) with associated choices of restriction coefficient and number of groups*

Often it is convenient to consider a further trimming step, which discards a proportion α_X of the units, after taking into account their degree of remoteness in the X space, among the observations which have survived the first trimming operation. The observations surviving to the two trimming steps are then used for updating the regression coefficients, weights and scatter matrices. This modification of the algorithm is usually referred in the literature as *adaptive TCLUST-REG*. In the sequel we contrast the performance of (adaptive) TCLUST-REG to a large data set to provide guidelines when complex big data are available.

3 Data, results and further research

In the data analysed, kept anonymous for confidentiality, there are shopping tracks of 24 month sales of non-food items. The number of customers is approximately 470000. The average sale of each customer in the time period is the response variable Y from which we perform the clustering regression approach. The set of explanatory variables is given by the number of visits, the number of items bought per visit, the percentage value bought with promotion/sales, age and the gender of the customer. The optimal number of groups, and the optimal constraint factor, are displayed for likelihoods 1 and 2 in Figure 1 (left and right panels respectively).

In all cases we obtained 3 groups with approximately 8% of customers identified as outliers and un-allocated. The outliers in the data (roughly 40000 customers) are mostly characterized by occasional shops and low revenue for the retailer. Broadly speaking in cluster 2 the customers spend more and buy more articles compared to the average. Customers in cluster 1 tend to buy on sales rather than full price and buy more compared to the other clusters. In

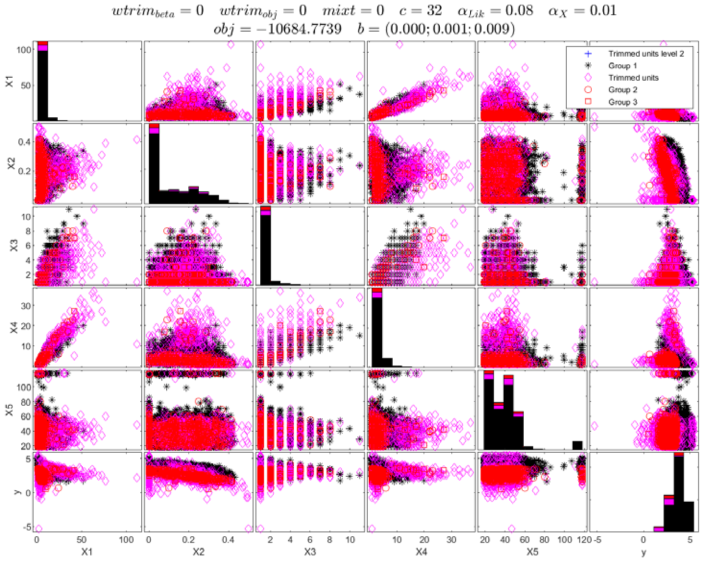


Figure 2. *Scatter plot of data and cluster membership for the optimal solution*

cluster 3 people buy less expensive articles, but often.

We want to remark that the identification of these outliers is fully automatic and not arbitrary, but comes as a by-product of an optimal model-based algorithm. The cluster membership is displayed in Figure 2 and the overlap of units would create troubles in many "standard" cluster methods. Further details and comments will be provided during the Conference.

References

García-Escudero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S., & Mayo-Iscar, A. 2017. Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *Statistics and Computing*, **27**, 377–402.

García-Escudero, L.A, Gordaliza, A., Mayo-Iscar, A., & San Martin, R. 2010. Robust clusterwise linear regression through trimming. *Computational Statistics & Data Analysis*, **54**, 3057 – 3069.

Gershenfeld, N., Schoner, B., & Metois, E. 1999. Cluster-Weighted Modelling for Time-Series Analysis. *Nature*, **397**(6717), 329–332.

A NONPARAMETRIC APPROACH FOR STATISTICAL MATCHING UNDER INFORMATIVE SAMPLING AND NONRESPONSE

Daniela Marella[1] and Danny Pfeffermann[2]

[1] Department of Education, University of Roma Tre,
(e-mail: daniela.marella@uniroma3.it)

[2] Central Bureau of Statistics and Hebrew University of Jerusalem, Israel; University of Southampton, UK, (e-mail: D.Pfeffermann@soton.ac.uk)

ABSTRACT: Statistical matching attempts to combine the information obtained from different, non-overlapping samples, selected from the same target population, to form a matched sample containing the data in the different samples. The aim of this paper is to propose a nonparametric approach of handling statistical matching under informative sampling and not missing at random (NMAR) nonresponse, by use of empirical likelihood.

KEYWORDS: calibration, empirical likelihood, informative sampling, NMAR nonresponse.

1 Introduction

Statistical matching is becoming more and more popular in recent years. Information on a set of variables is often obtained from different data sources related to the same target population, each containing only some of the variables, with no joint observations on all the variables.

Let A and B be two independent samples of size n_A and n_B respectively, selected from a population of N independent and identically distributed (*i.i.d.*) records, generated from some joint probability distribution function (*pdf*), $f_p(x,y,z)$ of variables (X,Y,Z). Only (X,Y) are observed for the units in sample A, and only (X,Z) are observed for the units in sample B. Because of the lack of joint information, the joint *pdf* $f_p(x,y,z)$ is not identifiable. Several alternative techniques have been proposed in the literature to overcome the identification problem. At first, techniques based on the conditional independence assumption (CIA) between Y and Z given X were considered. A second group of techniques uses external auxiliary information on the statistical relationship between Y and Z. Finally, a third approach consists of analysing the uncertainty regarding the joint distribution of (X,Y,Z), that is several alternative models for the joint distribution of (X,Y,Z), compatible with the distributions of (X,Y) and (X,Z)

in the samples A and B, are considered. See, D'Orazio *et al.* (2006), Conti *et al.* (2016) and references therein.

In practice, the sample selection in survey sampling involves complex sampling designs based on different levels of clustering and differential inclusion probabilities. When the inclusion probabilities are related to the value of the target outcome variable even after conditioning on the model covariates, the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data is then different from the model holding in the population. This, quite common phenomenon is known as *informative sampling*, see Pfeffermann and Sverchkov (2009). The case of informative sampling designs in the statistical matching problem in a parametric setting assuming complete response is analysed in Marella and Pfeffermann (2019). However, in practice, not all the sampled units respond. When the response probabilities are correlated with the missing target outcomes, even after conditioning on the observed data (often, the model covariates), the missing data are not missing at random (NMAR). Valid inference under NMAR nonresponse requires therefore modelling the response mechanism. The problem in applying standard inferential procedures, which ignore the sampling process and nonresponse, is that the distribution holding for the data observed for the responding units can be very different from the distribution holding for the population data, which may result in large bias of estimators and affect other aspects of the inference process.

The aim of this paper is to propose an approach of handling statistical matching under informative sampling and NMAR nonresponse, by use of empirical likelihood (EL). The main advantages of EL approach are: (i) it does not require to specify the population model; (ii) it facilities the use of calibration constraints.

2 Empirical likelihood approach for statistical matching

The empirical likelihood is essentially the likelihood of the multinomial distribution used in Hartley and Rao (1968), where the parameters are the point masses assigned to the distinct sample values. We assume that the sampling designs used for selecting the two samples A and B are informative for the corresponding joint population *pdf*, in the sense that the sample selection probabilities $\{\pi_{i,A}, \pi_{i,B}\}$ are correlated with at least some of the variables (X, Y, Z), implying that the joint sample *pdf* is different from the corresponding population *pdf*. Additionally to informative sampling, we assume that A and B are subject to NMAR unit nonresponse. Let $I_i^A\,(I_i^B)$ be the sample indicator taking the value 1 if the ith population unit is drawn to the sample $A(B)$ and 0 otherwise. Let $R_i^A\left(R_i^B\right)$ define the response indicator, taking the value 1 if sample unit $i \in A(i \in B)$ responds and 0 otherwise. The response process is assumed to be independent between units. We

assume that X can take K distinct values with probabilities $p_k^X = \mathrm{P}(X = x_k)$, while Y and Z are continuous.
The basic idea of the EL approach is to approximate the population distribution with a multinomial model which support is given by the empirical observations. Under the CIA, the probabilities $p_i^{XYZ} = P(x_i, y_i, z_i)$ can be factorized as,

$$p_i^{XYZ} = P(x_i, y_i, z_i) = P(x_i)P(y_i \mid x_i)P(z_i \mid x_i) = p_k^X p_i^{Y|X} p_i^{Z|X} . \tag{2.1}$$

The parameters $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$ are unknown and need to be estimated from the samples A and B. By Bayes rule,

$$p_{i,R_A}^{Y|X} = P(y_i \mid x_k, I_i^A = 1, R_i^A = 1) = \frac{P(R_i^A = 1 \mid x_k, y_i, I_i^A = 1)}{P(R_i^A = 1 \mid x_k, I_i^A = 1)} p_{i,A}^{Y|X} \tag{2.2}$$

$$p_{k,R_A}^X = P(x_i = x_k \mid I_i^A = 1, R_i^A = 1) = \frac{P(R_i^A = 1 \mid x_k, I_i^A = 1)}{P(R_i^A = 1 \mid I_i^A = 1)} p_{k,A}^X \tag{2.3}$$

where the sample models $p_{i,A}^{Y|X}$, $p_{k,A}^X$ are defined as,

$$p_{i,A}^{Y|X} = P(y_i \mid x_i, I_i^A = 1) = \frac{E_A(w_{i,A} \mid x_i)}{E_A(w_{i,A} \mid x_i, y_i)} p_i^{Y|X} , \tag{2.4}$$

$$p_{k,A}^X = P(x_i \mid I_i^A = 1) = \frac{E_A(w_{i,A} \mid x_i)}{\sum_{j \in A_{xk}} E_A(w_{j,A} \mid x_j) p_j^{Y|X}} p_k^X \tag{2.5}$$

and $A_{xk} = \left\{ i \in A : x_i = x_k \right\}$. Then, the sample models and the model for the response probabilities $P(R_i^A = 1 \mid x_k, y_i, I_i^A = 1)$ define the model holding for the outcomes of the responding units. Notice that unless $P(R_i^A = 1 \mid x_k, y_i, I_i^A = 1) = P(R_i^A = 1 \mid x_k, I_i^A = 1)$ for all (x_k, y_i), the model (2.2) is different from the sample model $p_{i,A}^{Y|X}$ (2.4), which in turn is different from the population model $p_i^{Y|X}$ under informative sampling. Analagous expressions to

(2.2)-(2.5) are obtained for the model holding for the responding units in B. Thus, assuming that the outcome, the sampling and the response are independent between units, the *empirical respondents likelihood* for the sample $A \cup B$ is given by,

$$ERL_{Obs}^{A\cup B} = \prod_{k=1}^{K} (p_{k,R_A}^{X})^{r_{k,A}^{X}} \prod_{i\in R_{A,k}} p_{i,R_A}^{Y|X} \prod_{k=1}^{K} (p_{k,R_B}^{X})^{r_{k,B}^{X}} \prod_{i\in R_{B,k}} p_{i,R_B}^{Z|X} \tag{2.6}$$

where $R_{A,k}$ ($R_{B,k}$) defines the group of respondents with $X = x_k$ in sample $A(B)$ of size $r_{k,A}^{X}(r_{k,B}^{X})$. The response probabilities in (2.6) are unknown and need to be modelled by a parametric model and estimated from the available data. Let γ_A, γ_B be the unknown response models parameters postulated in the two samples, the likelihood (2.6) must be maximized with respect to $[\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}, \gamma_A, \gamma_B]$ under the constraints,

$$p_k^X \geq 0, p_i^{Y|X} \geq 0, p_i^{Z|X} \geq 0, \sum_{k=1}^{K} p_k^X = 1, \sum_{j\in R_{A,k}} p_j^{Y|X} = 1, \sum_{j\in R_{B,k}} p_j^{Z|X} = 1. \tag{2.5}$$

An important advantage of the proposed approach is that it facilitates the use of calibration constraints. That is, auxiliary information on known population means for some auxiliary variables can be incorporated by placing additional constraints on the maximization process.

References

CONTI, P.L, MARELLA, D. & SCANU, M. 2016. Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, 111, 516, 1715-1725.

D'ORAZIO, M., DI ZIO, M., & SCANU, M. 2006. *Stastical Matching: Theory and Practice*. Chichester: Wiley.

HARTLEY, H.O & RAO, J.N.K., 1968. A new estimation theory for sample surveys. *Biometrika,* 55, 547-557.

MARELLA, D. & PFEFFERMANN, D. 2019. Matching information from two independent informative sampling. *Journal of Statistical Planning and Inference*, 203, 70-81.

PFEFFERMANN, D. & SVERCHKOV, M. 2009. Inference under informative sampling. In Handbook of Statistics 29B; Sample Surveys: Inference and Analysis (Eds. D.Pfeffermann and C.R.Rao). Amsterdam: North Holland.

INVESTIGATING MODEL FIT IN ITEM RESPONSE MODELS WITH THE HELLINGER DISTANCE

Mariagiulia Matteucci[1] and Stefania Mignani[1]

[1] Department of Statistical Sciences, University of Bologna,
(e-mail: m.matteucci@unibo.it, stefania.mignani@unibo.it)

ABSTRACT: Under the Bayesian approach, posterior predictive model checking has become a popular tool for fit assessment of item response theory models. In this study, we propose the use of the Hellinger distance to quantify the distance between the realized and the predictive distribution of the model-based covariance for item pairs. Specifically, the case of over-fitting is taken into account. The results of a simulation study show the effectiveness of the method.

KEYWORDS: posterior predictive model checking, Hellinger distance, MIRT models, goodness of fit.

1 Introduction

Bayesian estimation of item response theory (IRT) models via Markov chain Monte Carlo (MCMC) has been intensively applied due to its flexibility in arranging complex situations. Through posterior predictive model checking (PPMC; Rubin, 1984). it is possible to define tools for evaluating the fit of the model. Considerable advantages of the method are that it does not rely on distributional assumptions, and it is relatively easy to implement, given that the entire posterior distribution of all parameters of interest is obtained through MCMC algorithms.

The use of PPMC for IRT models received an increasing interest in assessing multidimensionality (Sinharay *et al.*, 2006; Levy and Svetina, 2011). The PPMC method is based on the comparison between the observed and the replicated data of a given discrepancy measure D. PPMC is implemented first with graphical analyses and then with the estimation of the posterior predictive p-values (PPP-values). However, the PPP-value simply counts the number of times the replicated D is equal or higher than the realized D without addressing the magnitude of the difference between the two distributions. To overcome these limitations, in a previous paper (Matteucci and Mignani, 2020) it is proposed to measure the difference between the predictive and the realized distribution via the Hellinger distance, a suitable measure for improving the interpretation of results in applied settings and useful for model comparison purposes.

The main objective of this paper is to deepen the performance of the Hellinger distance in an over-fitting situation to evaluate the potential misfit of a IRT multidimensional model when the data are generated by a unidimensional approach. We explore our proposal by simulation to enrich the previous results of an under-fitting scenario.

2 The discrepancy measures

PPMC techniques are based on the comparison of observed data with replicated data generated or predicted by the model by using a number of diagnostic measures that are sensitive to model misfit (Sinharay *et al.,* 2006). Substantial differences between the posterior distribution based on observed data and the posterior predictive distribution indicate poor model fit. Given the data y, let $p(y|\omega)$ and $p(\omega)$ be the likelihood for a model depending on the set of parameters ω and the prior distribution for the parameters, respectively.

From a practical point of view, one should define a suitable discrepancy measure $D(\cdot)$ and compare the posterior distribution of $D(y,\omega)$, based on observed data, to the posterior predictive distribution of $D(y\text{rep},\omega)$. Discrepancy measures should be chosen to capture relevant features of the data and differences among data and the model. It is possible to resort to the PPP-values defined as "the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity". The choice of a suitable discrepancy measure is crucial in PPMC. Effective diagnostic measures in checking for unidimensionality or multidimensionality are based on the association or on covariance/correlation among item pairs. In this paper we consider the model-based covariance (MBC; Reckase, 1997) that depends on both data and model parameters. The MBC is found to be effective as it measures the covariance among item pairs by explicitly conditioning on the latent variable. If the local independence assumption holds, the MBC is close to zero. If the local independence does not hold, the MBC is greater than zero for items loading on the same latent variable (PPP-values are close to zero) and smaller for items loading on different latent variables (PPP-values are close to one).

Lastly, Levy and Svetina (2011) proposed an overall measure, namely the generalized dimensionality discrepancy measure (GDDM) that is a unidirectional measure of average conditional covariance defined as the mean of the absolute values of MBC over unique item pairs. When the GDDM is equal to zero, a "weak" local independence for all the item pairs is assumed. If the assumption of local independence is violated, the GDDM is greater than zero and the PPP-value will be close to zero.

3 The Hellinger distance

To quantify the difference between the realized and the predictive distribution within PPMC, Matteucci and Mignani (2020) propose to use the Hellinger (H) distance which is symmetric, it does obey the triangle inequality and its range is 0-1. The direct calculation is computationally demanding and given the MCMC simulations, it is usually estimated by the normal kernel density. In order to check for local independence, we used the H distance with the MBC discrepancy measure (MBC-H) to take into account a fit measure for each item pair, and with the GDDM measure (GDDM-H) to evaluate the overall fit based on item pairs. It is proposed to investigate the assumption of local independence for 2PNO models by focusing on multidimensional data analyzed with the unidimensional model. The main strengths of the H distance, compared to traditional approaches rely on the possibility a) to directly quantify the amount of misfit; b) to be used for model comparison purposes, c) to make more informative analyses on item pairs. Furthermore, it is demonstrated that, in practical applications, the MBC-H can be used to: a) leave out the models that show serious misfit by using the threshold of 0.5; b) compare the amount of misfit of different competing models and choose the model which fits the data best; c) identify, also through graphical plots, critical items that may involve misfit which are associated to high MBC-H in several pairs.

In this paper we confirm the strength of our proposal through a simulation study in an over-fitting setting, where unidimensional data are analyzed through different multidimensional models.

4 The simulation

A simulation study is conducted to examine the performance of the proposed MBC-H and GDDM-H at detecting the misfit when data follow a two-parameter normal ogive (2PNO) unidimensional model and we fit a multi-unidimensional model and an additive model with two latent dimensions. Response data for tests with k=10 or k=20 items and a sample size of n=1,000 or n=2,000 are simulated. Two subtests are assumed for the multidimensional models (k_1=k_2=5 or k_1=k_2=10). The case of unidimensional data analyzed with the same model is also considered. A number of 5,000 MCMC iterations are conducted, where 1,000 are used for PPMC. Finally, 100 replications are done for each simulation condition. The parameters of the data-analysis model are estimated via the Gibbs sampler. The over-fitting scenario is particularly meaningful for its implications in real situations. Although unidimensionality is quite unrealistic, especially with a high number of items, there are situations addressing a different point of view. For example, in the educational context, a test could be arranged under the assumption that groups of items refer to different cognitive domains. In this situation, a multidimensional model should be estimated to investigate the different domains, but one predominant dimension should explain the most part of the variability.

The main results of the simulation study are reported in Table 1. We do not present PPP-values as they indicate lack of bad fit for all conditions. We found the more critical evidence for k=20 where the fitted models show, on average, MBC-H higher than 0.5 meaning bad fit. The additive model seems to be the more appropriate, even when data are unidimensional, as it also includes an overall latent trait. For k=10, the goodness of fit improves but again the results of the H-distance are towards the additive model. Classical Bayesian indicators such as DIC confirm these conclusions. The Hellinger distance seems to be an effective tool in highlighting the presence of possible misfit and determining plausible thresholds for classifying the misfit levels.

Table 1- Summary results for the 100 replications and for all item pairs.

	k	***n***	**DIC**	**MBC-H**				**GDDM-H**
				Mean	**Sd**	**Min**	**Max**	
Uni/multi-uni	10	1000	6600.27	0.427	0.086	0.275	0.605	0.352
	20	1000	15312.89	0.546	0.047	0.396	0.656	0.475
	10	2000	14671.31	0.446	0.068	0.312	0.618	0.376
	20	2000	25302.47	0.542	0.061	0.375	0.644	0.456
Uni/additive	10	1000	6433.11	0.383	0.081	0.246	0.551	0.292
	20	1000	15173.99	0.524	0.049	0.363	0.639	0.426
	10	2000	14410.26	0.373	0.060	0.276	0.518	0.279
	20	2000	25022.79	0.506	0.065	0.323	0.621	0.402
Uni/uni	10	1000	6614.53	0.450	0.082	0.295	0.619	0.376
	20	1000	12702.39	0.552	0.057	0.400	0.653	0.479
	10	2000	13180.96	0.447	0.083	0.283	0.643	0.339
	20	2000	25306.66	0.554	0.059	0.377	0.664	0.475

References

LEVY, R., & SVETINA, D. 2011. A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology,* **65**, 208-232.

MATTEUCCI, M., & MIGNANI, S. 2020. The Hellinger distance within posterior predictive assessment for investigating multidimensionality in IRT models. *Multivariate Behavioral Research*, DOI: 10.1080/00273171.2020.1753497

RECKASE, M. 2009. *Multidimensional Item Response Theory.* New York: Springer-Verlag.

RUBIN, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applies statistician. *Annals of Statistics*, **12**, 1151-1172.

SINHARAY, S., JOHNSON, M. S., & STERN, H. S. 2006. Posterior predictive assessment of item response theory models. *Applied Psychological Measurement,* **30**, 298-321.

PCA-BASED COMPOSITE INDICES AND MEASUREMENT MODEL

Matteo Mazziotta[1], Adriano Pareto[1]

[1] Italian National Institute of Statistics, Rome,
(e-mail: mazziott@istat.it, pareto@istat.it)

ABSTRACT: The measurement of complex phenomena, such as well-being, socio-economic development, and competitiveness, is very difficult because they are characterized by a multiplicity of aspects or dimensions. Principal Component Analysis (PCA) is probably the most popular multivariate statistical technique for reducing data with many dimensions. Thus, often, socio-economic indicators are reduced to a single index by using PCA. However, PCA is implicitly based on a reflective measurement model that is not suitable for all types of indicators. In this paper, we discuss the use and misuse of PCA for measuring complex phenomena.

KEYWORDS: PCA, data reduction, composite index, measurement model.

1 Introduction

Socio-economic indicators are often analysed by multivariate statistical technique, such as Principal Components Analysis (PCA), in order to summarize the data and to construct composite indices. However, a fundamental distinction must be made between reducing dimensionality and constructing composite indices.

Reducing dimensionality is a purely mathematical operation that consists in summarizing a set of individual indicators, so that most of the information in the data is preserved. Many techniques have been developed for this purpose, but PCA is one of the oldest and most widely used. Its idea is simple: reduce the dimensionality of a dataset, while preserving as much ‘variability’ as possible. This translates into finding new variables that are linear functions of the original ones, that successively maximize variance and that are uncorrelated with each other. Because the new variables are defined by the dataset at hand, and not a priori, PCA can be considered an adaptive data analysis tool.

Constructing a composite index (or composite indicator) is a conceptual, as well as mathematical, operation that consists in summarizing (or aggregating as it is termed) a set of individual indicators, on the basis of a well-defined measurement model: formative or reflective. Therefore, a composite indicator is formed when individual indicators are compiled into a single index, on the basis of an underlying model of the multi-dimensional concept that is being measured.

Obviously, a composite index can be obtained by reducing dimensionality (with an appropriate model of measurement), but not necessarily reducing dimensionality provides a composite index. In this paper, we discuss the use of PCA for studying socio-economic indicators and we explain how and why it can be improperly used as a method for constructing composite indices.

2 The measurement model

As it is known, a model of measurement can be conceived through two different approaches: reflective or formative.

The most popular approach is the reflective model, according to which individual indicators denote effects (or manifestations) of an underlying latent variable. Therefore, causality is from the concept to the indicators and a change in the phenomenon causes variation in all its measures. In this model, the construct exists independently of awareness or interpretation by the researcher, even if it is not directly measurable. Specifically, the latent variable R represents the common cause shared by all indicators X_i reflecting the construct, with each indicator corresponding to a linear function of the underlying variable plus a measurement error:

$$X_i = \lambda_i R + \varepsilon_i \tag{1}$$

where X_i is the indicator *i*, λ_i is a coefficient (loading) capturing the effect of R on X_i and ε_i is the measurement error for the indicator *i*. Measurement errors are assumed to be independent and unrelated to the latent variable. A typical example of reflective model is the measurement of the intelligence of a person. In this case, it is the 'intelligence level' that influences the answers to a questionnaire for measuring attitude, and not vice versa. Hence, if the intelligence of a person increased, this would be accompanied by an increase of correct answers to all questions.

The second approach is the formative model, according to which individual indicators are causes of an underlying latent variable, rather than its effects. Therefore, causality is from the indicators to the concept and a change in the phenomenon does not necessarily imply variations in all its measures. In this model, the construct is defined by, or is a function of, the observed variables. The specification of the formative model is:

$$R = \sum_i \lambda_i X_i + \zeta \tag{2}$$

where λ_i is a coefficient capturing the effect of X_i on R, and ζ is an error term. A typical example of formative model is the measurement of well-being of society. It depends on health, income, occupation, services, environment, etc., and not vice versa. So, if any one of these factors improved, well-being would increase (even if the other factors did not change). However, if well-being increased, this would not necessarily be accompanied by an improvement in all factors.

Note that (1) is a system of simple regression equations where each individual indicator is the dependent variable and the latent variable is the explanatory variable; whereas (2) represents a multiple regression equation where the latent variable is the dependent variable and the indicators are the explanatory variables.

Although the reflective approach dominates the psychological and management sciences, the formative view is common in economics and sociology.

3 How and when to use PCA

According to the "Handbook on Constructing Composite Indicators. Methodology and user guide" by OECD, PCA should be used to study the overall structure of the dataset, assess its suitability, and guide some methodological choices in constructing a composite index. Nevertheless, PCA can also be used for constructing composite indices. For this purpose, it is essential to define the model of measurement in order to describe relationships between the phenomenon to be measured (latent variable) and its measures (individual indicators). But above all, it is necessary to establish whether PCA is formative or reflective. To answer to this question it is important to distinguish between PCA and FA , since they are sometimes considered more or less interchangeable.

PCA is a pure data reduction technique that aggregates the observed variables (indicators) in order to reproduce the most amount of variance with fewer variables (principal components or factors). PCA works without an explicit hypothesis on the latent structure of the variables, so that the observed variables are themselves of interest. This makes PCA similar to multiple regression in some ways, in that it seeks to create optimized weighted linear combinations of variables.

FA is an explanatory model in which the observed variables (indicators) are assumed to be (linear) functions of a certain (fewer) number of unobserved variables (latent factors). FA hypothesizes an underlying latent structure of the variables and estimates latent factors influencing observed variables.

On the basis of these features, PCA is often views as formative, whereas FA is a reflective measurement model. However, the question whether PCA is formative or reflective is not trivial. Indeed, although the definition of principal component as weighted sum of individual indicators suggests a formative model, some important issues are involved. In particular:

1. In a PCA based index (e.g. the first factor), the weights depend on the correlations among indicators. But correlations among individual indicators are not relevant in a formative model and cannot be explained by it. Indeed, in a formative model, the indicators do not necessarily share the same theme and hence have no a preconceived pattern of intercorrelation.
2. Individual indicators aggregated by a PCA based index (e.g. the first factor) are – by construction – highly correlated. But in a multiple regression, such as Equation 2, individual indicators should have little or no correlation among themselves in order to avoid multicollinearity. Indeed, an excessive collinearity among indicators makes it difficult to separate the distinct influence of the individual indicators on the latent variable.
3. Under certain conditions, the principal components are equivalent to the factor scores obtained by FA and then they can be considered estimators of latent factors. But FA is a reflective measurement model, so PCA cannot be considered really formative.

In the light of the above, a composite index based on PCA looks more suited for a reflective approach than a formative one. In fact, PCA is commonly used for the evaluation of reflective measurement models and it is considered an appropriate method for examining the indicators' underlying factor structure in order to check the content validity.

4 Conclusions

The construction of composite indices for measuring multidimensional phenomena is a central issue in data analysis. Researcher cannot solve this question simply by using PCA or related methods, such as Factor Analysis, since they are typically used for a reflective approach.

Reducing dimensionality and constructing composite indicators are two separate issues that are repeatedly confused. Both the procedures aims to summarize a set of variables or individual indicators, but reducing dimensionality focuses on extracting the most important information from the data, whereas constructing composite indicators focuses on the use of a measurement model that can be reflective or formative. Extracting the most important information from the data translates in summarizing correlated indicators, but correlations can indicate causal, non-causal (spurious) and coincidental relationships, making the principal components meaningless or difficult to interpret. On the contrary, defining a measurement model means assuming a specific direction of causality between the measures (individual indicators) and the latent variable (phenomenon to be measured).

Measuring complex phenomena, such as development or well-being, requires a formative approach, where the index to be constructed does not exist as an independent entity, but it is a composite measure directly determined by a set of non-interchangeable individual indicators or pillars (e.g. the HDI by UNDP).

In such a context, PCA can be recommended for various reasons. Firstly, PCA is a powerful tool for reducing complexity and visualizing data, so that the researcher can identify clusters of units (regions, provinces or countries) that have the same characteristics. Secondly, it allows for comparing empirical dimensions (factors) with theoretical dimensions (pillars), in order to evaluate any differences and to detect possible dimensions that had not previously been taken into account. Lastly, PCA makes it easy to study correlations among many individual indicators in order to find redundant and non-redundant indicators and to assess linkages with other relevant measures, such as GDP. Nevertheless, the use of PCA for constructing formative composite indices is not recommended, since it can give very misleading information about the latent variable of interest, being based exclusively on the covariance structure between the individual indicators.

References

MAZZIOTTA, M., & PARETO, A. 2019. Use and Misuse of PCA for Measuring Well-Being. *Social Indicators Research*, **142**, 451–476.

GENDER INEQUALITIES FROM AN INCOME PERSPECTIVE

Marcella Mazzoleni[1], Angiola Pollastri[2] and Vanda Tulli[2]

[1] Center on Economic, Social and Cooperation dynamics (CESC), University of Bergamo, (e-mail: marcella.mazzoleni@unibg.it)

[2] Department of Statistics and Quantitative Methods, University of Milano-Bicocca (e-mail: angiola.pollastri@unimib.it, vanda.tulli@unimib.it)

ABSTRACT: The difference between females' and males' income is one of the main topics in the analysis of gender gap, as it is known that, even with a higher educational level, females earn less than males do. To inspect this, we analyse and estimate the distribution of the ratio of females' income over males' income using the methodology based on the distribution of the ratio of two Dagum with three parameters. We applied this method to the Bank of Italy Survey on Household Income and Wealth (SHIW) data to evaluate the deciles, the density functions, and the cumulative distribution functions of the ratio of the females' income over males' income in different age classes, Italian areas, and years.

KEYWORDS: Dagum distribution, distribution of the ratio of two Dagum random variables, distribution of the ratio of female and male income.

1 Introduction and method

It is well known that even with a higher educational level, women earn less than men do. The differences between men' and women' income on average are decreasing in the recent years but income parity has not yet been achieved.

The purpose of this paper is to estimate the distribution of the ratio of females' income over males' income. The methodology used to study the ratio is based on the distribution of the ratio of two Dagum with three parameters (Pollastri and Zambruno 2010). The distribution of this ratio studied in two different situations can reveal the gender inequality concerning income in different groups or times, accordingly the distribution of the ratio is analysed to reveal the gender inequality with applications to the income in different age classes, areas, and times.

In literature we have many examples which confirm that the model proposed in 1977 by Camilo Dagum fits very well to many distributions of economic variables. Supposing that X is a type I Dagum, then $X \sim D(a, b, p)$ with $a, b, p > 0$. The distribution function for $x > 0$ is defined as (Kleiber and Kotz 2003):

$$F_X(x) = \left[1 + \left(\frac{x}{b}\right)^{-a}\right]^{-p}$$

While the density function for $x > 0$ is:

$$f_X(x) = \frac{apx^{ap-1}}{b^{ap}\left[1+\left(\frac{x}{b}\right)^{a}\right]^{p+1}}$$

The Dagum distribution parameters are estimated using the function *dagum* implemented in the VGAM package in the software R. This function estimates the parameters using the maximum likelihood estimation method proposed by Kleiber and Kotz (2003). Domanski and Jedrzejczak (1998) showed, through a simulation study, that estimation method performance is good for a and p when $n > 2000$ or 3000, while for the scale parameter b the bias tends to 0 when $n > 4000$.

The purpose of this paper is to analyse the ratio:

$$U = \frac{X}{Y}$$

where $X \sim D(a_1, b_1, p_1)$ and $Y \sim D(a_2, b_2, p_2)$ with X and Y independent.

Following the definition of the density function of the ratio of two random variables in Mood, Graybill and Boes (1974), applying the independence of X and Y and the density function of a type I Dagum, it is possible to obtain the density function for the ratio U:

$$f_U(u) = \int_0^{+\infty} y \left\{ \frac{a_1 p_1 (uy)^{a_1 p_1 - 1}}{b_1^{a_1 p_1}\left[1+\left(\frac{uy}{b_1}\right)^{a_1}\right]^{p_1+1}} \right\} \times \left\{ \frac{a_2 p_2 y^{a_2\, p_2 - 1}}{b_2^{a_2 p_2}\left[1+\left(\frac{y}{b_2}\right)^{a_2}\right]^{p_2+1}} \right\} dy$$

Using the definition of the cumulative distribution function, it is possible to obtain:

$$F_U(u) = \frac{a_1 p_1 a_2 p_2}{b_1^{a_1 p_1} b_2^{a_2 p_2}} \int_0^u t^{a_1 p_1 - 1} \int_0^{\infty} y^{a_1 p_1 + a_2\, p_2 - 1} \left[1+\left(\frac{ty}{b_1}\right)^{a_1}\right]^{-p_1-1} \times$$

$$\left[1+\left(\frac{y}{b_2}\right)^{a_2}\right]^{-p_2-1} dydt$$

In Pollastri and Zambruno (2010) a graphical analysis of this method performance is exposed comparing the empirical and the computed density function, where the empirical one is created with the ratios of all the possible couples.

2 Applications to Survey on Household Income and Wealth

We apply this method to the individual net incomes in 2016 from the Bank of Italy Survey on Household Income and Wealth (SHIW). We compare the ratio of the females' and males' income in different groups:

- males and females divided in three age classes: *young* $(age < 40)$, *adult (*$40 \leq age < 70)$ and *old* $(age \geq 70)$

- males and females divided in three areas: *North, Centre,* and *South and Islands*
- males and females in two different years: 2016 and 1998

The dataset is composed of 11,844 subjects. Of these 50.98% are males, and 49.02% are females. Concerning the division by ages 15.21% are aged less than 40 years, 54.41% are aged between 40 and 70 years, and 30.39% are aged equal or more than 70 years. For the division by area, 43.90% of the subjects come from the North, 22.40% from the Centre, and 33.70% from the South and the Islands. The 1998 dataset, it is composed of 12,616 subjects, of these 56.52% are males and 43.48% are females.

After estimating the Dagum parameters, we evaluate the cumulative distribution function and the deciles of the ratio of the females' income over the males' income, comparing the results of the ratio distributions for different ages, areas, and times.

Comparing the ratio of the females' income over males' income in different age classes, we observe higher value of deciles for younger subject, lower for adult group, and even lower for the older group. This confirms that the income at the beginning of the career is similar between the two genders but increasing the age and the position achieved, the gap rises.

For the deciles of ratio of the females' income over males' income comparing different areas, we observe close and higher value for the subjects that live in North and Centre of Italy, and lower value for the subjects that live in South and Islands. This can be related to the different economical and social situation in the Islands and in the South of Italy.

We observe that the deciles of the ratio of the females' income over males' income are higher in 2016 with respect to the ratio in 1998. This confirms that the differences between men' and women' income are decreasing in the recent years, but income parity has not yet been achieved.

3 Conclusions

In this paper we propose to use the ratio of two type I Dagum random variables for analysing the difference of the income of females and males. We observe that this method gives us interesting conclusions and can be applied to different dataset comparing also the ratio of females' over males' income in different countries, in order to highlight the differences concerning gender gap.

This method is used to analyse the Italian situation and to compare the ratio of females' over males' income in different ages, areas and times. As a matter of fact, in the applications we observe less diversity for females' and males' income in the younger group, but the diversity rises increasing the subjects' age, passing from young to adult, and from adult to old group. In the division by areas, the deciles of the ratio of females' income over males' income for the North and Centre are close, while for the South and Islands a wider difference between genders is observed. The difference of the income for males' and females' is decreasing over the years. In fact, the deciles of the ratio of females' income over males' income are higher for 2016 dataset than the deciles of the 1998 dataset.

References

ALLEVA G. (2017). Indagine conoscitiva sulle politiche in materia di parità tra uomini e donne. Istat, Rome.

BANK OF ITALY (2016). I bilanci delle famiglie italiane nell'anno 2016. Supplementi al Bollettino Statistico, XVII, Centro Stampa Banca d'Italia, Roma.

BANK OF ITALY (1998). I bilanci delle famiglie italiane nell'anno 1998. Supplementi al Bollettino Statistico, XVII, Centro Stampa Banca d'Italia, Roma.

DAGUM C. (1977). A new model of personal income distribution: specification and estimation. *Economie Appliquée*, **30** (3), 413-437.

DAGUM C. (1990). Generation and properties of income distribution functions. [In:] *Studies in Contemporary economics. Income and wealth distribution, inequality and poverty*, C. Dagum, M. Zenga (Eds.), Springer, Berlin.

DOMANSKI, C., JEDRZEJCZAK A. (1998) Maximum likelihood estimation of the Dagum model parameters. *International Advances in Economic Research*, **4**, 243–252

KLEIBER, C., KOTZ S. (2003). Statistical Size Distributions in Economics and Actuarial Sciences. Hoboken, NJ, USA: Wiley-Interscience.

MOOD A., GRAYBILL F., BOES D. (1974). Introduction to the theory of statistics. Wiley, New York.

POLLASTRI A., ZAMBRUNO G. (2010) Distribution of the ratio of two independent Dagum random variables. *Operations Research and Decisions*, **3** (20), 95-102.

YEE T. (2019). VGAM: Vector Generalized Linear and Additive Models. R package.

TRANSFORMATION MIXTURE MODELING FOR SKEWED DATA GROUPS WITH HEAVY TAILS AND SCATTER

Yana Melnykov[1], Xuwen Zhu[1] and Volodymyr Melnykov[1]

[1] The University of Alabama, (e-mail: ymelnykov@cba.ua.edu, xzhu20@cba.ua.edu, vmelnykov@cba.ua.edu)

ABSTRACT: For decades, Gaussian mixture models have been the most popular mixtures in literature. However, the adequacy of the fit provided by Gaussian components is often in question. Various distributions capable of modeling skewness or heavy tails have been considered in this context recently. In this paper, we propose a novel contaminated transformation mixture model that is constructed based on the idea of transformation to symmetry and can account for skewness, heavy tails, and automatically assign scatter to secondary components.

KEYWORDS: finite mixture model, cluster analysis, transformation to normality, symmetry

UNCONDITIONAL M-QUANTILE REGRESSION

Luca Merlo[1], Lea Petrella[2] and Nikos Tzavidis[3]

[1] Department of Statistics, Sapienza University of Rome,
(e-mail: luca.merlo@uniroma1.it)

[2] MEMOTEF Department, Sapienza University of Rome,
(e-mail: lea.petrella@uniroma1.it)

[3] Department of Social Statistics and Demography and Southampton Statistical Sciences Research Institute, University of Southampton,
(e-mail: N.TZAVIDIS@soton.ac.uk)

ABSTRACT: In this paper we develop the unconditional M-quantile regression for modeling unconditional M-quantiles in the presence of covariates. Extending the paper by Firpo *et al.* (2009), we assess the impact of small changes in the explanatory variables on the M-quantile of the unconditional distribution of the dependent variable by running a mean regression of the recentered influence function of the unconditional M-quantile on the covariates. The proposed methodology is applied on the Survey of Household Income and Wealth (SHIW) 2016 conducted by the Bank of Italy.

KEYWORDS: Influence function, M-estimation, RIF regression, Robust method

1 Introduction

Quantile Regression (QR), as proposed by Koenker & Bassett Jr (1978), has proven to be a powerful tool to explore conditional distributions in many empirical applications. However, if one is interested in how the whole unconditional distribution of the dependent variable responds to changes in the covariates, using the well-known QR would yield misleading inferences (see Firpo *et al.* 2009 and Borah & Basu 2013). Motivated by this interest, Firpo *et al.* (2009) proposed the Unconditional Quantile Regression (UQR) approach for modeling unconditional quantiles of a dependent variable as a function of the explanatory variables. This method builds upon the concept of Recentered Influence Function (RIF) which originates from a widely used tool in robust statistics, namely the Influence Function (IF) discussed in Hampel *et al.* (2011). The RIF of a distributional statistic ν is obtained by adding back the statistic to the IF and it can be thought of as the contribution of an individual observation on ν. In the regression framework where covariates are available, Firpo *et al.* (2009) proposed to replace the dependent variable with the RIF to model the unconditional quantiles of the response and evaluate the effect of changes in the law of the covariates on unconditional quantiles. When the interest of the research is concentrated on the entire distribution of a response variable, in addition to the classical QR, a possible alternative is represented by the M-quantile regression (MQR) approach proposed by Breckling & Chambers (1988). This method provides a "quantile-like" generalization of the mean regression based on influence functions, combining in a common framework the robustness and efficiency properties of quantiles and expectiles (Newey & Powell 1987), respectively.

In this article, we extend the UQR of Firpo *et al.* (2009) to the M-quantile regression framework. We develop the Unconditional M-quantile Regression (UMQR) to model the M-quantiles of the unconditional distribution of the response variable. In order to analyze how the entire unconditional distribution of the outcome is affected by changes in the distribution of explanatory variables, we regress the RIF of the unconditional M-quantile on the covariates and denote such effect as Unconditional M-Quantile Partial Effect (UMQPE).

2 Methodology

Let Y denote a scalar random variable with absolutely continuous distribution function F_Y. The M-quantile of order $\tau \in (0,1)$ of Y is defined as the solution, $\theta_\tau \in \mathbb{R}$, of the following estimating equation:

$$\int \psi_\tau(y-\theta_\tau)dF_Y(y) = 0, \qquad (1)$$

where $\psi_\tau(u) = \mid \tau - \mathbf{1}_{(u<0)} \mid \psi(u/\sigma_\tau)$, with ψ being the first derivative of a convex loss function ρ and σ_τ is a suitable scale parameter. In this work, we consider the well-known Huber influence function (Huber (1964)):

$$\psi(u) = u\mathbf{1}_{(|u|\leq c)} + c\,\text{sign}(u)\mathbf{1}_{(|u|>c)}, \qquad (2)$$

where c denotes a tuning constant bounded away from zero that can be used to trade robustness for efficiency in the model fit. In particular, M-quantiles nicely include quantiles when $c \to 0$, $\psi(u) = \text{sign}(u)$, and expectiles when $c \to \infty$, $\psi(u) = u$.

To build the UMQR model, it follows from Firpo *et al.* (2009) and Hampel *et al.* (2011) that the RIF of the M-quantile θ_τ is defined as:

$$RIF(y;\theta_\tau) = \theta_\tau + IF(y;\theta_\tau) = \theta_\tau + \frac{\psi_\tau(y-\theta_\tau)}{\int \psi'_\tau(y-\theta_\tau)dF_Y(y)}, \qquad (3)$$

where $IF(y;\theta_\tau)$ is the IF of θ_τ and $\psi'(u) = \mathbf{1}_{(|u|<c)}$ is the derivative of ψ in (2). In a regression framework when covariates $\mathbf{X} \subset \mathbb{R}^k$ are available, from (3) we define the UMQR model as follows:

$$\mathbb{E}[RIF(Y;\theta_\tau) \mid \mathbf{X}=\mathbf{x}] = \theta_\tau + \mathbb{E}\Big[\frac{\psi_\tau(y-\theta_\tau)}{\int \psi'_\tau(y-\theta_\tau)dF_Y(y)}\Big|\mathbf{X}=\mathbf{x}\Big]. \tag{4}$$

Our objective is to identify how small changes in the distribution of $\mathbf{X}$ affect the M-quantile of the unconditional distribution of Y. From (4) and Firpo *et al.* (2009), the unconditional effect of the τ-th M-quantile, that we denote Unconditional M-quantile Partial Effect, α_τ, is formally defined as:

$$\alpha_\tau = \int \frac{d\mathbb{E}[RIF(Y;\theta_\tau) \mid \mathbf{X}=\mathbf{x}]}{d\mathbf{x}} dF_{\mathbf{X}}(\mathbf{x}) = \frac{1}{s_\tau}\int \frac{d\mathbb{E}[\psi_\tau(Y-\theta_\tau) \mid \mathbf{X}=\mathbf{x}]}{d\mathbf{x}} dF_{\mathbf{X}}(\mathbf{x}), \tag{5}$$

where $F_{\mathbf{X}}$ is the distribution function of $\mathbf{X}$ and $s_\tau = \int \psi'_\tau(y-\theta_\tau)dF_Y(y)$. As suggested by Firpo *et al.* (2009), we can estimate α_τ in (5) via a mean regression of the $RIF(Y;\theta_\tau)$ as dependent variable onto $\mathbf{X}$ by using a two-step procedure. Specifically, an estimate $\widehat{\theta}_\tau$ of θ_τ is obtained by solving (1) via Iterative Reweighted Least Squares, substitute $\widehat{\theta}_\tau$ in (3) and then regress the $RIF(Y;\widehat{\theta}_\tau)$ on $\mathbf{X}$.

3 Application

We investigate the effect of economic and socio-demographic characteristics on italian households' log-consumption using data from the SHIW 2016. We fit the UMQR at different points of the unconditional distribution of the response and compare the results with standard conditional M-quantile regressions. The tuning constant c in (2) has been set to 1.345 and 100. In the second case, we obtain the Unconditional Expectile Regression (UER). The results in Table 1 highlight that the impact of income, gender, age and education is very different on the conditional and unconditional distributions of consumption, especially in the tails. This demonstrates the ability of the UMQR to extend mean regression for estimating the effect of covariates, not only at the center, but also at different parts of the unconditional distribution of interest.

References

BORAH, BIJAN J, & BASU, ANIRBAN. 2013. Highlighting differences between conditional and unconditional quantile regression approaches through an

Variable	MQR			UMQR			ER			UER		
τ	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9	0.1	0.5	0.9
Log-Income	**0.570**	**0.595**	**0.442**	**0.447**	**0.391**	**0.429**	**0.483**	**0.413**	**0.263**	**0.450**	**0.413**	**0.436**
	(0.011)	(0.007)	(0.010)	(0.038)	(0.032)	(0.038)	(0.011)	(0.008)	(0.011)	(0.038)	(0.033)	(0.038)
Gender	−0.019	−0.011	**−0.043**	−0.011	**−0.024**	**−0.038**	−0.023	**−0.026**	**−0.046**	−0.010	**−0.026**	−0.035
	(0.016)	(0.009)	(0.014)	(0.018)	(0.012)	(0.018)	(0.016)	(0.011)	(0.016)	(0.017)	(0.012)	(0.018)
Age	−0.002	0.001	0.004	**−0.013**	**0.006**	**0.013**	−0.001	0.004	**0.008**	**−0.011**	0.004	**0.011**
	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)
Marital status												
never married	**−0.062**	**−0.084**	**−0.164**	**−0.094**	**−0.141**	**−0.187**	**−0.095**	**−0.138**	**−0.201**	**−0.101**	**−0.138**	**−0.176**
	(0.020)	(0.012)	(0.018)	(0.025)	(0.017)	(0.022)	(0.020)	(0.014)	(0.020)	(0.024)	(0.017)	(0.022)
separated	**−0.066**	**−0.056**	**−0.127**	**−0.102**	**−0.151**	**−0.155**	**−0.111**	**−0.137**	**−0.207**	**−0.105**	**−0.137**	**−0.141**
	(0.025)	(0.015)	(0.022)	(0.034)	(0.024)	(0.030)	(0.025)	(0.017)	(0.026)	(0.033)	(0.024)	(0.030)
widowed	−0.040	**−0.063**	**−0.119**	**−0.116**	**−0.136**	**−0.111**	**−0.074**	**−0.123**	**−0.193**	**−0.110**	**−0.123**	**−0.107**
	(0.022)	(0.013)	(0.020)	(0.029)	(0.019)	(0.025)	(0.022)	(0.015)	(0.022)	(0.028)	(0.019)	(0.025)
Education level												
elementary school	**0.175**	**0.120**	**0.151**	**0.488**	**0.125**	−0.037	**0.188**	**0.161**	**0.187**	**0.446**	**0.161**	−0.000
	(0.039)	(0.023)	(0.035)	(0.069)	(0.024)	(0.022)	(0.039)	(0.027)	(0.040)	(0.066)	(0.027)	(0.022)
middle school	**0.240**	**0.203**	**0.316**	**0.645**	**0.269**	**0.060**	**0.281**	**0.294**	**0.398**	**0.590**	**0.294**	**0.094**
	(0.041)	(0.024)	(0.037)	(0.070)	(0.028)	(0.029)	(0.041)	(0.028)	(0.042)	(0.067)	(0.030)	(0.028)
high school	**0.248**	**0.235**	**0.383**	**0.652**	**0.355**	**0.147**	**0.313**	**0.363**	**0.500**	**0.598**	**0.363**	**0.168**
	(0.042)	(0.025)	(0.038)	(0.072)	(0.033)	(0.037)	(0.042)	(0.029)	(0.043)	(0.069)	(0.034)	(0.036)
university	**0.298**	**0.297**	**0.521**	**0.631**	**0.440**	**0.506**	**0.391**	**0.484**	**0.705**	**0.608**	**0.484**	**0.515**
	(0.045)	(0.027)	(0.040)	(0.076)	(0.040)	(0.053)	(0.045)	(0.031)	(0.046)	(0.073)	(0.042)	(0.052)
Employment status												
self-employed	**−0.087**	0.010	**0.083**	**−0.060**	0.021	**0.121**	**−0.058**	0.023	**0.081**	**−0.046**	0.023	**0.107**
	(0.024)	(0.014)	(0.022)	(0.021)	(0.019)	(0.038)	(0.024)	(0.017)	(0.025)	(0.020)	(0.018)	(0.037)
not-employed	0.008	**0.027**	0.035	−0.046	**0.037**	0.037	−0.002	0.014	0.017	**−0.052**	0.014	0.031
	(0.021)	(0.013)	(0.019)	(0.025)	(0.016)	(0.025)	(0.021)	(0.015)	(0.022)	(0.024)	(0.015)	(0.024)

Table 1. *M-quantile and Expectile regression results at* $\tau = (0.1, 0.5, 0.9)$. *Parameter estimates are displayed in boldface when significant at the 5% level.*

application to assess medication adherence. *Health Economics*, **22**(9), 1052–1070.

BRECKLING, JENS, & CHAMBERS, RAY. 1988. M-quantiles. *Biometrika*, **75**(4), 761–771.

FIRPO, SERGIO, FORTIN, NICOLE M, & LEMIEUX, THOMAS. 2009. Unconditional quantile regressions. *Econometrica: Journal of the Econometric Society*, **77**(3), 953–973.

HAMPEL, FRANK R, RONCHETTI, ELVEZIO M, ROUSSEEUW, PETER J, & STAHEL, WERNER A. 2011. *Robust statistics: the approach based on influence functions*. Vol. 196. John Wiley & Sons.

HUBER, PETER J. 1964. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, **35**(1), 73–101.

KOENKER, ROGER, & BASSETT JR, GILBERT. 1978. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

NEWEY, WHITNEY K, & POWELL, JAMES L. 1987. Asymmetric least squares estimation and testing. *Econometrica*, 819–847.

MCMC COMPUTATIONS FOR BAYESIAN MIXTURE MODELS USING REPULSIVE POINT PROCESSES

Jesper Møller [1], Mario Beraha[2], Raffaele Argiento[3] and Alessandra Guglielmi[2]

[1] University of Aalborg, Department of Mathematics, Aalborg (Denmark), (e-mail: jm@math.aau.dk)

[2] Department of Mathematics, Politecnico di Milano, Milano (Italy)

[3] Università Cattolica del Sacro Cuore, Department of Statistical Sciences, Milano (Italy)

ABSTRACT: Repulsive mixture models have recently gained popularity for Bayesian cluster detection. Compared to more traditional mixture models, repulsive mixture models produce a smaller number of well separated clusters. The most commonly used methods for posterior inference either require to fix a priori the number of components or are based on reversible jump MCMC computation. We present a general framework for mixture models, when the prior of the 'cluster centres' is a finite repulsive point process depending on a hyperparameter, specified by a density which may depend on an intractable normalizing constant. By investigating the posterior characterization of this class of mixture models, we derive a MCMC algorithm which avoids the well-known difficulties associated to reversible jump MCMC computation. In particular, we use an ancillary variable method, which eliminates the problem of having intractable normalizing constants in the Hastings ratio. The ancillary variable method relies on a perfect simulation algorithm, and we demonstrate this is fast because the number of components is typically small. In several simulation studies and an application on sociological data, we illustrate the advantage of our new methodology over existing methods, and we compare the use of a determinantal or a repulsive Gibbs point process prior model.

KEYWORDS: birth-death Metropolis Hastings algorithm, cluster estimation, pairwise interaction point process, intractable normalizing constant, normalized infinitely divisible distribution, perfect simulation.

Infinite Mixtures of Infinite Factor Analysers

Keefe Murphy [1], Cinzia Viroli[2] and I. Claire Gormley[3]

[1] Department of Mathematics and Statistics, Maynooth University (e-mail: keefe.murphy@mu.ie)

[2] Department of Statistical Sciences, University of Bologna (e-mail: cinzia.viroli@unibo.it)

[3] School of Mathematics and Statistics, University College Dublin (e-mail: claire.gormley@ucd.ie)

Abstract: Factor-analytic Gaussian mixtures are often employed as a model-based approach to clustering high-dimensional data. Typically, the numbers of clusters and latent factors must be fixed in advance of model fitting. The pair which optimises some model selection criterion is then chosen. For computational reasons, having the number of factors differ across clusters is rarely considered.

Here the infinite mixture of infinite factor analysers (IMIFA) model is introduced. IMIFA employs a Pitman-Yor process prior to facilitate automatic inference of the number of clusters using the stick-breaking construction and a slice sampler. Automatic inference of the cluster-specific numbers of factors is achieved using multiplicative gamma process shrinkage priors and an adaptive Gibbs sampler. IMIFA is presented as the flagship of a family of factor-analytic mixtures.

Applications to benchmark data, metabolomic spectral data, and a handwritten digit example illustrate the IMIFA model's advantageous features. These include obviating the need for model selection criteria, reducing the computational burden associated with the search of the model space, improving clustering performance by allowing cluster-specific numbers of factors, and uncertainty quantification.

Keywords: model-based clustering, factor analysis, Pitman-Yor process, multiplicative gamma process, adaptive Markov chain Monte Carlo.

References

Murphy, K., Viroli, C., & Gormley, I. C. 2020. Infinite mixtures of infinite factor analysers. *Bayesian Analysis*, **15**(3), 937–963.

ANGULAR HALFSPACE DEPTH: COMPUTATION*

Stanislav Nagy[1], Petra Laketa[1] and Rainer Dyckerhoff[2]

[1] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
(e-mail: nagy@karlin.mff.cuni.cz, laketa@karlin.mff.cuni.cz)

[2] Institute of Econometrics and Statistics, University of Cologne, Köln, Germany
(e-mail: rainer.dyckerhoff@statistik.uni-koeln.de)

ABSTRACT: The angular halfspace depth is a nonparametric tool for the analysis of directional data. That depth was proposed already in 1987, but its widespread use has been hampered in practice by significant computational issues. We address these problems by considering a simple projection scheme that allows reducing the computation of the angular depth to the task of evaluating a variant of the usual halfspace depth in a linear space. Efficient algorithms for exact computation and approximation of the angular halfspace depth are thus developed.

KEYWORDS: angular depth, computation, directional data analysis, projection.

1 Angular halfspace depth

Nonparametric analysis of data living in non-linear spaces is an exciting and largely unexplored field of statistics. *Statistical depths* generalize quantiles, ranks, and orderings to multivariate and non-Euclidean data, by evaluating "centrality", or the depth, of points with respect to a probability measure.

We consider directional data (Ley & Verdebout, 2017), that is, observations naturally residing the unit sphere $\mathbb{S}^{d-1} = \left\{x \in \mathbb{R}^d \colon \|x\| = 1\right\}$ of the Euclidean space $\mathbb{R}^d$. For directional data, the *angular halfspace depth* was first introduced by Small, 1987, and later substantially elaborated on by Liu & Singh, 1992. Just as many other depths, the angular halfspace depth is, however, difficult to compute, and no efficient algorithms for its computation are available in dimensions $d > 2$. We use the gnomonic projection of $\mathbb{S}^{d-1}$ to reduce this problem to the computation of the usual halfspace depth in linear spaces $\mathbb{R}^{d-1}$, with respect to signed measures. This connection opens new possibilities for construction of efficient computational tools for directional data.

*This work was supported by the grant 19-16097Y of the Czech Science Foundation, and by the PRIMUS/17/SCI/3 project of Charles University. P. Laketa was supported by the OP RDE project "International mobility of research, technical and administrative staff at the Charles University" CZ.02.2.69/0.0/0.0/18_053/0016976.

2 The depth on spheres and gnomonic projection

The angular halfspace depth is a mapping that to each point on a sphere assigns the smallest probability of a hemisphere that contains that point. More precisely, denote by $\mathcal{H}_0$ the collection of all closed halfspaces in $\mathbb{R}^d$ whose boundary passes through the origin in $\mathbb{R}^d$. For a Borel probability measure P on $\mathbb{S}^{d-1}$ the angular halfspace depth of $x \in \mathbb{R}^d$ with respect to P is defined as

$$ahD(x;P) = \inf\left\{P(H) : H \in \mathcal{H}_0 \text{ and } x \in H\right\}. \tag{1}$$

In this short paper we assume for simplicity that P is absolutely continuous with respect to the spherical Lebesgue measure.[*] For $e_d = (0,\dots,0,1)$ we denote by

$$\mathbb{S}_+^{d-1} = \left\{x \in \mathbb{S}^{d-1} : \langle x, e_d\rangle > 0\right\}, \quad \mathbb{S}_-^{d-1} = \left\{x \in \mathbb{S}^{d-1} : \langle x, e_d\rangle < 0\right\},$$

the northern and the southern hemisphere of $\mathbb{S}^{d-1}$, respectively. We write

$$G = \left\{x \in \mathbb{R}^d : \langle x, e_d\rangle = 1\right\}$$

for the “horizontal” hyperplane that touches $\mathbb{S}^{d-1}$ at e_d.

We consider the *gnomonic projection* of $\mathbb{S}^{d-1}$ to G, that is a mapping that to each $x \in \mathbb{S}_+^{d-1}$ assigns a point $\pi(x) = x/\langle x, e_d\rangle$ from the hyperplane G. For $x \in \mathbb{S}_-^{d-1}$ we define $\pi(x) = \pi(-x)$; the mapping remains undefined if $\langle x, e_d\rangle = 0$. In the left panel of Figure 1 we present π in the plane $\mathbb{R}^2$ — two points, one from the northern (n) and one from the southern (s) halfcircle of $\mathbb{S}^1$, together with their gnomonic images are shown. A closed halfplane $H \in \mathcal{H}_0$ contains both n and s. The intersection $H \cap G$ is a closed halfline in G displayed as a thick line. One observes that $\pi(n) \in G \cap H$, while $\pi(s) \notin G \cap H$. A similar illustration with the sphere $\mathbb{S}^2$, the plane G and a halfspace from $\mathcal{H}_0$ in $\mathbb{R}^3$ is visualised in the right panel of Figure 1.

The gnomonic projection satisfies an important property — for any $H \in \mathcal{H}_0$ it holds true that

$$\pi\left(H \cap \mathbb{S}_+^{d-1}\right) = H \cap G, \quad \pi\left(H \cap \mathbb{S}_-^{d-1}\right) = G \setminus \operatorname{int}(H), \tag{2}$$

where $\operatorname{int}(H)$ is the interior of H. We define a signed measure $P_\pm$ on G by

$$P_\pm(H \cap G) = P\left(H \cap \mathbb{S}_+^{d-1}\right) - P\left(\mathbb{S}_-^{d-1} \setminus H\right). \tag{3}$$

[*] This assumption is not made without loss of generality; the general theory is technical and much more delicate. It will be presented elsewhere.

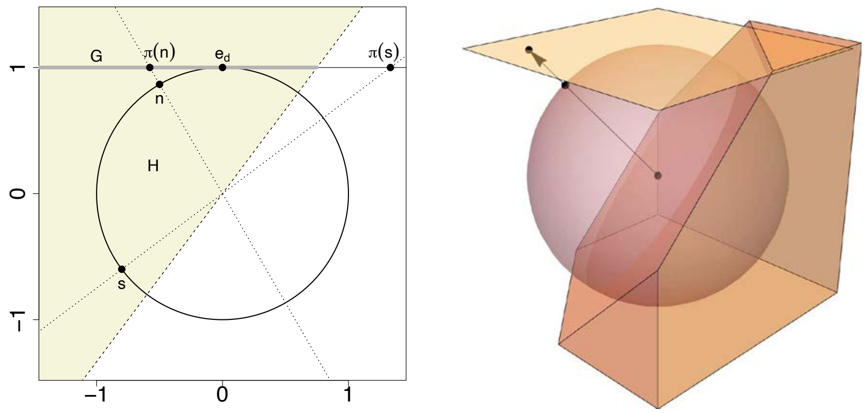


Figure 1. *Left: A halfplane H (coloured) that contains two points — n from the northern and s from the southern halfcircle. We see that $\pi(n) \in H \cap G$, while $\pi(s) \notin H \cap G$. Right: Analogous illustration for $d = 3$, with a halfspace from $\mathcal{H}_0$ and the plane G.*

The Cramér-Wold theorem asserts that $P_\pm$ is well defined. Due to the assumption of P being absolutely continuous, (2) and (3) imply that

$$P(H) = P\left(H \cap \mathbb{S}_+^{d-1}\right) + P\left(H \cap \mathbb{S}_-^{d-1}\right) = P\left(\mathbb{S}_-^{d-1}\right) + P_\pm(H \cap G). \qquad (4)$$

Equation (4) relates the probability of a halfspace $H \in \mathcal{H}_0$ with the value of the signed measure of its projection $H \cap G$ in G. Note that $H \cap G$ is a closed halfspace in space G, unless H is orthogonal to e_d. We denote by $\mathcal{H}$ the collection of all closed halfspaces in G. From (4) it is straightforward to see that

$$ahD(x;P) = P\left(\mathbb{S}_-^{d-1}\right) + \inf\left\{P_\pm(H) : H \in \mathcal{H} \text{ and } x \in H\right\}, \qquad (5)$$

for any $x \in \mathbb{S}_+^{d-1}$. This formula draws connections of the angular halfspace depth with the usual halfspace depth in linear spaces $hD(x;Q)$, defined for a point $x \in \mathbb{R}^{d-1}$ with respect to a given probability measure Q in $\mathbb{R}^{d-1}$ as the infimum of $Q(H)$ over all closed halfspaces in $\mathbb{R}^{d-1}$ that contain x.

The last term in (5) may be considered as the usual halfspace depth of a *signed measure* $P_\pm$ in $\mathbb{R}^{d-1}$. This connection is at the core of our approach. It opens ways of utilizing the highly developed algorithms for computing the usual halfspace depth, and applying it to the analysis of directional data. The main difference is that, for a probability measure Q, one may reduce attention only to those halfspaces that contain x on their boundary when computing

$hD(x;Q)$, which simplifies the computation substantially. The same is, however, not the case with signed measures, where all halfspaces that contain x must be considered. At a slight increase in the computational complexity, it is however possible to adopt the existing algorithms to resolve this issue.

3 Computation: An example

The depth $hD(x;P)$ can be written as an infimum of one-dimensional halfspace depths $hD(\langle x,u\rangle;P_u)$ of $x \in \mathbb{R}^d$ with respect to the projections P_u of P onto the lines given by all directions $u \in \mathbb{S}^{d-1}$. A standard approximation of hD then consists of computing the minimum over $hD(\langle x,u\rangle;P_u)$ for a collection $U \subset \mathbb{S}^{d-1}$ of randomly chosen directions $u \in U$. A halfspace depth of a signed measure $P_{\pm}$ has the same projection property, which may be used to compute (5). This rather naive approximate algorithm is extremely simple, but allows us to consider *ahD* also in dimensions $d > 3$.

For $d = 3$ we adopted a more sophisticated exact algorithm of Dyckerhoff & Mozharovskyi, 2016, generalized it from *hD* to *ahD*, and implemented the results in C++. As a benchmark, we use the implementation of *ahD* for $d = 3$ available as function `sdepth` from the R package `depth`. Detailed results of our comparison are omitted from the present note due to the space restrictions, but will be discussed during the conference talk. Here we only remark that compared to the currently available programs, our new algorithms compute *ahD* up to 10 000-times faster for standard datasets, deal with the exact depth for tens of thousands of observations in $\mathbb{S}^2$ within seconds, and approximate algorithms allow fast evaluation of *ahD* also for $d > 3$. All this illustrates the great potential of our projection method in the analysis of directional data.

References

DYCKERHOFF, RAINER, & MOZHAROVSKYI, PAVLO. 2016. Exact computation of the halfspace depth. *Comput. Statist. Data Anal.*, **98**, 19–30.

LEY, CHRISTOPHE, & VERDEBOUT, THOMAS. 2017. *Modern directional statistics*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Boca Raton, FL.

LIU, REGINA Y., & SINGH, KESAR. 1992. Ordering directional data: concepts of data depth on circles and spheres. *Ann. Statist.*, **20**(3), 1468–1484.

SMALL, CHRISTOPHER G. 1987. Measures of centrality for multivariate and directional distributions. *Canad. J. Statist.*, **15**(1), 31–39.

NONLINEAR INTERCONNECTEDNESS OF CRUDE OIL AND FINANCIAL MARKETS

Yarema Okhrin[1], Gazi Salah Uddin[2] and Muhammad Yahya[3]

[1] Department of Statistics, University of Ausgburg, Germany.
(e-mail: yarema.okhrin@uni-a.de)

[2] Department of Management and Engineering, Linköping University, Sweden.
(e-mail: gazi.salah.uddin@liu.se)

[3] Department of Safety, Economics and Planning, University of Stavanger, Norway.
(e-mail: muhammad.yahya@uis.no)

ABSTRACT: This paper investigates the heterogeneous and asymmetrical effect of COVID-19 on the crude oil, S&P 500 index, EUR/USD exchange rate, and various uncertainty measures. These assets reflect the overall health of the global financial and economic system. For instance, S&P 500 is the most liquid financial index and partly reflects the development of global financial system. Crude oil plays a fundamental role in the developmental and economic activities of a country. Elevated prices of energy commodities lead to a higher inflation and production cost, resulting in declined demand, output, and trade in the economy. The COVID-19 pandemic has contributed significantly to demand and supply shocks that has led to an unprecedented decline in crude oil price. In addition, global geopolitics is triggering the volatility of the crude oil market. The stability of the crude oil market is not only important for oil exporting countries but also oil importing and industrialized countries in order to maintain the price stability of goods. EUR/USD exchange rate is among the most liquid assets. This study adds to the literature by examining the heterogeneous and asymmetric impact of COVID-19 on these different asset classes. This would enable us to understand how different asset classes react to such unique shocks.

The contribution of this paper is fourfold. First, we evaluated the impact of the COVID-19 crisis on the interconnectedness of the financials, forex, and commodity markets with a specific focus on risk dynamics. Second, in contrast to the previous studies we consider high-frequency intraday data. This allows us to provide a deeper insight into the dependencies at a daily level. Third, we quantify the dependence and its dynamics using paired vine copulas. This class of copulas is highly flexible and can allows for a convenient visualization of the dependence. Forth, we put a particular focus on the crude oil returns as a function of several financial covariates using C- and D-vine regressions. This approach allows us to model the whole conditional distribution within a single day and to get insights the causal dependence in tails or at particular quantiles.

Detection of Internet Attacks with Histogram Principal Component Analysis

M. Rosário Oliveira [1], Ana Subtil[1] and Lina Oliveira[2]

[1] CEMAT and Mathematics Department, Instituto Superior Técnico, Universidade de Lisboa, Portugal (e-mail: rosario.oliveira@tecnico.ulisboa.pt, anasubtil@tecnico.ulisboa.pt)

[2] CAMGSD and Mathematics Department, Instituto Superior Técnico, Universidade de Lisboa, Portugal (e-mail: lina.oliveira@tecnico.ulisboa.pt)

Abstract: We propose symbolic unsupervised anomaly detection methods to identify Internet traffic redirection attacks based on histogram principal component analysis. We obtain histogram-valued scores, by applying Moore's based histogram algebraic structure. The symbolic means of the scores are the input for two unsupervised anomaly detection methods used to successfully signal Internet attacks.

Keywords: histogram principal component analysis, symbolic data analysis, Internet data.

1 Introduction

Internet security is a major concern for users and Internet Service Providers since successful attacks can produce substantial damage. These attacks may be aimed at gaining access to sensitive information from the victim, monitoring its online activity, causing network delay, among other motivations.

To identify traffic redirection attacks, we had access to measurements obtained from a worldwide probing platform, designed to detect routing variations based on round-trip-time (RTT)* measurements from multiple and disperse geographic locations (Salvador & Nogueira, 2014). At each timestamp, various measurements are made that are summarized by a histogram. Thus, we propose an anomaly detection method based on histogram principal component analysis. To do so, we consider linear combinations of histogram-valued data (according to a histogram algebraic structure, generalised from the Moore's interval algebraic structure, vide Moore *et al.*, 2009) and use the

*Round-trip-time (RTT) is the length of time since a data packet is sent until an acknowledgement of the packet is received back at the origin.

projected data on the first histogram principal component (PC) to successfully detect traffic redirection attacks.

2 Histogram Principal Component Analysis

Principal component analysis (PCA) is frequently used as a dimensionality reduction method. Given the importance of PCA, various generalisations have been proposed in the symbolic data analysis (SDA) framework. A common generalisation relies on the so-called symbolic-conventional-symbolic approach, where a symbolic covariance matrix is estimated, to which conventional PCA is applied, followed by rewriting the original symbolic data into the space spanned by the first eigenvectors of the covariance matrix. In the case of histogram PCA, Makosso-Kallyth & Diday, 2012 and Chen *et al.*, 2015 use the same definition of sample symbolic covariance, but differ on the way objects are represented in the reduced space. In Le-Rademacher & Billard, 2017, another definition of sample symbolic covariance matrix is used, and the original objects are represented in the reduced space relying on a geometric construction of polytopes. Other approaches to generalise PCA to histogram-valued data exist, but are not considered here.

To detect traffic redirection attacks, we project the original histogram-valued data in the direction of the first PC, whose loadings are determined by the first eigenvector of the chosen symbolic covariance matrix. In our case, we use the same covariance matrix definition as used in Makosso-Kallyth & Diday, 2012 and Chen *et al.*, 2015. The loadings define a weighted sum of the original observations leading to histogram-valued scores, by applying Moore's based histogram algebraic structure. The obtained symbolic means (vide Le-Rademacher & Billard, 2017, eq. (2)) of the scores are the input for the (conventional) unsupervised anomaly detection methods.

3 Detection of Traffic Redirection Attacks

The data set under analysis was gathered on a monitoring network that comprised 12 geographically dispersed servers (probes) that measured, at 120-second intervals, the RTT to two hosts under surveillance (targets: Frankfurt1 and Hong Kong). When an attack was being perpetrated, traffic from 12 probes to the target was diverted through an attacker (relay).

Each probe made 10 RTT measurements every 120 seconds, by sending 10 packets to the target, and the corresponding average, minimum, median, and maximum over the 10 RTT measurements were obtained. These summary

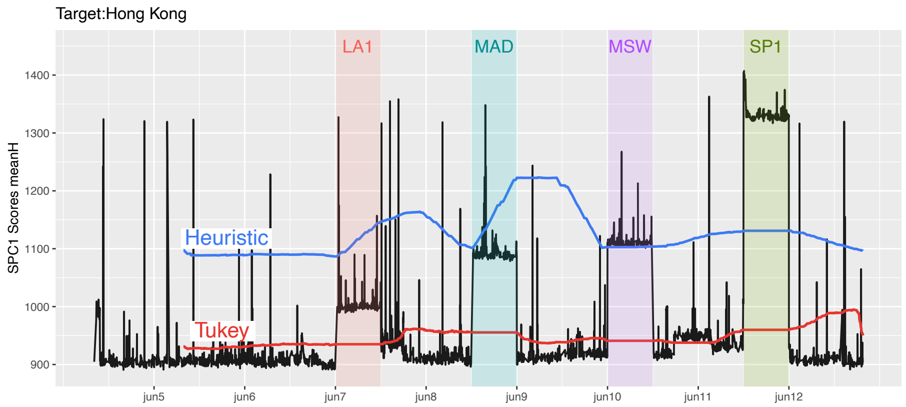


Figure 1. *Means of the first PC histogram-scores for target Hong Kong (black line) and thresholds for the heuristic (blue line) and Tukey's method (red line). Shaded background bands signal the attack periods, with the corresponding relays indicated.*

indicators can be taken as the features to be analysed, and conventional statistical methods may be applied (Salvador & Nogueira, 2014, Subtil *et al.*, 2018). Alternatively, SDA provides a framework to address this problem taking into account the intrinsic variability of the data. As such, we can consider, at every timestamp, for each probe monitoring a target, a histogram with two subintervals, whose bounds are the minimum, median, and maximum of the 10 RTT measurements. Therefore, for each target, we have a symbolic data set with $p = 12$ histogram-valued variables, with as many realisations as timestamps where measurements were made.

We apply the described histogram PCA to each target data set. Given the first PC scores, we calculate their symbolic means and use them as input for two anomaly detection methods: the heuristic proposed by Salvador & Nogueira, 2014 and Tukey's method for outlier detection.

Salvador & Nogueira, 2014 proposed a heuristic to discriminate between the RTTs of regular and redirected traffic. At every timestamp, the conventional average RTT is compared with a decision threshold set at 1.2 times the average of the past 480 observations that were not classified as attacks. Additionally, the heuristic requires a minimum sequence of 10 observations exceeding the threshold to signal attacks (rule-of-10). We apply this heuristic, replacing the average RTT by the means of the first PC histogram-scores. Tukey's method defines boundaries based on the quartiles of the data and identifies as outliers the observations that lie outside these boundaries. Since the first PC is an overall mean of the traffic volume going through the probes, we merely compare their absolute values with the upper boundary $Q3+3\times IQR$, where $IQR=Q3-Q1$ is the interquartile range, $Q1$ and $Q3$ are, respectively, the 1st and 3rd quartiles of the data. We also adopt the rule-of-10.

For the target Frankfurt, both the heuristic and Tukey's method detect all the attacks and no false positive results occur (recall=1, false positive rate=0, precision=1). For Hong Kong, the heuristic is unable to detect attacks perpetrated by the relays Los Angeles (LA1) and Madrid (MAD), as shown in Figure 1. The failure to detect two of the four attacks leads to recall=0.5. Moreover, for this target, the false positive rate is 0 and precision is 1. Tukey's method yields a small false positive rate (0.08), a recall of 1, and 0.79 precision.

4 Conclusions

This paper introduces novel symbolic unsupervised anomaly detection methods to identify Internet traffic redirection attacks based on histogram PCA, using histogram means of the first PC. Results point out the superiority of the symbolic Tukey's method over the symbolic heuristic in detecting the attacks. Overall, we show that PC histogram scores can be used as an interesting input for further statistical analysis (conventional or symbolic).

Acknowledgements: Work supported by FCT, Portugal, through projects UIDB/04621/2020, PTDC/EEI-TEL/32454/2017, and UID/MAT/04459/2020.

References

CHEN, M., WANG, H., & QIN, Z. 2015. Principal component analysis for probabilistic symbolic data: a more generic and accurate algorithm. *Adv. Data Anal. Classif.*, **9**, 59–79.

LE-RADEMACHER, J., & BILLARD, L. 2017. Principal component analysis for histogram-valued data. *Adv. Data Anal. Classif.*, **11**, 327–351.

MAKOSSO-KALLYTH, S., & DIDAY, E. 2012. Adaptation of interval PCA to symbolic histogram variables. *Adv. Data Anal. Classif.*, **6**, 147–159.

MOORE, R. E., KEARFOTT, R. B., & CLOUD, M. J. 2009. *Introduction to interval analysis*. SIAM, USA.

SALVADOR, P., & NOGUEIRA, A. 2014. Customer-side detection of Internet-scale traffic redirection. *Pages 1–5 of: 2014 16th Int. Telecom. Network Strategy and Planning Symp. (Networks)*.

SUBTIL, A., OLIVEIRA, M. R., VALADAS, R., PACHECO, A., & SALVADOR, P. 2018. Detecting Internet-Scale Traffic Redirection Attacks Using Latent Class Models. *Pages 370–380 of: Int. Conf. Soft Comp. and Pattern Recognit.* Springer.

SEMIPARAMETRIC IRT MODELS FOR NON-NORMAL LATENT TRAITS

Sally Paganin [1]Department of Biostatistics, Harvard School of Public Health, Harvard University (e-mail: spaganin@hsph.harvard.edu)

ABSTRACT: Item Response Theory models are widely used in many domains of applications to analyze questionnaires data, scaling categorical data into continuous construct. Interpretable inference is often obtained relying on a set of assumptions for the latent constructs, as for example normality for the unknown subject-specific latent traits. This assumption can often be unrealistic and lead to biased results, hence we consider more flexible models using Bayesian nonparametric mixtures for the individual latent traits. We study several identifiability constraints, and compare inferential results and different Markov chain Monte Carlo strategies for posterior sampling.

KEYWORDS: 2PL, Bayesian nonparametrics, Dirichlet Process, MCMC, NIMBLE.

1 IRT models for binary responses

Let y_{ij} denote the answer of an individual j to item i for $j = 1,\ldots,N$ and $i = 1,\ldots,I$, with $y_{ij} = 1$, when the answer is correct and 0 otherwise. Typically, different individuals are assumed to work independently, while responses from the same individuals are assumed independent conditional to the latent trait (local independence assumption). Hence each answer y_{ij}, conditionally to the latent parameters, is assumed to be a realization of a Bernoulli distribution, and the probability of a correct response is typically modeled via logistic regression.

2 Semiparametric 2PL models

In the two-parameter logistic (2PL) model, the conditional probability of a correct response is modeled as

$$\Pr(y_{ij} = 1|\lambda_i, \beta_i, \eta_j) = \frac{\exp\{\lambda_i(\eta_j - \beta_i)\}}{1 = \exp\{\lambda_i(\eta_j - \beta_i)\}}, i = 1,\ldots,I, \quad j = 1,\ldots,N. \quad (1)$$

where η_j represents the health status, or more in general latent trait, of the j-th individual, while β_i and λ_i encode item characteristics. The parameter $\lambda_i > 0$ is often referred to as *discrimination*, while β_i is called *difficulty* because for any fixed η_j the probability of a correct response to item i is decreasing in β_i. When $\lambda_i = 1$ for all $i = 1,\ldots,I$, the model in 1 reduces to the one-parameter logistic (1PL) model. Often, conditional log-odds in 1 are reparametrized as $\lambda_i\eta_i + \gamma_i$, with $\gamma_i = -\lambda_i \times \beta_i$. Sometimes this is reffered to as slope-intercept parameterization as opposed to the IRT parameterization in considered traditionally for interpretation.

Traditional literature assumes that $\eta_j \sim \mathcal{N}(0,1)$ for $j = 1,\ldots,N$, but there are situations in which such assumption can be too restrictive. We can extend the model in 1 to describe more flexible latent trait distributions using a Dirichlet Process (DP) mixture of normal distributions

$$\eta_j|G \sim G, \quad G \sim DP(\alpha, G_0),$$
$$G_0 \equiv \mathcal{N}(0,\sigma_0^2) \times \text{InvGamma}(\nu_1,\nu_2) \tag{2}$$

where α is the concentration parameter and G_0 the base measure. Alternative representations of the DP are known as the Chinese Restaurant Process (CRP) Blackwell *et al.*, 1973 or the truncated stick-breaking (SB) Sethuraman, 1994.

3 Model estimation

Estimation of the model parameters is carried out in the Bayesian framework via MCMC methods, using NIMBLE de Valpine *et al.*, 2017, a R software for hierarchical models. The NIMBLE system provides a suite of different sampling algorithms along with the possibility to code user-defined samplers. We compare results from the parametric and semiparametric 2PL model, using NIMBLE's default sampling configuration, that mixes conjugate samplers with adaptive Metropolis Hastings algorithm.

Typically parameters of the 2PL model are not identifiable, so constraints are either included in the model or one can post-process posterior samples to meet the constraints. This last approach is typical of parameter-expanded algorithms, which embed targeted models in a larger specification. We found this last option to be the most efficient in terms on both MCMC mixing and time.

In traditional literature on parametric 2PL model, identification is obtained constraining the discrimination parameters λ_i, for $i = 1,\ldots,I$ to be positive, when the latent trait distribution is assumed to be a standard normal. Since we are relaxing the normal assumption on the latent traits, we considered sum-to-zero constraints on the item parameters, i.e. $\sum_i \beta_i = 0$, $\sum_i \log(\lambda_i) = 0$.

4 Inferential results

We compare inferential results via simulation. We simulate data from two different scenarios changing the distribution generating the latent traits. We simulate responses from $N = 3,000$ individuals to $I = 20$ binary items. Values for the discrimination parameters $\{\lambda_i\}_{i=1}^{20}$ are sampled from a Uniform distribution over the interval $(0.5, 2)$, while values for difficulty parameters $\{\beta_i\}_{i=1}^{20}$ are sampled from a Normal distribution with mean zero and variance 2.

In particular, we considered two different generating distribution for the latent traits. A unimodal scenario, where η_j are i.i.d. draws from a $\mathcal{N}(0,1)$ and a multimodal scenario where

$$\eta_j \sim 0.4 \times \mathcal{N}(-3,1) + 0.2 \times \mathcal{N}(-2,4) + 0.4 \times \mathcal{N}(2,1). \qquad (3)$$

We chose moderately vague priors for the item parameters, $\beta_i \sim \mathcal{N}(0,3)$ and $\log(\lambda_i) \sim \mathcal{N}(0.5, 0.5)$. In the parametric model, η_js are assumed to follow $\mathcal{N}(0,1)$, while for DP we choose $G_0 \equiv \mathcal{N}(0,3) \times InvGamma(1.01, 2.01)$. We run the MCMC for $50,000$ iterations using a 10% burn-in of 5000 iterations, and check traceplots for convergence.

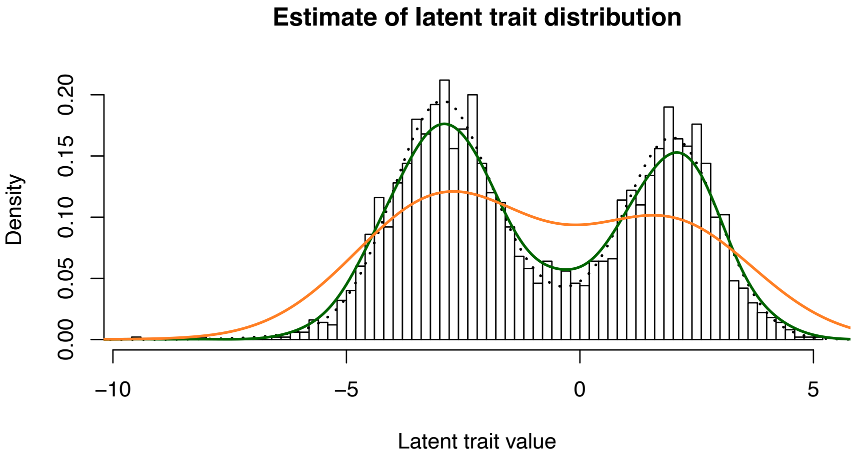


Figure 1. *Comparison of the latent trait density estimates, using a parametric 2PL model (orange line) and a semiparametric 2PL model (green line). The dotted black lines indicate the true distribution in (3).*

Figure 1 compares density estimates of the latent trait distribution from the parametric and semiparametric models, computed taking the posterior means of the η_js. It can be noticed that the parametric model leads to a flat distribution because of the underlying normal assumption, while the semiparametric specification recover the true density structure. Better estimation of the latent abilities helps to avoid bias in inference, for example when estimating item parameters or item characteristics curves (ICC).

References

BLACKWELL, DAVID, MACQUEEN, JAMES B, *et al.* 1973. Ferguson distributions via Pólya urn schemes. *The annals of statistics*, **1**(2), 353–355.

DE VALPINE, PERRY, TUREK, DANIEL, PACIOREK, CHRISTOPHER J., ANDERSON-BERGMAN, CLIFFORD, LANG, DUNCAN TEMPLE, & BODIK, RASTISLAV. 2017. Programming With Models: Writing Statistical Algorithms for General Model Structures With NIMBLE. *Journal of Computational and Graphical Statistics*, **26**(2), 403–413.

SETHURAMAN, J. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**(2), 639–650.

A GRAPHICAL DEPTH-BASED AID TO DETECT DEVIATION FROM UNIMODALITY ON HYPERSPHERES

Giuseppe Pandolfo [1]

[1] Department of Industrial Engineering, University of Naples Federico II, (e-mail: giuseppe.pandolfo@unina.it

ABSTRACT: A graphical tool for investigating unimodality of hyperspherical data is proposed. It is based on the notion of statistical data depth function for directional data. Then "standard" global depth is compared to its local version by means of a two-dimensional scatterplot. The proposal is illustrated on simulated data.

KEYWORDS: Data depth, distance measure, ranks, Von-Mises Fisher.

1 Setting

Testing unimodality of a sample $X_1,\ldots,X_n$ of a random vector X supported on the hypersphere $S^{q-1} := \{x \in \mathbb{R}^q : x'x = 1\}$, with $q > 1$, is one important step in multivariate data analysis for which only the directions (and not the magnitudes) are of interest – the so-called directional data. This kind of data arise in many applied disciplines, such as astronomy, biology, etc.

The inspiration for this contribution comes from the *center-outward ordering* provided by statistical depth functions, which can be intended as a multivariate generalization of standard univariate rank. Specifically, information about unimodality of hyperspherical distributions are obtained through data depths, and they can be displayed and visualized in a simple two-dimensional plot. Such graph is based on an analysis of the rankings derived from a data depth function and its local counterpart, so that they can offer an easy interpretable picture of the distributions.

The use of depth-induced rankings to investigate distributional features has been already used for analyzing data in $\mathbb{R}^q$ by means of graphical tools. Liu *et al.* , 1999 proposed the "sunburst plot" as a bivariate generalization of the box-plot and the DD-(depth versus depth) plots. Rousseeuw *et al.* , 1999 proposed the bagplot, a bivariate generalization of the univariate boxplot by exploiting the notion of halfspace location depth. Li *et al.* , 2012 used the DD-plot to perform classification of data in $\mathbb{R}^q$. A nonparametric classification procedure based on the DD-plot was introduced also by Lange *et al.* , 2014.

The concept of data depth was also used to build control charts for monitoring processes of multivariate quality measurement [11, 3]. However, despite the great and increasing interest for multivariate data analysis in $\mathbb{R}^q$, the adoption of depth-based visualizations for the analysis of directional data has been neglected so far, except for a recent work about the classification of data on the unit circle through the DD-plot (Liu, 1995, Pandolfo *et al.* , 2021).

2 Data depth

Data depth function is an important nonparametric tool for the analysis of complex data such as functional and directional data. It provides a center-outward ordering of the data and leads to a ranking of data which can be exploited for describing different features of the data distribution. Hence, a data depth function is any function $D(x,F)$ that measures the closeness or centrality of a point $x \in S^{q-1}$ d with respect to a distribution function F. Thus, a depth function assigns to each $x \in S^{q-1}$ a nonnegative score as its center-outward depth with respect to F. Observations close to the center of F receive high ranks whereas peripheral observations receive low ranks. Such notion is limited to data modeling with a unimodal distribution. For this reason, *local* versions of depth functions were derived in order to deal with multimodal distributions (see Agostinelli & Romanazzi, 2011 and Paindaveine & Van Bever, 2013).

Here, the notion of distance-based depths for directional data introduced by Pandolfo *et al.* , 2018 is adopted along with its local version which is derived by considering a neighborhood of each point x whose radius is the τ parameter, which cannot goes to ∞, as it occurs for data in $\mathbb{R}^q$, because of the boundedness of the space. The usual definition is recovered when τ approaches its maximum, thus local depth includes ordinary depth as a particular case. Hence local depth is a class of center-outward ranking functions serving multiple purposes, according to the value of the tuning parameter: low values describe centralness of the points of the space conditional on a small neighborhood around them, higher values lead to wider windows and therefore produce rankings which are more and more similar to "standard" global depth.

3 Plotting global and local depth rankings

The rankings produced by the notions of global and local depths can be compared by means of a two-dimensional scatterplot, which can be exploited to investigate unimodality of directional data. This is because for symmetric unimodal distributions the rankings of the data provided by global and local depths should be exactly the same. On the contrary, the more the distribution deviates from unimodality the greater the difference between the two rankings. Such difference can be easily understood by a two-dimensional plot where the *x*-coordinates are the global depth of the corresponding data point and the *y*-coordinates are the local depth of the corresponding data point. If the distribution is unimodal and thus the set of the deeper local points does not substantially differ from the corresponding set of the deeper global depth points, the plot will exhibit a concentration on the upper-right corner. In case of strong unimodality, the ranks of the two depth functions will coincide, and points on the plot will roughly form a straight diagonal line. On the other hand, departure from unimodality will show different scenarios, obviously depending on the kind of departure. Below, Figure 1 reports an example of the proposed tool, where the arc distance depth in its global and local were adopted for a unimodal von Mises-Fisher distribution in 5 dimensions with concentration parameter equal to 20 (a) and a bimodal distribution in 5 dimensions obtained trough a weighted mixture of two von Mises-Fisher distributions with means 90° far away from each other with 80% of the weight on the first component and different concentrations, i.e. 5 and 2 (b). The sample size was set equal to 250. In the first case one can see that points do not deviate too much from the straight line suggesting a strong unimodality. For the bimodal data, points are more scattered around, and the deepest sample points according to the global and local depth functions do not clearly lie on the upper-right quadrant, signaling a departure from unimodality.

References

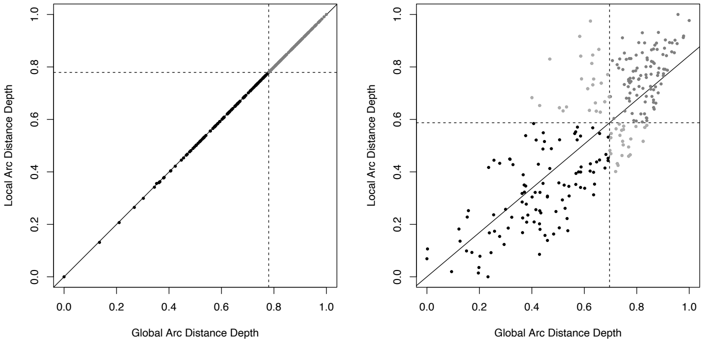
AGOSTINELLI, CLAUDIO, & ROMANAZZI, MARIO. 2011. Local depth. *Journal of Statistical Planning and Inference*, **141**(2), 817–830.

LANGE, TATJANA, MOSLER, KARL, & MOZHAROVSKYI, PAVLO. 2014. Fast nonparametric classification based on data depth. *Statistical Papers*, **55**(1), 49–69.

LI, JUN, CUESTA-ALBERTOS, JUAN A, & LIU, REGINA Y. 2012. DD-classifier: Nonparametric classification procedure based on DD-plot. *Journal of the American statistical association*, **107**(498), 737–753.

LIU, REGINA Y. 1995. Control charts for multivariate processes. *Journal of the American Statistical Association*, **90**(432), 1380–1387.

LIU, REGINA Y, PARELIUS, JESSE M, SINGH, KESAR, *et al.* 1999. Multivariate analysis by data depth: descriptive statistics, graphics and infer-



(a) (b)

Figure 1: Plot of global vs local depth-induced rankings of a von Mises-Fisher distribution in 5 dimensions with concentration parameter $\kappa = 5$ (a), and of a weighted mixture of two von Mises-Fisher distributions with means 90° far away from each other (b). The normalized global and local arc depths were used with τ equal to $\pi/2$.

ence. *The annals of statistics*, **27**(3), 783–858.

PAINDAVEINE, DAVY, & VAN BEVER, GERMAIN. 2013. From depth to local depth: a focus on centrality. *Journal of the American Statistical Association*, **108**(503), 1105–1119.

PANDOLFO, GIUSEPPE, PAINDAVEINE, DAVY, & PORZIO, GIOVANNI C. 2018. Distance-based depths for directional data. *Canadian Journal of Statistics*, **46**(4), 593–609.

PANDOLFO, GIUSEPPE, IORIO, CARMELA, STAIANO, MICHELE, ARIA, MASSIMO, & SICILIANO, ROBERTA. 2021. Multivariate process control charts based on the Lp depth. *Applied Stochastic Models in Business and Industry*, **37**(2), 229–250.

ROUSSEEUW, PETER J, RUTS, IDA, & TUKEY, JOHN W. 1999. The bagplot: a bivariate boxplot. *The American Statistician*, **53**(4), 382–387.

NETWORKS OF NETWORKS

Panos Pardalos[1]

[1] Department of Industrial and Systems Engineering, University of Florida,
(e-mail: pardalos@ufl.edu)

ABSTRACT: Many complex systems in nature (or man made) are represented not by single networks but by sets of interdependent networks. Such networks of networks (NoN) include the internet, airline alliances, biological networks, and smart city networks. There is no doubt that NoN will be the next frontier in network sciences. In my lecture I will address some recent developments (robustness, diversity) and discuss some challenging problems in NoN.

KEYWORDS: interdependent networks; network robustness; network diversity

PAIRWISE LIKELIHOOD ESTIMATION OF LATENT AUTOREGRESSIVE COUNT MODELS

Xanthi Pedeli [1] and Cristiano Varin[2]

[1] Department of Statistics, Athens University of Economics and Business, Athens, Greece (e-mail: xpedeli@aueb.gr)

[2] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University, Venice, Italy (e-mail: cristiano.varin@unive.it)

ABSTRACT: Latent autoregressive models are often employed for the analysis of infectious disease data. However, the likelihood function of latent autoregressive models is intractable and it is usually approximated by simulation-based methods. Through such approximations the inferential problem becomes feasible, but at the price of a high computational cost and difficulties in the assessment of the the quality of the numerical approximation. We consider instead a weighted pairwise likelihood approach and explore several computational and methodological aspects including estimation of robust standard errors and the role of numerical integration. The suggested approach is illustrated on monthly cases of invasive meningococcal disease in Italy.

KEYWORDS: latent autoregressive model, numerical integration, pairwise likelihood.

1 Pairwise Likelihood Inference for Latent Autoregressive Models

Let $y_1,\dots,y_n$ be an observed time series of length n and let $u_t = \phi u_{t-1} + \varepsilon_t$ be an unobserved autoregressive Gaussian model with $\varepsilon_t \sim N(0,\sigma^2)$ and $|\phi| < 1$. Latent autoregressive models assume that conditionally on the unobserved u_t, the observed counts y_t are independent Poisson random variables with conditional expectation $\mathrm{E}(y_t|u_t) = \exp(\boldsymbol{x}_t^T \boldsymbol{\beta} + u_t)$, where $\boldsymbol{x}_t$ is a vector of regressors and $\boldsymbol{\beta} = (\beta_0,\dots,\beta_p)^T$ the corresponding vector of regression coefficients. The inclusion of the latent variable u_t in the linear predictor induces both serial correlation and overdispersion which is frequently observed in time series of disease counts. Likelihood inference for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T,\sigma^2,\phi)^T$ of latent autoregressive models requires to approximate the n-fold integral

$$L(\theta) = \int_{\mathbb{R}^n} p(y_1|u_1;\beta)p(u_1;\sigma^2,\phi)\prod_{t=2}^{n} p(y_t|u_t;\beta)p(u_t|u_{t-1};\sigma^2,\phi)du_1\dots du_n. \quad (1)$$

Alternatively, the likelihood can be expressed as a series of n nested one-dimensional integrals using the filtering algorirthm described in Cagnone &

Bartolucci, 2017. The algorithm is based on recursive evaluation of the nested integrals, a process that can introduce propagation of the numerical error, which is the main drawback of the filtering algorithm approach. Various simulation strategies for approximation of the likelihood (1) have been suggested under the frequentist and Bayesian frameworks (Davis & Dunsmuir, 2016). In this paper, we consider a weighted pairwise likelihood approach (Varin & Vidoni, 2009) which is based on replacement of the high-dimensional integral in (1) with a limited set of double integrals. Consequently, a significant reduction of the computational cost related to ordinary likelihood is achieved.

The pairwise log-likelihood of order d is defined as a weighted sum of the form $\ell_d(\boldsymbol{\theta}) = \sum_{t=m_d+1}^{n} \sum_{i=1}^{m_d} w_i \log p(y_{t-i}, y_t; \boldsymbol{\theta})$, where m_d is a window length parameter, w_i are some non-negative weights, that are normalized, so that $\sum_{i=1}^{m_d} w_i = 1$ (see Section 2 and Pedeli & Varin, 2020 for more details) and

$$p(y_{t-i}, y_t; \boldsymbol{\theta}) = \int_{\mathbb{R}^2} p(y_t|u_t; \boldsymbol{\beta}) p(y_{t-i}|u_{t-i}; \boldsymbol{\beta}) p(u_{t-i}, u_t; \sigma^2, \phi) du_{t-i} du_t.$$

The maximum pairwise likelihood estimator of order d is denoted as $\hat{\boldsymbol{\theta}}_d$ and is the solution of the pairwise score equations $\psi_d(\hat{\boldsymbol{\theta}}_d) = \sum_{t=m_d+1}^{n} \psi_{d,t}(\hat{\boldsymbol{\theta}}_d) = \mathbf{0}$, where $\psi_{d,t}(\boldsymbol{\theta}) = \sum_{i=1}^{m_d} w_i \frac{\partial}{\partial \boldsymbol{\theta}} \log p(y_{t-i}, y_t; \boldsymbol{\theta})$ are the averaged pairwise scores.

It can be shown (Davis & Yau, 2011) that the limiting distribution of $\hat{\boldsymbol{\theta}}_d$ is normal with mean equal to the true value, $\boldsymbol{\theta}_*$, and asymptotic variance equal to the inverse of the Godambe information $G_d(\boldsymbol{\theta}_*) = H_d(\boldsymbol{\theta}_*) J_d(\boldsymbol{\theta}_*)^{-1} H_d(\boldsymbol{\theta}_*)$, where $H_d = \mathrm{E}\left\{-\frac{\partial}{\partial \boldsymbol{\theta}} \psi_{d,t}(\boldsymbol{\theta}_*)\right\}$ and $J_d = \sum_{k=-\infty}^{\infty} \mathrm{E}\left\{\psi_{d,t-k}(\boldsymbol{\theta}_*) \psi_{d,t}(\boldsymbol{\theta}_*)^T\right\}$ are referred as the sensitivity and variability matrices, respectively. For the estimation of H_d one can work with either the observed pairwise likelihood information or an outer-product estimator which derives from the second-order Bartlett identity that holds for each specific pair of observations. Estimation of J_d is more demanding. We consider an heteroskedasticity and autocorrelation consistent (HAC) estimator (Newey & West, 1994) of the form

$$\hat{J}_d = \sum_{k=-r}^{r} \left(1 - \frac{|k|}{r}\right) \sum_{t=m_d+1}^{n} \left\{\frac{1}{n} \psi_{d,t-k}(\hat{\boldsymbol{\theta}}_d) \psi_{d,t}(\hat{\boldsymbol{\theta}}_d)^T\right\},$$

where the weights $(1 - |k|/r)$ correspond to the Bartlett kernel, although other types of kernels might also be used. Empirical evidence suggests that the default lag length considered by autocorrelation functions of standard statistical softwares can serve as a reliable choice for the window semi-length r. We thus adopt the rule $r = \lfloor 10 \log_{10} n \rfloor$ corresponding to the number of lags used in the acf() function of the R software (R Core Team, 2020).

2 Application

This section illustrates the proposed approach with an update of the application considered in Pedeli & Varin, 2020. Data on the monthly number of meningococcal disease cases in Italy for the years 1999-2018 have been obtained from the Surveillance Atlas of the European Center of Disease Control (ECDC). Thereafter, a latent autoregressive Poisson model is fitted to the period 1999-2017 and then it is used to predict the disease cases in 2018. In the time series plot of the data (left panel of Figure 1), it can be observed that the main feature of the series is a level shift corresponding to a reduction of the monthly number of cases after March 2005. We therefore consider the latent autoregressive model $E(y_t|u_t) = \exp(\eta_t + u_t)$ with $\eta_t = \beta_0 + \beta_1 x_t + \beta_2 \sin(2\pi t/12) + \beta_3 \cos(2\pi t/12)$, where x_t is a binary indicator for observations before ($x_t = 1$) and after ($x_t = 0$) March 2005. The Pearson residuals obtained by a standard Poisson regression model with linear predictor η_t are non-spuriously autocorrelated at the first two lags. We thus fit the latent autoregressive model with the pairwise likelihood of order two and trapezoidal weights, as suggested by simulation results discussed in Pedeli & Varin, 2020. The trapezoidal weights have a window length parameter $m_d = 2d$ and are defined as

$$w_i \propto \begin{cases} 1, & 1 \leq i < d, \\ (2d - i)/d, & d \leq i < 2d, \\ 0, & i \geq 2d. \end{cases}$$

Numerical integration for computation of the pairwise likelihood is performed through Gauss-Hermite quadrature with 5, 10 and 20 nodes per dimension giving the same estimates and standard errors up to two decimal digits. Maximum pairwise likelihood estimates are in close agreement with integrated nested Laplace approximation (INLA) (Rue et al., 2009) and confirm the significant level shift in the invasive meningitis cases after March 2015. The maximum pairwise likelihood estimates and corresponding standard errors were obtained after 0.164, 0.379 and 1.439 CPU seconds, with five, 10 and 20 quadrature nodes per dimension, respectively, while INLA required 5.256 CPU seconds.

The observed and predicted cases of meningococcal infections in Italy and the corresponding 95% upper bounds are illustrated in the right panel of Figure 1. Predictions were computed with 10,000 simulations from the fitted model. The comparison of in-sample predictions with the observed disease counts indicates a realistic model fitting and retrospectively identifies some periods of excess disease cases. Out-of-sample predictions are also close to the observed miningitis cases for the year 2018 indicating a good predictive ability.

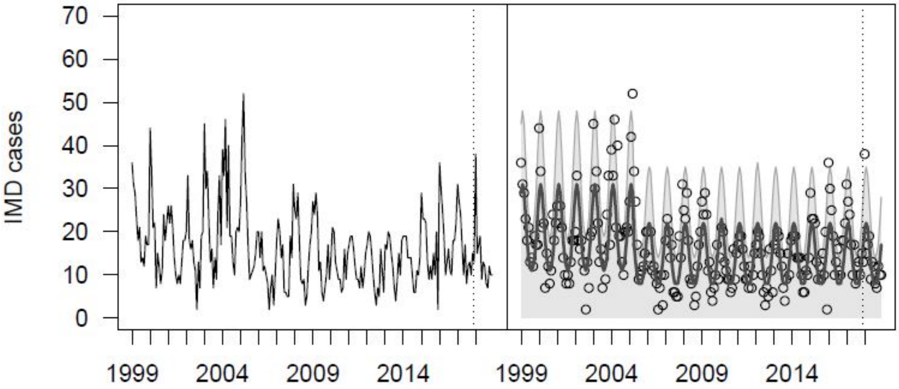


Figure 1. *Left panel: Time series of the monthly number of invasive meningococcal disease (IMD) cases in Italy for the years 1999–2018. Right panel: observed (○) and predicted (–) number of IMD cases. The vertical dotted line separates the data used for model fitting from the data used for the prediction exercise.*

References

CAGNONE, S. & BARTOLUCCI, F. 2017. Adaptive quadrature for maximum likelihood estimation of a class of dynamic latent variable models. *Computational Economics*, **49**, 599–622.

DAVIS, R.A. & DUNSMUIR, W.T.M. 2016. *State Space Models for Count Time Series*. In Handbook of discrete-valued time series. CRC Press.

DAVIS, R.A. & YAU, C.Y. 2011. Comments on pairwise likelihood in time series models. *Statistica Sinica*, **21**, 255–77.

NEWEY, W.K. & WEST, K.D. 1994. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, **61**, 631–653.

PEDELI, X. & VARIN, C. 2020. Pairwise likelihood estimation of latent autoregressive count models. *Statistical Methods in Medical Research*, **29**, 3278–3293.

R CORE TEAM, 2020. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RUE, H., & MARTINO, S., & CHOPIN, N. 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. R. Statist. Soc. B*, **71**, 319–392.

VARIN, C. & VIDONI, P. 2009. Pairwise likelihood inference for general state space models. *Econometric Reviews*, **28**, 170–85.

A Study of Lack-of-Fit Diagnostics for Models Fit to Cross-classified Binary Variables

Mark Reiser[1] and Maduranga Dassanayake[2]

[1] School of Mathematical and Statistical Sciences, Arizona State University, USA,
(e-mail: `mark.reiser@asu.edu`,)

[2] Department of Statistics, University of Georgia, USA,
(e-mail: `maduranga@uga.edu`)

Abstract: In this paper, a new version of the *GFfit* statistic is compared to other lack-of-fit diagnostics for models fit to cross-classified binary variables. The new *GFfit* statistic is obtained by decomposing the Pearson statistic from the full table into orthogonal components defined on marginal distributions. The new version of the *GFfit* statistic can be applied to a variety of models for cross-classified tables. Simulation results show that $GFfit_{\perp}^{(ij)}$ has good Type I error performance even if the joint frequencies are very sparse and has higher power for detecting the source of lack of fit compared to other diagnostics on bivariate marginal tables.

Keywords: Item response model, goodness of fit, orthogonal components

1 Introduction

For a multi-way contingency table, the traditional Pearson's chi-square statistic is obtained by comparing observed frequencies to the expected frequencies under the null hypothesis. For a composite null hypothesis where the null distribution depends on a vector of g unknown parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_g)^T$, requires the Pearson-Fisher statistic, $X_{PF}^2 = \sum_s z_s^2$, where $z_s = \sqrt{n}(\pi_s(\hat{\boldsymbol{\beta}}))^{-\frac{1}{2}}\left(\hat{p}_s - \pi_s(\hat{\boldsymbol{\beta}})\right)$. Fisher (1924) gave the degrees of freedom, $T - g - 1$. Orthogonal components of X_{PF}^2 have been studied by many authors, including Lancaster (1969). Reiser, Cagnone, and Zhu (2021) propose a new *GFfit* statistic for the purpose of detecting lack of fit. The new statistic, $GFfit_{\perp}^{(ij)}$, is obtained by decomposing the Pearson statistic from the full table into orthogonal components defined on marginal distributions. $GFfit_{\perp}^{(ij)}$ are the squared elements of $\hat{\boldsymbol{\gamma}} = n^{\frac{1}{2}}\widehat{\boldsymbol{F}}'\mathbf{e}$, where $\mathbf{e}$ is a vector of residuals on marginals distributions, such as bivariate residuals, $\boldsymbol{F} = (\boldsymbol{C}')^{-1}$, where $\boldsymbol{C}$ is the Cholesky factor of $\boldsymbol{\Omega}_{\mathbf{e}}$, and $\boldsymbol{\Omega}_{\mathbf{e}}$ is the covariance matrix of $\sqrt{n}\mathbf{e}$.

In this paper, the performance of $GFfit_{\perp}^{(ij)}$ is compared to adjusted residuals (Reiser, 1996) and $\bar{\chi}_{ij}^{2}$ (Liu & Maydeu-Olivares, 2014) using simulations to assess Type I error rate and power for models fit to binary cross-classified variables. The adjusted residual k for the second-order marginal is $z_{ij} = n^{\frac{1}{2}}\mathrm{e}^{(k)}/\hat{\sigma}_{\mathrm{e}}^{(k)}$, where $k = 1, 2, \cdots, \binom{q}{2}$ and corresponds to item pair ij, $\mathrm{e}^{(k)}$ is an element of $\mathbf{e}$, and $\hat{\sigma}_{\mathrm{e}}^{(k)}$ is the square root of a diagonal element of $\mathbf{\Sigma_e}$ where $\mathbf{\Sigma_e} = n^{-1}\widehat{\mathbf{\Omega}}_{\mathbf{e}}$. $\bar{\chi}_{ij}^{2} = 2\frac{\hat{\mu}_1}{\hat{\mu}_2}\chi_{ij}^{2}$, where $\hat{\mu}_1$ and $\hat{\mu}_2$ are the first and second asymptotic moments of χ_{ij}^{2}, and χ_{ij}^{2} is the Pearson chi-square statistic calculated on a bivariate table.

2 Type I Error Study

The first simulation included eight manifest variables. One thousand data sets were generated using Monte-Carlo methods related to a one factor model where β_1' = (0.1, 0.1, 0.1, 0.9, 0.9, 0.9, 0.2, 0.2). Three intercept settings were used. Only results for a simulation with intercepts symmetric around zero are shown below. A 2 PL item response model with one latent dimension was estimated for each of these datasets, and empirical Type I error rates of the individual orthogonal components were calculated. Since each individual orthogonal component is distributed approximately as chi-square with one degree of freedom, to calculate the empirical Type I error rate for each component, the sum of the number of cases that exceed the chi-square critical value (at 5% significance level) with one degree of freedom was divided by the number of datasets. Similar process was used to calculate the Type I error rates of the adjusted residual and $\bar{\chi}_{ij}^{2}$. This simulation was repeated for sample sizes 300 and 500.

Table 1 below indicates the empirical Type I error rates for $q = 8$ manifest variables for symmetric intercept model. The Type I error rates outside of the Monte-Carlo error interval $0.05 \pm \sqrt{0.05(0.95)/1000} = (0.0365, 0.0635)$ are bolded. When n=300, Type I error rates related to $GFfit_{\perp}^{(ij)}$ (4,5) and (5,6) were outside the Monte-Carlo error interval. Given that there are twenty eight individual $GFfit_{\perp}^{(ij)}$, it is possible that one or two components may randomly fall slightly outside the Monte-Carlo error interval. However, five $\bar{\chi}_{ij}^{2}$ and four adjusted residuals were outside the Monte-Carlo error interval. This suggests, when n=300, orthogonal components have better Type I error rates compared to $\bar{\chi}_{ij}^{2}$ and adjusted residuals for $q = 8$ manifest variables for symmetric intercept model. When n=500, all the $GFfit_{\perp}^{(ij)}$ and most of the $\bar{\chi}_{ij}^{2}$ and adjusted residuals were inside the Monte-Carlo error interval $(0.0365, 0.0635)$.

Table 1. *Type I Error Study for Symmetric Intercept Model*

	n=300			n=500		
Pair (i,j)	GFfit⊥	Std. Residuals	$\bar{\chi}^2_{ij}$	Gffit⊥	Std. Residuals	$\bar{\chi}^2_{ij}$
(1,2)	0.046	0.055	0.052	0.052	0.0590591	0.056
(1,3)	0.048	0.048	0.047	0.044	0.046046	0.046
(1,4)	0.044	0.057	0.054	0.051	0.0510511	0.047
(1,5)	0.044	**0.034**	**0.03**	0.042	0.043043	0.043
(1,6)	0.049	0.048	0.044	0.053	0.049049	0.047
(1,7)	0.051	0.063	0.06	0.043	0.045045	0.045
(1,8)	0.057	0.053	0.052	0.051	0.0530531	0.053
(2,3)	0.051	0.055	0.054	0.041	0.042042	0.043
(2,4)	0.039	0.046	0.049	0.038	0.047047	0.046
(2,5)	0.043	0.054	0.052	0.049	0.0500501	0.05
(2,6)	0.052	0.063	0.059	0.048	0.042042	0.042
(2,7)	0.043	0.059	0.06	0.048	0.049049	0.047
(2,8)	0.047	0.048	0.048	0.057	0.0530531	0.054
(3,4)	0.05	0.058	0.058	0.05	0.0520521	0.051
(3,5)	0.042	0.038	0.038	0.043	0.044044	0.042
(3,6)	0.049	0.06	0.056	0.051	0.046046	0.046
(3,7)	0.043	0.048	0.049	0.056	0.0500501	0.048
(3,8)	0.041	0.043	0.043	0.047	0.039039	0.04
(4,5)	**0.074**	**0.08**	**0.079**	0.064	**0.07**	**0.069**
(4,6)	0.062	**0.079**	**0.077**	0.057	**0.068**	**0.067**
(4,7)	0.037	0.054	0.052	0.037	0.0510511	0.049
(4,8)	0.05	0.042	0.042	0.048	0.042042	0.042
(5,6)	**0.07**	**0.074**	**0.073**	0.062	0.0630731	0.064
(5,7)	0.039	0.044	0.043	0.039	0.037	0.038
(5,8)	0.052	0.05	0.052	0.037	0.037	0.037
(6,7)	0.045	0.045	0.048	0.054	0.048048	0.05
(6,8)	0.037	0.044	0.044	0.049	0.037	0.038
(7,8)	0.052	0.04	**0.036**	0.041	0.037	0.04

3 Power Study for Eight Variables

Asymptotic and empirical power comparison for symmetric intercept models are given in Table 2. Higher values for slopes were allocated to items 4, 5, and 6 on a second latent dimension, and higher power is expected for components related to those item pairs. By examining the highlighted values in Table 2, it is clear that the empirical power of second order marginal components (4,5), (4,6) and (5,6) are significantly higher compared to other components. Thus, these second order components were successful in detecting the source of a poorly fit model. This process was repeated for n=300 and n=500. By the results in these tables, it is clear that the empirical power will increase with the sample size and the components were more successful in detecting the lack-of-fit for larger sample sizes. However, when n=300, empirical power results were somewhat lower compared to asymptotic power results. This indicates when sample size is smaller empirical distribution may not close to the hypothesized theoretical distribution. When n=500, empirical power results and asymptotic power results were fairly close. This indicates when sample size increases the empirical distribution approaches hypothesized theoretical distribution.

Table 2. *Asymptotic and Empirical Power Comparison for Symmetric Intercept Model*

	n=300				n=500			
Pair (i,j)	GFfit⊥	Std. Residuals	$\bar{\chi}^2_{ij}$	Asymptotic power*	GFfit⊥	Std. Residuals	$\bar{\chi}^2_{ij}$	Asymptotic power*
(1,2)	0.068	0.07	0.073	0.05077	0.065	0.063	0.064	0.05128
(1,3)	0.073	0.073	0.071	0.05233	0.082	0.076	0.076	0.05389
(1,4)	0.054	0.064	0.063	0.05796	0.071	0.08	0.081	0.06331
(1,5)	0.064	0.072	0.072	0.05861	0.062	0.064	0.066	0.06439
(1,6)	0.072	0.064	0.064	0.05866	0.077	0.062	0.065	0.06448
(1,7)	0.056	0.058	0.06	0.05	0.051	0.055	0.055	0.05
(1,8)	0.063	0.063	0.062	0.05001	0.059	0.059	0.057	0.05002
(2,3)	0.076	0.079	0.077	0.05699	0.085	0.063	0.062	0.06169
(2,4)	0.06	0.058	0.055	0.06535	0.088	0.08	0.081	0.07572
(2,5)	0.064	0.062	0.067	0.06642	0.079	0.073	0.072	0.07752
(2,6)	0.067	0.053	0.054	0.06648	0.085	0.065	0.065	0.07763
(2,7)	0.041	0.061	0.061	0.05	0.03	0.052	0.05	0.05
(2,8)	0.051	0.052	0.048	0.05014	0.055	0.049	0.049	0.05023
(3,4)	0.08	0.064	0.066	0.08986	0.109	0.074	0.075	0.11717
(3,5)	0.104	0.056	0.055	0.09223	0.116	0.075	0.074	0.12118
(3,6)	0.121	0.048	0.049	0.09236	0.15	0.075	0.075	0.1214
(3,7)	0.062	0.056	0.057	0.05157	0.068	0.062	0.062	0.05262
(3,8)	0.044	0.07	0.064	0.05004	0.052	0.061	0.06	0.05007
(4,5)	0.531	0.601	0.599	0.60186	0.757	0.826	0.824	0.81689
(4,6)	0.482	0.553	0.553	0.56285	0.717	0.781	0.781	0.78068
(4,7)	0.044	0.047	0.049	0.05011	0.046	0.068	0.069	0.05019
(4,8)	0.055	0.065	0.065	0.05005	0.053	0.06	0.06	0.05008
(5,6)	0.539	0.562	0.562	0.62046	0.803	0.803	0.803	0.83304
(5,7)	0.048	0.059	0.057	0.05	0.035	0.054	0.055	0.05
(5,8)	0.054	0.054	0.056	0.05009	0.043	0.061	0.06	0.05015
(6,7)	0.046	0.056	0.055	0.05001	0.064	0.064	0.064	0.05002
(6,8)	0.043	0.066	0.064	0.05009	0.048	0.065	0.066	0.05015
(7,8)	0.064	0.07	0.072	0.05001	0.054	0.057	0.058	0.05001

Asymptotic power was calculated only for the orthogonal components.

References

FISHER, R.A. 1924. The conditions under which chi square measures the discrepancy between observation and hypothesis. *Royal Statistical Society.*, **87**, 19–43.

LANCASTER, H. 1969. The chi-squared distribution. New York: Wiley

LIU, Y., & MAYDEU-OLIVARES, A. 2014. Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research.*, **49**, 354–371.

REISER, M. 1996. Analysis of residuals for the multinomial item response model. *Psychometrika.*, **61**, 509–528

REISER, M., CAGNONE, S., & ZHU, J. 2021. An extended GFfit statistic defined on orthogonal components of Pearson's chi-square. *paper under review.*

ASSESSING FOOD SECURITY ISSUES IN ITALY: A QUANTILE COPULA APPROACH

Giorgia Rivieccio [1], Jean-Paul Chavas [2], Giovanni De Luca [1], Salvatore Di Falco [3] and Fabian Capitanio [4]

[1] Department of Management Studies and Quantitative Methods, University of Naples Parthenope, Italy (e-mail: giorgia.rivieccio@uniparthenope.it, giovanni.deluca@uniparthenope.it)

[2] Taylor Hall, University of Wisconsin, Madison, USA, (e-mail: jchavas@wisc.edu)

[3] Geneva School of Economics and Management, University of Geneva, Switzerland (e-mail: Salvatore.DiFalco@unige.ch)

[4] Department of Veterinary Medical and Animal Technology Production, University of Naples Federico II, Italy, (e-mail: fabian.capitanio@gmail.com)

ABSTRACT: The study investigates the dependence structure of the Italian crop yields to provide better insights about the role of climate changes and the crop rotation effects on agricultural productivity. Modeling such dependence, in contemporaneous and serial framework, attempts to explain climate changes as possible engine for co-dependency in the tails of the joint distribution as well as the crop diversification. We used a quantile copula approach to estimate the multivariate distribution of yields across 7 Italian provinces per crop (wheat and corn) over the last 116 years. Findings show a possible dependence by climate for some provinces. Northern regions show higher dependence with a lower crop diversification, thus resulting more exposed to risk of climate effects.

KEYWORDS: QAR models, copulas, tail dependence

1 Introduction

Food security issues are going to focus our attention on the lower tail of the yield distributions. At the regional level, the issue is co-dependence across crop yields of the closest locations (Chavas *et al.* [2019]). At the national level, the issue is co-dependence across locations per each crop where the role of climate could emerge. Therefore, we attempt to explore the possible effect of climate, modeling the joint tail behavior of yields among 7 Italian provinces for each crop (corn and wheat). To estimate the yield distribution we propose a two-step estimation method which involves the use of copulas in a multivariate framework. The first step relies on estimating a Quantile AutoRegressive (QAR) model to shape the yield dynamics of each crop and location, taking into account the dependence structure in terms of lagged quantiles. The second step involves the parametric estimation of multivariate copulas among the conditional quantiles of QAR model estimated in the first step, to measure the whole dependence structure, that is the contemporaneous and the serial dependence as well as the tail dependence of yields across locations per crop. Copulas can be considered as the suitable tool to model both co-dependence and extreme dependence (see Nelsen [2006]). Findings reveal that tail dependence coefficients are high among locations per crop, and such result induces to consider climate as the possible common factor of yield joint behavior, specifically both for higher and lower co-movements in some areas. The paper is organized as follows. In Section 2, we provide a brief review of the QAR model and discuss the use of copulas. Finally, Section 3 develops the empirical analysis.

2 Methodology

Let Y_t be the random variable denoting a crop yield at time t, y_t $(t = 1,\dots,T)$ a sample of T observations and $q_t(\theta)$ the corresponding quantiles at θ (with $0 < \theta < 1$). The Quantile AutoRegression (QAR) model describes the dynamics of the θ-th quantile as:

$$q_t(\theta) = c + \sum_{k=1}^{K} a_k(q_{t-k}(\theta)) + \sum_{m=1}^{M} b'_m x_{t-m}. \tag{1}$$

where M and K are the possibly different number of lags and x_{t-m} is a vector of exogenous lagged values which affect y_t. The parameters of the model are estimated by regression quantiles, as introduced by Koenker [2005]. The conditional quantiles of QAR model of each crop yield is then the input margin of the joint distribution described by the copula function. Copulas allow to better describe the dependence structure among variables and here among quantiles, providing a flexible and well-suited specification of the joint distribution (see Nelsen [2006]). According to the Sklar's theorem (Sklar [1959]), the joint distribution function H of $q_1,\dots,q_p$ can be expressed by a copula function C defined in the unit interval as

$$H(q_1(\theta),\dots,q_p(\theta)) = C(F_1(q_1(\theta)),\dots,F_p(q_p(\theta)))$$

where $F_i(q_i(\theta))$ $(i = 1,...,p)$ is the distribution function of the conditional quantile in Eq. (1) and C is uniquely determined if F are continuous. An important feature of copulas relies on modeling general form of dependencies, also nonlinear as well as focused on the extreme values of variables. The association between extreme values, known as tail dependence, is defined, respectively, in the lower and upper tails, as the limit for a copula C of some h variables with respect to remaining $p-h$ (De Luca & Rivieccio [2012] for details). The specific behavior in the tails of the joint distribution can suggest the copula to select among the parametric families. This feature allows to consider nonlinear association among conditional quantiles.

3 Empirical Application

Data cover period from 1901 to 2017 and concern yearly crop yields of 7 Italian provinces (see Table 1) and 2 crops (wheat and corn). It also includes the variables $x_t = (t_0, t_1, t_2)$ where t_0 is an overall time trend starting at 0 in 2000, t_1 is a time trend starting at 0 in 1940, t_2 is a time trend starting at 0 in 1980. The time trends capture technological and structural changes taking place during the sample period. The best univariate QAR fitting model has 3 lags for each province and for both crop variety, wheat and corn, according to BIC. Elliptical, Archimedean and mixture copulas were applied to model the whole dependence structure of the selected seven provinces per each crop. The most suitable 7-variate copula across provinces per each crop is the mixture of Normal and Student-t copula (see Hu [2006] for details), where the mixture weights are, respectively, 0.216 and 0.784 for wheat and 0.751 and 0.249 for corn. Accordingly, each tail dependence coefficient (λ) is a weighted average of the coefficients of the two copulas, estimated by following Demarta & McNeil [2005] (Table 1). Findings reveal that tail dependence coefficients are high among locations of the same area, and such result induces to consider climate as the common factor of yield joint behavior. In particular, northern regions, generally characterized by a lower crop diversification, highlight higher tail dependence, thus resulting more exposed to risk of climate effects.

References

CHAVAS, J.P., DI FALCO, S., ADINOLFI, F., & CAPITANIO, F. 2019. Weather effects and their long-term impact on the distribution of agricultural yields: evidence from Italy. *European Review of Agricultural Economics*, **46**, 29–51.

DE LUCA, G, & RIVIECCIO, G. 2012. *Multivariate tail dependence coefficients for Archimedean Copulae*. New York: in A. Di Ciaccio et al. (eds.), Advanced Statistical Methods for the Analysis of Large, Studies in Theoretical and Applied Statistics, Springer.

DEMARTA, S., & MCNEIL, A.J. 2005. The t copula and related copulas. *International Statistical Review*, **73**, 111–129.

HU, L. 2006. Dependence Patterns Across Financial Markets: A Mixed Copula Approach. *Applied Financial Economics*, **16**, 717–729.

KOENKER, R. 2005. *Quantile Regression*. Cambridge University Press.

NELSEN, R.B. 2006. *An Introduction to Copulas*. New York: Springer.

SKLAR, A. 1959. Fonctions de répartition à *n* dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229–231.

Table 1. *Tail dependence estimates from a Normal-Student−t copula mixture across the Italian provinces: Milan (1), Venice (2), Bologna (3), Florence (4), Rome (5), Naples (6), Palermo (7).*

Wheat		Corn	
Coefficient	Estimate	Coefficient	Estimate
λ_{12}	0.5693	λ_{12}	0.5571
λ_{13}	0.4828	λ_{13}	0.5763
λ_{14}	0.4801	λ_{14}	0.5539
λ_{15}	0.4798	λ_{15}	0.5515
λ_{16}	0.4346	λ_{16}	0.4585
λ_{17}	0.3789	λ_{17}	0.3460
λ_{23}	0.5002	λ_{23}	0.5702
λ_{24}	0.5124	λ_{24}	0.6023
λ_{25}	0.5245	λ_{25}	0.5842
λ_{26}	0.4626	λ_{26}	0.4552
λ_{27}	0.3578	λ_{27}	0.3138
λ_{34}	0.4776	λ_{34}	0.5835
λ_{35}	0.4754	λ_{35}	0.5511
λ_{36}	0.4969	λ_{36}	0.4749
λ_{37}	0.3724	λ_{37}	0.3153
λ_{45}	0.5023	λ_{45}	0.6050
λ_{46}	0.4403	λ_{46}	0.4511
λ_{47}	0.3496	λ_{47}	0.3136
λ_{56}	0.4751	λ_{56}	0.4497
λ_{57}	0.3658	λ_{57}	0.3214
λ_{67}	0.4422	λ_{67}	0.3693

CO-CLUSTERING FOR HIGH DIMENSIONAL SPARSE DATA

Nicoleta Rogovschi[1]

[1] LIPADE, Université de Paris (e-mail:nicoleta.rogovschi@u-paris.fr)

ABSTRACT: Data anonymization is the process of de-identifying sensitive data while preserving its format and data type (Venkataramanan & Shriram, 2016 Raghunathan, 2013), generally this procedure is achieved by masking one or multiple values in order to hide some aspects of the data. In this paper, we propose a co-clustering model for data anonimization based on topological co-clustering. Co-clustering which is a simultaneous clustering of rows and columns of data matrix consists in interlacing row clusterings with column clusterings at each iteration Govaert, 1995; co-clustering exploits the duality between rows and columns which allows to effectively deal with high dimensional data.

KEYWORDS: anonymization, co-clustering, self-organizing maps, sparse data

1 Introduction

To mine collected data without security breaching, some rules related especially to the privacy of the people on the dataset have to be respected. The process of preserving data privacy is called data anonymization and was used for quite a while to statistical purposes.

k-anonymity is a global framework to evaluate the amount of privacy in some dataset, as the elimination of key identifiers was proven to be inefficient, microdata was disclosed using the microaggregation technique
(Domingo-Ferrer & Torra, 2001).

Li et al. Li *et al.*, 2006 introduced the first algorithm that combines clustering and anonymization. The algorithm forms equivalence classes from the database by finding an equivalence class with records' number less than *k*. It measures the distance between the found equivalence class and the other equivalence classes and merges it with the nearest equivalence class in order to form a cluster of at least *k* elements with minimum information distortion. This method gives good computational results but it is very time consuming.

The topological co-clustering approaches leads to a simultaneous clustering on the rows and columns of data matrix, as well as the projection of the clusters on a two-dimensional grid while preserving the topological order of the initial data.

The co-clustering implicitly performs an adaptive dimensionality reduction at each iteration, leading to better document clustering accuracy compared to one side clustering methods (Dhillon, 2001). Co-clustering is also preferred when there is an association relationship between the data and the features (i.e., the columns and the rows) Ding *et al.*, 2006.

In text mining field, (Dhillon, 2001) has proposed a spectral block clustering method by exploiting the duality between rows (documents) and columns (words). In the analysis of microarray data, where data are often presented as matrices of expression levels of genes under different conditions, block clustering of genes and conditions has been used to overcome the problem of choosing the similarity on the two sets found in conventional clustering methods (Cheng & Church, 2000). The aim of block clustering is to try to summarize this matrix by homogeneous blocks.

2 The proposed algorithm

We propose to use the topological co-clustering in order to k-anonimyze a large sparse dataset. This way, the curse of dimensionality is implicitly dealt with, as the algorithm treats each part simultaneously and the results are proved to be more accurate.

The proposed k-coTCA approach take in input the dataset OT to anonimyze and performs in output an anonymized datset AT having the same size. The algorithm is composed from two steps : co-clustering and anonimization.

Topological co-clustering step:

1. Form the affinity matrix A
2. Define D_r and D_c to be the diagonal matrices

$$D_r = diag(A\mathbf{1}) \text{ and } D_c = diag(A^t\mathbf{1})$$

3. Find U,V the $(g-1)$ left-right largest eigenvectors of

$$\tilde{A} = D_r^{-\frac{1}{2}} A D_c^{-\frac{1}{2}}$$

4. From U and V, form the matrices $\tilde{U}$, $\tilde{V}$ and
$$\mathbf{D} = \begin{pmatrix} \tilde{U} \\ \tilde{V} \end{pmatrix}$$
5. Cluster the rows of $\mathbf{D}$ into g clusters by using SOM and compute the prototypes $w^{[ii]}$
6. Assign object i to cluster R_k if and only if the corresponding row $\mathbf{d}_i$ of the matrix $\mathbf{D}$ was assigned to cluster R_k and assign attribute j to cluster C_k if and only if the corresponding row $\mathbf{d}_j$ of the matrix $\mathbf{D}$ was assigned to cluster C_k.

Anonymization step

For each co-cluster C_k :

- Find the BMU of each object j in R_k using corresponding w_{jc} where c is the matching neuron:
$$(X_i^{[ii]} - w_{jc}^{[ii]})$$
- Code each element j with its corresponding vector: $X_j' \leftarrow [w_{jc_{(1)}}^{[1]}, w_{jc_{(2)}}^{[2]}, \ldots, w_{jc_{(q)}}^{[P]}]$, where $c_{(q)}$ is the index of the cell associated with element j.

To evaluate the co-clustering results, we use the Davies Bouldin index which is a clustering evaluation indicator that reflects the quality of the clustering, as a stopping criterion.

In order to compare the performances of our approach with other traditional unsupervised clustering algorithms, we use many text datasets, which represent the frequency of words in documents. We used eight datasets for document clustering. ”Classic30”, ”Classic150”, ”Classic300”, ”Classic400” are an extract of Classic3 Dhillon, 2001 which contains three classes denoted Medline, Cisi, Cranfield as their original database source.

The impact of co-clustering on the utility of anonymized data is quantified as the resulting accuracy of a machine learning model (RodríGuez-Hoyos *et al.*, 2018). To quantify the utility of the dataset for further study and since all the datasets used are labelled we thought that the best way to evaluate the proposed approaches is to use an external evaluation i.e. the classification. For this purpose, we designed a decision tree model and used it to see how the anonymized data was classified by this model. We then compared the accuracy of the results of both approaches to understand how much data *quality* have we traded for the sake of anonymization. The obtained results, shows that the accuracy doesn’t decrease after the anonymization and allows to maintain the initial structure of the data.

3 Conclusions

In this paper, we introduced a new data anonymization approach based on topological co-clustering which allows to use the prototypes as new values for the anonymized data. The experiences shows that using an classification model on the anonymized dataset, the accuracy doesn't deacrease which means that there is no loose of knowledge from the initial data.

References

CHENG, Y., & CHURCH, G. 2000. Biclustering of expression data. *Pages 93–103 of: ISMB2000, 8th International Conference on Intelligent Systems for Molecular Biology*. Philadelphia, PA: KDD'06.

DHILLON, I. 2001. Co-clustering documents and words using bipartite spectral graph partitioning. *ACM SIGKDD International Conference*, 269–274.

DING, C., LI, T., PENG, W., & PARK, H. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

DOMINGO-FERRER, JOSEP, & TORRA, VICENC. 2001. Disclosure control methods and information loss for microdata. *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies*, 91–110.

GOVAERT, G. 1995. Simultaneous Clustering of Rows and Columns. *Control and Cybernetics*, **24**, 437–458.

LI, JIUYONG, WONG, RAYMOND CHI-WING, FU, ADA WAI-CHEE, & PEI, JIAN. 2006. Achieving k-anonymity by clustering in attribute hierarchical structures. *Pages 405–416 of: International Conference on Data Warehousing and Knowledge Discovery*. Springer.

RAGHUNATHAN, BALAJI. 2013. *The Complete Book of Data Anonymization: From Planning to Implementation*. CRC Press.

RODRÍGUEZ-HOYOS, A., ESTRADA-JIMÉNEZ, J., REBOLLO-MONEDERO, D., PARRA-ARNAU, J., & FORNÉ, J. 2018. Does *k* -Anonymous Microaggregation Affect Machine-Learned Macrotrends? *IEEE Access*, **6**, 28258–28277.

VENKATARAMANAN, NATARAJ, & SHRIRAM, ASHWIN. 2016. *Data Privacy: Principles and Practice*. Chapman & Hall/CRC.

MALARIA RISK DETECTION VIA MIXED MEMBERSHIP MODELS

Massimiliano Russo [1]

[1] Harvard-MIT Center for Regulatory science, Harvard medical school, & Department of Data Science, Dana-Farber Cancer institute (e-mail: m_russo@hms.harvard.edu

ABSTRACT: The diffusion of malaria is a complex phenomenon evolving over time and space, driven by several aspects that include economical biological, behavioral and environmental factors, which act and interact together. We consider as a case study the Machadinho settlement project in Brazil, and provide a risk classification for the households in the area. To accomplish this goal we estimate survey based environmental and behavioral risk profiles via a mixed membership model. We then validate the model comparing the predictive ability of the estimated risk profiles for the crude malaria rate with the performances of standard machine learning (ML) tools.

KEYWORDS: Malaria risk, mixed membership models, multivariate categorical variables.

1 Introduction

The risk of malaria infection is favored by multiple and interacting causes that are largely driven by human behaviors and their interaction with the surrounding environment. To evaluate the risk of malaria infection in a certain geographical area, biological and economical aspects juxtaposed with behavioral and environmental factors should be evaluated. We focus on these last two aspects providing a risk classification for the Machadinho Settlement Project, located in the Rondônia state, Western Brazilian Amazon. The project was approved in 1982, with occupation starting in late 1984. The area was previously a forest sparsely inhabited by rubber tappers (Castro *et al.*, 2006).

Since the early phases of the settlement, malaria diffusion became a problem because of the proliferation of the *Anopheles Darlingi* mosquito, the main malaria vector in the Amazon area. Spread of malaria in frontier settlements can profoundly impact the ecosystem at different levels, and its quantification is of primary importance to design effective measures of mitigation and prevention.

Crude malaria risk measures (e.g., number of malaria cases reported in an households) can be adopted to determine risk profiles through regression or classification models. However, for the Machadinho settlement these indicators can lead to misleading results because of the presence of transient individuals, responsible for high malaria rates in zones presenting low risk conditions (see for example Castro *et al.*, 2006). Unsupervised analysis is not affected by this bias being only based on household features, and can lead reliable findings for targeted stable populations in the settlement project.

We consider the risk classification provided in Russo *et al.*, 2019 and provide a validation of their method. Specifically, we consider the prediction ability of the estimated risk classification of the crude malaria rate and compare them with popular ML tools.

2 Model specification

We follow the model proposed in Russo *et al.*, 2019. We observe categorical variables $X_{ij} \in \{1,\ldots,d_j\}$ for household $i = 1,\ldots,n$ and variable $j = 1,\ldots,p$. These variable are naturally partitioned in two groups: behavioral and environmental variables. We indicate with $\boldsymbol{g} = (g_1,\ldots,g_p)^\intercal$ the group for each of the p variables, where $g_j \in \{1,2\}$; the 1 codifies behavioral variables and the 2 the environmental ones. All households are endowed with 2 membership score vectors $(\boldsymbol{\lambda}_i^{(1)},\boldsymbol{\lambda}_i^{(2)})^\intercal$ such that $\sum_{h=1}^{H}\lambda_{ih}^{(g)} = 1$ for $g = 1,2$.

The proposed model can be expressed in the following hierarchical form:

$$\begin{aligned} X_{ij} \mid Z_{ij} = h, \boldsymbol{\psi}_h^{(j)} &\sim \mathrm{Cat}(\psi_{h1}^{(j)},\ldots,\psi_{hd_j}^{(j)}), \\ Z_{ij} \mid \boldsymbol{\lambda}_i^{(g_j)} &\sim \mathrm{Cat}(\lambda_{i1}^{(g_j)},\ldots,\lambda_{iH}^{(g_j)}), \\ (\boldsymbol{\lambda}_i^{(1)},\boldsymbol{\lambda}_i^{(2)}) &\sim \mathrm{MLND}(\boldsymbol{\mu},\boldsymbol{\Sigma}). \end{aligned} \tag{1}$$

Here $Z_{ij} \in \{1,\ldots,H\}$ are discrete latent variables that identify H latent groups. The notation $X \sim \mathrm{Cat}(\pi_1,\ldots,\pi_d)$ indicates a d-dimensional categorical random variable, i.e. $\mathrm{pr}(X = k) = \pi_k$, for $k = 1,\ldots,d$, while $\mathrm{MLND}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the multivariate logistic normal distribution introduced in Russo *et al.*, 2019.

Model (1) provides a low dimensional summary of the analyzed variables. In fact, for a certain latent group h, $\{\psi_h^{(j)}$ for $j = 1,\ldots,p\}$ describe the h-th group in terms of the observed variables. The scores $\boldsymbol{\lambda}_i^{(1)}$ and $\boldsymbol{\lambda}_i^{(2)}$ can be interpreted as a degree of similarity of the i-th individual for the latent group h (see figure 1 for a representation). Hence, partitioning the variables in behavioral and environmental domains, and choosing $H = 2$, the scores $\boldsymbol{\lambda}_i^{(1)}$ and $\boldsymbol{\lambda}_i^{(2)}$

can be interpreted as a summary of behavioral and environmental risks. We refer to Russo *et al.*, 2019 for additional details, prior specification and for a description of the algorithm to approximate the posterior of model (1).

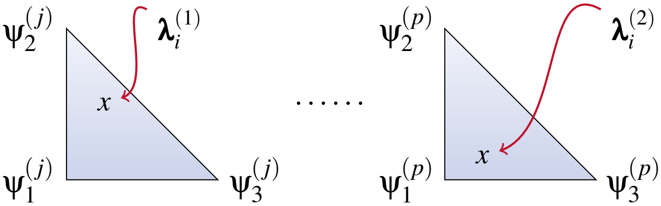


Figure 1. *Representation of the risk profiles for subject i and the p variables. In this example variable* 1 *is associated with domain* 1 *(behavioral) and variable p with domain* 2 *(environmental)*

3 Model validation

We consider an external validation of model (1), checking its performances in the out-of-sample prediction of the crude malaria rate. We use 5-fold cross validation, comparing the results with random forest, lasso prediction, and PCA regression on the raw data. For the random forest and lasso we used the R packages `randomForest` and `glmnet`, respectively. We refer to Castro *et al.*, 2006 for a detailed description of the data.

Note that crude malaria rates and the risk profiles estimated from model (1) measure different aspect of risk. Therefore, they are expected to be related but they are not expressed in the same scale. To overcome this issue, we assume that the estimated risk are proportional to the malaria rate r_i via:

$$r_i = c_1\hat{\lambda}_i^{(1)} + c_2\hat{\lambda}_i^{(2)},$$

where $\hat{\lambda}_i^{(1)}$ and $\hat{\lambda}_i^{(2)}$ are the posterior mean of the behavioral and environmental risks estimated via (1). The coefficients c_1 and c_2 are estimated via least squares on a training sample. To avoid over fitting in the considered ML models, for each fold we select tuning parameters with an additional cross validation.

Based on Table 1, multivariate mixed membership scores can predict malaria rates with comparable out-of-sample predictive performance to black box machine learning algorithms, such as random forests. The ML algorithms and

Table 1. *5 fold cross validation mean square prediction error (standard deviation) divided by year.*

	1985	1986	1987	1995
Random Forest	0.104(0.026)	0.131(0.023)	0.108(0.011)	0.021(0.006)
lasso regression	0.104(0.026)	0.169(0.017)	0.123(0.005)	0.021(0.006)
PCA regression	0.101(0.030)	0.161(0.021)	0.121(0.005)	0.021(0.006)
Russo *et al.*, 2019	0.099(0.024)	0.148(0.015)	0.105(0.003)	0.021(0.006)

other supervised approaches are subject to the selection bias issue mentioned in Section 1. This result gives evidence that model (1) provides a reasonable low dimensional representation in the considered context.

References

CASTRO, MARCIA CALDAS, MONTE-MÓR, ROBERTO L., SAWYER, DIANA O., & SINGER, BURTON H. 2006. Malaria risk on the Amazon frontier. *Proceedings of the National Academy of Sciences*, **103**(7), 2452–2457.

RUSSO, MASSIMILIANO, SINGER, BURTON H, & DUNSON, DAVID B. 2019. Multivariate mixed membership modeling: Inferring domain-specific risk profiles. *arXiv preprint arXiv:1901.05191*.

NONPARAMETRIC ESTIMATION OF THE NUMBER OF CLUSTERS FOR DIRECTIONAL DATA

Paula Saavedra-Nieves [1] and Rosa M. Crujeiras[1]

[1] Department of Statistics, Mathematical Analysis and Optimization,
Universidade de Santiago de Compostela,
(e-mails: paula.saavedra@usc.es, rosa.crujeiras@usc.es)

ABSTRACT: Set estimation is focused on the reconstruction of a set (or the estimation of any of its features such as its volume or its boundary) from a random sample of points. Target sets to be estimated may appear in different contexts, but from a distribution-based perspective, level set estimation is a problem of interest. Actually, this theory is also linked to clustering methods: Hartigan (1975) defines the number of population clusters as the number of connected components of density level sets. This topic has received some attention in the literature specially for densities supported on an Euclidean space. However, just as density level sets, this clustering approach can be easily extended to more general settings such as the circle or the sphere.

The rationale for establishing the definition of cluster provided by Hartigan (1975) is quite related with the notion of mode. In fact, several cluster algorithms are based on the detection of modes noting that the number of modes (local maxima of the density function) is rarely smaller than the number of clusters. Nevertheless, the concept of cluster is easier to handle, since it has a global and geometrical nature, whereas the local maxima depend on analytical properties.

In this work, we derive some methodology for estimating the number of directional clusters as the number of connected components of directional level sets. From an empirical perspective, directional level sets are estimated using a nonparametric plug-in reconstruction (see, for instance, Saavedra-Nieves and Crujeiras, 2020). An extensive simulation study shows the performance of this estimator for densities supported on the unit circle and the sphere. Additionally, this methodology is applied to analyse a real data set.

KEYWORDS: Connected components, density level sets, directional data.

References

HARTIGAN, J. A. 1975. *Clustering algorithms*. John Wiley & Sons, Inc.

SAAVEDRA-NIEVES, P., & CRUJEIRAS, R. M. 2020. Nonparametric estimation of directional highest density regions. arXiv preprint arXiv:2009.08915.

TENSOR-VARIATE FINITE MIXTURE MODEL FOR THE ANALYSIS OF UNIVERSITY PROFESSOR REMUNERATION

Shuchismita Sarkar[1], Volodymyr Melnykov[2] and Xuwen Zhu[2]

[1] Bowling Green State University, (e-mail: ssarkar@bgsu.edu)

[2] University of Alabama, (e-mail: vmelnykov@cba.ua.edu, xzhu20@cba.ua.edu)

ABSTRACT: Finite mixture modeling and model-based clustering of matrix- and tensor-variate data has recently gained a lot of attention. In this paper a novel tensor regression mixture model has been proposed to analyze salary data collected by the American Association of University Professors over a span of thirteen years at several faculty rank and gender levels. Most of the studies involving faculty remuneration employs linear regression models intended for predicting individual salaries. Such models, however, are not suitable for developing strategies and policies at institutional level. The tensor regression mixture framework adopted in this paper allows for an university level analysis of faculty remuneration by considering the heterogeneous, skewed, multi-way, and temporal nature of the data. The developed model addresses several important issues related to gender equity and peer institution comparison.

KEYWORDS: finite mixture model, model-based clustering, EM algorithm, tensor regression mixture model

Specifying composites in structural equation modeling: the Henseler-Ogasawara specification

Florian Schuberth [1]

[1] Department of Design, Production and Management, University of Twente, The Netherlands, (e-mail: f.schuberth@utwente.nl))

Abstract: Structural equation modeling (SEM) is a versatile statistical method that can deal in principle with latent variables and composites. In practice, however, researchers using SEM encounter problems incorporating composites into their models. To overcome this problem, I present a specification for SEM that was recently sketched by Henseler (2021) to incorporate composites in structural models. It draws from the same idea that was proposed by Ogasawara (2007) to conduct a canonical correlation analysis in SEM. Therefore, the specification is dubbed Henseler-Ogasawara (H-O) specification. In the H-O specification, a set of observed variables forming a composite is expressed by a set of synthetic variables, which are labeled as emergent and excrescent variables. An emergent variable is a linear combination of variables that is related to other variables in the structural model, whereas an excrescent variable is a linear combination of variables that is unrelated to all other variables in the structural model. This approach is advantageous over existing approaches, as it offers the same flexibility in terms of model specification for modeling with composites as SEM provides for modeling with latent variables. As a consequence, the H-O specification makes all existing developments in SEM available for modeling with composites, such as testing parameter estimates, testing for overall model fit and dealing with missing values.

Keywords: model specification, composite model, emergent variable, excrescent variable, components

References

Henseler, J. 2021. *Composite-Based Structural Equation Modeling: Analyzing Latent and Emergent Variables.* Guilford Press.

Ogasawara, H. 2007. Asymptotic expansions of the distributions of estimators in canonical correlation analysis under nonnormality. *Journal of Multivariate Analysis*, **98**, 1726–1750.

NETWORK ANALYSIS IMPLEMENTING A MIXTURE DISTRIBUTION FROM BAYESIAN VIEWPOINT

Jarod Smith[1], Mohammad Arashi[2] and Andriëtte Bekker[1]

[1] Department of Statistics, University of Pretoria, South Africa (e-mail: jarodsmith706@gmail.com, andriette.bekker@up.ac.za)

[2] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran (e-mail: arashi@um.ac.ir)

ABSTRACT: Differential networks (DN) are important tools for modeling the changes in conditional dependencies between multiple samples. A Bayesian approach for estimating DNs, from the classical viewpoint, is introduced with a computationally efficient threshold selection for graphical model determination. The algorithm separately estimates the precision matrices of the DN using the Bayesian adaptive graphical lasso procedure. Synthetic experiments illustrate that the Bayesian DN performs exceptionally well in numerical accuracy and graphical structure determination in comparison to state of the art methods. The proposed method is applied to South African COVID-19 data to investigate the change in DN structure between various phases of the pandemic.

KEYWORDS: Bayesian graphical lasso, differential network, double-exponential distribution, Gaussian graphical model, precision matrix

MEASUREMENT ERRORS IN MULTIPLE SYSTEMS ESTIMATION

Paul A. Smith[1], Peter G.M. van der Heijden[12] and Maarten Cruyff[2]

[1] Department of Social Statistics & Demography, University of Southampton, UK, (e-mail: p.a.smith@soton.ac.uk)

[2] Dept of Social Sciences, Methodology and Statistics, Utrecht University, Netherlands, (e-mail: P.G.M.vanderHeijden@uu.nl, m.cruyff@uu.nl)

ABSTRACT: Dual and multiple system estimation use the presence ('capture') of people in different data sources as the basis for estimation of the population size. Where further characteristics are also available, these can be used to provide estimates of the population size classified by these characteristics. We consider the situation that there are measurement errors in these classifying variables, but not in the linkage of people between data sources. We consider strategies to produce estimates of the population size and breakdown using a consistent, adjusted definition taking account of all the evidence in the collected data sources.

KEYWORDS: capture-recapture, latent class analysis, ethnicity, population size estimation.

1 Introduction

Dual and multiple system estimation have a long history of use to estimate the size of populations which cannot be completely observed, and in recent years there have been many applications to estimating the size of human populations. In the simplest cases this may result from observing people on two sources, and using an assumption of independence between the sources to obtain an estimated population size. When there are more sources, interactions between the sources can be fitted, and an appropriate model needs to be fitted to the (implied) contingency table formed from the presence or absence of people in each source. In general this procedure assumes no errors of observation, and that no errors are made in linking people on the different sources. If an independent estimate of the linkage errors can be obtained, it can be used in an adjusted estimator (Zult *et al.* 2021). However, in this paper we work with the usual framework that assumes that linkage is made perfectly.

Auxiliary information is often available on the different sources, in addition to the existence of a person (or record), and this information may be used in linkage where it corresponds to a stable characteristic. Other variables are of more substantive interest, and may be expected to vary between sources for a number of reasons: they may be characteristics which vary in time, or they may be measured in different ways in different data sources, leading to variations in the measurement. In this latter case we may consider that there is an underlying ‘true’ variable, and that one or more of the sources that we are using observe a version of this variable with some added measurement error. The process of linking datasets means that some variables will not be observed for some records, and that no variables are observed for records in none of the sources, the number of which will nevertheless be estimated during population size estimation.

In this paper we consider strategies for dealing with population size estimation, broken down using variables measured in one or more sources and subject to measurement error in this way. Section 2 deals with solutions based on explicit decisions about which measure is the best, and with simple combinations of variables, and Section 3 with the use of a latent class model to derive an underlying measure, which we consider to estimate the ‘true’ measure based on the available data.

2 Population size estimation with a preferred covariate source

First we consider that there are two sources, and both sources contain what is conceptually the same covariate, though we know or suspect that they are measured differently, or that their resulting quality is different because of the way they are collected. Van der Heijden *et al.* (2018) present an example where characteristics of accidents, specifically whether a motorised vehicle was involved, are recorded both by the police and by hospitals. It would be possible to treat these as the same variable, but investigation of the data where the ‘motorised vehicle’ variable is available from both sources shows that about 5% of cases have discrepancies. Instead we treat them as two different variables, and construct a four-way contingency table formed from presence/absence on the two sources and the motorised/non-motorised variables in the two sources. We then use loglinear modelling to choose a suitable model for this contingency table, and use this model together with the EM algorithm to produce a completed table (Table 1), which provides an estimate of the missing part of the population, and also estimates of the population sizes in each cell of the contingency table (where they are not observed). This allows us to add up in any way we want to achieve a set of consistent estimates.

In the accidents example, we have reasonable confidence that the police register is better at recording whether a motorised vehicle is involved, as gathering this

		B = 1		B = 0		Total
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	
A = 1	$X_1 = 1$	5970.0	287.0	1289.0	62.0	7608.0
	$X_1 = 0$	28.0	256.0	6.9	63.1	354.0
A = 0	$X_1 = 1$	2933.2	2177.6	633.3	470.2	6214.3
	$X_1 = 0$	13.8	1942.4	3.4	478.8	2438.4
Total		8945.0	4663.0	1932.6	1074.1	16614.7

Table 1: Completed road accidents table. A is the police register, B the hospital register, X_1 is the police record of motor vehicle involvement, and X_2 the hospital record.

information is part of the police function. So we consider the classification of the total according to the variable X_1 in the police register (whether observed or estimated) to be the correct one. And since the full dataset, cross-classified by both police and hospital versions of the motorised vehicle variable is available, we can make inferences about the measurement error in the hospital version.

In a situation where the relative merits of the measurements are less clear, we could pragmatically use the average of the population size estimates under the different versions of the auxiliary variable.

3 Latent class models

A further approach is to treat the different measurements separately in the population size estimation, but then to embed them in a latent class model (LCM), which postulates an underlying, unobserved parameter related to all the separate measurements, and which can be interpreted as the true parameter. This approach can be considered when there are at least three measurements. It is conceptually different from using LCMs to deal with heterogeneity in the capture probabilities (as in Stanghellini & van der Heijden 2004). Van der Heijden et al. (2021) apply this approach in analysing four linked data sources in New Zealand – the population census, the health register, the birth registration register, and an education register (covering largely, but not only, tertiary education). Each of these sources includes an ethnicity variable, which we consider in a simplified version recoded to Māori or all other ethnicities. We would like to estimate both the size of the population in New Zealand and the size of the Māori population based on these sources.

Two approaches are possible. In the first, we treat the four sources using multiple system estimation, fitting a loglinear model to the eight-way table formed by the inclusion or not in each source and Māori ethnicity or not. Some of the estimates from a saturated model go to infinity, so a reduced form of the model is needed to obtain parameter estimates with reasonable interpretation and stability. The estimates arising from this model (including the estimates of the size of the unobserved part of the population) are then used as the inputs to a latent class model with two latent classes. This gives a two-part procedure which has the advantage of being close to the original model for the 8-way contingency table. The model produces estimates of the size of the Māori population from one of the latent classes, which can be interpreted as the true Māori variable. It can also be used to give estimates of the errors in the four observed variables in measuring this underlying Māori concept.

The second approach aims to include the latent class model directly in the modelling of the 8-way contingency table. The assumption of the latent class model is that the (unobserved) interactions between the observed variables and the latent variable explain all the interactions within the observed variables in the original data. Therefore we replace [abcd] in the original model with [aX][bX][cX][dX] with the latent variable X (where a, b, c and d label the ethnicity variables in the four data sources). This replaces all interactions of a, b, c, and d, so any terms containing two or more of these parameters are dropped from the model (which serves to make the loglinear model hierarchical with respect to interactions, and therefore more easily

			census	DIA	MOH	DOE
		π_x	$\pi^a_{r=1\|x}$	$\pi^b_{r=1\|x}$	$\pi^c_{r=1\|x}$	$\pi^d_{r=1\|x}$
two-step	class 1	0.827	0.004	0.016	0.003	0.015
	class 2	0.173	0.937	0.937	0.826	0.922
LCMSE	class 1	0.834	0.007	0.014	0.004	0.016
	class 2	0.166	0.957	0.958	0.857	0.959

Table 2: Estimates of probabilities from latent class models with two latent classes. Class $r = 1$ is interpreted as non-Māori, and class 2 as Māori.

interpretable). This leaves a latent class model embedded in the multiple system estimation, and van der Heijden *et al.* (2021) call this the latent class multiple system estimation (LCMSE) model.

In the application to data from the New Zealand Integrated Data Infrastructure (IDI-ERP), the LCMSE has a slightly lower estimate of the number of Māori and a slightly higher overall population estimate than the two-step procedure based on latent class estimation using the multiple system estimation results. The LCMSE therefore takes a more conservative approach to the definition of Māori in this dataset.

The population census has been generally held to be the best measure of Māori ethnicity among the different sources available in New Zealand, and it has low values for measurement error in both Māori and non-Māori in both approaches (Table 2). The Health register has a low error for non-Maori, but the largest measurement error for Māori. The births and education registers are similar to the census in the estimated measurement error in the Māori class, but have more error in estimating the non-Māori class. Therefore overall our results support the conclusion that the census is the best overall measure of ethnicity.

References

STANGHELLINI, E. & VAN DER HEIJDEN, P.G.M. (2004). A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. *Biometrics*, **60**, 510–516.

VAN DER HEIJDEN, P.G.M., CRUYFF, M., SMITH, P.A., BYCROFT, C., GRAHAM, P. & MATHESON-DUNNING, N. 2021 (in press). Multiple system estimation using covariates having missing values and measurement error: estimating the size of the Māori population in New Zealand. *Journal of the Royal Statistical Society, Series A*.

VAN DER HEIJDEN, P.G.M., SMITH, P.A., CRUYFF, M. & BAKKER, B. 2018. An overview of population size estimation where linking registers results in incomplete covariates, with an application to mode of transport of serious road casualties. *Journal of Official Statistics*, **34**, 239–263.

ZULT, D., DE WOLF, P.-P., BAKKER, B.F.M. & VAN DER HEIJDEN, P.G.M. 2021 (in press). A general framework for multiple-recapture estimation that incorporates linkage error correction. *Journal of Official Statistics*.

ROBUST CLASSIFICATION IN HIGH DIMENSIONS USING REGULARIZED COVARIANCE ESTIMATES

Valentin Todorov[1] and Peter Filzmoser[2]

[1] United Nations Industrial Development Organization, AUSTRIA,
(e-mail: valentin@todorov.at)

[2] Vienna University of Technology, (e-mail: p.filzmoser@tuwien.ac.at)

ABSTRACT: High-dimensional highly correlated data exist in many application domains which makes the classical classification methods like LDA and QDA practically useless because they will suffer from the singularity problem if the number of observed variables p exceeds the number of observations n. A number of regularization techniques with the purpose to stabilize the classifier and to achieve an improved classification performance have been developed and there exist several studies comparing various regularization techniques trying to facilitate the choice of a method. However, these methods are vulnerable to the presence of outlying observations (outliers) in the training data set which can influence the obtained classification rules and make the results unreliable. On the other hand, the proposed in the literature high breakdown point versions of discriminant analysis do not work or are not reliable in high dimensions. We propose to utilize the recently introduced regularized versions of the minimum covariance determinant (MCD) estimator – RMCD and MRCD - and thus to combine high robustness to outliers, the possibility to be computed for high dimensions and readily available software in R. Simulated and real data examples show that the proposed method performs better than, or at least as well as, the existing methods in a wide range of settings.

KEYWORDS: robust classification, regularization, outliers, MCD

CLUSTERING VIA NEW PARSIMONIOUS MIXTURES OF HEAVY TAILED DISTRIBUTIONS

Salvatore D. Tomarchio[1], Luca Bagnato[2] and Antonio Punzo[1]

[1] Department of Economics and Business, University of Catania, (e-mail: daniele.tomarchio@unict.it, antonio.punzo@unict.it)

[2] Department of Economic and Social Sciences, Università Cattolica del Sacro Cuore, (e-mail: luca.bagnato@unicatt.it)

ABSTRACT: Two families of parsimonious mixture models are used for model-based clustering. They are based on the multivariate shifted exponential normal and the multivariate tail-inflated normal distributions, heavy tailed generalizations of the multivariate normal. Parsimony is achieved via the eigen-decomposition of the component scale matrices, as well as by imposing a constraint on the tailedness parameter. Two variants of the expectation-maximization algorithm are used for parameter estimation. Identifiability conditions are illustrated, and the advantages of our models with respect to other existing parsimonious heavy-tailed mixture models are commented. Our models are firstly tested via simulation studies, and then compared to some competing models in real data applications.

KEYWORDS: mixture models, model-based clustering, parsimony, heavy-tailed distributions

A GENERAL BI-CLUSTERING TECHNIQUE FOR FUNCTIONAL DATA

Agostino Torti[1 2] , Marta Galvani[1], Alessandra Menafoglio[1], Piercesare Secchi[1 2] and Simone Vantini[1]

[1] MOX - Department of Mathematics, Politecnico di Milano, Italy

[2] Center for Analysis Decisions and Society, Human Technopole, Milano, Italy

ABSTRACT: The problem of bi-clustering in the Functional Data Analysis framework is considered, with the aim of simultaneously clustering the rows and columns of a data matrix whose entries are functions, possibly taking values in a multidimensional space. A definition of bi-cluster for functional data is given and a novel bi-clustering method - called Functional Cheng and Church (FunCC) - is developed. The FunCC method is a non parametric and very flexible technique able to discover bi-clusters, based on a flexible modeling depending on the problem at hand.

KEYWORDS: Bi-clustering, Clustering, Functional Data

Nowadays, many systems are able to collect information with high frequency recording multiple phenomena at the same time in an almost continuous way. For this reason, researchers have put a lot of efforts into the development of new statistical methods able to deal with this new type of data. In particular, functional data analysis (FDA) is the branch of statistics that deals with random variables taking values into an infinite dimensional functional space, see Ramsay (2004) for more details.

In this contribution, we consider the problem of bi-clustering functional data which has recently been addressed in the literature and we describe the methods we proposed in Galvani *et al.* (2021) and Torti *et al.* (2021). Bi-clustering methods, commonly known thanks to the work of Cheng & Church (2000)), allow to discover subgroubs of observations behaving in a similar way on a subset of features or vice-versa. This is of particular interest when the data are intrinsically ordered in a matrix structure and the aim is to simultaneously group the rows and the columns of the data matrix without constraining the rows (or the columns) of a data matrix to belong to only one group over all the features (or the observations) as in the classical clustering methods. In the FDA framework, there are just few works dedicated to the problem of bi-clustering functional data framed in a matrix structure. Bouveyron *et al.* (2018) developed a parametric bi-clustering technique, based on the functional latent block model, to co-cluster different electricity consumption curves on different days. Although, this approach needs to rely on strong modelling assumptions of the data, which are hardly verified in the FDA framework, and only detect exhaustive bi-clusters, i.e. discovering a checkerboard structure that does not always fit with real data, for uni-variate functional data. An alternative extension of bi-clustering to the functional realm is proposed by Di Iorio & Vantini (2019): given a set of functions, they propose an algorithm to identify sub-domains of the original functional domain where a subset of functions shows similar patterns. In our work, we proceed along the same line introduced by Bouveyron *et al.* (2018) and go a step further developing a non parametric algorithm able to discover non exhaustive bi-clusters in a data matrix whose entries are functions, possibly taking values in a multidimensional space. First, we introduce a novel methodology based on the extension of the Cheng and Church algorithm, called FunCC, by proposing an iterative procedure based on a non parametric approach which allows to find flexible and non exclusive bi-clusters for univariate functional data. Then, the FunCC algorithm is extended to the general case of multivariate data, therefore bi-clustering data matrices whose entries in each cell are multivariate functional data. In this way, we are able to deal with bi-clustering problems where multiple aspects are observed at the same time for each observation. For more details about the developed methodology and the implemented algorithm see Galvani *et al.* (2021) and Torti *et al.* (2021).

Given a dataset arranged in a matrix A composed by n rows and m columns, the aim of a bi-clustering technique is to find a submatrix $B(I,J) \in A$, corresponding to a subset of rows I and a subset of columns J, with a *similar behavior*. In particular, in the Cheng and Church algorithm (Cheng & Church (2000)), an ideal bi-cluster is a set of rows I and a set of columns J such that each element in the bi-cluster can be represented as the average value in the bi-cluster plus a row and column components. A particular measure of goodness is evaluated for each sub-matrix $B(I,J)$ considering a similarity score - which is the *Mean Squared Residual* between all the real values and their representative values in the bi-cluster - and the sub-matrix $B(I,J)$ is selected as bi-cluster if its similarity score is lower than a threshold value.
Extending these concepts in the FDA framework, in each cell of the data matrix A a function $f_{ij}(t)$ defined on a continuous domain T is contained.

Definition 0.1 *Given a data matrix A, an ideal bi-cluster is a sub-matrix $B(I,J) \subseteq A$, s.t. each element f_{ij} with $i \in I$ and $j \in J$ can be expressed as*

$$f_{ij}(t) = \mu(t) + a\alpha_i(t) + b\beta_j(t), \ \ \forall i \in I \ , \ \forall j \in J \ \ with \ \ t \in T$$

with $(a,b) \in \{0,1\}^2$ fixed by the analyst, μ defined for the bi-cluster $B(I,J)$ as $\mu(t) = \frac{1}{|I||J|}\sum_{i\in I}\sum_{j\in J} f_{ij}(t)$ for $t \in T$, and α_i and β_j being the rows and columns components, respectively, s.t. $\sum_{i\in I}\alpha_i = 0$ and $\sum_{j\in J}\beta_j = 0$.

Starting from Definition 0.1 (Galvani *et al.* (2021)), it is possible to obtain different kinds of ideal bi-clusters, associated to different application contexts, by differently considering a and b. For example, setting $(a,b) = (0,0)$ in the Definition 0.1, only the average value in the bi-cluster is considered, hence the ideal bi-cluster is composed by a group of functions f_{ij} all equal to the average value μ of the bi-cluster. Moreover, while μ is evaluated as the average function of the functions contained in $B(I,J)$, the computation of the row and column components α_i and β_j depends on their functional form. If α_i and β_j are assumed to be functional objects, then, they can be evaluated as the average functional residuals of rows and columns, respectively, with respect to the average function μ, i.e. $\alpha_i(t) = \frac{1}{|J|}\sum_{j\in J} f_{ij}(t) - \mu(t)$ and $\beta_j(t) = \frac{1}{|I|}\sum_{i\in I} f_{ij}(t) - \mu$. If instead, α_i and β_j are assumed to be constant, then, they can be consistently evaluated as the average value of the functional residuals of rows and columns, respectively, with respect to the average function μ, i.e. $\alpha_i = \frac{1}{|T|}\int_T \left(\frac{1}{|J|}\sum_{j\in J} f_{ij}(t) - \mu(t)\right)dt$ and $\beta_j = \frac{1}{|T|}\int_T \left(\frac{1}{|I|}\sum_{i\in I} f_{ij}(t) - \mu(t)\right)dt$. In practice, we want to find sub-matrices $B(I,J)$ as similar as possible to an ideal bi-cluster, i.e. sub-matrices $B(I,J)$ which minimize a specific objective function. The so-called H-score measures the deviation of the selected elements from an ideal bi-cluster (Cheng & Church (2000)). In our case, we define the H-score of a sub-matrix $B(I,J)$ as

$$H(I,J) = \frac{1}{|I||J|}\sum_{i\in I}\sum_{j\in J} \left\|f_{ij} - f_{ij}^0\right\|_{L^2}^2$$

with $f_{ij}^0(t) = \mu(t) + a\alpha_i(t) + b\beta_j(t)$ being the template function of the bi-cluster, where $(a,b), \mu, \alpha_i$ and β_j are defined as in Definition 0.1.

Notice that, the definition just mentioned above can be generalised also in the multivariate case, e.g. dealing with data matrices A whose entries are multivariate functional data $\boldsymbol{f_{ij}} = (f_{ij}^1, ..., f_{ij}^P)$ with one-dimensional domain and a P-dimensional codomain with $P \geq 1$. In this case, the definition of ideal bi-cluster is re-defined in way such that each element of the bi-cluster can be expressed on each p-dimension, with $p \in \{1, ..., P\}$, as in Definition 0.1. Consistently, a measure of goodness of the bi-cluster can be trivially evaluated by estimating the H-score of a sub-matrix $B(I,J)$ as the average value of the single H-score on each p-dimension. In both uni-variate and multivariate functional cases, the implemented algorithm starts considering the whole dataset and try to find the biggest bi-cluster with a H-score value lower then a given threshold δ by adding/removing rows/columns. Each time a row/column is added/removed, the H-score is updated. For more details about the steps of the algorithm and the choice of the treshold parameter δ see Galvani *et al.* (2021) and Torti *et al.* (2021).

To bi-cluster a data matrix whose entries are functions possibly taking values in a multidimensional space, a bi-clustering technique - called Functional Cheng and Church (FunCC) - is developed. The presented approach is non parametric, thus no assumptions are made on the distribution generating the data, and very flexible, allowing to discover non-exhaustive and different biclusters depending on the problem at hand. During the presentation of this contribution, we will show the performance of the developed methodology both on simulated data and on real case studies stimulated by challenging research questions related to mobility infrastructures.

Acknowledgement: the authors gratefully acknowledge Trenord for providing the data that will be shown during the presentation of this contribution.

References

BOUVEYRON, CHARLES, BOZZI, LAURENT, JACQUES, JULIEN, & JOLLOIS, FRANÇOIS-XAVIER. 2018. The functional latent block model for the co-clustering of electricity consumption curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**(4), 897–915.

CHENG, YIZONG, & CHURCH, GEORGE M. 2000. Biclustering of expression data. *Pages 93–103 of: Ismb*, vol. 8.

DI IORIO, JACOPO, & VANTINI, SIMONE. 2019. funBI: a Biclustering Algorithm for Functional Data. *MOX-Report No. 46/2019.*

GALVANI, MARTA, TORTI, AGOSTINO, MENAFOGLIO, ALESSANDRA, & VANTINI, SIMONE. 2021. FunCC: a new bi-clustering algorithm for functional data with misalignment. *Computational Statistics Data Analysis.*

RAMSAY, JAMES O. 2004. Functional data analysis. *Encyclopedia of Statistical Sciences*, **4**.

TORTI, AGOSTINO, GALVANI, MARTA, MENAFOGLIO, ALESSANDRA, SECCHI, PIERCESARE, & VANTINI, SIMONE. 2021. A General Biclustering Algorithm for Hilbert Data: Analysis of the Lombardy Railway Service. *Mox-Report No. 21/2021.*

DEVELOPING A MULTIDIMENSIONAL AND HIERARCHICAL INDEX FOLLOWING A COMPOSITE-BASED APPROACH

Laura Trinchera[1]

[1] Department of Information Systems, Supply Chain Management and Decision-making, NEOMA Business School, (e-mail: laura.trinchera@neoma-bs.fr)

ABSTRACT: The development of a measurement instrument involves establishing a link between the concepts (theoretical world) and the data (empirical world) (Zeller & Carmines, 1980) . When modelling multidimensional concepts on a higher level of abstraction is common practice to include higher-order constructs as a proxy of such concepts. Higher-order construct are defined as constructs whose indicators are not directly observable but again constructs (Henseler, 2021). Following Henseler's (2021) classification we can specify high-order constructs according to 4 different structures:

- Type I: Latent variables measured by latent variables.
- Type II: Emergent variables made of latent variables.
- Type III: Latent variables measured by emergent variables.
- Type IV: Emergent variables made of emergent variables.

Each of these four specifications needs to apply a different validation process of the measurement instrument. During this presentation I will discuss the more recent advances on higher-order construct validation (Schuberth *et al.*, 2020).

KEYWORDS: measurement model, PLS path modelling, formative index

References

HENSELER, J. 2021. *Composite-Based Structural Equation Modeling: Analyzing Latent and Emergent Variables.* Guilford Press.

MASSIERA, P, TRINCHERA, L, & RUSSOLILLO, G. 2018. Evaluating the presence of marketing capabilities: a multidimensional, hierarchical index. *Recherche et Applications en Marketing*, **33**, 30–52.

SCHUBERTH, F, RADEMAKER, M E, & J, HENSELER. 2020. Estimating and assessing second-order constructs using PLS-PM: the case of composites of composites. *Industrial Management and Data Systems*, **120**, 2211–2241.

ZELLER, R A, & CARMINES, E G. 1980. *Measurement in the social sciences: The link between theory and data.* CUP Archive.

A GENERALISED CLUSTERWISE REGRESSION FOR DISTRIBUTIONAL DATA

Rosanna Verde [1], Francisco de A. T. de Carvalho [2] and Antonio Balzanella [1]

[1] Department of Mathematics and Physics, University of Campania ”Luigi Vanvitelli”, (e-mail: rosanna.verde@unicampania.it, antonio.balzanella@unicampania.it)

[2] CIN-UFPE, Av. Jornalista Anibal Fernandes, s/n - Cidade Universitária 50.740-560, Recife, PE, Brasil, (e-mail: fatc@cin.ufpe.br)

ABSTRACT: This paper deals with a cluster-wise regression method for distributional data. The objects to cluster are observed on a dependent character and on a set of explanatory variables. A dependence relation is then assumed, which can be improved by considering local structures among the data. The proposed algorithm is based on the K-means clustering algorithm: the centroids of the clusters are linear regression models and the objects are assigned to the clusters according to minimum sum of squared errors. The generalised CR algorithm is based on a linear regression model for distributional variables and on a K-means algorithm developed for similar data; both the methods use a L_2 Wasserstein distance.

KEYWORDS: Distributional data, Clusterwise regression, K-means, Wasserstein distance.

1 Introduction

In this paper we propose a cluster-wise regression strategy for distributional-valued data. Clusterwise Regression (CR) methods aim at identifying both the partition of a set of data in a fixed number of clusters and regression models as representative elements of the clusters. A pioneering paper for the search of local models for clustered data is the Typological Principal Component Analysis (Diday, 1974). It carries out K sub-spaces of maximal inertia assigning elements to the clusters according to the minimum distances from the local factorial planes, until the convergence to a stable partition and to K final subspaces. Späth (Späth, 1979, Späth, 1991) introduced a criterion for partitioning a set of objects into K classes establishing a regression model within each class. Preda & Saporta, 2005 use PLS regression for solving an ill-posed problem in clusterwise regression. Morever, DeSarbo & Cron, 1988, Hennig, 2000 proposed mixture-model-based clusterwise regression. They assume that the response variable estimations, related to the clusters, are obtained as mixtures of K conditional density distributions.

In this framework, we propose a Cluster-wise Regression method for Distributional data (CRD). The latter are a particular kind of multi-valued Symbolic Data, like: intervals, multi-categoricals, histograms or continue distributions (Bock & Diday, 2000). Many exploratory statistical methods have been extended to such data, especially considering them as suitable aggregated data. These are assuming more and more relevance for the treatment of high dimensional data. Among the methods proposed in Symbolic Data Analysis context, a CR method for interval data was presented by De Carvalho *et al.*, 2010. It performs a double regression on the centers and on the radii of the intervals, recalling a suitable strategy for interval data analysis. Recently De Carvalho *et al.*, 2021 have developed a non linear clusterwise regression which extends the previous proposal. A prediction model based on CR for data aggregated as empirical distributions was proposed by Suresh *et al.*, 2020.

Our method aims at clustering distributional-valued data in K clusters according to a local dependence structure between distributional variables. Consistently with the K-means algorithm the centroids of the clusters are expressed as ordinary least squares (OLS) regression models and the objects are assigned to the clusters assuming as criterion the minimum increasing of sum of the squared errors. Related to the type of variables, the generalised CR algorithm is based on a linear regression model (Irpino & Verde, 2015) and on a K-means algorithm (Irpino & Verde, 2007) for distributional data; both these methods use the L_2 Wasserstein distance (Wasserstein, 1969) as measure of distance between distributions. Moreover, we propose to determine the optimal number of clusters K according to a criterion of global best fitting of the cluster regression models. In the same way, a selection of the best explanatory variables, for each cluster regression model, is carried out in order to improve the prediction of the dependent variable in each cluster. The final achieved cluster regression models are evaluated using root-mean-square error (RMSE), goodness of fit R^2 index and the Pseudo-R^2 index. For sake of brevity, we have omitted some promising results obtained on real and synthetic distributional data sets.

2 Clusterwise Regression for Distributional-valued data (CRD)

Let $W = \{w_1, \dots, w_N\}$ be a set of N objects described by $p+1$ distributional-valued variables. We assume that one of the $p+1$ distributional-valued variables, denoted by Y, is a dependent variable from the p explanatory variable X_j $(j = 1, \dots, p)$. Each object w_i $(1 \leq i \leq N)$ is represented by $p+1$ distributions (or distributional-valued data): f_i^y, f_{ij}^x $(j = 1, \dots, P)$. The CRD method looks for clustering the data set W into K clusters according to the best fitting regression model for each cluster. The regression model used to fit clustering distributional data was introduced by Irpino & Verde, 2015, as follows:

$$y_i(t) = \beta_0 + \sum_{j=1}^{p} \beta_j \bar{x}_{ij} + \sum_{j=1}^{p} \gamma_j x_{ij}^c(t) + e_i(t), \quad \forall t \in [0,1] \tag{1}$$

where: β_0 is the constant, β_j are the coefficients associated with the vectors of the averages $\bar{x}_{ij}$ of each distribution f_{ij}; γ_j are the coefficients of the centred quantile functions x_{ij}^c $(j = 1, \dots, p)$.

The Sum of Square Errors function (SSE), like in LS method, is computed using the L_2 Wasserstein distance.
Fixed the number K of clusters, CR algorithm seeks the better partition $P_k = \{C_1, \dots, C_K\}$ and the best fitting models $\hat{y}^k$ for each cluster C_k by minimising:

$$SSE(\beta_0^k, \beta_j^k, \gamma_j^k | P_k) = \sum_{k=1}^{K} \sum_{i \in C_k} \int_0^1 [y_i^k(t) - (\beta_0^k + \sum_{j=1}^{p} \beta_j^k \bar{x}_{ij} + \sum_{j=1}^{p} \gamma_j^k x_{ij}^c(t))]^2 dt \tag{2}$$

An element w_i is assigned to a cluster C_k according to the minimum squared distance $\hat{e}_{ik}^2$ from the estimated regression model $\hat{y}^k$:

$$min_k: \ \hat{e}_{ik}^2 = \int_0^1 [y_i(t) - (\hat{\beta}_0^k + \sum_{j=1}^{p} \hat{\beta}_j^k \bar{x}_{ij} + \sum_{j=1}^{p} \hat{\gamma}_j^k x_{ij}^c(t))]^2 dt \tag{3}$$

The convergence of the algorithm is guaranteed by the criterion decreasing related to the improvement of the best fitting of the cluster regression models.

We consider two indexes to evaluate the goodness of fit of the clusterwise regressions: the Ω index proposed by Dias & Brito, 2015, and the $RMSE_W$ (Root Mean Square Error, according to the L_2 Wasserstein distance), computed for each cluster (denoted as Ω^k and $RMSE_W(C_k)$), and the total $RMSE_W(P_k)$ for the entire partition P_k. To determine the best number K of clusters of the partition P_k, we consider the Root Mean Square Error $RMSE_W(P_k)$ as a measure of total within variability of the clusters. According to the elbow method, we choose the number of clusters such that adding another cluster does not lead to an important decrease of the total $RMSE_W(P_k)$. Finally, a forward selection of the explanatory variables allows of defining the best cluster regression models as well as the variables which affect the prediction of the response variable the most. It is worth noticing that the regression models can differ in the importance of the predictors from one cluster to another. The more different are the estimated cluster regression models the more the linear relations in the clusters of the partition are different for distinct observed data subsets

References

Bock, H.-H., & Diday, E. 2000. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Heidelberg: Springer Verlag.

De Carvalho, F. A.T., Saporta, G., & Queiroz, Danilo N. 2010. A Clusterwise Center and Range Regression Model for Interval-Valued Data. *In: Proc. of COMPSTAT'2010*.

De Carvalho, F. A.T., Lima Neto, Eufrásio de A., & da Silva, Kassio C.F. 2021. A clusterwise nonlinear regression algorithm for interval-valued data. *Information Sciences*, **555**, 357–385.

DeSarbo, W.S., & Cron, W.L. 1988. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, **5**, 249–282.

Dias, S., & Brito, P. 2015. Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining*, **8**(2), 75–113.

Diday, E. 1974. Introduction aĺ'analyse factorielle typologique. *Revue de Statistique Appliqueé*, **22**(4), 29–38.

Hennig, C. 2000. Identifiability of models for Clusterwise linear regression. *Journal of Classification*, **17**, 273–296.

Irpino, A., & Verde, R. 2007. *Dynamic Clustering of Histogram Data: Using the Right Metric*. Berlin: Springer Verlag.

Irpino, A., & Verde, R. 2015. Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein Distance. *Advances in Data Analysis and Classification*, **9**(1), 81–106.

Preda, C., & Saporta, G. 2005. Clusterwise PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, **49**(1), 99–108.

Späth, H. 1979. Clusterwise linear regression. *Computing*, **22**, 367–373.

Späth, H. 1991. Agorithm 48: A fast algorithm for clusterwise linear regression. *Computing*.

Suresh, N., Brito, P., & Dias, S. 2020. Prediction of pollution levels from atmospheric variables A study using clusterwise symbolic regression. *In: Proc. RECPAD'20*.

Wasserstein, L. 1969. Markov processes over denumerable products of spaces describing large systems of automata. *rob. Inf. Transm.*, **5**, 47–52.

A MACHINE LEARNING APPROACH FOR EVALUATING ANXIETY IN NEUROSURGICAL PATIENTS DURING THE COVID-19 PANDEMIC

Vezzoli M.[1], Doglietto F.[2], Renzetti S.[1], Fontanella M.M.[2], Calza S.[1]

[1] Department of Molecular and Translational Medicine, University of Brescia,

(e-mail: marika.vezzoli@unibs.it, stefano.renzetti@unibs.it, stefano.calza@unibs.it)

[2] Neurosurgery, Department of Medical and Surgical Specialties, Radiological Sciences and Public Health, University of Brescia,

(e-mail: francesco.doglietto@unibs.it, marco.fontanella@unibs.it)

ABSTRACT: In 2020, the COVID-19 pandemic has forced many countries into lockdown postponing nonurgent neurosurgical procedures. After the lockdown, neurosurgical patients admitted to eastern Lombardy hospitals, filled pre- and postoperative questionnaires which measured anxiety (State Anxiety Inventory) related to COVID-19, and safety perception during hospital admission. These data were merged with information on age, sex, primary pathology, and time on surgical waiting list. By means of Random Forest, Variable importance measure and Partial Dependence Plots, we identified which variables had a strong impact on anxiety, and safety perception. Results highlighted that worry about positivity to SARS-CoV-2 was associated with anxiety. Bed distance and hand sanitizer were associated with a feeling of safety.

KEYWORDS: COVID-19, random forest, variable importance measure, partial dependence plot

1 Introduction

In 2020, the COVID-19 pandemic forced Italy and many other countries over the world into lockdown. In that period, in Lombardy nonurgent neurosurgical procedures were rescheduled from the end of May 2020.

Although stress and anxiety during the COVID-19 pandemic is being investigated in general population (Gasteiger, 2021), no studies investigated anxiety in patients whose neurosurgery has been postponed.

The aim of this study was to investigate anxiety in neurosurgical patients undergoing nonurgent surgical procedures in the post-lockdown phase of the COVID-19 pandemic. Moreover, we inspected safety perception from SARS-CoV-2 infection during hospitalization. Data of various nature (qualitative and quantitative), including state anxiety, were collected in hospitals mainly located in eastern Lombardy, an area in Italy extremely affected by COVID-19. Since during COVID-19 period the percentage of anxious patients that must undergo surgery is 25%, the study will require 100 patients for estimating the expected proportion with 8.5% accuracy (95%

CI). The study was approved by the local ethics committee (Study n. 4290; COVID-SAFENSG).

2 Methods

2.1 Inclusion criteria, questionnaires and clinical Data

Inclusion criteria for the study were: adult patients (>18) undergoing nonurgent neurosurgical procedures who consented to study participation. Each patient filled in 3 questionnaires: 2 before surgery and 1 after. The first questionnaire collected demographic data (age, sex, and highest academic degree), days of postponement of the surgery, fear related to disease, COVID-19 and hospitalization (measured on a Likert scale from 1 (not at all) to 10 (very)).

The second questionnaire, widely used and validated in many languages, was the State Anxiety Inventory (STAI-State) (Spielberger, 2010), which contains 20 questions on a 4-point Likert scale. It measures the latent constructs of state anxiety related to an event in a specific moment, such as a surgical procedure. Each item belonging to this questionnaire has a range from a minimum of 1 to a maximum of 4 points, hence the score ranges from a minimum of 20 to a maximum of 80. In detail:

- $20 \leq$ STAI-State score < 48: Normal
- $48 \leq$ STAI-State score ≤ 52: Mild
- $52 <$ STAI-State score ≤ 80: Severe

The last questionnaire collected patients' impressions (Likert scales from 1 (not at all) to 10 (very)) on safety from SARS-CoV-2 infection during hospitalization. First and third questionnaires were tested at the beginning of June 2020 on an external and independent sample of 30 subjects in order to improve the questions' semantics and their comprehension. Answers were collected with REDCap, a secure web application for building and managing online surveys and databases. Clinical data, provided by the neurosurgeon in charge of the patient, included among others, prolongation of time on the waiting list and postponement of hospital admission.

2.2 Machine learning approach

Two different models were used to identify which covariates (X) have the greatest impact on the outcomes (Y_1 *and* Y_2 which are ordinal variables). Since variables were qualitative and quantitative, mostly asymmetrical, and related to Y by nonlinear relationships, the Random Forest (RF; Breiman, 2001) was applied, and, for each model, 10,000 regression trees were grown. In detail:

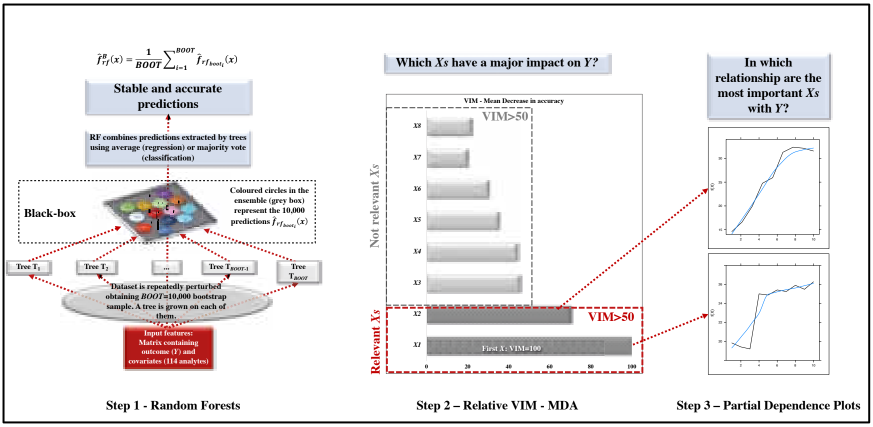
1. **RF1**: STAI-State (Y_1) was modeled to investigate which concerns (Table 1, column 1), in the preoperative questionnaire, have a primary role on it.
2. **RF2**: The question "*How much did you feel protected from the risk of being infected with Coronavirus during your hospital admission in Neurosurgery?*" (Y_2), collected on a Likert scale from 1 (not at all) to 10 (very), was modeled using items in the post-operative questionnaire (Table 1, column 2).

To highlight the relationships between covariates and outcomes, two additional methods were used: (*i*) relative Variable Importance Measure (relVIM: Vezzoli, 2011) which identifies covariates that most impact on the prediction of *Y* (VIM > 50); (*ii*) partial dependence plots[5] (PDPs; Friedman, 2001) which visualize the functional relationship between the selected covariates at point (*i*) and the RF predictions. Figure 1 provides a visual summary of this three-step procedure. Analyses were performed with R 4.0.1.

Table 1. Covariates used in RF1 (left), RF2 (right) models

RF1 covariates	RF2 covariates
Worried for positivity to Coronavirus	Feeling of safety due to distance between beds
How worried are you about the pathology for which you have been admitted?	Feeling of safety due to hand sanitizer gel available in hospital
How much are you worried about the surgical procedure?	Feeling of safety due to health personnel following security protocols
How anxious were you about a possible worsening of your condition?	Feelings of safety due to the procedures to prevent infection from COVID-19
How stressed were you during the waiting time to admission?	Feeling of safety due to measure body temperature at hospital entrance
Age	Feeling of safety due to masks
Becoming positive to COVID-19 during hospitalization	Feeling of safety due to a reassuring behavior of health personnel
How many days would you have been willing to post-pone your admission?	Feeling of safety due to sanitization of hospital environments
How safe do you feel in Neurosurgical ward?	In the operating room, did you feel safe from Coronavirus
Perception of time from neurosurgical evaluation to admission	Did the health personnel seem prepared for the post-operative period
How much COVID-19 increased concern about Neurosurgery admission?	
How useful is the screening on COVID-19 performed pre-operatively?	
How safe is the screening on COVID-19 performed pre-operatively?	

Figure 1. Three-step procedure based on Random Forest, VIM and PDP

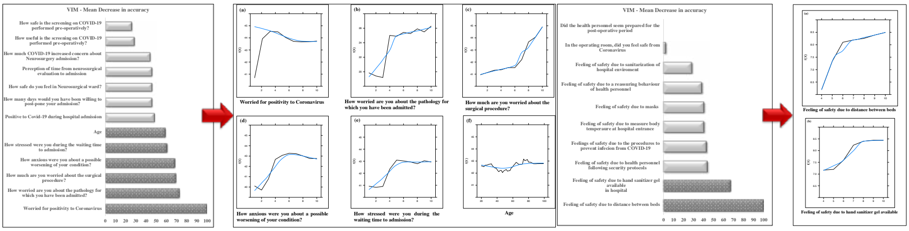

$\hat{f}^{B}_{rf}(x) = \frac{1}{BOOT}\sum_{i=1}^{BOOT} \hat{f}_{rf_{boot_i}}(x)$
Stable and accurate predictions
RF combines predictions extracted by trees using average (regression) or majority vote (classification)
Black-box
Coloured circles in the ensemble (grey box) represent the 10,000 predictions $\hat{f}_{rf_{boot_i}}(x)$
Tree T_1
Tree T_2
...
Tree T_{BOOT-1}
Tree T_{BOOT}
Dataset is repeatedly perturbed obtaining *BOOT*=10,000 bootstrap sample. A tree is grown on each of them.
Input features: Matrix containing outcome (*Y*) and covariates (114 analytes)
Step 1 - Random Forests
Which *Xs* have a major impact on *Y?*
VIM - Mean Decrease in accuracy
VIM>50
Not relevant *Xs*
Relevant *Xs*
X8
X7
X6
X5
X4
X3
X2
X1
First *X*: VIM=100
VIM>50
0
20
40
60
80
100
Step 2 – Relative VIM - MDA
In which relationship are the most important *Xs* with *Y*?
f(X)
30
25
20
15
2
4
6
8
10
f(X)
26
24
22
20
2
4
6
8
10
Step 3 – Partial Dependence Plots

3 Results and discussion

After exclusion of 11 patients due to significant missing data, 123 subjects (M/F, 64/59; mean age 60.28 (SD=15.08) years were included in the study, for 114 variables. Modeling state anxiety (STAI-State, RF1 in Fig. 2 on the left), the patients' condition was significantly associated with the worry of being positive for SARS-CoV-2. This was the first variable identified by VIM, followed by intuitive ones such as the concern for the primary pathology, surgery, and worsening of their condition, as well as waiting time. In fact, hospital admission to neurosurgery was postponed in mean of 49.72 days and it was due to organizational issues (83%) or, rarely, for positivity to SARS-CoV-2 (1.6%). Our data confirm that psychological support should be enhanced during outbreaks, possibly using novel solutions to provide follow-up care remotely during waiting times.

This study also investigated the feeling of safety conveyed by different features that were activated in all Italian hospitals during the pandemic (RF2, in Fig. 2 on the right). Interestingly, the increased distance between surgical beds was the first factor associated with a feeling of safety from SARS-CoV-2, followed by the availability of hand sanitizers. These data might be interpreted as a result of the ongoing social media communication on the importance of social distancing; we believe they might be important for hospital managers and to optimize communication with patients during this pandemic.

Figure 2: Results from RF1 and RF2



References

BREIMAN, L. 2001. Random Forests. *Mach. Learn.* **45**, 5-32.

FRIEDMAN, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189-1232.

GASTEIGER, N. *et al.* 2021. Depression, anxiety and stress during the COVID-19 pandemic: results from a New Zealand cohort study on mental well-being. *BMJ Open*, **11**, 1-16.

SPIELBERGER, C.D. 2010. State-Trait Anxiety Inventory. in *The Corsini Encyclopedia of Psychology*. John Wiley & Sons, Inc, Hoboken.

VEZZOLI, M. 2011. Exploring the facets of overall job satisfaction through a novel ensemble learning. *Electron. J. Appl. Stat. Anal.* **4**, 23-38.

PREDICTION OF LARGE OBSERVATIONS VIA BAYESIAN INFERENCE FOR EXTREME-VALUE THEORY

Isadora Antoniano Villalobos[1], Simone Padoan[2] and Boris Beranger[3]

[1] Ca' Foscari University of Venice, (e-mail: isadora.antoniano@unive.it)

[2] Bocconi University, (e-mail: simone.padoan@unibocconi.it)

[3] University of New South Wales, (e-mail: b.beranger@unsw.edu.au)

ABSTRACT: In many applications placing interest on large observations, usual inferential methods may fail to reproduce the heavy tail behaviour of the quantities involved. Recent literature has proposed the use of multivariate extreme value theory to predict an unobserved component of a random vector given large observed values of the rest. This is achieved through the estimation of the angular measure controlling the dependence structure in the tail of the distribution. The idea can be extended and used for effective data imputation and prediction of multiple components at adequately large levels, provided the model used for the angular measure is flexible enough to capture complex dependence structures. A Bayesian nonparametric model based on constrained Bernstein polynomials ensures such flexibility. Tractable inference for both the dependence structure and the marginal parameters of the model is achieved via a trans-dimensional MCMC algorithm for posterior simulation.

KEYWORDS: Bernstein polynomials, extremal dependence, multivariate regular variation, trans-dimensional MCMC

COMMUNITY DETECTION IN TRIPARTITE NETWORKS OF UNIVERSITY STUDENT MOBILITY FLOWS

Vitale Maria Prosperina[1] , Vincenzo Giuseppe Genova[2], Giuseppe Giordano[1]
and Giancarlo Ragozini[3]

[1] Department of Political and Social Studies, University of Salerno, (e-mail: mvitale@unisa.it, ggiordano@unisa.it)

[2] Department of Economics, Business, and Statistics, University of Palermo, (e-mail: vincenzogiuseppe.genova@unipa.it)

[3] Department of Political Science, Federico II University of Naples, (e-mail: giragoz@unina.it)

ABSTRACT: The purpose of this study is to explore how the multimode network approach can be used to analyse network patterns derived from student mobility flows. We define a tripartite network based on a three-mode data structure, consisting of Italian provinces of residence, universities and fields of study, with student exchanges representing the links between them. A comparison of algorithms for detecting communities from tripartite networks based on modularity optimization is provided, revealing relevant information about the phenomenon under analysis over time. The findings are applied to a real dataset containing micro-level longitudinal information on Italian university students' careers.

KEYWORDS: student mobility, tripartite networks, modularity optimisation

1 Introduction

The analysis of intra- and international student mobility has become a vibrant research field in migration literature and a key concern for national policy-making on tertiary education systems (Van Mol & Timmerman, 2014; Riaño *et al.*, 2018). Usually, European mobility in higher education is described by considering the dynamics of the Erasmus programme. From a national perspective, Italian student mobility from high school to bachelor and master degrees is analysed as a crucial step in determining future migration choices. Such analysis shows an unbalanced migration of students from the southern to the northern regions of the country (Genova *et al.*, 2019), which is influenced by the attractiveness of universities, related to the socio-economic characteristics and the job market opportunities in the geographic areas where they are located (Giambona *et al.*, 2017; Impicciatore & Panichella, 2019). Given the nature of the student mobility data (i.e. flows of students connecting provinces of residence and universities of destination), network analysis has been adopted as one of the most appropriate methodological approach to interpret this phenomenon (Santelli *et al.*, 2019; Genova *et al.*, 2019; Columbu *et al.*, 2021). Based on this theoretical framework and the intrinsic complexity of student mobility flows, this study analyses the data at hand using the framework of multimode networks (Fararo & Doreian, 1984). More specifically, we define a tripartite network based on a three-mode data structure, consisting of Italian provinces of residence, universities and fields of study, with student exchanges representing the links between them. A comparison of algorithms for detecting communities from tripartite networks or k-partite modularity (Neubauer & Obermayer, 2009; Ikematsu & Murata, 2013; Melamed *et al.*, 2013; Ignatov *et al.*, 2017; Feng *et al.*, 2019), mainly based on modularity optimisation, is applied to reveal relevant information about the phenomenon under analysis. The algorithms are applied to the MOBYSU.IT dataset which contains micro-level longitudinal information on university students' careers from 2008 to 2017 in Italy.[*]

2 Community detection algorithms in tripartite networks

Many real-world networks have a natural multimode network structure in which vertices of different types are linked together. Without reducing generalisability, in the case of tripartite networks, three types of vertices are defined and links can be present only between vertices of distinct types (Fararo & Doreian, 1984). Several approaches can be pursued to disentangle the inherent complexity of such kinds of data. Recently, Everett & Borgatti (2019) suggested that, in the case of multimode data, the collection of all bipartite networks should be examined.

In our case study, a tripartite network is considered in which $\mathscr{V}_P$ is the set of provinces of residence of Italian students enrolled in the first academic year of any bachelor/master degree, $\mathscr{V}_U$ is the set of public and private universities, and $\mathscr{V}_F$ is the set of educational fields of study. The tripartite network $\mathscr{T}$ can be defined as consisting of a pair $(\mathscr{V}, \mathscr{E})$, being $\mathscr{V} = \{\mathscr{V}_P, \mathscr{V}_U, \mathscr{V}_F\}$ the collection of three sets of vertices, one for each mode, and being $\mathscr{E} = \{\mathscr{E}_{PUF}\}$, $\mathscr{E}_{PUF} \subseteq \mathscr{V}_P \times \mathscr{V}_U \times \mathscr{V}_F$, with $\mathscr{E}_{PP}, \mathscr{E}_{UU}, \mathscr{E}_{FF} = \emptyset$, the collection of links among

[*]This study was supported by the Italian Ministerial grant PRIN 2017 'From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide', n. 2017HBTK5P - CUP B78D19000180001.

the vertices belonging to the three modes. Given $\mathscr{T}$, a unique supra-adjacency matrix $\mathbb{A}$ could be defined by combining the sociomatrices in a block matrix $\mathbf{A}_{PU}$, $\mathbf{A}_{UF}$, and $\mathbf{A}_{PF}$, where the links are the number of students enrolled, and the corresponding bipartite networks are weighted. Thus, the related supra-adjacency matrix is:

$$\mathbb{A} = \begin{bmatrix} \mathbf{0} & \mathbf{A}_{PU} & \mathbf{A}_{PF} \\ \mathbf{A}_{PU}^T & \mathbf{0} & \mathbf{A}_{UF} \\ \mathbf{A}_{PF}^T & \mathbf{A}_{UF}^T & \mathbf{0} \end{bmatrix}.$$

Over the past two decades, a growing number of studies have been devoted to community detection algorithmic solutions in tripartite graphs. The first and simplest proposed method consists of applying on the matrix $\mathbb{A}$, or on its version built up after matrices' transformation, the usual community detection algorithms (Melamed *et al.*, 2013; Everett & Borgatti, 2019). Other methods adopting an optimisation of tripartite networks (Neubauer & Obermayer, 2009; Ikematsu & Murata, 2013), extending the idea of bipartite modularity.

Given the nature of our data, the approaches which maximise the bipartite modularity seem more appropriate. A detailed comparison of proposed algorithms could be of interest in understanding how tripartite community detection can be used to interpret the network patterns underlying the Italian student mobility phenomenon.

References

COLUMBU, SILVIA, PORCU, MARIANO, & SULIS, ISABELLA. 2021. University choice and the attractiveness of the study area: Insights on the differences amongst degree programmes in Italy based on generalised mixed-effect models. *Socio-Economic Planning Sciences*, **74**, 100926.

EVERETT, MARTIN G, & BORGATTI, STEPHEN P. 2019. Partitioning multimode networks. *Pages 251–265 of: Advances in network clustering and blockmodeling*. John Wiley and Sons.

FARARO, THOMAS J, & DOREIAN, PATRICK. 1984. Tripartite structural analysis: Generalizing the Breiger-Wilson formalism. *Social Networks*, **6**(2), 141–175.

FENG, LIANG, ZHAO, QIANCHUAN, & ZHOU, CANGQI. 2019. An efficient method to find communities in K-partite networks. *Pages 534–535 of: 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE.

GENOVA, VINCENZO GIUSEPPE, TUMMINELLO, MICHELE, ENEA, MARCO, AIELLO, FABIO, & ATTANASIO, MASSIMO. 2019. Student mobility in higher education: Sicilian outflow network and chain migrations. *Electronic Journal of Applied Statistical Analysis*, **12**(4), 774–800.

GIAMBONA, FRANCESCA, PORCU, MARIANO, & SULIS, ISABELLA. 2017. Students mobility: Assessing the determinants of attractiveness across competing territorial areas. *Social indicators research*, **133**(3), 1105–1132.

IGNATOV, DMITRY I, SEMENOV, ALEXANDER, KOMISSAROVA, DARIA, & GNATYSHAK, DMITRY V. 2017. Multimodal clustering for community detection. *Pages 59–96 of: Formal Concept Analysis of Social Networks*. Springer.

IKEMATSU, KYOHEI, & MURATA, TSUYOSHI. 2013. A fast method for detecting communities from tripartite networks. *Pages 192–205 of: International Conference on Social Informatics*. Springer.

IMPICCIATORE, ROBERTO, & PANICHELLA, NAZARENO. 2019. Internal migration trajectories, occupational achievement and social mobility in contemporary Italy. A life course perspective. *Population, Space and Place*, **25**(6), e2240.

MELAMED, DAVID, BREIGER, RONALD L, & WEST, A JOSEPH. 2013. Community structure in multi-mode networks: Applying an eigenspectrum approach. *Connections*, **33**(1), 1823.

NEUBAUER, NICOLAS, & OBERMAYER, KLAUS. 2009. Towards community detection in k-partite k-uniform hypergraphs. *Pages 1–9 of: Proceedings of the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs*.

RIAÑO, YVONNE, VAN MOL, CHRISTOF, & RAGHURAM, PARVATI. 2018. New directions in studying policies of international student mobility and migration. *Globalisation, Societies and Education*, **16**(3), 283–294.

SANTELLI, FRANCESCO, SCOLORATO, CONCETTA, & RAGOZINI, GIANCARLO. 2019. On the determinants of student mobility in an interregional perspective: a focus on Campania region. *Statistica Applicata - Italian Journal of Applied Statistics*, **31**(1), 119–142.

VAN MOL, CHRISTOF, & TIMMERMAN, CHRISTIANE. 2014. Should I stay or should I go? An analysis of the determinants of intra-European student mobility. *Population, Space and Place*, **20**(5), 465–479.

CAUSAL REGULARIZATION

Ernst C. Wit [1], Lucas Kania[1]

[1] Institute of Computing, Università della Svizzera italiana, (e-mail: wite@usi.ch, lucas.kania@usi.ch)

ABSTRACT: When predicting a response variable from a set of covariates, the ordinary least squares estimator (OLS) provides the best in-sample risk but with limited out-of-sample guarantees. Conversely, the causal parameters provide the best out-of-sample guarantees but the worst in-sample risk. Based on the causal Dantzig and Anchor Regression, we develop a *causal regularization* approach that interpolates between then the OLS and the causal Dantzig solutions. As the regularization is increased, we prove that causal regularization provides a solution that has better out-of-sample risk guarantees at the cost of increasing the in-sample risk. Moreover, we provide an efficient algorithm to recover the regularized solution for every tuning parameter.

KEYWORDS: causal regularization, causal Dantzig, anchor regression, out-of-sample risk.

1 Introduction

We will consider a causal graphical model, for example expressed by Figure 1a (Pearl, 2009). As we are interested in uncovering the causal structure involving a particular *target variable* Y, in particular, in identifying the causal parents of Y and the associated causal parameters β_{PA}.

Besides having access to observational data on the system, we will also assume that we have data on the some intervened version of the same system. We will refer to such intervened system as an *environment*. Formally, given a causal DAG D, such as in Figure 1a, for a probability distribution P over random variables (X,Y). The tuple (D,P^e,X^e,Y^e,A^e) for $e \in \varepsilon$ is called an environment, where A^e is the set of shift-intervention variables in D^e, the extended intervention graph of D for environment e, such as for example in Figure 1b.

For simplicity, we focus on a particular structure of the distribution P, described by means of a linear structural equation model (SEM), also known as linear structural causal model. In particular, for $e \in \varepsilon$, let the distribution P^e of

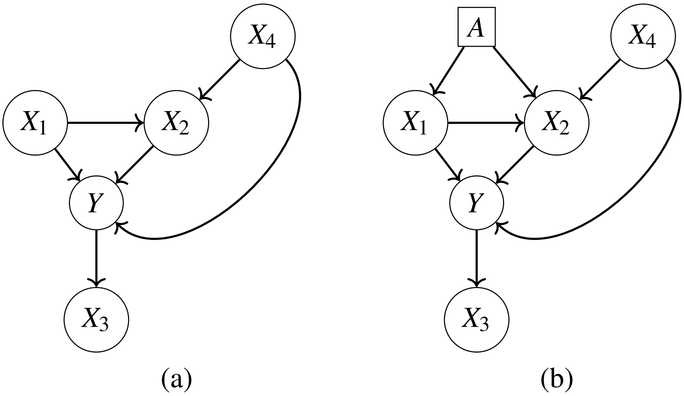


Figure 1. *(a) Causal directed acyclic graph D associated with a causal graphical model. (b) An extended intervention graph D^e associated with this causal GM.*

(X^e, Y^e, A^e) be determined by the solution of the system

$$\begin{bmatrix} Y^e \\ X^e \end{bmatrix} = \underbrace{B}_{\substack{\text{unknown}\\\text{constant}\\\text{structure}}} \cdot \begin{bmatrix} Y^e \\ X^e \end{bmatrix} + \underbrace{\varepsilon^e}_{\text{noise}} + \underbrace{A^e}_{\substack{\text{shift}\\\text{intervention}}} \tag{1}$$

where $B \in \mathbb{R}^{(p+1)\times(p+1)}$ is a constant matrix, and $X^e \in \mathbb{R}^p$, $Y^e \in \mathbb{R}$, $\varepsilon \in \mathbb{R}^{p+1}$ and $A^e \in \mathbb{R}^{p+1}$ are random vectors. We require $A^eY \equiv 0$, i.e., $A^e = \begin{bmatrix} 0 \\ A^eX \end{bmatrix}$, that the target variable is not intervened on. Interventions and noise variables must be uncorrelated, $\mathrm{Cor}[A^e, \varepsilon^e] = 0$, and have finite second moments $E[A^eA^{eT}] < \infty$ and $\mathrm{Cor}[\varepsilon^e] < \infty$. Furthermore, ε^e is assumed to have zero-mean, i.e. $E[\varepsilon^e] = 0$. Additionally, the noise random variables are assumed to be identically distributed across environments, i.e., $\varepsilon^e \sim \zeta$. Moreover, for the distribution to be well defined, we ask for the existence of $(I-B)^{-1}$ so that $\begin{bmatrix} Y^e \\ X^e \end{bmatrix} = (I-B)^{-1}(\varepsilon^e + A^e)$. This is guaranteed if the underlying graph D is a directed acyclic graph.

Given that the structure B is fixed across environments, we can talk about $X_S \subseteq X$ being a descendant or ancestor of Y without referring to the environmental variables Y^e and X^e. Moreover, since we are interested is estimating the structural equation corresponding to Y^e, it is useful to split B into

$$B = \begin{bmatrix} 0 & \beta_{PA}^T \\ \beta_{CH} & Bx \end{bmatrix} \tag{2}$$

Consequently, the structural equation of Y^e would be

$$Y^e = \beta_{PA}^T X^e + \varepsilon_Y^e \tag{3}$$

where β_{PA} are called the *causal parameters* since they are non-zero only for $X^e_{pa(Y)}$. In the SEM context, the components of A^e are called shift-interventions or interventions. If $A^e X_i \not\equiv 0$ for $i \in \{1, \ldots, p\}$, we say that X_i is intervened. Thus, note that assuming $A^e Y \equiv 0$ means that no interventions is performed on the target, which is the equivalent of assuming $E \notin pa(Y)$. When $A^e \equiv 0$, the environment is called *observational*.

2 Discovering causes from inner-product invariance

Under the SEM in equation (3), the following *distribution invariance* holds (Peters *et al.*, 2016) $\forall e \in \varepsilon: \; Y^e - \beta_{PA}^T X^e = \varepsilon_Y^e \sim \zeta$, which we call *residual invariance*. Furthermore, by left multiplying with X^e and taking the expectation, we obtain,

$$\begin{aligned} \forall e \in \varepsilon : E[X^e(Y^e - \beta_{PA}^T X^e)] &= P_X(I-B)^{-1}(E[\varepsilon^e \varepsilon_Y^e] + E[A^e \varepsilon_Y^e]) \\ &= P_X(I-B)^{-1}\text{Cor}[\zeta, \zeta_Y] \quad \text{constant over } e \end{aligned}$$

which yields *inner-product invariance*. By taking the difference between the expected inner-product of an interventional environment (X^e, Y^e) and an observational one (X^o, Y^o), we obtain

$$E[Z - G\beta_{PA}] = E[Z] - E[G]\beta_{PA} = 0 \tag{4}$$

where $Z = X^e Y^e - X^o Y^o$ and $G = X^e X^{eT} - X^o X^{oT}$. Since $||\alpha||_\infty = 0 \iff \alpha = 0$, we get $||E[Z] - E[G]\beta_{PA}||_\infty = 0$. Thus, equation (4) gives a plausible method for identifying β_{PA} without having search over all possible subsets of X. That is, to solve the following linear regression problem,

$$\beta_{CS} \in \arg\min_{\beta \in \mathbb{R}^p} ||E[Z] - E[G]\beta||_\infty, \tag{5}$$

which is referred to as the unregularized *causal Dantzig problem* (Rothenhäusler *et al.*, 2019). Although β_{PA} is a solution, depending on the rank of $E[G]$, the solution β_{CS} may not be unique. We call $R_{inv}(\beta) = ||E[Z - G\beta]||_\infty$ the invariance risk for β. Let $R^e(\beta) = E[(Y^e - \beta^T X^e)]$ be the risk in environment e and $R_{pred}(\beta) = R^e(\beta) + R^o(\beta)$ the pooled risk of the in-sample environments, then we remind the reader that the OLS problem minimizes the in-sample risk $\beta_{OLS} \in \arg\min_{\beta \in \mathbb{R}^p} R_{pred}(\beta)$.

3 Causal regularization

We define *causal regularization* as an estimator that provides the best possible in-sample risk for a certain out-of-sample risk guarantee, as follows:

$$\beta_{CR}(t) = \arg\min_{\beta \in \mathbb{R}^p} R_{pred}(\beta) \text{ such that } R_{inv}(\beta) \leq t \tag{6}$$

Note that for $t \to \infty$ we recover the OLS solution β_{OLS}, whereas for $t \to 0$ we obtain the Causal Dantzig solution β_{CS}.

Given the in-sample shift environment (X^e, Y^e, A^e), we define a set of environments C_γ such that their interventions only differ in magnitude to the ones contained in the in-sample environment e,

$$C_\gamma = \{f \in \varepsilon : E[A^f A^{fT}] \preceq \gamma E[A^e A^{eT}]\}.$$

The causal regularizer has strong out-of-sample risk guarantees within G_γ.

Theorem. *Causal regularization out-of-sample risk guarantees*
For any CR estimator $\beta \in \beta_{CR}(t)$, we have the following risk bound

$$\sup_{f \in C_{1+\tau}} R^f(\beta) \leq R_{pred}(\beta) + \tau t ||\beta_{PA} - \beta||_1, \tag{7}$$

in particular, $\forall \beta \in \beta_{CR}(t): \ \sup_{f \in C_{1+1/t}} R^f(\beta) \leq \underbrace{R_{pred}(\beta) + ||\beta_{PA} - \beta||_1}_{\text{Constant}}.$

The theorem tells us that if we expect out-of-sample environments to have interventions that are τ times stronger than in the in-sample environment e, then setting $t = \tau^{-1}$ would provide an estimator that guarantees a bounded risk on such environments. In other words, $\beta \in \beta_{CR}(t)$ guarantees a bounded out-of-sample risk for environments in $C_{1+1/t}$. Particularly, β_{CS} provides a bounded out-of-sample risk for the *biggest* set of environments, i.e., C_∞, while β_{OLS} guarantees a bounded out-of-sample risk for environments whose interventions are at most as strong as the intervention present in environment e, i.e., C_1.

References

PEARL, JUDEA. 2009. *Causality*. Cambridge university press.

PETERS, J, BÜHLMANN, P, & MEINSHAUSEN, N. 2016. Causal inference by using invariant prediction. *JRSS-B (Statistical Methodology)*, 947–1012.

ROTHENHÄUSLER, D, BÜHLMANN, P, & MEINSHAUSEN, N. 2019. Causal dantzig: fast inference in linear structural equation models with hidden variables. *The Annals of Statistics*, **47**(3), 1688–1722.

MINIMIZING CONFLICTS OF INTEREST: OPTIMIZING THE JSM PROGRAM

Qiuyi Wu[1] and David Banks[2]

[1] University of Rochester, (e-mail: jqiuyi_wu@urmc.rochester.edu)

[2] Duke University, (e-mail: dlbanks@duke.edu)

ABSTRACT: Sometimes the Joint Statistical Meetings (JSM) are frustrating to attend, because multiple sessions on the same topic are scheduled at the same time. This paper uses seeded Latent Dirichlet Allocation and a scheduling optimization algorithm to very significantly reduce overlapping content in the 2020 program. Of course, since the pandemic forced the 2020 JSM to be held virtually, our superior schedule was made moot. Nonetheless, this approach may assist in organizing future meetings, both for statistics and for other disciplines.

KEYWORDS: latent Dirichlet allocation, topic modeling, greedy algorithm, scheduling

Contributed Papers

Giovanni C. Porzio, University of Cassino and Southern Lazio, Italy, porzio@unicas.it, 0000-0003-1208-6991
Carla Rampichini, University of Florence, Italy, carla.rampichini@unifi.it, 0000-0002-8519-083X
Chiara Bocci, University of Florence, Italy, chiara.bocci@unifi.it, 0000-0001-8189-4445

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Giovanni C. Porzio, Carla Rampichini, Chiara Bocci (edited by), *CLADAG 2021 Book of abstracts and short papers. 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-340-6 (PDF), DOI 10.36253/978-88-5518-340-6

MODEL SELECTION PROCEDURE FOR MIXTURE HIDDEN MARKOV MODELS

A. Abbruzzo [1], M.F. Cracolici [1] and F. Urso[1]

[1] Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy, (e-mail: antonino.abbruzzo@unipa.it, mariafrancesca.cracolici@unipa.it, furio.urso@unipa.it)

ABSTRACT: This paper proposes a model selection procedure to identify the number of clusters and hidden states in discrete Mixture Hidden Markov models (MHMMs). The model selection is based on a step-wise approach that uses, as score, information criteria and an entropy criterion. By means of a simulation study, we show that our procedure performs better than classical model selection methods in identifying the correct number of clusters and hidden states or an approximation of them.

KEYWORDS: model selection, clusters, hidden states, entropy-based scores, information criteria

1 Introduction

In many research fields, we deal with data whose independent units present one or more categorical sequences that represent the evolution of a specific feature over time (longitudinal data). Thus, it is necessary to define suitable methods capable of modelling an evolving process by describing some unknown variables that influence the observed sequences. Latent class models such as MHMMs can be used to analyse longitudinal data under the assumptions that (i) the sequences follows a latent Markov process and that (ii) the population is heterogeneous (Vermunt *et al.*, 2008; Bartolucci & Pandolfi, 2015). These models present two latent levels: one related to the hidden states of the discrete-time Markov chain and one representing the population's subgroups. The identification of the number of clusters and hidden states can be achieved, according to the literature on Mixture and Hidden Markov models, by fitting different models to the data and then selecting the model by using the results of information criteria (IC) such as AIC and BIC or classification criteria based on entropy (Dias *et al.*, 2009; Crayen *et al.*, 2012). However, these criteria tend to underestimate or overestimate these numbers (Wang & Chan, 2011). Here, we define a model selection procedure that combines IC and an entropy criterion to balance their limitations. Performing a simulation study, we show that the proposed procedure exhibits promising results compared to the classical techniques.

2 Mixture Hidden Markov models

Let $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iT})$ be the generic i-th sequence of length T with card$|Y_i| = R$, $U_i = (U_{i1}, U_{i2}, \dots, U_{iT})$ the i-th hidden random vector with card$|U_i| = S$ and assume n independent sequences. Let $M = \{M^1, M^2, \dots, M^K\}$ be a set of Hidden Markov Models, where $\Theta^k = \{\pi^k, A^k, B^k\}$ is the set of parameters for each sub-models M^k, related to each sub-population $k = 1, \dots, K$. For each sequence Y_i, we define the prior cluster probabilities that the model parameters are the ones related to the k-th sub-model M^k as $P(M^k) = w_k$. Then, the log-likelihood is

$$\ell(\Theta; Y) = \sum_{i=1}^{n} \log P(Y_i|\Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} w_{ik} \sum_{u} \pi^k_{u_1} b^k_{u_1}(y_{i1}) \prod_{t=2}^{T} a^k_{u_{t-1},u_t} b^k_{u_t}(y_{it}) \right), \tag{1}$$

where the hidden state sequences $u = (u_1, u_2, \dots, u_T)$ take all possible combinations of values in the hidden state space S and where y_{it} are the observations of subject i at time t, $\pi^k_{u_1} = P(u_1 = s|\Theta^k)$ with $s \in \{1, \dots, S^k\}$ is the initial probability of the hidden state at time $t = 1$ in sequence u for cluster k; $a^k_{u_{t-1},u_t} = P(u_t = j|u_{t-1} = i, \Theta^k)$ with $i, j \in \{1, \dots, S\}$ is the transition probability from the hidden state at time $t-1$ to the hidden state at t in cluster k; and $b^k_{u_t}(y_{it}) = P(y_{it} = r|u_t = s, \Theta^k)$ with $s \in \{1, \dots, S\}$ and $r \in \{1, \dots, R\}$ is the probability that the hidden state of subject i at time t emits the observed state at t in cluster k. Parameters can be estimated by means of the Expectation-Maximization; and the log-likelihood is calculated by using the forward-backward algorithm.

3 Proposed model selection procedure

Our proposed procedure combines IC and entropy for identifying MHMMs models on the basis of both goodness-of-fit and degree of class separation. Hence, the procedure consists of two stages. Firstly, we estimate models with different number of clusters and states, for each model the IC value is calculated and the models having these values below a predetermined threshold (the mean of the IC) are selected. At the second stage, an entropy criterion is used to identify among the models selected at the first-stage the one with the best degree of separation between classes (clusters and states). At the second stage, when dealing with MHMMs, it is necessary to define a criterion that takes into account two levels of entropy: the first $\mathrm{En}_1(S)$ relating to the classification of observations in latent states and the second $\mathrm{En}_2(K)$ concerning the degree of separation between clusters.

$$\mathrm{E}_{new}(S,K) = 1 - \frac{1}{2n}\left[\frac{\mathrm{En}_1(S)}{T\log S} + \frac{\mathrm{En}_2(K)}{\log K}\right] \tag{2}$$

where

$$\mathrm{En}_1(S) = \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{s=1}^{S_k} P(u_{it} = s|Y_i, M^k)\log P(u_{it} = s|Y_i, M^k),$$

$$\mathrm{En}_2(K) = \sum_{i=1}^{n}\sum_{k=1}^{K} P(M^k|Y_i)\log P(M^k|Y_i).$$

$P(M^k|Y_i)$ is the posterior probability that the given i-th observed sequence has been generated by the k-th model; $P(u_{it} = s|Y_i, M^k)$ is the posterior probability that the t-th element in the i-th hidden sequence takes the s-th hidden states given the observed sequence Y_i and that the sequence has been generated by the model related to the k-th cluster. The $S = \sum_{k=1}^{K} S^k$ is the total number of hidden states in all the K clusters. E_{new} takes value from 0 to 1. Values close to 1 are related to low entropy and a good degree of class separation, values close to 0 are related to a high entropy level and unreliable classification.

4 Simulation study

We compare our procedure of modeling selection to other methods such as AIC, BIC, sample-size adjusted BIC (ssBIC) through a Monte Carlo simulation study. We define 24 scenarios considering 4 models having different number of clusters K and latent states $(S^1, S^2, \dots, S^K)$, by varying the number of sequences $n \in \{200, 2000\}$ and the state-dependent conditional probabilities $b^k_{u_t}(y_{it})$ to represent low, medium, and high levels of uncertainty in hidden states classification of observations. We generate 100 longitudinal datasets for each scenario for a total of 2400 datasets, the analysis is carried out by using the *R* package “seqHMM” (Helske & Helske, 2017). In Table 1 we report methods’ success rate for $n = 2000$, where success means identifying a model having the correct number of clusters K, and number of hidden states equal to the exact number or one from this number. The last column report the results of our procedure considering the AIC as the IC used at the first stage as it showed better results than other IC.

classification uncertainty	$(S^1,S^2,\dots,S^K)$	BIC	AIC	ssBIC	E_{new}	Our Procedure
LOW	$(2,3)$	**1.00** -	0.91 (0.029)	0.96 (0.020)	0.85 (0.036)	0.85 (0.036)
	$(2,2,3)$	0.62 (0.048)	0.40 (0.049)	0.58 (0.049)	0.62 (0.048)	**0.66** (0.047)
	$(2,2,3,3)$	0.30 (0.046)	0.37 (0.048)	0.34 (0.047)	0.21 (0.041)	**0.60** (0.049)
	$(2,2,3,3,2)$	0.19 (0.039)	0.38 (0.048)	0.24 (0.043)	0.19 (0.039)	**0.48** (0.050)
MEDIUM	$(2,3)$	**1.00** -	0.86 (0.035)	**1.00** -	0.49 (0.050)	0.73 (0.044)
	$(2,2,3)$	0.20 (0.040)	0.40 (0.049)	0.29 (0.045)	0.41 (0.049)	**0.59** (0.049)
	$(2,2,3,3)$	0.02 (0.014)	0.35 (0.048)	0.08 (0.027)	0.10 (0.042)	**0.53** (0.049)
	$(2,2,3,3,2)$	0.00 (0.000)	0.12 (0.032)	0.00 (0.000)	0.29 (0.045)	**0.31** (0.046)
HIGH	$(2,3)$	**1.00** -	0.58 (0.049)	**1.00** -	0.46 (0.050)	0.62 (0.048)
	$(2,2,3)$	0.10 (0.030)	0.40 (0.049)	0.14 (0.035)	0.42 (0.049)	**0.59** (0.049)
	$(2,2,3,3)$	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.10 (0.030)	**0.28** (0.045)
	$(2,2,3,3,2)$	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.19 (0.039)	**0.25** (0.043)

Table 1: Results of the Monte Carlo study for $n = 2000$. Low, medium and high level of uncertainty in hidden states classification scenario

As we can see, the proposed procedure has a better performance than the classic IC-based model selection methods when the number of clusters is $K > 2$. We also note how, unlike these methods, it is less affected by an increase in the uncertainty of hidden states' classification.

References

Bartolucci, F., & Pandolfi, S. 2015. LMest: Latent Markov Models with and without Covariates. *R package version.*, **2**.

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., & Vermunt, J. K. 2012. Exploring dynamics in mood regulation—mixture latent Markov modeling of ambulatory assessment data. *Psychosomatic medicine.*, **74**(4), 366–376.

Dias, J. G., Vermunt, J. K., & Ramos, S. 2009. Mixture hidden Markov models in finance research. *Pages 451–459 of: Advances in data analysis, data handling and business intelligence*. Springer.

Helske, S., & Helske, J. 2017. Mixture hidden Markov models for sequence data: The seqHMM package in R.

Vermunt, J. K., Tran, B., & Magidson, J. 2008. Latent class models in longitudinal research. *Handbook of longitudinal research: Design, measurement, and analysis*, 373–385.

Wang, M., & Chan, D. 2011. Mixture latent Markov modeling: Identifying and predicting unobserved heterogeneity in longitudinal qualitative status change. *Organizational Research Methods*, **14**(3), 411–431.

A FULL MIXTURE OF EXPERTS MODEL TO CLASSIFY CONSTRAINED DATA

Ascari Roberto [1] and Migliorati Sonia[1]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, (e-mail: roberto.ascari@unimib.it, sonia.migliorati@unimib.it)

ABSTRACT: This contribution proposes a model-based classifier developed for compositional data. A full mixture of experts model with Dirichlet components is used to incorporate information both on the composition and on a set of covariates. Estimation issues are dealt with by a Bayesian approach, allowing the researcher to use the posterior distribution of the parameters to measure the classification uncertainty.

KEYWORDS: Dirichlet, mixture model, Bayesian, simplex.

1 Introduction

Many fields have witnessed the increasing popularity of compositional data (i.e., vectors representing parts of a whole), which are defined on the D-part simplex $\mathcal{S}^D = \left\{\mathbf{y} = (y_1, \ldots, y_D)^\intercal : y_d > 0, \sum_{d=1}^{D} y_d = 1\right\}$ (Ongaro *et al.*, 2020). Due to the unit-sum constraint imposed by $\mathcal{S}^D$, standard statistical methods are often unsuitable to deal with compositional data. Several ad-hoc proposals have been prompted by mapping the simplex into a different (unconstrained) space, but leaving the simplex often results in interpretative difficulties, especially when the relationship among variables is of interest. This is particularly true in the classification context, where methods for compositional data still present many unsolved issues (Gu & Cui, 2021). In this work, we define a full mixture of experts model (fmem, Bouveyron *et al.*, 2019) and use it to implement a supervised classification algorithm for compositional data in the presence of covariates. Since we adopt a Bayesian approach to inference, we take advantage of posterior samples to measure the classification uncertainty.

2 Full mixture of experts model

A fmem is a generalization of a finite mixture model with G components where the mixing weights $\mathbf{p}$ and (some of) the component-specific parameters can be linked to a set of covariates through proper link functions. Since the random vector $\mathbf{Y}$ belongs to the simplex $\mathcal{S}^D$, a mixture with Dirichlet components displaying different means $\boldsymbol{\mu}_j \in \mathcal{S}^D$ $(j = 1,\dots,G)$ and a common precision parameter $\phi > 0$ is a proper choice. Thus, we can define the fmem probability density function (pdf) as

$$f_{\mathbf{Y}}(\mathbf{y}_i|\mathbf{x}_i,\boldsymbol{\mu}(\mathbf{x}_i),\mathbf{p}(\mathbf{x}_i),\phi) = \sum_{j=1}^{G} p_j(\mathbf{x}_i) f^D\left(\mathbf{y}_i;\boldsymbol{\mu}_j(\mathbf{x}_i),\phi\right), \quad i = 1,\dots,n \tag{1}$$

where $f^D(\cdot;\cdot,\cdot)$ is the Dirichlet pdf, $\mathbf{p}(\mathbf{x}_i) = (p_1(\mathbf{x}_i),\dots,p_G(\mathbf{x}_i))^\intercal \in \mathcal{S}^G$, $\boldsymbol{\mu}_j(\mathbf{x}_i) \in \mathcal{S}^D$ for any fixed $\mathbf{x}_i$, $\mathbf{x}_i$ is the $(K+1)$-dimensional vector of covariates, and n is the sample size. Since both $\mathbf{p}$ and $\boldsymbol{\mu}_j$ belong to the simplex, we suggest to take advantage of the multinomial logit link function, so that:

$$p_j(\mathbf{x}_i) = \frac{\exp\left(\mathbf{x}_i^\intercal \boldsymbol{\gamma}_j\right)}{1+\sum_{r=1}^{G-1}\exp\left(\mathbf{x}_i^\intercal \boldsymbol{\gamma}_r\right)}, \quad \mu_{d,j}(\mathbf{x}_i) = \frac{\exp\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}_{d,j}\right)}{1+\sum_{r=1}^{D-1}\exp\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}_{r,j}\right)},$$

$(j = 1,\dots,G; d = 1,\dots,D)$, where $\boldsymbol{\gamma}_j$ and $\boldsymbol{\beta}_{d,j}$ are $(K+1)$-dimensional vectors, with $\boldsymbol{\gamma}_G = \boldsymbol{\beta}_{D,j} = \mathbf{0}$. Although considering a common (and constant) ϕ keeps the model simple, one can further generalize the model linking it to some covariates through a proper link function. Note that the above proposed approach allows to avoid any transformation of compositional data, so that regression coefficients deserve an easy and meaningful interpretation.

3 Estimation and classification issues

Let us consider a supervised classification problem, where we want to learn a classifier on a training set, so that we can assign a label to new observations. More specifically, suppose we have observed a compositional vector $\mathbf{Y}_i \in \mathcal{S}^D$, a vector of covariates $\mathbf{x}_i$, and a discrete variable S_i, $i = 1,\dots,n$. S_i can assume G different labels, denoted by $1,\dots,G$, and $S_i = j$ if the i-th observation belongs to the j-th group/label. Therefore, S_i is the target in the classification task. Our training set consists in a vector $\mathbf{S} = (S_1,\dots,S_n)^\intercal$ and two matrices $\mathbf{Y}$ and $\mathbf{X}$, with generic i-th row $\mathbf{Y}_i$ and $\mathbf{x}_i$, respectively. Here, the mixture components represent the G groups encoded by $\mathbf{S}$. This means that we know which mixture component generated a specific training data point, and thus we can resort to the complete-data likelihood, which can be written as

$$L_C(\boldsymbol{\eta};\mathbf{y},\mathbf{x},\mathbf{s}) = \left[\prod_{j=1}^{G}\prod_{i:S_i=j} f^D\left(\mathbf{y}_i;\boldsymbol{\mu}_j(\mathbf{x}_i),\phi\right)\right] \cdot \left[\prod_{j=1}^{G}\prod_{i:S_i=j} p_j(\mathbf{x}_i)\right], \tag{2}$$

where $\boldsymbol{\eta} = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_D^*, \boldsymbol{\gamma}^*, \phi)^\intercal$, and $\boldsymbol{\beta}_j^*$ and $\boldsymbol{\gamma}^*$ are matrices obtained concatenating by row the vectors $\boldsymbol{\beta}_{1,j}, \ldots, \boldsymbol{\beta}_{D,j}$ and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_G$, respectively. Following a Bayesian approach to inference, we have to specify a joint prior distribution for $\boldsymbol{\eta}$. We select a multivariate normal with zero mean vector and diagonal covariance matrix with "large" values of the variances as non-informative prior for the regression parameters $\boldsymbol{\beta}_{d,j}$ and $\boldsymbol{\gamma}_j$, for any proper choice of d and j. For the precision parameter ϕ, we adopt a Gamma(g,g) prior distribution, with rate parameter g "small" enough to induce a large variability. We simulate samples from the posterior distribution through the Hamiltonian Monte Carlo algorithm in the Stan language. Please note that we do not face label switching problems because we know the true allocations of training observations to the mixture components. Once we have drawn B samples from the simulated posterior distribution of $\boldsymbol{\eta}$ (namely, $\boldsymbol{\eta}^{(1)}, \ldots, \boldsymbol{\eta}^{(B)}$), we can use them to classify a new observation for which we observe only $(\mathbf{y}_u, \mathbf{x}_u)$, $u > n$. Indeed, Bayes' theorem enables to compute the posterior probability that unit u arises from group j given its observed value $\mathbf{y}_u$ and $\mathbf{x}_u$, $b = 1, \ldots, B$, that is:

$$\hat{z}_{u,j}^{(b)} = P\left(S_u = j | \mathbf{Y}_u = \mathbf{y}_u, \mathbf{x}_u; \boldsymbol{\eta}^{(b)}\right) = \frac{p_j^{(b)}(\mathbf{x}_u) \cdot f^D\left(\mathbf{y}_u; \boldsymbol{\mu}_j^{(b)}(\mathbf{x}_u), \phi^{(b)}\right)}{\sum_{l=1}^{G} p_l^{(b)}(\mathbf{x}_u) \cdot f^D\left(\mathbf{y}_u; \boldsymbol{\mu}_l^{(b)}(\mathbf{x}_u), \phi^{(b)}\right)}, \quad (3)$$

where $\boldsymbol{\mu}_j^{(b)}$ and $p_j^{(b)}$ are computed based on $\boldsymbol{\eta}^{(b)}$. Although a Bayesian classification rule can be defined by allocating to group j whenever the mean of the simulated $\hat{z}_{u,j}^{(b)}$ is the highest ($j = 1, \ldots, G$), the purpose of this contribution is to take advantage of the (simulated) posterior distribution of the probability of each category to measure the classification uncertainty, as we discuss in the next section.

4 Application on plants data

We consider an application based on a compositional dataset regarding $n = 500$ plants (Douma & Weedon, 2019). The composition is defined by the proportion of biomass in roots (RMF), stems (SMF), and leaves (LMF). We aim to classify the species of a plant (*D. flexuosa* or *H. lanatus*, so that $G = 2$) based on the biomass composition, as well as two covariates represented by the nitrate supply level (high or low), and a measure of the total amount of biomass (TDM). Since we have neither a validation nor a test set, we use V-fold cross-validation to assess the performance of the classification rule. Thus, we randomly divide the dataset into $V = 4$ parts and classify each fold using the remaining three parts as the training set. The estimated overall misclassification error rate (MER) (defined as the average of the fold-specific MERs.) resulted in 0.238. Fig. 1 shows the simulated distribution of the posterior probability of being classified as *D. flexuosa* for eight randomly selected plants. Classifying as *D. flexuosa* every plant with a mean (or median) posterior probability greater than 0.5, we misclassify two plants (2/8 $\approx$ 0.238). The range of each subject-specific posterior probability distribution helps in assessing the classification uncertainty. For example, the distribution of the posterior probability for plant 6 is very wide and centered close to 0.5, suggesting that its classification could be unreliable, while the reverse holds for the other plants.

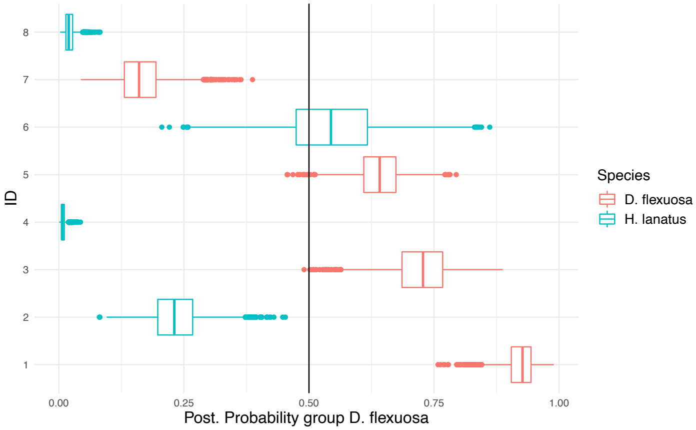


Figure 1. *Boxplots of the posterior probability of being classified as D. flexuosa for 8 randomly selected plants. Colors represent the true label of each plant.*

References

Bouveyron, C., Celeux, G., Brendan Murphy, T., & Raftery, A.E. 2019. *Model-based Clustering and Classification for Data Science*. 1

Douma, J.C., & Weedon, J.T. 2019. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods Ecol Evol*, **10**, 1412–1430. 3

Gu, J., & Cui, B. 2021. A classification framework for multivariate compositional data with Dirichlet feature embedding. *Knowl Based Syst*. 1

Ongaro, A., Migliorati, S., & Ascari, R. 2020. A new mixture model on the simplex. *Stat Comp*, **30**, 749–770. 1

SPARSE INFERENCE IN COVARIATE ADJUSTED CENSORED GAUSSIAN GRAPHICAL MODELS

Luigi Augugliaro[1], Gianluca Sottile[1] and Angelo M. Mineo[1]

[1] Dep. of Economics, Business and Statistics, University of Palermo, Italy,
(e-mail: luigi.augugliaro@unipa.it, angelo.mineo@unipa.it,
gianluca.sottile@unipa.it)

ABSTRACT: The covariate adjusted glasso is one of the most used estimators for inferring genetic networks. Despite its diffusion, there are several fields in applied research where the limits of detection of modern measurement technologies make the use of this estimator theoretically unfounded, even when the assumption of a multivariate Gaussian distribution is satisfied. In this paper we propose an extension to censored data.

KEYWORDS: censored data, censored glasso estimator, Gaussian graphical model, glasso estimator.

1 Introduction

An important aim in genomics is to understand interactions among genes, characterized by the regulation and synthesis of proteins under internal and external signals. These relationships can be represented by a genetic network, i.e., a graph where nodes represent genes and edges describe the interactions among them. Gaussian graphical models (GGM, Lauritzen (1996)) have been widely used for reconstructing a genetic network from expression data. The reason of such diffusion relies on the statistical properties of the multivariate Gaussian distribution which allow the topological structure of a network to be related with the non-zero elements of the concentration matrix, i.e., the inverse of the covariance matrix. Thus, the problem of network inference can be recast as the problem of estimating a concentration matrix. The covariate adjusted glasso estimator (Yin & Li, 2011) is a popular method for estimating a sparse concentration matrix, based on the idea of adding an ℓ_1-penalty function to the likelihood function of the multivariate Gaussian distribution. Despite the widespread literature on the covariate adjusted glasso estimator, there is a great number of fields in applied research where the use of the graphical model is theoretically unfounded. For example in some cases data are left- or right-censored. In this paper we propose an extension of the covariate adjusted glasso estimator that takes into account the censoring mechanism of the data explicitly.

2 The covariate adjusted censored Gaussian graphical model

Let $\boldsymbol{Y} = (Y_1, \dots, Y_p)^\top$ be a p-dimensional random vector. Graphical models allow to represent the set of conditional independencies among these random variables by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the set of nodes associated to $\boldsymbol{Y}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$

is the set of ordered pairs, called edges, representing the conditional dependencies among the p random variables (Lauritzen (1996)). The covariate adjusted Gaussian graphical model (CGGM) is an extension of the classical GGM based on the assumption that the conditional distribution of $\boldsymbol{Y}$ given a q-dimensional vector of predictors, say $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$, follows a multivariate Gaussian distribution with expected value: $\boldsymbol{\mu}(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \boldsymbol{x}$, where $\boldsymbol{\beta} = (\beta_{hk})$ is a matrix $q \times p$ coefficient matrix, and covariance matrix denoted by $\Sigma = (\sigma_{hk})$. Denoting with $\Theta = (\theta_{hk})$ the concentration matrix, i.e., the inverse of the covariance matrix, the conditional density function of $\boldsymbol{Y}$ can be written as follows:

$$\phi(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta) = (2\pi)^{-p/2} |\Theta|^{1/2} \exp[-1/2\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}^\top \Theta \{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}]. \qquad (1)$$

As shown in Lauritzen (1996), the off-diagonal elements of the concentration matrix are the parametric tools relating the pairwise Markov property to the factorization of the density (1), i.e., two random variables, say Y_h and Y_k, are conditionally independent given all the remaining variables if and only if θ_{hk} is equal to zero.

As done in Augugliaro *et al.* (2020), we assume that $\boldsymbol{Y}$ is a (partially) latent random vector with density function (1). In order to include the censoring mechanism inside our framework, let us denote by $\boldsymbol{l} = (l_1, \ldots, l_p)^\top$ and $\boldsymbol{u} = (u_1, \ldots, u_p)^\top$, with $l_h < u_h$ for $h = 1, \ldots, p$, the vectors of known left and right censoring values. Thus, Y_h is observed only if it is inside the interval $[l_h, u_h]$ otherwise it is censored from below if $Y_h < l_h$ or censored from above if $Y_h > u_h$. Using the approach for missing data with nonignorable mechanism (Little & Rubin (2002)) we denote the quantity $R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u})$, to encode the censoring patterns, such that the hth element of $R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u})$ is defined as $R(Y_h; l_h, u_h) = I(Y_h > u_h) - I(Y_h < l_h)$, where $I(\cdot)$ denotes the indicator function. By construction $R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u})$ is a discrete random vector with support the set $\{-1, 0, 1\}^p$ and probability function $\Pr\{R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r}\} = \int_{D_{\boldsymbol{r}}} \phi(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta) d\boldsymbol{y}$, where $D_{\boldsymbol{r}} = \{\boldsymbol{y} \in \mathbb{R}^p : R(\boldsymbol{y}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r}\}$. Given a censoring pattern, we can simplify our notation by partitioning the set $I = \{1, \ldots, p\}$ into $o = \{h \in I : r_h = 0\}, c^- = \{h \in I : r_h = -1\}$ and $c^+ = \{h \in I : r_h = +1\}$ and, in the following of this paper, we shall use the convention that a vector indexed by a set of indices denotes the corresponding subvector. As done in Augugliaro *et al.* (2020), the probability distribution of the observed data, denoted by $\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta)$, can be defined as follows:

$$\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta) = \int \phi(\{\boldsymbol{y}_o, \boldsymbol{y}_c\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta) \Pr\{R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r} | \boldsymbol{Y} = \boldsymbol{y}\} d\boldsymbol{y}_c, \qquad (2)$$

where $c = c^- \cup c^+$. Density (2) can be simplified by observing that $\Pr\{R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r} \mid \boldsymbol{Y} = \boldsymbol{y}\}$ is equal to one if the censoring pattern encoded in $\boldsymbol{r}$ is equal to the pattern observed in $\boldsymbol{y}$, otherwise it is equal to zero, hence $\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta)$ can be rewritten as

$$\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta) = \int_{D_c} \phi(\{\boldsymbol{y}_o, \boldsymbol{y}_c\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta) d\boldsymbol{y}_c I(\boldsymbol{l}_o \leq \boldsymbol{y}_o \leq \boldsymbol{u}_o), \qquad (3)$$

where $D_c = (-\infty, \boldsymbol{l}_{c^-}) \times (\boldsymbol{u}_{c^+}, +\infty)$. Using density (3), the covariate adjusted censored Gaussian graphical model (CCGGM) is defined as the set $\{\boldsymbol{Y}, R(\boldsymbol{Y}; \boldsymbol{l}, \boldsymbol{u}), \varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} \mid \boldsymbol{x}; \boldsymbol{\beta}, \Theta), \mathcal{G}\}$, where $\varphi(\{\boldsymbol{y}_o, \boldsymbol{r}\} | \boldsymbol{x}; \boldsymbol{\beta}, \Theta)$ factorizes according to the undirected graph $\mathcal{G}$.

3 The covariate adjusted censored glasso estimator

Suppose we have a sample of size n independent observations drawn from a CCGGM. For ease of exposition, we shall assume that $\boldsymbol{l}$ and $\boldsymbol{u}$ are fixed across the n observations. To simplify our notation the set of indices of the variables observed in the ith observation is denoted by $o_i = \{h \in I : r_{ih} = 0\}$, while $c_i^- = \{h \in I : r_{ih} = -1\}$ and $c_i^+ = \{h \in I : r_{ih} = +1\}$ denote the sets of indices associated to the left and right-censored data, respectively. Denoting by $\boldsymbol{r}_i$ the realization of the random vector $R(\boldsymbol{Y}_i; \boldsymbol{l}, \boldsymbol{u})$, the ith observed data is the vector $(\boldsymbol{y}_{io_i}^\top, \boldsymbol{x}_i^\top, \boldsymbol{r}_i^\top)^\top$. Using the density function (3), the observed log-likelihood function can be written as

$$\ell(\boldsymbol{\beta}, \Theta) = \sum_{i=1}^{n} \log \int_{D_{c_i}} \phi(\{\boldsymbol{y}_{io_i}, \boldsymbol{y}_{ic_i}\} | \boldsymbol{x}_i; \boldsymbol{\beta}, \Theta) d\boldsymbol{y}_{ic_i} = \sum_{i=1}^{n} \log \varphi(\{\boldsymbol{y}_{io_i}, \boldsymbol{r}_i\} | \boldsymbol{x}_i; \boldsymbol{\beta}, \Theta), \quad (4)$$

where $D_{c_i} = (-\infty, \boldsymbol{l}_{c_i^-}) \times (\boldsymbol{u}_{c_i^+}, +\infty)$ and $c_i = c_i^- \cup c_i^+$. Although inference about the parameters of this model can be carried out via the maximum likelihood method, the application of this inferential procedure to real datasets is limited.

We propose to estimate the parameters of the CCGGM by generalizing the approach proposed in Yin & Li (2011), i.e., by maximizing a new objective function defined by adding two lasso-type penalty functions to the observed log-likelihood (4). The resulting estimator, called covariate adjusted censored glasso estimator, is formally defined as

$$\{\hat{\boldsymbol{\beta}}^\lambda, \widehat{\Theta}^\rho\} = \arg\max_{\boldsymbol{\beta}, \Theta \succ 0} \frac{1}{n} \sum_{i=1}^{n} \log \varphi(\{\boldsymbol{y}_{io_i}, \boldsymbol{r}_i\} | \boldsymbol{x}_i; \boldsymbol{\beta}, \Theta) - \lambda \sum_{h,k} |\beta_{hk}| - \rho \sum_{h \neq k} |\theta_{hk}|, \quad (5)$$

where λ and ρ are two non-negative tuning parameters. The lasso penalty on $\boldsymbol{\beta}$ introduces sparsity in $\hat{\boldsymbol{\beta}}^\lambda$, while the tuning parameter ρ controls the amount of sparsity in the estimated concentration matrix $\widehat{\Theta}^\rho = (\hat{\theta}_{hk}^\rho)$.

4 Simulation study

In this section, we compare our proposed estimator with MissGlasso (Städler & Bühlmann, 2012), which performs ℓ_1-penalized estimation under the assumption that the censored data are missing at random, and with the covariate adjusted glasso estimator (Yin & Li, 2011), where the empirical covariance matrix is calculated by imputing the missing values with the censoring values. These estimators are evaluated in terms of both recovering the structure of the true graph. We use the method implemented in the R package `huge` (Zhao *et al.*, 2020), to simulate a sparse concentration matrix with a random structure for $\boldsymbol{Y}$. We set the probability of observing a link between two nodes to k/p, where p is the number of responses and k is used to control the amount of sparsity in $\boldsymbol{\Theta}$. Moreover, we set the right censoring value to 40 for any variable and the sample size n to 100. The predictors matrix $\boldsymbol{X}$ is sampled from a multivariate gaussian distribution with zero expected value and sparse covariance matrix simulated as done for $\boldsymbol{Y}$. Each column of the true matrix of predictors $\boldsymbol{\beta}$ contains only two non-zero regression coefficients sampled from a uniform distribution on the interval $[0.3, 0.7]$. The values of the intercepts are chosen in such a way that H response variables are right censored with probability equal to 0.40. The quantities k, p, q and H are chosen according to the following cases:

- **Scenario 1**: $k = 3$, $p = 50$, $q = 10$ and $H = 25$. This setting is used to evaluate the effects of the number of censored variables on the behavior of the proposed estimators when $n > p$.
- **Scenario 2**: $k = 3$, $p = 150$, $q = 10$ and $H = 75$. This setting is used to evaluate the impact of the high dimensionality on the estimators ($p > n$).

For each scenario, we simulate 50 samples and in each simulation, we compute the coefficients path using cglasso, MissGlasso, and glasso. Each path is computed using an equally spaced sequence of ρ and λ-values. Moreover, the precision-recall curves and the area under the curves (AUCs) are computed for each Scenarios. Table 1 shows how cglasso gives a better estimate of the concentration and coefficient matrices in terms of AUCs, for any given value of the tuning parameters. We report only five evenly spaced values of λ and ρ.

Table 1. *Mean area under the curves across the sequence of* ρ *and* λ*-values under the specification of the two Scenarios (see row blocks). The first column block refers to the concentration matrix (*$\boldsymbol{\Theta}$*) when* λ *is fixed and the second refers to the coefficient matrix (*$\boldsymbol{\beta}$*) when* ρ *is fixed. In the first column (1), (2) and (3) refer to cglasso, MissGlasso and glasso algorithms, respectively.*

	$\lambda/\lambda_{\max}$					$\rho/\rho\text{max}$				
	0.00	0.25	0.50	0.75	1.00	0.00	0.25	0.50	0.75	1.00
(1)	0.546	0.429	0.139	0.103	0.101	0.844	0.877	0.883	0.882	0.885
(2)	0.239	0.199	0.086	0.073	0.073	0.745	0.764	0.766	0.767	0.768
(3)	0.414	0.218	0.097	0.092	0.091	0.813	0.847	0.864	0.866	0.866
(1)	0.418	0.094	0.037	0.035	0.035	0.794	0.930	0.931	0.929	0.933
(2)	0.329	0.098	0.033	0.031	0.030	0.753	0.830	0.831	0.830	0.831
(3)	0.321	0.040	0.033	0.032	0.031	0.751	0.902	0.906	0.907	0.907

References

AUGUGLIARO, L., ABBRUZZO, A., & V., VINCIOTTI. 2020. ℓ_1-Penalized censored Gaussian graphical model. *Biostatitistics.*, **21**(2), e1–e16.

LAURITZEN, S.L. 1996. *Graphical Models.* Oxford University Press, Oxford.

LITTLE, R.J.A., & RUBIN, D.B. 2002. *Statistical Analysis with Missing Data.* John Wiley & Sons, Inc., Hoboken.

STÄDLER, N., & BÜHLMANN, P. 2012. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing.*, **22**(1), 219–235.

YIN, J., & LI, H. 2011. A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Annals of Applied Statistics.*, **5**(4), 2630–2650.

ZHAO, T., LI, X., LIU, H., ROEDER, K., LAFFERTY, J., & WASSERMAN, L. 2020. *huge: High-Dimensional Undirected Graph Estimation.* R package version 1.3.4.1.

SEMI-SUPERVISED LEARNING THROUGH DEPTH FUNCTIONS

Simona Balzano[1], Mario R. Guarracino,[1] and Giovanni C. Porzio [1]

[1] Department of Economics and Law, University of Cassino and Southern Lazio
(e-mail: s.balzano@unicas.it, mario.guarracino@unicas.it, porzio@unicas.it)

ABSTRACT: Depth functions have been exploited in supervised learning since years. Given that the depth of a point is somehow a distribution-free measure of its distance from the center of a distribution, their use in supervised learning arose naturally and it has seen a certain degree of success. Particularly, DD-classifers and their extensions have been extensively studied and applied in many applied fields and statistical settings. What has not been investigated so far is their use within a semi-supervised learning framework. That is, in case some labeled data are available along with some unlabeled data within the same training set. A case which arises in many applications and where it has been proved that combining information from labeled and unlabeled data can improve the overall performance of a classifier. For this reason, this work aims at introducing semi-supervised learning techniques in association with DD-classifiers and at investigating to what extent such technique is able to improve DD-classifier performances. Performances will be evaluated by means of an extensive simulation study and illustrated on some real data sets.

KEYWORDS: DD-classifiers, labeled and unlabeled data, supervised learning.

A COMBINED TEST OF THE BENFORD HYPOTHESIS WITH ANTI-FRAUD APPLICATIONS

Lucio Barabesi [1], Andrea Cerasa[2], Andrea Cerioli[3] and Domenico Perrotta[2]

[1] University of Siena, Department of Economics and Statistics, Siena, Italy, (e-mail: lucio.barabesi@unisi.it)

[2] European Commission, Joint Research Centre (JRC), Ispra, Italy, (e-mail: andrea.cerasa@ec.europa.eu, domenico.perrotta@ec.europa.eu)

[3] University of Parma, Department of Economics and Management, Parma, Italy, (e-mail: andrea.cerioli@unipr.it)

ABSTRACT: In this work we describe a combined test of the null hypothesis that the significant digits in a random sample of numbers follow Benford's law. We also show the potential of the method for the purpose of fraud detection in international trade.

KEYWORDS: Anomaly detection, Benford's law, sum-invariance, customs data.

1 Motivating framework of data analysis

Most unsupervised fraud detection methods look for anomalies in the data. Therefore, all of these techniques assume that the available data have been generated by an appropriate contamination model. Any parameter of the distribution that models the "genuine" part of the data, say F_0, must then be estimated in a robust way, in order to avoid the well-known masking and swamping effects due to the anomalies themselves (Cerioli *et al.*, 2019b). In the context of fraud detection in international trade, where the value of an individual import transaction X originates from the product of the traded amount v with the unit price β, the available anti-fraud tools are derived from the theory of outlier identification in robust regression; see, e.g., Perrotta *et al.*, 2020b. Under this approach it is then assumed that non-fraudulent transactions for a specific good are generated according to the distribution function

$$F_0(x) = \Phi\left(\frac{x - \beta v}{b}\right), \tag{1}$$

where Φ is the distribution function of a standard Normal random variable. In model (1), the regression slope β corresponds to unit price and $b > 0$ defines the (usually unknown) model variability, which is taken to be constant.

Robust and efficient estimation of β in model (1) may lead to the definition of a "fair" unit price for the good under consideration, against which individual or aggregate transaction prices can be contrasted. Transactions well below the "fair" price may correspond to revenue frauds leading to substantial undervaluation of goods imported into the European Union; see, e.g., European Anti-Fraud Office, 2018, p. 26. The normality assumption in model (1) has proven to be satisfactory in the case of monthly-aggregated trade data (Perrotta *et al.*, 2020b). However it may become less adequate when analyzing individual customs declarations, where multiple populations often occur and a skew distribution may seem more appropriate for the definition of F_0 (Perrotta *et al.*, 2020a). An alternative contamination model based on Benford's law then becomes very useful in such a framework: see Cerioli *et al.*, 2019a.

2 Benford's law

Benford's law (BL, for short) is a fascinating phenomenon which rules the pattern of the leading digits in many types of data. Informally speaking, the law states that the digits follow a logarithmic-type distribution in which the leading digit 1 is more likely to occur than the leading digit 2, the leading digit 2 is more likely than the leading digit 3, and so on. Indeed, the first-digit form of BL gives the probability that the first leading digit equals d, for $d = 1, \ldots, 9$, as

$$\log_{10}\left(1 + \frac{1}{d}\right). \tag{2}$$

Another, perhaps even less intuitive, property of Benford's law concerns sum invariance. Given an absolutely-continuous random variable X, in the first digit setting of (2), this property states that, for $d = 1, \ldots, 9$,

$$\mathrm{E}[S(X)\mathrm{I}_{[d,d+1[}(S(X))] = C, \tag{3}$$

where I_E is the indicator function of the set E, while

$$S(x) = 10^{\langle \log_{10}|x| \rangle} \tag{4}$$

is the *significand* of the non-null real number x, $\langle x \rangle = x - \lfloor x \rfloor$ and $C = \log_{10} \mathrm{e}$. First-digit sum invariance thus means that the expected value (3) does not depend on d when X is a Benford random variable. Although (2) and (3) are not equivalent when only the first digit is concerned, they are both implied by the full form of BL, which states that

$$S(X) \overset{\mathcal{L}}{=} 10^U, \tag{5}$$

with U a Uniform random variable on $[0,1[$. We refer to Berger & Hill, 2020 for a recent survey of the mathematical properties of BL and to Barabesi *et al.*, 2021 for a thorough study of the relationship between (2) and (3).

3 Tests of the Benford hypothesis

In the motivating framework sketched in §1, Cerioli *et al.*, 2019a investigate the conditions under which Benford's law may yield a reasonable approximation for the first-digit distribution of customs declarations. If Benford's law is expected to hold for genuine transactions, then deviations from the law can be taken as evidence of possible data manipulation. Several exact tests of the Benford hypothesis exist according to which characterization is considered. Those that follow have proven to be useful under a variety of circumstances:

- The chi-square test of the first-digit distribution (2) considered by Barabesi *et al.*, 2018, say χ^2;
- The Hotelling-type test of the sum-invariance property (2) proposed by Barabesi *et al.*, 2021, say Q;
- The Kolmogorov-Smirnov test of the Benford property (4) described in Barabesi *et al.*, 2021, say KS.

Barabesi *et al.*, 2021 show that the combination of χ^2 and Q provides a test which is consistently close to the best solution provided by either χ^2 or Q. We further develop this strategy in two directions. First, we derive the asymptotic joint distribution of χ^2 and Q under Benford's law. This result gives theoretical substance to the observed empirical behavior of the combined test. We then extend our combination strategy to include KS. The proposed extension is extremely relevant in view of the motivating framework of §1, since the performance of the individual tests may vary considerably according to the actual digit generating process when Benford's law does not hold. Our combined test thus provides a powerful, yet robust, solution when the type of departure from Benford's law is unknown, as it happens in anti-fraud applications. Some preliminary simulation results for a sample size of $n = 100$ observations are shown in Table 1, where $L_{\chi^2,Q,KS}$ denotes the newly developed combined test. The alternative data generating models for X are a Lognormal random variable of scale parameter 1 and shape parameter 0.5, and a Generalized Benford random variable of parameter -0.6.

Table 1. *Estimated power of tests of the Benford hypothesis for sample size $n = 100$.*

Alternative	χ^2	Q	KS	$L_{\chi^2,Q,KS}$
Lognormal	0.903	0.926	0.899	0.940
Generalized Benford	0.446	0.466	0.853	0.785

References

Barabesi, L., Cerasa, A., Cerioli, A., & Perrotta, D. 2018. Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud. *Journal of Business and Economic Statistics*, **36**, 346–358.

Barabesi, L., Cerasa, A., Cerioli, A., & Perrotta, D. 2021. On Characterizations and Tests of Benford's Law. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2021.1891927.

Berger, A., & Hill, T. P. 2020. The mathematics of Benford's law: a primer. *Statistical Methods and Applications*. https://doi.org/10.1007/s10260-020-00532-8.

Cerioli, A., Barabesi, L., Cerasa, A., Menegatti, M., & Perrotta, D. 2019a. Newcomb-Benford law and the detection of frauds in international trade. *PNAS*, **116**, 106–115.

Cerioli, A., Farcomeni, A., & Riani, M. 2019b. Wild adaptive trimming for robust estimation and cluster analysis. *Scandinavian Journal of Statistics*, **46**, 235–256.

European Anti-Fraud Office. 2018. *The OLAF report 2017*. Tech. rept. Publications Office of the European Union, Luxembourg. https://doi.org/10.2784/652365.

Perrotta, D., Checchi, E., Torti, F., Cerasa, A., & Novau, X. A. 2020a. *Addressing Price and Weight heterogeneity and Extreme Outliers in Surveillance Data: The Case of Face Masks*. Tech. rept. JRC121650, EUR 12345 EN. Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/817681.

Perrotta, D., Cerasa, A., Torti, F., & Riani, M. 2020b. *The Robust Estimation of Monthly Prices of Goods Traded by the European Union*. Tech. rept. JRC120407, EUR 30188 EN. Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/635844.

UNBALANCED CLASSIFICATION OF ELECTRONIC INVOICING

Chiara Bardelli [1]

[1] Department of Mathematics, University of Pavia, (e-mail: chiara.bardelli01@universitadipavia.it)

ABSTRACT: Real world classification problems may present a high number of classes to predict which are not equally distributed in the dataset. We propose a two-step approach to address this problem analyzing data from the accounting world. Electronic invoices have a hierarchical structure which is exploited in the first step of our model. A classifier, then, is trained on the lines of the invoices for each subset generated by the cluster analysis. The results obtained show a higher recall values for the least frequent classes of the dataset with respect to the adoption of a single classification model.

KEYWORDS: unbalanced classification, text mining, prediction model

1 Introduction

The issue of unbalanced classification is known to affect different domains of application when a machine learning classifier is trained on real world data. If a dataset presents a lot of classes to predict, these classes often are not equally distributed, leading to an unbalanced problem which needs ad hoc analyses. Different methods have been proposed in literature to implement the possible solutions to this problem (Santos *et al.*, 2018; Ganganwar, 2012). However, in case of large datasets and a high number of classes, one suggested methodology is the possibility of splitting data in smaller subsets which contain similar observations and develop a single classifier for each subset of data to have simpler classification algorithms to manage and a lower number of classes per each cluster to predict (Tsoumakas *et al.*, 2008).

The classification of electronic invoices in the accounting process, using accounts which are part of Chart of Accounts, is a multiclass classification problem characterized by a high number of classes and unbalanced distributions. Nowadays, the interest in the automation of this task is really high (Bardelli *et al.*, 2020; Beļskis *et al.*, 2020). This is considered a repetitive and monotonous routine activity which can be easily replaced by a machine learning model giving the possibility for the accountants to focus on more stimulating projects. This classification task presents different challenges, due to the nature of the problem and to the high number of accounts employed in the classification. In this work we address this issue exploiting the hierarchical structure of the invoices documents to cluster them into smaller subsets easier to manage in terms of classification task. A single classifier is then trained on each subset. The results of this two-step methodology are compared with the performance of a single classifier trained on the entire dataset.

2 Data and method

The dataset analyzed in this work consists of 13.605 supplier and customer invoices for a total of 121.946 lines of invoices to classify. This data are part of the accountability database of a single business company. The total number of different classes to predict is 42, with the frequency of the majority class equal to 87.513.

Given the information about a single line of an invoice and the generic characteristics of the invoice, the aim is to predict the accounting code related to the line. For the sake of simplicity, we assume that lines inside an invoice are independent ignoring the grouping term which could influence the predictive output. In future works, this grouping information can be included in the features space too. In our classification task, we construct the prediction rule given the training sample $\{(y_i, \mathbf{x}_i)\}_{i=1}^{N}$ with:

- y_i, $i = 1,...,N$, categorical observations which represent the accounting codes associated to the i-th line of invoice
- $\mathbf{x}_i$, the vector of predictors related to the content of the invoice and the line

Thanks to the hierarchical structure of the invoice, which is composed by an header and the description lines, we apply a two-step approach: first of all, we exploited the information of the header of the invoice to cluster data in smaller datasets, and secondly, we train a classifier on the lines of the invoices for each cluster of data. We compare this two-step approach with a direct approach that develops a single classification model on the entire original dataset combining predictors both from the header and lines of the invoice.

The classification algorithm used in this work is the xgboost model, known to be very efficient in case of large dataset. To process the textual description of lines of invoices we apply the Word2vec model. Clusters of invoices are computed using the k-means algorithm. The number of clusters suggested by the Elbow method and used in the analysis is 4.

3 Results

We describe the results of the two different approaches providing some indexes of accuracy computed on the test set (the original dataset has been split into 80% for training set and 20% for test set preserving the class proportions in the split). In Table 1 we report the values of macro and weighted recall. As we can observe, the two approaches show similar values in terms of recall, meaning that splitting the original dataset into small clusters of data does not affect the overall performance of the model.

Table 1. *Macro and weighted recall computed on the test set fort the two approaches*

Methodology	Macro recall	Weighted recall
Direct approach	82.5%	98.2%
Two-step approach	83.3%	98.5%

The most interesting result is obtained for the accuracy of some classes which have low frequencies in the original data. Figure 1 reports the values of recall of the least 10 frequent classes of the dataset. Most of the classes improves its recall values, in particular the accounting codes 680203024 and 680203010 show recall values from 0 of the direct approach to 80% and 62% with the two-step approach, respectively. The invoices related to these accounts belong all to the same cluster which can be identified as group of supplier invoices related to the purchase of materials. On the other hand, high frequencies classes preserve same values of accuracy in both the approaches.

4 Conclusions

The challenge of unbalanced dataset with high number of classes is to develop accurate classification model able to correctly predict classes with low frequencies (minority classes). The problem we addressed allows us to exploit the hierarchical structure of the invoice document to divide the original dataset in smaller clusters, based on characteristics of invoice headers. Thanks to this procedure, the original classification problem has been split into simpler classification tasks with a smaller number of classes in each cluster.

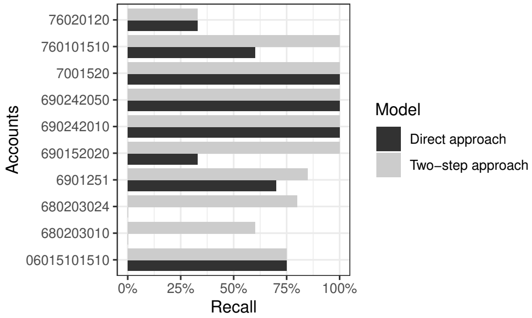


Figure 1. *Recall values for the 10 least frequent classes of the dataset.*

The results obtained in this analysis encourage deeper studies, trying to completely automate the classification of invoices into accounting codes which is an expensive and demanding task for the accountant work.

References

Bardelli, Chiara, Rondinelli, Alessandro, Vecchio, Ruggero, & Figini, Silvia. 2020. Automatic electronic invoice classification using machine learning models. *Machine Learning and Knowledge Extraction*, **2**(4), 617–629.

Beļskis, Zigmunds, Zirne, Marita, & Pinnis, Mārcis. 2020. Features and Methods for Automatic Posting Account Classification. *In: International Baltic Conference on Databases and Information Systems*. Springer.

Ganganwar, Vaishali. 2012. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, **2**(4), 42–47.

Santos, Miriam Seoane, Soares, Jastin Pompeu, Abreu, Pedro Henrigues, Araujo, Helder, & Santos, Joao. 2018. Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. *IEEE Computational Intelligence Magazine*, **13**(4), 59–76.

Tsoumakas, Grigorios, Katakis, Ioannis, & Vlahavas, Ioannis. 2008. Effective and efficient multilabel classification in domains with large number of labels. *In: Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, vol. 21.

PREDICTIVE POWER OF BAYESIAN CAR MODELS ON SCALE FREE NETWORKS: AN APPLICATION FOR CREDIT RISK

Claudia Berloco [12] , Raffaele Argiento[34] and Silvia Montagna[14]

[1] Dipartimento di Scienze Economico-sociali e Matematico-statistiche, Università degli Studi di Torino, Corso Unione Sovietica, 218/bis, 10134 Torino, Italy, (e-mail: claudia.berloco@unito.it, silvia.montagna@unito.it)

[2] Intesa Sanpaolo, Piazza San Carlo, 156, 10121 Torino, Italy

[3] Dipartimento di Scienze statistiche, Università Cattolica Sacro Cuore, Largo A. Gemelli, 1, 20123 Milano, Italy, (e-mail: raffaele.argiento@unicatt.it)

[4] Collegio Carlo Alberto, Piazza Vincenzo Arbarello, 8, 10122 Torino, Italy

ABSTRACT: The monitoring of loans' life-cycle has received the increasing attention of the scientific community after the 2008 global financial crisis. A number of aspects of this broad topic have been addressed by means of several regulatory, statistical and economical tools. However, many issues still require further investigation. In this work, we are interested in the monitoring phase of granted loans to anticipate possible defaults and to investigate whether there is evidence of a liquidity contagion effect within a trade network of firms. To this end, we apply a Bayesian spatial model to a proprietary dataset, and assess its out-of-time predictive performance.

KEYWORDS: Bayesian modelling, spatial modelling, credit risk, CAR model.

1 Introduction

The European Central Bank requires banks to adapt their organization, processes and IT infrastructure in order to give an integrated answer to the non-performing loans problem. Banks can mitigate their credit risk in several steps of the loan life-cycle, for example by foreseeing liquidity problems for those customers which already have a debt to the bank. A timely detection of the transition to financial distress is pivotal, and it will be addressed it in this work leveraging on statistical models and bank data.

Recently, a number of contributions (see, e.g., Dolfin *et al.* , 2019) focused on introducing information on the supply chain connections in credit risk models based on the evidence of trade credit use in European markets. The main idea is that liquidity distress can flow along these connections, and a firm experienc-

ing a period of liquidity distress can delay payments towards its commercial partners, that can consequently experience liquidity distress. The supply chain is seen as a complex network in these studies, but it can also be represented as an adjacency matrix with proper assumptions (Lamieri & Sangalli, 2019).
In this work, we set up a predictive model leveraging Bayesian conditionally auto-regressive (CAR) models for areal data (Banerjee *et al.* , 2003). Specifically, inference is based on a sample of firms from a trade network in a given month, and the predictive performance of a CAR model is tested by estimating the probability of default for both a different sample of firms and for the same sample in the future. Although spatial CAR models have been widely used in ecology, environmental science, biology and medicine, to the best of our knowledge they have not yet been fully exploited in econometrics when dealing with hundreds of thousands of data points interacting in a dynamic complex network (e.g., firms or natural persons).

2 Methodology

With some due simplifications, the monthly goal for a lending bank is to red flag those borrowing firms that have the greatest probability of default (delay in paying their debts to the bank) in the following 3 months. In this paper, we analyse a proprietary dataset of Intesa Sanpaolo collected in a given month, for a total of $n = 944$ firms. Our response variable is a binary indicator such that $Y_k = 1$ if firm k switches to a liquidity distress state in the next 3 months.

The trade network can be represented as a link matrix $W \in \mathbb{R}^n \times \mathbb{R}^n$, with binary entries $w_{kj} = 1$ if $k \neq j$ and k supplier, j customer in the previous year. The link matrix W represents a complex network with a scale free structure (Barabási & Albert, 1999). Further, the Bank database stores several credit and trend information on each specific customer firm, but for the sake of simplicity here we only consider two possible covariates $\boldsymbol{x}_k$ for each firm k. The first covariate, x_k^1, represents the used amount of credit over the granted amount among all Italian financial institutions, while the second, x_k^2, represents the maximum number of days of payment delay recorded in the past 3 months.

We fit a proper CAR specification (Banerjee *et al.* , 2003) to our data as follows:

$$\begin{aligned} Y_k &\sim Bernoulli(\theta_k) \\ logit(\theta_k) &= \boldsymbol{\beta x}_k + \phi_k \\ \phi_k|\phi_{-k}, \alpha, \tau, W &\sim N\left(\alpha \frac{\sum_{i=1}^n w_{ki}\phi_i}{\sum_{i=1}^n w_{ki}}, \tau^{-1}\right), \end{aligned} \tag{1}$$

Here ϕ_k is a firm-specific spatial random effect incorporating the information contained in the network of relationships W. Conditionally on W, ϕ_k is modelled as a Markov random field, meaning that the value of ϕ_k only depends on the value of its neighbours. Indeed, we expect the probability of default of firm k to increase (decrease) if one of more firms connected with k are (not) in default. Parameters α and τ represent the strength and the precision of the autocorrelation, respectively. The CAR specification is chosen because the information arising from the network (incorporated through ϕ_k) can help explain those default events that are not ubiquitously captured by the linear covariates. Standard priors are placed on α, τ, and $\beta_0, \beta_1, \beta_2$, and estimation of model parameters proceeds via MCMC (Banerjee *et al.* , 2003).

3 Results and conclusions

Testing model (1) on real data, we notice that the posterior distributions of the linear parameters obtained with the CAR model are coherent with those of a standard GLM, which considers covariates $\boldsymbol{x}_k$ only. The overlap between the credible intervals of the linear parameters from the two models implies that the spatial random effects estimated by the CAR model contribute to explain a part of the default phenomenon not entirely captured by firm-specific information. Further, we record very good in-sample performance in terms of area under the curve (AUC), as the GLM has a 0.79 AUC while the CAR specification reaches a 0.89 AUC. Furthermore, model (1) helps in identifying defaulted firms through the spatial random effects. Indeed, Figure 1 (left panel) shows that, for most truly defaulted firms (red dots), the estimated probability that the spatial effect is positive, computed as $\widehat{\mathbb{P}}(\phi_k > 0) = \frac{1}{T-B} \sum_{g=B+1}^{T} \mathbb{1}(\phi_k^g > 0)$, is strictly greater than 50%. Here T is the total number of MCMC iterations, and B denotes the burn-in.

Further, we test the predictive power of the model on a disjoint sample drawn from the network seen at the same timestamp of the training sample (out-of-sample set composed of unseen firms), and on the training dataset but seen six months later (out-of-time set composed of future observations of the same firms used in training). In line with the original aim of spatial CAR models, which are intended to fit data referring to static maps, the model does not generalise in the out-of-sample case. This is an unfortunate result for our credit risk application, as one can instead expect the liquidity distress contagion dynamics to spread with similar strength (α) and precision (τ) in different areas of the trade network. In the out-of-time case, the CAR model shows slightly better predictive performance with respect to the simple GLM, as shown in

Figure 1 (right panel).

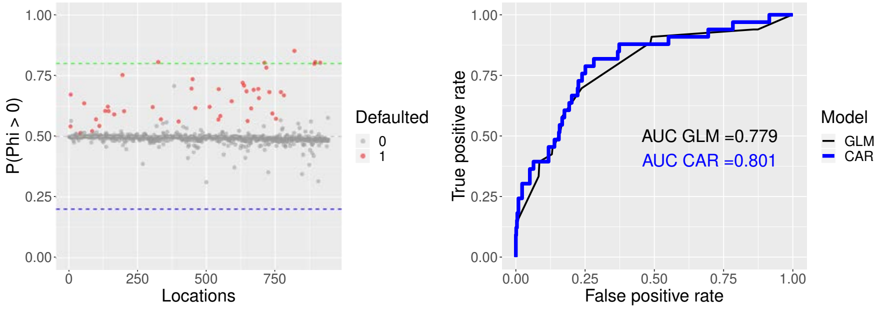


Figure 1. *Left: Estimated probability of a strictly positive spatial effect (i.e., $\widehat{\mathbb{P}}(\phi_k > 0)$) for each firm. Red dots are defaulted firms ($Y_k = 1$) with estimated probability of strictly positive spatial effects greater than 50%. Black dots indicate all other firms. Right: ROC curves and AUC for a GLM considering only covariates $\boldsymbol{x}_k$ (black) and CAR model (blue) for the prediction six-months ahead with respect to training.*

To conclude, the application of disease mapping methods to a scale free network represents a novelty at present. The encouraging results on the out-of-time set suggest to further investigate spatial modelling of trade networks.

References

BANERJEE, SUDIPTO, CARLIN, BRADLEY P, & GELFAND, ALAN E. 2003. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

BARABÁSI, ALBERT-LÁSZLÓ, & ALBERT, RÉKA. 1999. Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

DOLFIN, MARINA, KNOPOFF, DAMIAN, LIMOSANI, MICHELE, & XIBILIA, MARIA GABRIELLA. 2019. Credit risk contagion and systemic risk on networks. *Mathematics*, **7**(8), 713.

LAMIERI, MARCO, & SANGALLI, ILARIA. 2019. The propagation of liquidity imbalances in manufacturing supply chains: evidence from a spatial auto-regressive approach. *The European Journal of Finance*, **25**(15), 1377–1401.

SEMIPARAMETRIC FINITE MIXTURE OF REGRESSION MODELS WITH BAYESIAN P-SPLINES

Marco Berrettini [1], Giuliano Galimberti [1]and Saverio Ranciati [1]

[1] Department of Statistical Sciences , University of Bologna,
(e-mail: marco.berrettini2@unibo.it,giuliano.galimberti@unibo.it,
saverio.ranciati2@unibo.it)

ABSTRACT: A semiparametric finite mixture of regression models is defined, with concomitant information assumed to influence both the component weights and the conditional means. The contribution of a concomitant variable is flexibly specified as a smooth function represented by cubic splines. A Bayesian estimation procedure is proposed and an empirical analysis of the baseball salaries dataset is illustrated.

KEYWORDS: mixture of experts models, Gibbs sampling, data augmentation

1 Introduction

Mixture models provide a useful tool to account for unobserved heterogeneity. In order to gain additional flexibility, some model parameters can be expressed as functions of concomitant covariates, introducing the Mixture of Experts (MoE) framework. In this Paper, a semiparametric MoE regression model is proposed, where component weights and conditional means are smooth functions of a univariate covariate. Estimation is carried out within the Bayesian paradigm: a new Gibbs sampler algorithm is developed, exploiting data augmentation to express the effect of the covariate on the component weights, as in Früwirth-Schnatter et al. (2012). Bayesian P-splines (Lang & Brezger, 2004) are used to achieve a parsimonious representation of the smooth functions.

2 Model specification

Suppose that $\{y_i\}, i = 1, \ldots, n$ is a random sample from a population clustered into G components. It is assumed that the conditional distribution of y_i, given a concomitant covariate x_i, is represented by the following MoE model:

$$f(y_i|x_i) = \sum_{g=1}^{G} \pi_g(x_i) f_{\mathcal{N}}\left(\mu_g(x_i), \sigma_g^2\right). \tag{1}$$

Each component $g = 1, \dots, G$ is modelled by a normal density function $f_{\mathcal{N}}(\cdot)$, and has weight $\pi_g(x_i) > 0$, such that $\sum_{g=1}^{G} \pi_g(x_i) = 1$, for $i = 1, \dots, n$. The conditions for identifiability of Model (1) can be deduced by the ones Huang et al. (2013) provide for their nonparametric mixture of regression models. Jacobs et al. (1991) model the component weights $\pi_g(x_i)$ using a multinomial logistic regression model, thus expressing the log-odds of these probabilties, with respect to the reference one (e.g., the G-th), as linear functions of the covariate x_i. In this Paper, each of these $G-1$ linear predictors is replaced with an additive structure, defined as a linear combination of m cubic B-spline bases $B_\rho(\cdot)$ and coefficients $\gamma_{g\rho}$:

$$\log \frac{\pi_g(x_i)}{\pi_G(x_i)} = \eta_g(x_i) = \sum_{\rho=1}^{m} B_\rho(x_i)\gamma_{g\rho}, \quad \text{for } i = 1, \dots, n. \tag{2}$$

In the Bayesian framework, Lang & Brezger (2004) suggest a high number of knots to ensure enough flexibility, and to define priors for the regression parameters $\gamma_{g1}, \dots, \gamma_{gm}$ in terms of a random walk:

$$\gamma_{g\rho} = \gamma_{g,\rho-1} + w_{g\rho}, \quad w_{g\rho} \sim N(0, \delta_g^2). \tag{3}$$

The amount of smoothness is controlled by the additional variance parameters δ_g^2. Their presence protect against possibile overfitting when a large number of knots is chosen. The multinomial model in Equation (2) can be conveniently represented as a binary formulation in the partial difference random utility model (dRUM) representation proposed by Früwirth-Schnatter et al. (2012), conditional on knowing each $\lambda_g(x_i) = \exp(\eta_g(x_i))$:

$$z_{gi} = \eta_g(x_i) - \log\left(\sum_{l \neq g} \lambda_l(x_i)\right) + \varepsilon_{gi}, \quad D_{gi} = \mathbf{1}(z_{gi} > 0); \tag{4}$$

where z_{gi} is a latent variable, D_{gi} is the allocation indicator and $\varepsilon_{gi}, i = 1, \dots, n$, are i.i.d. errors following a logistic distribution. To avoid any Metropolis-Hastings (MH) step, Früwirth-Schnatter et al. (2012) approximate the logistic distribution of the error terms ε_{gi} by a finite scale mixture of normal distributions with parameters drawn with fixed probabilities.

Regarding the components' normal densities, each mean $\mu_g(\cdot)$ is assumed to be an unknown smooth function of covariate x, represented by Bayesian P-splines:

$$\mu_g(x_i) = \sum_{\rho=1}^{m} B_\rho(x_i)\beta_{g\rho}, \quad \beta_{g\rho} = \beta_{g,\rho-1} + u_{g\rho}, \quad u_{g\rho} \sim N(0, \tau_g^2). \tag{5}$$

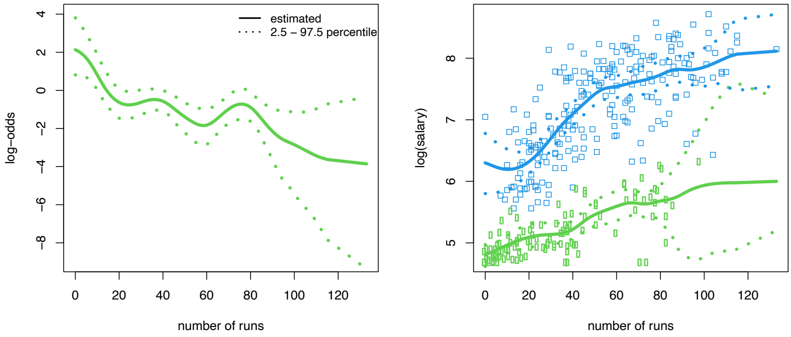


Figure 1. *Estimated posterior effects (and pointwise 95% posterior credible bands) of the number of runs on the log-odds $\eta_1(x)$ (left plot) and conditional means $\mu_1(x)$ and $\mu_2(x)$ (right plot), in green and blue respectively.*

The proposed Gibbs sampler requires the number of components G to be fixed. The optimal number of components can be selected according to the Akaike's Information Criterion for MCMC samples (AICM) proposed by Raftery et al. (2007). Finally, to obtain a hard clustering, observations can be allocated into the G components using the maximum-a-posteriori (MAP) rule once the algorithm completes the prefixed number of iterations.

3 Application: Baseball salaries

Watnik (1998) provides a dataset consisting of information about players for the 1992 Major League Baseball season. The following analysis evaluates the effect of the number of runs (x), taken as a measure of a player's contribution to the team, on the log-salary (y). Number of components G ranging from 1 to 4 has been considered; the optimal value resulted to be equal to 2 for the proposed model, according to AICM. The left plot of Figure 1 shows a lack of monotonicity in the effect of the number of runs on the log-odds of the mixture weight $\eta_1(x)$. However, the overall decreasing trend indicates a lower prior probability of belonging to Cluster 1, rather than Cluster 2 (i.e. the reference one), for players providing better performances in terms of number of runs. Players' allocations, with respect to x and y, are depicted in the right plot of the same Figure, where it can also be noticed a nonlinear effect of the number of runs on the log-salary for Cluster 2 (the upper one, in blue), while the bands does not exclude a linear effect for Cluster 1 (the lower one, in green). These two clusters appear quite well separated, apart from the region with low values of both x and y. Group 1 might be broadly interpreted as the cluster of "underrated" (or "underpaid") baseball players. In fact, while it is obvious that players with better performances get paid more, as is comfirmed by the increasing trends of both means, there seems to be a group of players whose salary is substantially lower than that of players with similar performances (in terms of number of runs), belonging to the upper group (in blue). Indeed, the two estimated function for $\mu_1(x)$ and $\mu_2(x)$ in the right plot of Figure 1 appear almost parallel.

References

FRÜHWIRTH-SCHNATTER, S., PAMMINGER, C., WEBER, A, & WINTER-EBMER, R. 2012. Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *Journal of Applied Econometrics*, **27**, 1116–1137.

HUANG, MIAN, LI, RUNZE, & WANG, SHAOLI. 2013. Nonparametric mixture of regression models. *Journal of the American Statistical Association*, **108**(503), 929–941.

JACOBS, R.A., JORDAN, M.I., NOWLAN, S.J., & HINTON, G.E. 1991. Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87.

LANG, S., & BREZGER, A. 2004. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

RAFTERY, A.E., NEWTON, M.A., SATAGOPAN, J.M., & KRIVITSKY, P.N. 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Pages 1–45 of:* BERNARDO, J.M., BAYARRI, M.J., BERGER, J.O., DAWID, A.P., HECKERMAN, D., SMITH, A.F.M., & WEST, M. (eds), *Bayesian Statistics*, vol. 8. Oxford University Press.

WATNIK, MITCHELL R. 1998. Pay for play: Are baseball salaries based on performance? *Journal of Statistics Education*, **6**(2).

A SUBJECT-SPECIFIC MEASURE OF INTERRATER AGREEMENT BASED ON THE HOMOGENEITY INDEX

Giuseppe Bove[1]

[1] Dipartimento di Scienze della Formazione, Università degli Studi Roma Tre,
(e-mail: giuseppe.bove@uniroma3.it)

ABSTRACT: Interrater agreement for classifications on nominal scales is usually evaluated by overall measures across subjects like the Cohen's kappa index. In this paper, the homogeneity index for a qualitative variable is proposed to evaluate the agreement between raters for each single case (subject or object), and to obtain also a global measure of the interrater agreement for the whole group of cases evaluated. The subject-specific and the global measures proposed do not depend on a particular definition of agreement (simultaneously between two, three or more raters) and are not influenced by the marginal rater distributions of the scale like most of the kappa-type indexes.

KEYWORDS: nominal classification scales, interrater agreement, homogeneity index.

1 Introduction

In behavioral and biomedical sciences classifications of subjects or objects into predefined classes or categories and the analysis of their agreement are a rather common activity. For instance, agreement between clinical diagnoses provided by more physicians (raters) is considered for identifying the best treatment for the patient, and the extent to which the diagnoses coincide, the rating procedure (or scale) can be used with confidence. Hence, in this type of applications it is important to analyse interrater absolute agreement, that is the extent that raters assign the same (or very similar) values on the rating scale.

Agreement between two raters who rate each of a sample of subjects (objects) on a nominal scale is usually assessed with Cohen's kappa (Cohen 1960). Generalizations of kappa for the case of more than two raters and for the case where raters assessing one subject are not always the same have been proposed by many authors (e.g., Fleiss, 1971, Conger 1980). These indexes are used to analyse the agreement between multiple raters for a whole group of subjects. Moreover, methods to detect subsets of raters who demonstrate a high level of interobserver agreement were considered, for instance, by Landis & Koch (1977). Less frequently agreement on a single subject has been considered (O'Connell & Dobson, 1984), in spite of the fact that having evaluations of the agreement on the single case is particularly useful, for example, in situations where the rating scale is being tested, and it is necessary to identify any changes to improve it, or to request the raters for a specific comparison on the single case in which agreement is poor.

In the next sections an index to measure the interrater agreement on a single subject is proposed based on a measure of dispersion for nominal variables. Furthermore, a global measure of agreement on the whole group of subjects obtained as the arithmetic average of the subject values of the index will be also considered and applied to a data set concerning the cause of death of 35 hypertensive patients.

2 Method

O'Connell and Dobson (1984) proposed a chance-corrected measure of agreement for several raters using nominal (or ordinal) categories on a single subject i (i=1,2, ….,N), given by

$$S_i = 1 - D_i/\Delta,$$

where D_i is the overall disagreement on the whole response profile i and Δ is the disagreement expected by chance (see O'Connell and Dobson (1984), equation (6)). The measure takes the value 1 when there is perfect agreement; it is positive when the agreement is better than chance, and negative otherwise. Besides, an overall measure of agreement across subjects S_{av} can be obtained as the arithmetic average of the S_i individual values. The index S_i has some drawbacks: 1) it cannot be computed for only one observation, because in that case the disagreement expected by chance Δ is not defined; 2) it is formulated in terms of agreement statistics based on all pairs of raters, but some authors argued that simultaneous agreement among three or more raters can be alternatively considered (e.g., see Warrens, 2012); 3) agreement expected by chance depends on the observed proportions of subjects allocated to the categories of the scale by each rater, and this imply that the measure of agreement depends on the marginal distributions of the categories of the scale observed for each rater (for this aspect see, e.g., Marasini *et al.*, 2016).

A different approach is proposed here that is based on the largely known homogeneity index to measure the dispersion of a qualitative variable (e.g., Leti 1983). For a classification in K categories the index is given by

$$O = \sum_{j=1}^{K} f_j^2,$$

where f_j is the proportion of ratings in category j (j=1,2,…,K). The index is equal to 1 in the case of maximum homogeneity (perfect agreement), and $1/K$ in the case of maximum heterogeneity (total disagreement, for each category j is $f_j = 1/K$). O depends on the number of categories, so the normalization in the interval [0,1] is given by

$$O_{rel} = (K\,O - 1)/(K - 1).$$

Thus, O_{rel} : 1) is equal to zero for total disagreement and one for perfect agreement; 2) can be computed also for only one observation; 3) does not depend on the definition of pairwise agreement; 4) does not depend on the observed proportions of subjects allocated to the categories of the scale.

A global measure of agreement on the whole group (indicated with $\bar{O}_{rel}$) can be easily obtained as the arithmetic average of the individual values of O_{rel}.

3 Application

Data considered are about a study with seven nosologists assessing the cause of death of 35 hyperthensive patients by using the death certificates (Woolson, 1987). The scores were assigned by the following categories: 1=Arteriosclerotic disease, 2= cerebrovascular disease, 3=other hearth disease, 4=renal disease, 5=other disease. The marginal proportions of ratings for the five categories were 0.21, 0.17, 0.19, 0.27 and 0.16, respectively. Some preliminary results are presented for the method based on the O_{rel} index.

The subjective values of O_{rel} allowed to detect low level of agreement for many evaluations (28.6% of the O_{rel} values less than 0.4), that call for a possible revision of the assessment procedure. It can be also interesting to analyse some descriptive statistics provided in Table 1 for the comparison of S_i and O_{rel}. The mean values for the global agreement are S_{av}=0.48 and $\bar{O}_{rel}$=0.56. S_i values show higher dispersion respect to the O_{rel} values. The measures are almost perfectly correlated (*r*=0.99).

Table 1: Some descriptive statistics for S_i and O_{rel} values

	N	Mean	Std. Dev.	CV
S_i	*35*	*0.48*	*0.27*	*56.5*
O_{rel}	*35*	*0.56*	*0.23*	*42.1*

We also add that the value of the average Cohen's kappa coincides with S_{av} and the value of Fleiss kappa (Fleiss, 1971) is also approximately equal to 0.48.

It is interesting to point out that if we increase the level of agreement between raters by collapsing the five categories in the two strongly unbalanced categories cerebrovascular disease (marginal proportion 0.17) and all other diseases (marginal proportion 0.83), the values of S_{av}, average Cohen's kappa and Fleiss kappa remain almost the same, while the new value of $\bar{O}_{rel}$ increases to 0.75, accordingly to the new high level of agreement. It is not uncommon in applications to have highly unbalanced categories, this happens, for example, when a diagnostic category is rare or when for some reasons the raters use almost exclusively very few levels of the scale.

4 Conclusion

A descriptive approach to the analysis of absolute interrater agreement has been proposed that presents some advantages respect to the approach by kappa-type measures based on pairwise agreement between raters. The index proposed is mainly considered as a measure of size of the interrater agreement, therefore future developments may concern the definition of reliable thresholds useful in the application. Finally, we notice that a measure of interrater agreement for ordinal data recently proposed and applied in educational studies follow an approach similar to the present proposal (Bove *et al.* 2020), where a measure of dispersion for ordinal variables is considered instead of the homogeneity index.

References

BOVE, G., CONTI, P.L., & MARELLA, D. 2020. A measure of interrater absolute agreement for ordinal categorical data. *Statistical Methods & Applications*, doi.org/10.1007/s10260-020-00551-5.

COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.,* **20**, 213–220.

CONGER, A.J. 1980. Integration and generalization of kappas for multiple raters. *Psychol. Bull.,* **88**, 322–328.

FLEISS, J.L. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, **76**, 378–382.

LANDIS, J.R., & KOCH, G.G. 1977. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, **33**, 363-374.

LETI, G. 1983. *Statistica descrittiva*. Bologna: Il Mulino.

MARASINI, D., QUATTO P., & RIPAMONTI E. 2016. Assessing the interrater agreement for ordinal data through weighted indexes. *Statistical Methods in Medical Research*, **25**, 2611-2633.

O'CONNELL, D.L., & DOBSON, A.J. 1984. General Observer-Agreement Measures on Individual Subjects and Groups of Subjects. *Biometrics,* **40**, 973-983.

WARRENS, M. J. 2012. Equivalences of weighted kappas for multiple raters. *Statistical Methodology*, **9**, 407-422.

WOOLSON, R. F. 1987. *Statistical methods for the analysis of biomedical data*. New York: John Wiley and Sons.

ESTIMATING LATENT LINEAR CORRELATIONS FROM FUZZY CONTINGENCY TABLES

Antonio Calcagnì [1]

[1] DPSS, University of Padova, Italy (e-mail: antonio.calcagni@unipd.it)

ABSTRACT: In this contribution, we describe a method to estimate polychoric correlations when data are available in the form of fuzzy frequency tables. A simulation study is used to assess the characteristics of the proposed approach. Fuzzy polychoric correlations can be of particular utility, for instance, in studies involving covariance structural analysis (e.g., CFA) and dimensionality reduction techniques (e.g., EFA).

KEYWORDS: fuzzy frequencies, polychoric correlations, fuzzy classification

1 Introduction

The latent linear correlation (LLC), also called polychoric correlation, is a measure of linear association which is usually adopted when dealing with categorical variables or statistics such as frequency or contingency tables. Given a set of J variables, LLC is computed pairwise for each pair (j,k) of variables by considering their joint frequencies $\mathbf{N}^{(j,k)}_{R\times C} = (n^{(j,k)}_{11},\ldots,n^{(j,k)}_{rc},\ldots,n^{(j,k)}_{RC})$ over a $R_{jk} \times C_{jk}$ partition space of the variables' domain. The general idea is to adopt a bivariate Gaussian distribution with correlation ρ_{jk} as a latent statistical model underlying the observed frequency table $\mathbf{N}^{(j,k)}_{R\times C}$, which maps the $R_{jk} \times C_{jk}$ space to the real domain of the bivariate density via a threshold-based approach. There are several contexts in which LLCs have been applied, including covariance structural analysis (e.g., CFA) and dimensionality reduction techniques (e.g., PCA, EFA). In this contribution, we generalize the problem of estimating polychoric correlations from fuzzy frequency tables, which are of particular utility when observed data are classified using fuzzy categories as done, for example, in socio-economic studies, images/videos classification, and content analysis. In all these cases, the $R_{jk} \times C_{jk}$ space of the variables' domain constitutes a fuzzy partition and observed counts in $\mathbf{N}^{(j,k)}_{R\times C}$ are no longer natural numbers. In order to deal with this issue, in this paper we describe a novel way to compute fuzzy frequency tables and provide a way to estimate ρ_{jk} when observed frequencies are fuzzy. In what follows, we will set $R = C$ and $J = 2$ for the sake of simplicity.

2 Fuzzy frequencies

A fuzzy subset $\tilde{A}$ of a universal set $\mathcal{A}$ is defined by means of its characteristic function $\xi_{\tilde{A}} : \mathcal{A} \to [0,1]$. Let $\mathcal{A} \subset \mathbb{R}$ without loss of generality and consider (X,Y) a pair of random variables taking values on $\mathcal{A}$. Then $\mathcal{A}$ can conveniently be partitioned into a collection of fuzzy subsets, namely $\mathcal{C}_j = \{\tilde{C}_1,\ldots,\tilde{C}_r,\ldots,\tilde{C}_R\}$ and $\mathcal{C}_k = \{\tilde{C}_1,\ldots,\tilde{C}_c,\ldots,\tilde{C}_C\}$. The random realizations $\mathbf{x} = (x_1,\ldots,x_I)$ and $\mathbf{y} = (y_1,\ldots,y_I)$ can partially or fully be classified into $\mathcal{C}_j$ or $\mathcal{C}_k$. The evaluation of the amount of sample realizations over $\tilde{C}_j$ or $\tilde{C}_k$ is called *cardinality*. This is a natural number or crisp count (i.e., $n_{rc} \in \mathbb{N}_0$) when the observations fully belong to subsets of $\tilde{C}_j$ or $\tilde{C}_k$. On the opposite case, it is a fuzzy natural number $\tilde{n}_{rc} \in \mathcal{F}(\mathbb{N})$, with $\mathcal{F}(\mathbb{N})$ being the set of all *generalized natural numbers* (Bodjanova & Kalina, 2008). Let $\tilde{C}_{rc}$ be an element of the fuzzy Cartesian product $\tilde{C}_j \;\tilde{\times}\; \tilde{C}_k$. Then a fuzzy count $\tilde{n}_{rc}$ is a fuzzy set with membership function $\xi_{\tilde{n}_{rc}} : \mathbb{N}_0 \to [0,1]$ being computed as follows: $\xi_{\tilde{n}_{rc}}(n) = \min(\nu_{rc}(n), \mu_{rc}(n))$, with $\nu_{rc}(n) = \mathrm{FGC}(\boldsymbol{\varepsilon}_{rc})$ and $\mu_{rc}(n) = \mathrm{FLC}(\boldsymbol{\varepsilon}_{rc})\ \forall\, n \in \{0,1,\ldots,I\} \subset \mathbb{N}_0$. In this context, FGC(.) and FLC(.) are the fuzzy counting functions as defined by Zadeh (1983) whereas $\boldsymbol{\varepsilon}_{rc} = \min(\xi_{\tilde{C}_r}(\mathbf{x}_j), \xi_{\tilde{C}_c}(\mathbf{y}_k))$ contains the joint degrees of inclusion of the sample observations $\mathbf{x}$ and $\mathbf{y}$ w.r.t. the fuzzy categories. More details can be found in Bodjanova & Kalina (2008). Finally, the fuzzy frequency table $\widetilde{\mathbf{N}}_{R\times C}$ can be computed by applying the above calculus over $r = 1,\ldots,R$ and $c = 1,\ldots,C$.

3 LLCs for fuzzy frequency tables

The latent statistical model underlying the sample realizations is bivariate Gaussian $(X^*,Y^*) \sim \mathcal{N}(\mathbf{0},\rho)$ under the constraints that $(X \in \tilde{C}_r) \wedge (Y \in \tilde{C}_c)$ iif $(X^*,Y^*) \in (\tau^X_{r-1},\tau^X_r] \times (\tau^Y_{c-1},\tau^Y_c] \subset \mathbb{R}^2$ for all $r = 1,\ldots,R$ and $c = 1,\ldots,C$. The thresholds $\boldsymbol{\tau}^X$ and $\boldsymbol{\tau}^Y$ are defined so that $\tau_0 = -\infty$ and $\tau_R = \infty$ for both X and Y variables. Note that (X^*,Y^*) are unobserved pairs of latent variables. Following Olsson (1979), the parameters $\boldsymbol{\theta} = \{\rho,\boldsymbol{\tau}^X,\boldsymbol{\tau}^Y\} \in [-1,1] \times \mathbb{R}^{R-1} \times \mathbb{C}^{C-1}$ can be estimated using a two step-approach. In particular, given the filtered counts at the current iteration, thresholds are estimated using the cumulative marginals of $\widehat{\mathbf{N}}_{R\times C}$ (first step). Then, ρ is estimated by maximizing the log-likelihood implied by the model conditioned on $\hat{\boldsymbol{\tau}}^X$ and $\hat{\boldsymbol{\tau}}^Y$ (second step):

$$\ln \mathcal{L}(\boldsymbol{\theta};\mathbf{N}) \propto \sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc} \ln \int_{\tau^X_{r-1}}^{\tau^X_r} \int_{\tau^Y_{c-1}}^{\tau^Y_c} \phi(x,y;\rho)\; dxdy \qquad (1)$$

with $\phi(x,y;\rho)$ being the bivariate Gaussian density centered at zero. In what follows, we will focus on estimating ρ as estimation of thresholds follows straightforwardly from Olsson (1979). As we observe fuzzy frequencies $\widetilde{\mathbf{N}}_{R\times C}$, we solve the maximization problem via the fuzzy EM algorithm proposed by Denoeux (2011), which in this case requires the computation of the following quantity:

$$\mathbb{E}_{\boldsymbol{\theta}'}\left[\ln\mathcal{L}(\boldsymbol{\theta};\mathbf{N})|\widetilde{\mathbf{N}}\right] \propto \sum_{r=1}^{R}\sum_{c=1}^{C}\mathbb{E}_{\boldsymbol{\theta}'}\left[N_{rc}|\tilde{n}_{rc}\right]\ln\int_{\tau^X_{r-1}}^{\tau^Y_r}\int_{\tau^Y_{c-1}}^{\tau^Y_c}\phi(x,y;\rho)\;dxdy \qquad (2)$$

given a candidate estimate $\boldsymbol{\theta}'$. The quantity $N_{rc}|\tilde{n}_{r,c}$ is a random variable conditioned on a fuzzy event:

$$\mathbb{E}_{\boldsymbol{\theta}'}\left[N_{rc}|\tilde{n}_{rc}\right] = \sum_{n\in\mathbb{N}_0}\frac{\xi_{\tilde{n}_{rc}}(n)f_{N_{rc}}(n;\pi_{rc}(\boldsymbol{\theta}))}{\sum_{n\in\mathbb{N}_0}\xi_{\tilde{n}_{rc}}(n)f_{N_{rc}}(n;\pi_{rc}(\boldsymbol{\theta}))}\,n \qquad (3)$$

where $f_{N_{rc}}(n;\pi_{rc}(\boldsymbol{\theta})) = \mathcal{B}in(n;\pi_{rc}(\boldsymbol{\theta}))$, with $\pi_{rc}(\boldsymbol{\theta}) = \int_{\tau^X_{r-1}}^{\tau^Y_r}\int_{\tau^Y_{c-1}}^{\tau^Y_c}\phi(x,y;\rho)\;dxdy$. Note that $\hat{n}_{rc} = \mathbb{E}_{\boldsymbol{\theta}'}\left[N_{rc}|\tilde{n}_{rc}\right]$ denotes the reconstructed rc-th count. The fuzzy EM algorithm proceeds by alternating between the computation of Eq. (3) and the maximization of Eq. (1) once $\hat{n}_{rc}$ has been obtained.

4 Simulation study

The aim of this Monte Carlo study is twofold. First, we will evaluate the performances of fuzzy-EM estimator for ρ_{jk} when fuzzy frequency data are available. Second, we will assess whether the standard maximum likelihood estimator for polychoric correlations performs as good as the proposed method if applied on max-based and mean-based defuzzified data. The case $J = 2$ was considered for the sake of simplicity.

Design. The design involved two factors, namely (i) $I \in \{150, 250, 500\}$, and (ii) $\rho \in \{0.15, 0.50, 0.85\}$, which were varied in a complete factorial design. For each combination, $B = 5000$ samples were generated.

Data generation. For each condition of the simulation design, data were generated according to a two-step procedure. First, a crisp frequency table $\mathbf{N}_{R\times C}$ was computed using the approximation $n_{rc} = I\cdot\pi_{rc}$ ($r = 1,\dots,R$; $c = 1,\dots,C$), with $\boldsymbol{\tau}^X = \boldsymbol{\tau}^Y = (-2,-1,0,1,2)$. Second, each element of $\mathbf{N}_{R\times C}$ was fuzzified via the following probability-possibility transformation: $\boldsymbol{\xi}_{\tilde{n}_{rc}} = f_{\mathcal{G}_\text{d}}(\mathbf{n};\alpha_{rc},\beta_{rc})\big/\max f_{\mathcal{G}_\text{d}}(\mathbf{n};\alpha_{rc},\beta_{rc})$, $\alpha_{rc} = 1 + m_1\beta_{s_1}$, $\beta_{s_1} = 1 + (m_1 + m_1^2 +$

$4s_1^2)^{\frac{1}{2}}/2s_1^2$, $\beta_{rc} = (m_1 + m_1^2 + 4s_1^2)^{\frac{1}{2}}/2s_1^2$, $m_1 \sim \mathcal{G}amma_{\mathrm{d}}(\alpha_{m_1}, \beta_{m_1})$ where $\alpha_{m_1} = 1 + n_{rc}\beta_{m_1}$, $\beta_{m_1} = (n_{rc} + n_{rc}^2 + 4s_1^2)^{\frac{1}{2}}/2s_1^2$, $s_1 \sim \mathcal{G}amma_{\mathrm{d}}(\alpha_{s_1}, \beta_{s_1})$, $\alpha_{s_1} = 1 + m_0\beta_{s_1}$, $\beta_{s_1} = (m_0 + m_0^2 + 4s_0^2)^{\frac{1}{2}}/2s_0^2$, $m_0 = 1$ and $s_0 = 0.15$. Note that $f_{\mathcal{G}_{\mathrm{d}}}$ is the density of the discrete Gamma random variable $\mathcal{G}amma_{\mathrm{d}}$.

Outcome measures. For each condition of the simulation design, sample results were evaluated using bias of estimates and root mean square error.

Results. Table 1 shows the results of the study. As expected, fEM outperformed standard ML applied on both max-based and mean-based defuzzified data in terms of bias and root mean square errors. This is mainly due to the fact that ρ_{fEM} estimator weights the observed fuzzy data $\xi_{\widetilde{n}_{rc}}$ with the probabilistic model for the unobserved n_{rc}.

	fEM		dML (max)		dML (mean)	
	bias	*rmse*	*bias*	*rmse*	*bias*	*rmse*
$\rho = 0.15$						
$I = 150$	0.0358	0.0881	-0.0105	0.1142	-0.0402	0.0846
$I = 250$	0.0043	0.0514	-0.0284	0.0817	-0.0403	0.0683
$I = 500$	0.0099	0.0297	0.0020	0.0416	-0.0082	0.0335
$\rho = 0.50$						
$I = 150$	0.0103	0.0747	-0.0933	0.1545	-0.1797	0.1956
$I = 250$	-0.0363	0.0626	-0.1216	0.1488	-0.1706	0.1800
$I = 500$	-0.0006	0.0264	-0.0457	0.0689	-0.0828	0.0903
$\rho = 0.85$						
$I = 150$	0.0013	0.0441	-0.2150	0.2525	-0.3274	0.3354
$I = 250$	-0.0028	0.0269	-0.1707	0.1967	-0.2580	0.2642
$I = 500$	-0.0009	0.0145	-0.1034	0.1211	-0.1630	0.1672

Table 1. *Monte Carlo study: Estimating* ρ *via fuzzy-EM (fEM) and standard ML (dML) on max-based and mean-based defuzzified frequency tables.*

References

BODJANOVA, SLAVKA, & KALINA, MARTIN. 2008. Cardinalities of Granules of Vague Data. *Pages 63–70 of:* MAGDALENA, L., OJEDA-ACIEGO, M., & VERDEGAY, J.L. (eds), *Proceedings of IPMU2008, Torreliminos (Malaga), June 22-27 2008.*

DENOEUX, THIERRY. 2011. Maximum likelihood estimation from fuzzy data using the EM algorithm. *Fuzzy sets and systems*, **183**(1), 72–91.

OLSSON, ULF. 1979. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, **44**(4), 443–460.

ZADEH, LOTFI A. 1983. A computational approach to fuzzy quantifiers in natural languages. *Pages 149–184 of: Computational linguistics*. Elsevier.

MODEL-BASED CLUSTERING WITH SPARSE MATRIX MIXTURE MODELS

Andrea Cappozzo [1], Alessandro Casa[2] and Michael Fop[2]

[1] Deparment of Mathematics, Politecnico di Milano
(e-mail: andrea.cappozzo@polimi.it)

[2] School of Mathematics and Statistics, University College Dublin
(e-mail: alessandro.casa@ucd.ie, michael.fop@ucd.ie)

ABSTRACT: In recent years we are witnessing to an increased attention towards methods for clustering matrix-valued data. In this framework, matrix Gaussian mixture models constitute a natural extension of the model-based clustering strategies. Regrettably, the overparametrization issues, already affecting the vector-valued framework in high-dimensional scenarios, are even more troublesome for matrix mixtures. In this work we introduce a sparse model-based clustering procedure conceived for the matrix-variate context. We introduce a penalized estimation scheme which, by shrinking some of the parameters towards zero, produces parsimonious solutions when the dimensions increase. Moreover it allows cluster-wise sparsity, possibly easing the interpretation and providing richer insights on the analyzed dataset.

KEYWORDS: model-based clustering, penalized likelihood, sparse matrix estimation, EM-algorithm

1 Introduction

Model-based clustering represents a well established framework to cluster multivariate data. When dealing with continuous data, the generative mechanism is routinely described by means of Gaussian Mixture Models (GMMs). Partitions are obtained by exploiting the one-to-one correspondence between the groups and the components of the mixture. This approach has been used in many different applications; nonetheless GMMs tend to be over-parameterized in high-dimensional settings where their usefulness might be jeopardized.

This problem complicates even further in three-way data scenarios, where multiple variables are measured on different occasions for the considered units. Here matrix-variate distributions have often been used and embedded in the mixtures framework, thus providing a valid solution when partitions of matrices are required (Viroli, 2011). In spite of its strenght points, this approach is dramatically over-parameterized even in moderate dimensions. Therefore, we propose a penalized model-based clustering strategy in the matrix-variate framework. Our approach reduces the number of parameters to be estimated, by shrinking some of them towards zero, and possibly leads to a gain in terms of interpretability. The rest of the paper is organized as follows. In Section 2 we introduce matrix Gaussian mixture models (MGMMs) and we outline our proposal. An application to real world data is reported in Section 3 alongside with some concluding remarks and possible future research directions.

2 Penalized matrix-variate mixture model

Let $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ be a set of n matrices with $\mathbf{X}_i \in \mathbb{R}^{p\times q}$. MGMM provides an extension of the GMM when clustering of matrices are needed. The density of $\mathbf{X}_i$ is then expressed as follows

$$f(\mathbf{X}_i;\Theta) = \sum_{k=1}^{K} \tau_k \phi_{(p\times q)}(\mathbf{X}_i; M_k, \Omega_k, \Gamma_k) \tag{1}$$

where $\Theta = \{\tau_k, M_k, \Omega_k, \Gamma_k\}_{k=1}^{K}$, τ_k's are the mixing proportions, with $\tau_k > 0$ and $\sum_k \tau_k = 1$. On the other hand $\phi_{(p\times q)}(\mathbf{X}_i; M_k, \Omega_k, \Gamma_k)$ denotes the density of a $p \times q$ matrix normal distribution where $M_k \in \mathbb{R}^{p\times q}$ is the mean of the *k-th* component while $\Omega_k \in \mathbb{R}^{p\times p}$ and $\Gamma_k \in \mathbb{R}^{q\times q}$ represent respectively the rows and the columns component precision matrices.

In (1) the number of parameters to estimate scales quadratically with both p and q, endangering the pratical usefulness of the model. Recently some solutions have been proposed, trying to overcome this issue (see Wang & Melnykov, 2020 and Sarkar *et al.* , 2020). These approaches present some drawbacks as they are computationally intensive and as they implement a rigid way to induce parsimony. Therefore in this work we take a different path, adopting a penalized estimation approach which implicitly assumes that M_k, Ω_k, Γ_k, for $k = 1, \dots, K$, possess some degree of sparsity.

To this aim, we introduce a penalized likelihood strategy to obtain $\hat{\Theta}$. The log-likelihood function to be maximized is defined as

$$\ell(\Theta;\mathbf{X}) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X}_i; M_k, \Omega_k, \Gamma_k)\right\} - p_{\lambda_1,\lambda_2,\lambda_3}(M_k, \Omega_k, \Gamma_k) \tag{2}$$

with the penalization term $p_{\lambda_1,\lambda_2,\lambda_3}(M_k, \Omega_k, \Gamma_k)$ equals to

$$p_{\lambda_1,\lambda_2,\lambda_3}(M_k, \Omega_k, \Gamma_k) = \sum_{k=1}^{K} \lambda_1 ||P_1 * M_k||_1 + \sum_{k=1}^{K} \lambda_2 ||P_2 * \Omega_k||_1 + \sum_{k=1}^{K} \lambda_3 ||P_3 * \Gamma_k||_1$$

Table 1. *Adjusted Rand Index (ARI) and number of free estimated parameters for three clustering procedures.*

	Sparsemixmat	Sarkar *et al.* , 2020	GMM
ARI	0.7883	0.7772	0.3841
# of parameters	218	275	850

P_1, P_2, P_3 are matrices with non-negative entries, $||A||_1 = \sum_{jh} |A_{jh}|$, $\lambda_1, \lambda_2, \lambda_3$ are the penalization parameters while $*$ denotes the element-wise product.
To estimate Θ, we devise an ad-hoc EM-algorithm which maximizes the *penalized complete data log-likelihood* associated with (2). The E-step computes class membership a posteriori probabilities via the standard updating formula. On the other hand the M-step consists of three partial optimization cycles. An estimate for M_k is obtained by means of a cell-wise coordinate ascent algorithm while, to estimate Ω_k and Γ_k, we propose a suitable modification of the graphical LASSO (Friedman *et al.* , 2008). The resulting model, inducing sparsity in the precision matrices, accounts for cluster-wise conditional independence patterns, which might ease the interpretation of the results, and possibly provides indications about irrelevant variables. Moreover the number of parameters is reduced without imposing rigid structures.

3 Application and concluding remarks

We employ the procedure outlined in Section 2 to obtain a partition of the Landsat satellite data, where $n = 845$ matrices, with dimensions 4×9, coming from three different classes are available (see Viroli, 2011 for a detailed description). In Table 1 we report the results obtained with the proposed procedure (Sparsemixmat) and with two plausible competitors being the approach by Sarkar *et al.* , 2020 and the standard GMM applied to the unfolded two-way representation of the data. Our model outperforms the competitors, when recovering the true clustering structure is the aim. Furthermore, we provide the most parsimonious solution, displaying the lowest number of non zero estimated parameters. The retrieved sparse matrix structures are graphically displayed, for the three classes, in Figure 1. While the clustering is mainly driven by the different patterns in M_k's, the Γ_k's are the ones showing the highest degree of sparsity, with different intensities for the three classes.

The promising results obtained in the application demonstrate how the penalized matrix-variate mixture model proposed in this work might alleviate the flaws of standard three-way data clustering in high-dimensional scenarios.

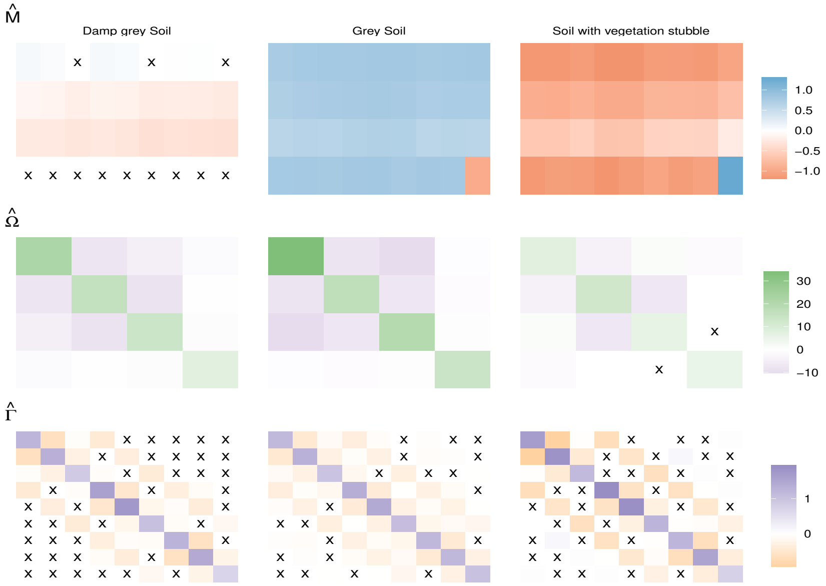


Figure 1. *Sparsely estimated M_k (upper plots), Ω_k (middle plots) and Γ_k (lower plots) for $k = 1,2,3$. Entries that are shrunk to 0 by the estimator are highlighted with an $\times$.*

Our proposal is able to effectively reduce the number of parameters to estimate while, at the same time, flexibly accounting for different relationships among the variables and for different level of sparsity across the groups. Future research directions would focus on the derivation of an appropriate model selection procedure, determining jointly reasonable values for the penalty coefficients as well as for the number of mixture components.

References

FRIEDMAN, J., HASTIE, T., & TIBSHIRANI, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.

SARKAR, S., ZHU, X., MELNYKOV, V., & INGRASSIA, S. 2020. On parsimonious models for modeling matrix data. *Computational Statistics & Data Analysis*, **142**, 106822.

VIROLI, C. 2011. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, **21**(4), 511–522.

WANG, Y., & MELNYKOV, V. 2020. On variable selection in matrix mixture modelling. *Stat*, **9**(1), e278.

EXPLORING SOLUTIONS VIA MONITORING FOR CLUSTER WEIGHTED ROBUST MODELS

Andrea Cappozzo [1], Luis Angel García Escudero[2], Francesca Greselin [3] and Agustín Mayo-Iscar[2]

[1] Department of Mathematics, Politecnico di Milano, (e-mail: andrea.cappozzo@polimi.it)

[2] Departamento de Estadística e Investigación Operativa, Facultad de Ciencias, Universidad de Valladolid, (e-mail: lagarcia@uva.es, agustin.mayo.iscar@uva.es)

[3] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, (e-mail: francesca.greselin@unimib.it)

ABSTRACT: Depending on the selected hyper-parameters, cluster weighted modeling may produce a set of diverse solutions. Particularly, the user can manually specify the number of mixture components, the degree of heteroscedasticity of the clusters in the explanatory variables and of the errors around the regression lines. In addition, when performing robust inference, the level of impartial trimming enforced in the estimation needs to be selected. This flexibility gives rise to a variety of "legitimate" solutions. To mitigate the problem of model selection, we propose a two stage monitoring procedure to identify a set of "good models". An application to the benchmark tone perception data showcases the benefits of the approach.

KEYWORDS: Cluster-weighted modeling, Outliers, Trimmed BIC, Eigenvalue constraint, Monitoring, Constrained estimation, Model-based clustering.

1 Introduction and model preliminaries

Assume to have observed a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$ of n i.i.d. samples, where the regression on Y varies across G groups, based on a vector $\mathbf{X}$ of explanatory variables with values in $\mathbb{R}^d$. Within this framework, the Gaussian Cluster Weighted Robust Model (García-Escudero *et al.*, 2017) is based on the constrained maximization of the *trimmed* log-likelihood:

$$\ell_{trimmed}(\Theta|\mathbf{X},Y) = \sum_{i=1}^{n} z(\mathbf{x}_i, y_i) \log \left[\sum_{g=1}^{G} \pi_g \phi(y_i; \mathbf{b}_g' \mathbf{x}_i + b_g^0, \sigma_g^2) \phi_d(\mathbf{x}_i; \mu_g, \Sigma_g) \right], \tag{1}$$

subject to: $\lambda_{l_1}(\Sigma_{g_1}) \leq c_X \lambda_{l_2}(\Sigma_{g_2})$ for every $1 \leq l_1 \neq l_2 \leq d$, $1 \leq g_1 \neq g_2 \leq G$ and $\sigma^2_{g_1} \leq c_y \sigma^2_{g_2}$ for every $1 \leq g_1 \neq g_2 \leq G$. The 0-1 trimming indicator function $z(\cdot,\cdot)$ tells us whether observation $(\mathbf{x}_i, y_i)$ is trimmed off, with trimming level $\alpha\%$ of observations being left unassigned by setting $\sum_{i=1}^{n} z(\mathbf{x}_i, y_i) = \lfloor n(1-\alpha) \rfloor$. The set $\{\lambda_l(\Sigma_g)\}_{l=1,\dots,d}$ denotes the eigenvalues of the scatter matrices Σ_g and the constants c_X and c_y are respectively finite real numbers such that $c_X \geq 1$ and $c_y \geq 1$.

2 Tone perception data application

The tone perception dataset (De Veaux, 1989) is employed as a case study to illustrate the proposed two-step monitoring procedure. In the first step, dedicated graphical and exploratory tools are employed for determining one or more plausible values for the trimming level α. Specifically, group proportion (black bars denote the trimmed units), total sum of squares decomposition (Ingrassia & Punzo, 2020), regression coefficients, standard deviations, cluster volumes and Adjusted Rand Index (ARI) between consecutive cluster allocations are monitored within a grid of αs, as reported in Figure 1. For each trimming level, the best model is selected according to a novel penalized likelihood criterion tailored for the CWRM framework, building upon the proposal developed in Cerioli *et al.*, 2018 for Gaussian mixtures. As it is clearly visible for the plots in Figure 1, model parameters stabilize as soon as α is set higher than 0.08, a value sufficient to trim off the level of contamination known to be present in this dataset (García-Escudero *et al.*, 2017).

In the second stage, conditioning on the α selected in the previous step, solutions stability and validity are fully investigated varying hyper-parameters in $\mathcal{E}_0 = \{(G, c_X, c_y) : G = 1, \dots, 4, c_X, c_y = 2^1, \dots, 2^5\}$, as reported in Figure 2. Darker and lighter opacity cells respectively indicate the sets of $\mathcal{B}_t$ best and $\mathcal{S}_t$ stable solutions, for each optimal solution $t, t = 1 \dots, 4$, where optimality is in the sense of the penalized criterion. The former set includes solutions ARI-similar to the optimal and not worse than the next optimal, while the latter encompasses all solutions ARI-similar to the optimal, such that $\mathcal{B}_t \subseteq \mathcal{S}_t$. In this example, solutions are assumed to be ARI-similar if the ARI between the estimated partitions is higher than 0.7. It is interesting to notice that the CWRM favors models with higher number of clusters with respect to the accepted truth of $G = 2$ (fourth optimal solution, stable in the entire grid of c_X and c_y). The reason being that, contrarily to the standard mixture of regression, the CWRM treats the covariate as random, thus allowing the learning of group-wise different distributions in the explanatory variable (Figure 3).

We have demonstrated the adequacy of our monitoring procedure in aiding

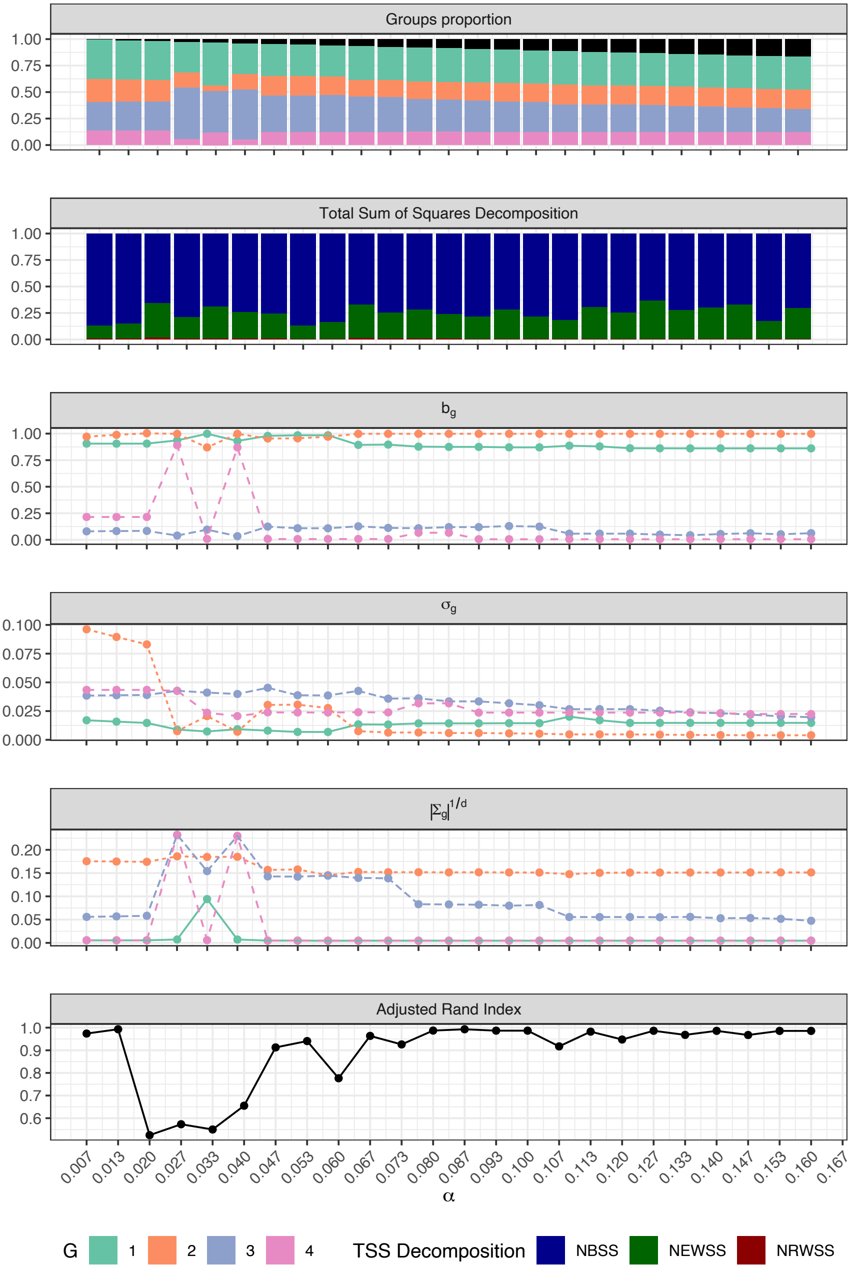


Figure 1: Step 1, monitoring the choice of a plausible trimming level α, tone perception data.

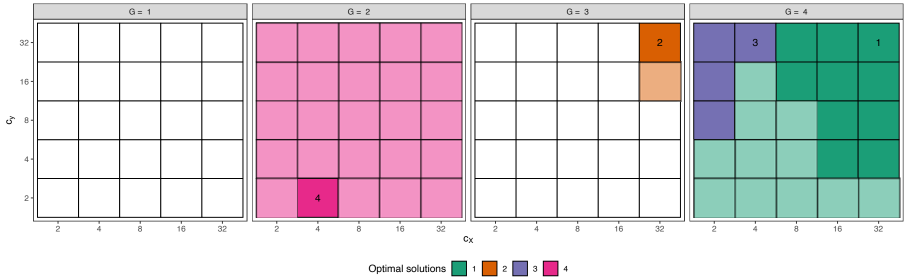


Figure 2: Step 2: monitoring optimal solutions, in terms of validity and stability. Trimming level $\alpha = 0.08$, tone perception data.

practitioners in the hyper-parameters selection when fitting CWRM. Furthermore, by exploring the space of solutions a deeper understanding of the data structure is achieved, uncovering sometimes unexpected yet valuable results.

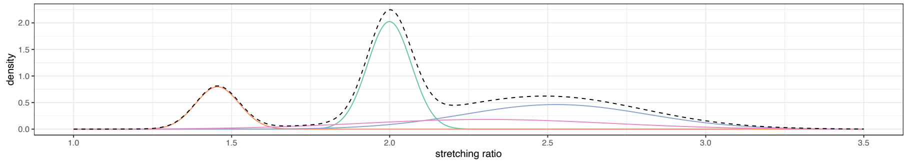


Figure 3: Estimated density on the explanatory variable, first optimal solution. Trimming level $\alpha = 0.08$, tone perception data.

References

CERIOLI, ANDREA, GARCÍA-ESCUDERO, LUIS ANGEL, MAYO-ISCAR, AGUSTÍN, & RIANI, MARCO. 2018. Finding the number of normal groups in model-based clustering via constrained likelihoods. *Journal of Computational and Graphical Statistics*, **27**(2), 404–416.

DE VEAUX, RICHARD D. 1989. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, **8**(3), 227–245.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., GRESELIN, F., INGRASSIA, S., & MAYO-ISCAR, A. 2017. Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *Statistics and Computing*, **27**(2), 377–402.

INGRASSIA, SALVATORE, & PUNZO, ANTONIO. 2020. Cluster Validation for Mixtures of Regressions via the Total Sum of Squares Decomposition. *Journal of Classification*, **37**(2), 526–547.

CATEGORICAL CLASSIFIERS IN MULTI-CLASS CLASSIFICATION PROBLEMS

Maurizio Carpita [1], Silvia Golia [1]

[1] Department of Economics and Mnagement, University of Brescia, (e-mail: maurizio.carpita@unibs.it, silvia.golia@unibs.it)

ABSTRACT: This paper shows the preliminary results of a simulation study devoted to comparing, in a multi-class classification setting, three classifiers that transform the probabilities produced by a probabilistic classifier into a single class: the usual Bayes Classifier and the new Max Difference Classifier and Max Ratio Classifier. As well known, the Bayes Classifier has some limits with rare classes, whereas the proposed Max Difference and Max Ratio Classifiers seem to represent better alternatives.

KEYWORDS: categorical classifier, polytomous variable, Bayes classifier

1 The proposed categorical classifiers and preliminary results

In machine learning, when dealing with a classification problem, it is possible to distinguish two aspects. The first one concerns the identification of a so-called *probabilistic classifier*, which corresponds to a suitable method that assigns a probability to all the categories that can be assumed by the target variable. The second one regards the so-called *categorical classifier*, which transforms the probabilities produced by the probabilistic classifier into a single category. There is a large literature concerning how to find the best probabilistic classifier in both the dichotomous and polytomous contexts, whereas less attention was paid to the choice of the criterion to be used to pass from the probabilistic to the categorical classifier. The *Bayes Classifier* (BC), which assigns, based on the probabilistic classifier, a unit to the most likely category, minimizes, on average, the test error rate (James *et al.*, 2013), so it is the optimal criterion if one is interested in the accuracy of the classification. Nevertheless, this classifier favors the prevalent category most and in situations in which there is not a category of interest but all the categories have the same relevance, the BC can not be the best choice. In previous papers (see, for example, Golia & Carpita, 2020) the authors investigated the performances of different categorical classifiers and they found one of them promising. In this study this classifier, called Maximum Difference Classifier, is considered jointly to a new proposal, denoted as Maximum Ratio Classifier. Both classifiers are based on the comparison between the predicted probabilities and the sample frequencies. Let pr_i be the predicted probability of the category a_i ($i = 1, 2, \ldots, k$) of the variable A, and let fr_i be the corresponding frequency computed from observed data. The *Maximum Difference Classifier* (MDC) computes the deviations of pr_i from fr_i and takes the category corresponding to the maximum difference, that is: $\arg\max_{i \in (a_1, a_2, \ldots, a_k)}(pr_i - fr_i)$. This classifier represents the extension of what proposed by Cramer (1999) for the dichotomous case. The *Maximum Ratio Classifier* (MRC) computes the relative deviations of pr_i from fr_i and takes the category corresponding to the maximum ratio, that is: $\arg\max_{i \in (a_1, a_2, \ldots, a_k)}(pr_i / fr_i)$. In order to evaluate the predictive performance of a classifier, some indicators computed from the confusion matrix can be used. In this study they are: the *Sensitivity* (Sen) and the *Specificity* (Spe) of each category, the *Maximum Distance Between Sensitivities* (MDBSen) and the *Maximum Distance Between Specificities* (MDBSpe), the *Overall Accuracy* (OvAc) and the *Macro Average F1 score* (MAF1) (Raschka & Mirjalili, 2019). Sen_i (Spe_i) expresses how well the classifier recognizes a unit belonging (not belonging) to the category a_i. MDBSen and MDBSpe highlight the balanced or unbalanced ability of the classifier to assign a unit to the right category, the lower the MDBSen and MDMSpe, the more balanced the classification. The OvAc is the rate of correct classification and it is the indicator maximized by BC. The MAF1 is another indicator to measure the accuracy of the classifier and it is obtained as the average of the F1 scores class-by-class. The choice of MAF1 instead of the weighted average F1 score, is linked to the will to attribute the same relevance to all classes.

This study originates from a real classification problem related to the prediction of the result of a soccer match from the home-team side (Golia & Carpita, 2020), so the target variable admits three possible categories, which own a natural order. However, wanting to consider a more general framework, the variable to be predicted, considered in the present study, will be nominal. In order to simulate the probability distribution of this nominal variable, the trivariate Dirichlet random variable (r.v.) was used. This r.v. is determined by three parameters, α_A, α_B and α_C. Table 1 reports the values chosen in the simulation study which is described below and the values of the mean and the skewness of the marginals. The chosen sample size is 1200, which implies that in the balanced case each category gets around 400 units. This sample size is high, and the reason lies in the aim to investigate the performance of the analyzed categorical classifiers in a less problematic framework, given that they are working in a context of big samples. For each of these 1200 units the

Table 1. *Parameters of the Dirichlet r.v. and mean and skewness of the marginals*

Condition	α_A	α_B	α_C	M. X_A	M. X_B	M. X_C	Sk. X_A	Sk. X_B	Sk. X_C
C1	10	10	10	0.333	0.333	0.333	0.246	0.246	0.246
C2	2	5	10	0.118	0.294	0.588	1.060	0.404	-0.160
C3	5	5	10	0.250	0.250	0.500	0.481	0.481	0.000
C4	2	10	10	0.091	0.455	0.455	1.137	0.073	0.073

probability distribution of the target variable is simulated from the trivariate Dirichlet r.v. The use of the set of these three probabilities was twofold; first, a realization of this random variable was extracted and it represents the actual (observed) value of the target variable, second, the same set of probabilities was considered as the output of a probabilistic classifier for the target variable. Then, the BC, MDC and MRC were applied, the predicted classifications for the 1200 units were obtained and the performance indicators previously described were calculated. This scheme was repeated 1000 times and the mean values of the indicators, with standard deviation in parenthesis, are reported in Table 2. When all the three categories are equally represented in the population, as in condition C1, the three categorical classifiers perform in the same way. When one category is rare, as in conditions C2 and C4, BC is not able to recognize it, whereas MDC and MRC have a certain ability in doing it, and in general, Sen and Spe are more balanced when MDC and MRC are used. Moreover, as expected, BC performs better for OvAc, but MDC and MRC have higher MAF1. So, concluding, this first simulation study reveals that, in a multi-class setting, giving equal importance to all the classes (i.e. different types of mis-classification do not involve different costs) both MDC and MDR are preferable to BC.

References

CRAMER, J. S. 1999. Predictive performance of the binary logit model in unbalanced samples. *The Statistician*, **48**(1), 85–94.

GOLIA, S., & CARPITA, M. 2020. Comparing classifiers for ordinal variables. In A. Pollice, N. Salvati and F. Schirripa Spagnolo (Eds.). *Book of short papers SIS 2020*, 1160–1165.

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. 2013. *An introduction to statistical learning with applications in R*. New York: Springer.

RASCHKA, S., & MIRJALILI, V. 2019. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Birmingham: Packt Publishing.

Table 2. *Mean values of the indicators, with standard deviation in parenthesis*

		Sen_A	Sen_B	Sen_C	MDBSen
	BC	0.425 (0.03)	0.424 (0.03)	0.425 (0.03)	0.042 (0.02)
	MDC	0.423 (0.06)	0.424 (0.06)	0.424 (0.06)	0.120 (0.06)
	MRC	0.423 (0.07)	0.424 (0.06)	0.424 (0.06)	0.131 (0.07)
		Spe_A	Spe_B	Spe_C	MDBSpe
	BC	0.712 (0.02)	0.713 (0.02)	0.712 (0.02)	0.030 (0.02)
C1	MDC	0.712 (0.05)	0.711 (0.05)	0.712 (0.05)	0.103 (0.05)
	MRC	0.712 (0.05)	0.711 (0.05)	0.712 (0.05)	0.112 (0.06)
		OvAc	MAF1		
	BC	0.425 (0.01)	0.424 (0.01)		
	MDC	0.422 (0.01)	0.421 (0.01)		
	MRC	0.421 (0.01)	0.420 (0.02)		
		Sen_A	Sen_B	Sen_C	MDBSen
	BC	0.015 (0.01)	0.150 (0.02)	0.939 (0.01)	0.924 (0.01)
	MDC	0.444 (0.06)	0.489 (0.05)	0.483 (0.05)	0.109 (0.06)
	MRC	0.572 (0.07)	0.458 (0.05)	0.403 (0.05)	0.189 (0.08)
		Spe_A	Spe_B	Spe_C	MDBSpe
	BC	0.997 (0.00)	0.936 (0.01)	0.139 (0.02)	0.858 (0.02)
C2	MDC	0.789 (0.03)	0.703 (0.04)	0.699 (0.04)	0.122 (0.05)
	MRC	0.700 (0.04)	0.720 (0.04)	0.759 (0.04)	0.100 (0.05)
		OvAc	MAF1		
	BC	0.598 (0.01)	0.331 (0.02)		
	MDC	0.479 (0.02)	0.434 (0.02)		
	MRC	0.437 (0.03)	0.413 (0.02)		
		Sen_A	Sen_B	Sen_C	MDBSen
	BC	0.135 (0.02)	0.136 (0.02)	0.895 (0.01)	0.770 (0.02)
	MDC	0.443 (0.05)	0.445 (0.05)	0.459 (0.05)	0.107 (0.06)
	MRC	0.470 (0.06)	0.472 (0.06)	0.405 (0.05)	0.134 (0.07)
		Spe_A	Spe_B	Spe_C	MDBSpe
	BC	0.942 (0.01)	0.942 (0.01)	0.205 (0.02)	0.742 (0.02)
C3	MDC	0.734 (0.04)	0.732 (0.04)	0.702 (0.05)	0.090 (0.05)
	MRC	0.711 (0.05)	0.710 (0.05)	0.744 (0.04)	0.101 (0.05)
		OvAc	MAF1		
	BC	0.514 (0.02)	0.359 (0.02)		
	MDC	0.449 (0.02)	0.436 (0.02)		
	MRC	0.436 (0.02)	0.429 (0.02)		
		Sen_A	Sen_B	Sen_C	MDBSen
	BC	0.004 (0.01)	0.586 (0.02)	0.588 (0.02)	0.595 (0.02)
	MDC	0.409 (0.06)	0.476 (0.05)	0.482 (0.06)	0.131 (0.06)
	MRC	0.590 (0.07)	0.401 (0.05)	0.407 (0.05)	0.224 (0.09)
		Spe_A	Spe_B	Spe_C	MDBSpe
	BC	0.999 (0.00)	0.574 (0.02)	0.573 (0.02)	0.438 (0.02)
C4	MDC	0.806 (0.03)	0.681 (0.05)	0.675 (0.05)	0.166 (0.05)
	MRC	0.682 (0.05)	0.736 (0.05)	0.731 (0.04)	0.107 (0.06)
		OvAc	MAF1		
	BC	0.534 (0.01)	0.320 (0.09)		
	MDC	0.471 (0.02)	0.422 (0.02)		
	MRC	0.419 (0.03)	0.393 (0.02)		

MODEL-BASED CLUSTERING FOR ESTIMATING CETACEANS SITE-FIDELITY AND ABUNDANCE

Gianmarco Caruso[1], Greta Panunzi[1], Marco Mingione[1], Pierfrancesco Alaimo di Loro[1], Stefano Moro[2], Edoardo Bompiani[1], Caterina Lanfredi[1], Daniela Silvia Pace[2], Luca Tardella[1] and Giovanna Jona Lasinio[1]

[1] Department of Statistical Sciences, Sapienza University of Rome, Italy (e-mail: gianmarco.caruso@uniroma1.it, greta.panunzi@gmail.com, marco.mingione@uniroma1.it, pierfrancesco.alaimodiloro@uniroma1.it, edoardo.bompiani@uniroma1.it, lanfredicaterina@gmail.com, luca.tardella@uniroma1.it, giovanna.jonalasinio@uniroma1.it)

[2] Department of Environmental Biology, Sapienza University of Rome, Italy (e-mail: stefano.moro@uniroma1.it, danielasilvia.pace@uniroma1.it)

ABSTRACT: Estimating the size of animal populations in a given area is of particular interest in ecological studies on wildlife conservation, and this task is commonly handled via capture-recapture methods. A recent work (Pace *et al.*, 2021) adopts a two-step approach for identifying groups of animals with similar site-fidelity patterns - according to specific metrics - and estimating the abundance of bottlenose dolphins between 2017 and 2020 at the Tiber Estuary (Mediterranean Sea, Rome, Italy). In this work, we aim at simultaneously classifying individuals and estimating their abundance in the study area, by introducing finite mixtures within the *Open-Population Jolly-Seber* framework. In capture-recapture analyses, finite mixture models allow to account for groups heterogeneity and to reduce the bias in the final abundance estimates (Pledger, 2005).

KEYWORDS: Capture-recapture analysis, Wildlife population, Finite mixture models, Unsupervised classification, Applied statistics

1 Introduction

Capture-recapture methods are widely employed in estimating the size of wildlife populations, whose units are subject to multiple captures across several occasions. We will use the terms *capture* and *recapture* in accordance with the classical literature (Seber, 1986), but animals are not necessarily *captured*: nowadays, non-invasive ways of keeping trace of a wild animal over time are successfully employed. In that spirit, for example, Pace *et al.* (2021) employs photo-identification for identifying bottlenose dolphins from natural markings present on their bodies. In the same paper an interesting characteristic of this type of animals is illustrated: marked individuals may show a different level of *site-fidelity*. This point introduces the need of defining a statistical protocol or a specific model accounting for the different probabilities of *capture* among the categories. Here, we propose a method that allows both to differentiate between *resident* and *non-resident* individuals and to estimate the population abundance in a common modelling framework. This improves on the original multi-step protocols (see Pace *et al.*, 2021) in guaranteeing the correct uncertainty propagation of the two estimation processes.

2 The model

We consider the Schwarz & Arnason (1996)'s formalization of the Jolly-Seber model, which assumes the existence of a *super-population*, representing the set of individuals potentially available in the study area between the first and the last sampling period. In Jolly-Seber-type models, captures are assumed to be independent across individuals and along time. Moreover, the population is assumed to be *open*, meaning that individuals can either enter (e.g. birth or immigration) or exit (e.g. death or emigration) the population during the study. Notably, we assume that individuals leaving the population cannot come back in it. Here, we adopt the Bayesian framework illustrated by Royle & Dorazio (2012), where the super-population size (N_{super}) is provided with a discrete uniform distribution in the interval $\{0,...,M\}$, with M sufficiently large. The hyperparameter M can be seen as an upper bound for the super-population size and it implies the use of an augmented dataset of M individuals. Moreover, we consider a sampling scheme divided in T periods and, for each time $t = 1,\dots,T$, a number J_t of capture sessions. Thus, the augmented data matrix $\mathbf{Y} = [y_{it}]$ has M rows and T columns and contains the capture frequency of each individual in each period. If D is the number of individuals that have been observed at least once, the matrix contains $M-D$ rows of zeros: among them, $N_{\text{super}} - D$ rows correspond to individuals which belong to the super-population but have never been captured, while $M - N_{\text{super}}$ correspond to *pseudo*-individuals which do not belong to the super-population.

Recruitment and survival process Population dynamics consisting in recruitment and survival can be expressed through the following latent binary variables:

- r_{it} which is equal to 1 iff individual i is recruitable at time t;

- z_{it} which is equal to 1 iff individual i belongs to the population at time t. Let ϕ_t be the probability of remaining in the population at time t, being in the population at time $t-1$, and let ρ_t be the probability of belonging to the super-population *and* being recruited into the population at time t. Without loss of generality, in this context we assume these two parameters to be constant over time, i.e. $\phi_t = \phi$ and $\rho_t = \rho$. Following Royle & Dorazio (2012), it can be proved that, for $i = 1,\dots,M$, $r_{i1} = 1$ and $z_{i1} \sim Bern(\rho)$, and

$$r_{it} = \min\{r_{i,t-1}, 1 - z_{i,t-1}\}, \qquad t > 1$$

$$z_{it}|z_{i,t-1}, r_{it} \sim Bern(\phi \cdot z_{i,t-1} + \rho \cdot r_{it}), \qquad t > 1.$$

Notice that when an individual becomes part of the population, it cannot be recruitable any more: for $t > 1$, r_{it} and z_{it} cannot simultaneously be equal to 1.

Detection process In this work, we consider a finite mixture model in order to model the different propensity to the capture among different groups of individuals. The generic element of the augmented data matrix is such that

$$y_{it}|z_{it}, c_i = g \sim Binom(J_t, p_g \cdot z_{it}), \qquad g = 1\dots,G \quad ,$$

with p_g being the capture probability of individuals in group g and $P(c_i = g) = w_g$ being the probability that the i-th individual belongs to the g-th mixture component. Notice that $y_{it} = 0$ almost surely when $z_{it} = 0$, so that the previous model corresponds to a finite mixture of zero-inflated binomial distributions.

Abundance estimation The population size at time t and the super-population size can be estimated through the latent variables z's, namely, $N_t = \sum_{i=1}^{M} z_{it}$ and $N_{\text{super}} = \sum_{i=1}^{M} \mathbb{1}_{\{\sum_{t=1}^{T} z_{it} > 0\}}$.

3 Illustration

A graphical visualization of the main components of the model is provided by the DAG in *Figure* 1. The model is used to estimate the abundance of bottlenose dolphins between 2017 and 2020 at the Tiber Estuary (Mediterranean Sea, Rome, Italy) and identifying groups of animals with different propensities to the capture: individuals with a low detection probability are considered *non-resident*, while the others are considered *resident*. The model is implemented using JAGS (Plummer, 2003) and the results will be shown in details during the conference.

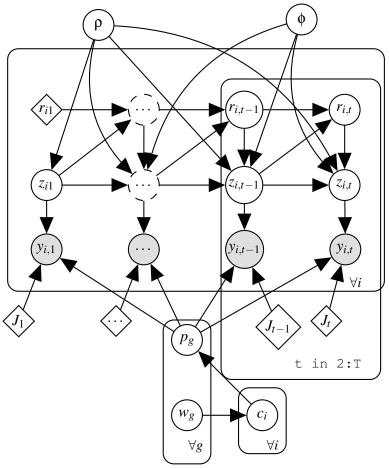


Figure 1. *Bayesian DAG with the main components of the model. White rhombi represent deterministic variables. White circles represent latent variables and parameters. Grey circles represents observable variables.*

Acknowledgements This work was partially supported by project "Joint Cetacean Database and Mapping (JCDM) in Italian waters: a tool for knowledge and conservation", Sapienza University of Rome (nr. RM1201729F23D51B).

References

PACE, DANIELA SILVIA, *et al.* 2021. Capitoline Dolphins: Residency Patterns and Abundance Estimate of Tursiops truncatus at the Tiber River Estuary (Mediterranean Sea). *Biology*, **10**(4), 275.

PLEDGER, SHIRLEY. 2005. The performance of mixture models in heterogeneous closed population capture–recapture. *Biometrics*, **61**(3), 868–873.

PLUMMER, MARTYN. 2003. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*.

ROYLE, J ANDREW, & DORAZIO, ROBERT M. 2012. Parameter-expanded data augmentation for Bayesian analysis of capture–recapture models. *Journal of Ornithology*, **152**(2), 521–537.

SCHWARZ, CARL JAMES, & ARNASON, A NEIL. 1996. A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics*, 860–873.

SEBER, GEORGE ARTHUR FREDERICK. 1986. A Review of Estimating Animal Abundance. *Biometrics*, **42**(2), 267–292.

Model-Based Clustering with Parsimonious Covariance Structure

Carlo Cavicchia[1] , Maurizio Vichi[2] and Giorgia Zaccaria[2]

[1] Econometric Institute, Erasmus University Rotterdam, Rotterdam, The Netherlands, (e-mail: `cavicchia@ese.eur.nl`)

[2] Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy, (e-mail: `maurizio.vichi@uniroma1.it`, `giorgia.zaccaria@uniroma1.it`)

Abstract: Complex multidimensional concepts are often explained by a tree-shape structure by considering nested partitions of variables, where each variable group is associated with a specific concept. Recalling that relations among variables can be detected by their covariance matrix, this paper introduces a covariance structure that reconstructs hierarchical relationships among variables highlighting three features of the variable groups. We finally present an application of the latter covariance structure to the model-based clustering.

Keywords: Gaussian mixture model, hierarchical latent concepts, partition of variables

1 Introduction

The main goal of Factor Analysis (FA, Spearman, 1904) is to reconstruct the covariance matrix of variables by computing a reduced number of factors while preserving as much information as possible. However, since FA is unable to reconstruct hierarchical relations, a model with a hierarchical form is therefore required. Among several models based on the sequential application of FA addressing the same problem, Cavicchia *et al.* (2020) proposed a model to reconstruct a nonnegative correlation matrix via an ultrametric one. The model results in a simultaneous procedure which is able both to detect the best variable partition in a reduced number of groups and build the hierarchy upon them. The latter model ensues particularly suitable for complex hierarchical multidimensional concepts due to the one-to-one relation between a hierarchy of concepts and an ultrametric correlation matrix (Dellacherie *et al.*, 2014). Our paper overcomes the limitations of the model presented by Cavicchia *et al.* (2020) extending the same idea to a general covariance matrix and applies this special covariance structure in the Gaussian Mixture Models (GMMs) framework.

Since GMMs can easily fall into the so-called "curse of dimensionality" because of the large number of parameters dedicated to covariance structures, in the specialized literature several different parametrizations are present. One of the most used is the eigen-decomposition (Banfield & Raftery, 1993) of the form $\mathbf{\Sigma} = \lambda\mathbf{DAD}'$, where λ is a scalar determining the cluster volume, $\mathbf{A}$ is a diagonal matrix controlling the cluster shape, and $\mathbf{D}$ is an orthogonal matrix which specifies the cluster orientation. Another parameterization is proper of the mixture of factor analyzers (Ghahramani & Hilton, 1997) and assumes a cluster covariance structure of the form $\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$, where p is the number of variables, Q is the number of factors, $\mathbf{\Lambda}$ is the $p \times Q$ factor loading matrix and $\mathbf{\Psi}$ is the p-dimensional diagonal covariance matrix of the error. Our proposal aims to implement a new parameterization of a covariance matrix via a hierarchical covariance one for each cluster that can be extremely parsimonious.

2 Features of the covariance structure

Multidimensional phenomena are often composed of nested dimensions characterized by distinct levels of abstraction. Each dimension is uniquely connected to a group of variables and represents a specific concept. Merging two dimensions together gives rise to a broader dimension up to the general one such that the hierarchical structure underlying a multidimensional phenomenon is detected. In order to model the hierarchical relationships among the dimensions, we introduce three main features of a variable group: the variance of the variable group, the covariance within the variable group, which measures the internal concordance among variables belonging to the same group, and the covariance between concepts associated with the variable groups. These features are constrained to be "ordered" such that the variance of the groups is greater (in the absolute sense) than the covariance within or between groups, whereas the covariance within groups must be in turn larger than the covariance between groups. These constraints allow to define a hierarchical structure of concepts, from the most concordant to the most discordant. The last aggregations in the hierarchy may occur between: (i) concordant concepts defining a general one; (ii) discordant concepts with negative between-group covariance; (iii) uncorrelated concepts.

Given the number of specific dimensions Q which underlie the multidimensional phenomenon, each level $q = Q, \ldots, 1$ of the hierarchy is characterized by: (i) the $p \times q$ membership matrix $\mathbf{V}_q$, which pinpoints the membership of each variable to a group; (ii) the diagonal matrix $\mathbf{S}_q^V$ of order q, whose main diagonal represents the variance of each group; (iii) the diagonal matrix $\mathbf{S}_q^W$

of order q, whose main diagonal represents the covariance within each group; (iv) the ultrametric matrix $\mathbf{S}_q^B$ of order q, whose diagonal entries are set to zero and off-diagonal ones represent the hierarchical relationships between pairs of concepts. Given $\mathbf{V}_q$, the estimates of the matrices $\mathbf{S}_q^V$, $\mathbf{S}_q^W$ and $\mathbf{S}_q^B$ are

$$\widehat{\mathbf{S}}_q^V = (\widehat{\mathbf{V}}_q'\widehat{\mathbf{V}}_q)^{-1}\widehat{\mathbf{V}}_q'\text{diag}(\mathbf{S})\widehat{\mathbf{V}}_q, \tag{1}$$

$$\widehat{\mathbf{S}}_q^W = [(\widehat{\mathbf{V}}_q'\widehat{\mathbf{V}}_q)^2 - \widehat{\mathbf{V}}_q'\widehat{\mathbf{V}}_q]^{-1}\text{diag}\Big[\widehat{\mathbf{V}}_q'\Big(\mathbf{S} - \text{diag}(\widehat{\mathbf{V}}_q\widehat{\mathbf{S}}_q^V\widehat{\mathbf{V}}_q')\Big)\widehat{\mathbf{V}}_q\Big], \tag{2}$$

$$\widehat{\mathbf{S}}_q^B = \widehat{\mathbf{V}}_q^+\mathbf{S}(\widehat{\mathbf{V}}_q')^+, \tag{3}$$

respectively, where $\mathbf{S}$ represents the $p \times p$ observed covariance matrix, $\mathbf{I}_p$ is the identity matrix of order p and $\text{diag}(\cdot)$ denotes the diagonal matrix whose diagonal elements are those of a parenthesized one.

We implement the parameterization of the covariance matrix based on the aforementioned quantities into the GMMs in order to simultaneously detect homogeneous clusters of units and a hierarchical definition of a multidimensional phenomenon.

3 Application

Our proposal is applied on the "Human Development Index" dataset[*] which consists of 167 countries and 9 variables. The optimal model in terms of Bayesian Information Criterion (BIC, Schwarz, 1978) considers 3 clusters of countries (Fig. 1) and 3 groups of variables. It is worth highlighting that the model requires 71 parameters to be estimated, of which only 14 for each covariance structure. The first cluster is characterized by the countries with high income, gdp per capita and very low child mortality. The second cluster is constituted by the poorest countries with low life expectancy and income, whereas the third one is composed by countries with median performances. Each cluster is characterized by a different hierarchy of the latent concepts associated with the three groups of variables. The group made by the economic variables (income, gdp, exports and imports) in Cluster 1 is the one with the highest value of internal variance, whereas the same group in Cluster 3 is merged with the group considering child mortality and fertility and has the highest covariance within the group. Notwithstanding the latent concepts and their hierarchical relationships are specific per cluster, all the hierarchies end

[*] https://www.kaggle.com/rohan0301/unsupervised-learning-on-country-data

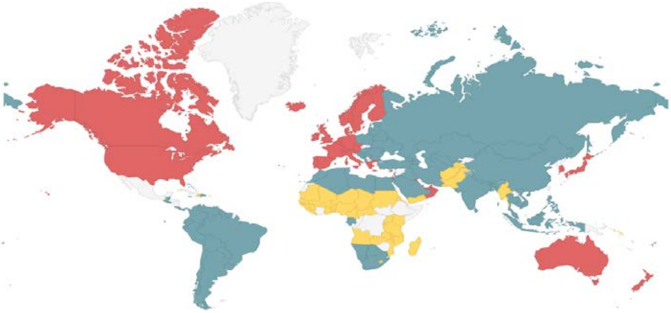


Figure 1: Clusters of countries: Cluster 1 (red), Cluster 2 (yellow) and Cluster 3 (blue)

with a negative between-group covariance highlighting the absence of a unique concordant general concept.

4 Conclusions

This paper proposes a parsimonious GMM which aims at modeling multidimensional phenomena, usually defined by hierarchically nested latent concepts. The application of the method on real data shows its potentialities.

References

BANFIELD, J.D., & RAFTERY, A.E. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**(3), 803–821.

CAVICCHIA, C., VICHI, M., & ZACCARIA, G. 2020. The ultrametric correlation matrix for modelling hierarchical latent concepts. *Advances in Data Analysis and Classification*, **14**(4), 837–853.

DELLACHERIE, C, MARTINEZ, S, & MARTIN, J SAN. 2014. *Inverse M-matrices and ultrametric matrices*. Lecture Notes in Mathematics. Springer International Publishing.

GHAHRAMANI, Z., & HILTON, G.H. 1997. The EM algorithm for factor analyzers. Technical report CRG-TR-96-1, University of Toronto, Toronto.

SCHWARZ, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461–464.

SPEARMAN, C. E. 1904. "General intelligence,' objectively determined and measured. *The American Journal of Psychology*, **15**(2), 201–293.

CLUSTERING INCOME DATA BASED ON SHARE DENSITIES

Francesca Condino [1]

[1] Department of Economics, Statistics and Finance ”Giovanni Anania”, University of Calabria, Italy, (e-mail: francesca.condino@unical.it)

ABSTRACT: Different measures, generally used to analyse income inequality, refer to the context of information theory and the concept of entropy. In particular, Theil's T index can be interpreted in terms of entropy and related to the well-known Lorenz curve. Indeed, Lorenz curve and its derivative, the so-called share density, provides different information regarding inequality. Starting from this evidence, the aim of this work is to compare income inequality of different subgroups, by using a proper dissimilarity measure between parametric share densities, and to use these information for their clustering. Preliminary results regarding data from Survey on Households Income and Wealth (SHIW) by Bank of Italy are shown.

KEYWORDS: tail inequality, dissimilarity measure, income concentration.

1 Lorenz curve and share density

In economic literature, Lorenz curve is a well-known and widely used tool for analysing income inequality. Since its proposal, in 1905 (Lorenz, 1905), a lot of investigation has been suggested among statisticians and economists, generating a fertile field of study. Conversely, Lorenz density is rarely explicitly mentioned. One of the few reference to Lorenz density can be found in Farris, 2010, where this curve is referred as share density. Afterwards, the concept of Lorenz density is resumed in Zizler, 2014, in Kämpke & Radermacher, 2015 and Shao, 2021. Actually, it is known that each Lorenz curve $L(u)$ ($u \in [0,1]$) can be viewed as a distribution function on the unit interval, therefore it is possible to consider it's derivative with respect to u, $l(u) = L'(u)$, as a density function. It is worth to note that, this density function furnishes different information regarding income inequality, as suggested by Rohde, 2008, who has shown that the two well-known Theil's inequalities indexes, L and T, can be directly obtained from $l(u)$. In particular, Theil's T index coincides with Shannon entropy, changed in sign, of $l(u)$. In this perspective, it arises natural to compare different groups of income earners in terms of inequality, by quantifying the dissimilarity between share densities through a proper measure.

2 Jensen-Shannon divergence between share densities

The Jensen-Shannon divergence (JSD), also called total divergence to the average, is a well known measure of dissimilarity among probability distributions. It can be obtainined starting from the Kullback–Leibler divergence, considering K densities $f_1,...,f_K$ and their mixture $m = \sum_{k=1}^{K} \pi_k \cdot f_k$ with $\pi_k \in [0,1]$, as follows:

$$D_{JS}(f_1,...,f_K) = \sum_{k=1}^{K} \pi_k \cdot D_{KL}(f_k||m) \tag{1}$$

where

$$D_{KL}(f_k||m) = \int_X f_k(x) \log \frac{f_k(x)}{m(x)} dx. \tag{2}$$

Alternatively, expression (1) can be rewritten in terms of Shannon entropy H, as follows:

$$D_{JS}(f_1,...,f_K) = H(m) - \sum_{k=1}^{K} \pi_k H(f_k) \tag{3}$$

where $H(f_k) = -\int_X f_k(x) \log f_k(x)\ dx$. It is easy to prove that $D_{JS}(f_1,...,f_K) \geq 0$ and equality holds when $f_1 = f_2 = ... = f_K$. In addition, for two densities, it is symmetric, i.e. $D_{JS}(f_1||f_2) = D_{JS}(f_2||f_1)$, and then it is a bonafide measure of dissimilarities between $f_1(\cdot)$ and $f_2(\cdot)$. Now, with the aim to analyse existing differences among various groups of income earners, this dissimilarity measure will be considered in connection with the Lorenz density. Let $L_1,...,L_K$ be the Lorenz curves corresponding to K different groups of income earners and $l_1,...,l_K$ the corresponding derivatives with respect to u. Hence, the JSD among l_k densities ($k = 1,...,K$) is given by:

$$D_{JS}(l_1,...,l_K) = H(l_m) - \sum_{k=1}^{K} \pi_k H(l_k) \tag{4}$$

where $l_m = \sum_{k=1}^{K} \pi_k l_k(u)$. To define the l_m mixture density, the decomposition of Lorenz curve proposed by Bishop *et al.*, 2003 is considered, so that π_k represents the income share for the $k-th$ group. From (4), it is evident that the JSD takes into account, for each share density, the whole function and its entropy, so that it will be influenced by existing differences in tail inequality among groups, as well as in concentration around the center of income distribution. Therefore, clustering procedures based on JSD will exploit these discrepancies.

3 Clustering income data: an application

In this section, data from the Survey on Households Income and Wealth (SHIW), carried out by Bank of Italy in 2016, are considered. To take into account the composition of households, equivalent income are obtained, using the OECD-modified equivalent scale. The Dagum distribution (Dagum, 1977) is used to model income and to obtain the expressions of Lorenz curves, $L_k(u)$, and share densities, $l_k(u)$, for each region ($k = 1,...,20$). For this model, a closed expression for $H(l_k)$ function is obtained. This result (not reported for space reason) agrees, unless the sign, with that reported in Chotikapanich *et al.*, 2018 for Theil's T index, confirming the relation between T and $H(l)$. Table 1 shows, for each Italian region, the estimates for average income ($\hat{\mu}_k$, in tens of thousands of euros), entropy ($\hat{H}(l_k)$) and Gini index ($\hat{G}_k$). Furthermore, the mem-

Table 1. *Fitted means, entropy, Gini index and membership cluster for Italian regions*

Regions	$\hat{\mu}_k$	$-\hat{H}(l_k)$	$\hat{G}_k$	Cluster
Piedmont	2.0797	0.1277	0.2751	1
Aosta Valley	2.3134	0.1195	0.2663	1
Veneto	1.9739	0.1244	0.2706	1
Friuli	2.3140	0.1226	0.2693	1
Emilia Romagna	2.3971	0.1141	0.2602	1
Tuscany	2.3648	0.1131	0.2588	1
Abruzzo	1.9542	0.1237	0.2717	1
Calabria	1.3472	0.1425	0.2910	1
Sardinia	1.5772	0.1361	0.2843	1
Lombardy	2.4798	0.1632	0.3064	2
Molise	1.7789	0.1873	0.3294	2
Campania	1.3461	0.1815	0.3252	2
Apulia	1.4558	0.1557	0.3026	2
Basilicata	1.5191	0.1850	0.3287	2
Sicily	1.4610	0.1633	0.3071	2
Trentino	2.2247	0.1008	0.2408	3
Liguria	2.2482	0.1356	0.2782	3
Umbria	1.9897	0.1024	0.2456	3
Marche	2.1809	0.1053	0.2475	3
Lazio	1.9972	0.1437	0.2883	3

bership cluster is reported. In order to obtain this partition, elements $D_{JS}(l_i, l_j)$ $(i, j = 1, ..., 20)$ of dissimilarity matrix D are computed from expression (4). Here, numerical integration method is used to compute $H(l_m)$. Then, a hierarchical clustering based on matrix D has been conducted, considering complete agglomeration method and a final number of groups equal to 3. As we can see from the results, clusters seem clearly characterized, with regions having generally lower concentration of income belonging to cluster 1 and 3 and regions with higher concentrations levels included in cluster 2. Furthermore, by analysing more in depth the obtained results, it appears that this method allows to gather together regions with similar behaviour in tail inequality (results not reported), as well as similar values of Theil's and Gini indexes.

References

BISHOP, J.A., CHOW, K.V., & ZEAGER, L.A. 2003. Decomposing Lorenz and Concentration Curves. *International Economic Review*, **44**, 965–978.

CHOTIKAPANICH, D., GRIFFITHS, W.E., HAJARGASHT, G., KARUNARATHNE, W., & RAO, D.S.P. 2018. Using the GB2 Income Distribution. *Econometrics*, **6**, 21.

DAGUM, C. 1977. A new model of personal distribution: specification and estimation. *Economie Appliquée*, **30**, 413–437.

FARRIS, F.A. 2010. The Gini index and measures of inequality. *American Mathematical Monthly*, **117**(10), 851–864.

KÄMPKE, T., & RADERMACHER, F. 2015. *Income Modeling and Balancing: A Rigorous Treatment of Distribution Patterns*. Switzerland: Springer.

LORENZ, M.O. 1905. Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association*, **9**(70), 209–219.

ROHDE, N. 2008. *Lorenz Curves and Generalised Entropy Inequality Measures. In Chotikapanich D. (eds) Modeling Income Distributions and Lorenz Curve. Economic Studies in Equality, Social Exclusion and Well-Being, vol 5*. New York: Springer.

SHAO, B. 2021. Decomposition of the Gini index by income source for aggregated data and its applications. *Computational statistics*, **Epub ahead of print**, 1–25.

THEIL, H. 1967. *Economics and Information Theory*. Amsterdam: North Holland.

ZIZLER, P. 2014. Gini indices and the moments of the share density function. *Applications of Mathematics*, **59**, 167–175.

GROUP-DEPENDENT FINITE MIXTURE MODEL

Paula Costa Fontichiari[1], Miriam Giuliani[1], Raffaele Argiento[1] and Lucia Paci[1]

[1] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, (e-mail: paula.costafontichiari01@icatt.it, miriam.giuliani01@icatt.it, raffaele.argiento@unicatt.it, lucia.paci@unicatt.it)

ABSTRACT: We present a Bayesian nonparametric group-dependent mixture model for clustering. This is achieved by building a hierarchical structure, where the discreteness of the shared base measure is exploited to cluster the data, between and within groups. We study the properties of the group-dependent clustering structure based on the latent parameters of the model. Furthermore, we obtain the joint distribution of the clustering induced by the hierarchical mixture model and define the complete posterior characterization of interest. We construct a Gibbs sampler to perform Bayesian inference and measure performances on simulated and a real data.

KEYWORDS: Bayesian analysis, clustering, Gibbs sampling, EPPF.

1 Introduction

In several statistical settings there is the need to model data organized in groups, allowing for sharing of information across them. In the Bayesian framework, this is achieved by hierarchical modeling, where the joint distribution of group-specific parameters accounts for such dependence. For instance, in Bayesian nonparametrics, the seminal work of Teh *et al.* , 2006 considered a mixture model within each group j, where the group-specific parameter is the mixing measure P_j and whose joint law is defined by an extra layer of hierarchy, yielding to the hierarchical Dirichlet process. This approach has been extended to the class of NRMI (Regazzini *et al.* , 2003) by Camerlenghi *et al.* , 2019 and Argiento *et al.* , 2020. In the cited works, the mixing measure is infinite dimensional.

In this work, we propose a hierarchical model where the group-specific mixing distribution belongs to the class of almost surely finite dimensional distributions introduced by Argiento & Iorio, 2019. We assign the joint law of the group-specific parameter such that the random measures within each group share the same support. In this framework, it is possible to define a group-dependent clustering as follows. First, an latent parameter $\boldsymbol{\theta}_{j,i} \sim P_j$ for individual i and group j is introduced. Second, since P_j is almost surely discrete, ties within are expected, leading to a group-specific clustering. Finally, since the P_j's share same support, we expect also ties between groups, providing a global clustering. We are able to derive the joint law of the group-specific clustering as well as the one of the global clustering. Such results allows to build up a posterior sampling strategy based on the Gibbs sampler.

2 Model developments

Let y_{ji} be the observed variable for group j, $j = 1,\dots,d$, and individual i, $i = 1,\dots,n_j$. We assume that the data in each group j come from a mixture of M components, that is

$$y_{j1},\dots,y_{jn_j} \mid w_{jl}, \boldsymbol{\tau}_l, M \sim \sum_{l=1}^{M} w_{jl} f(y_{ji} \mid \boldsymbol{\tau}_l), \qquad (1)$$

where $f(y_{ji} \mid \boldsymbol{\tau}_l)$ is called kernel and is a parametric density over the sampling space, w_{jm} are the group-specific mixing weights and $\boldsymbol{\tau}_l$ are the kernel parameters that are shared across groups. We assign a prior distribution on the mixing weights by normalization, namely we define $w_{jl} = \frac{S_{jl}}{T_j}$, where $T_j = \sum_{l=1}^{M} S_{jl}$. Also, we assume a prior distribution on the number of components, i.e., $M \sim q(m)$. Conditionally on M, S_{jl} are independent positive random variables with distribution $h_j(s)$, while $\boldsymbol{\tau}_l$ follows a prior distribution over Θ, the parameter space of the kernel, that we denote $p_0(\boldsymbol{\tau})$.

As in Argiento & Iorio, 2019, the model can be framed in a Bayesian nonparametric fashion. Indeed, $q(M)$, $h_j(s)$ and $p_0(\boldsymbol{\tau})$ define the joint distribution of a vector of almost sure discrete random measures $P_1,\dots,P_d$ with support $\boldsymbol{\Theta}$, where

$$P_j = \sum_{l=1}^{M} \frac{S_{jl}}{T_j} \delta_{\boldsymbol{\tau}_l}(\boldsymbol{\theta}), \quad j = 1,\dots,d \qquad (2)$$

with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. We refer the joint distribution of $P_1,\dots,P_d$ to as the Vector Normalized Independent weights, i.e., $V - NIw(q, h_j, p_0)$. Model (1) and the priors described above can be rewritten in a hierarchical form as follows:

$$\begin{aligned} &y_{ji} \mid \boldsymbol{\theta}_{ji} \overset{\text{ind}}{\sim} f(y_{ji} \mid \boldsymbol{\theta}_{ji}) \\ &\boldsymbol{\theta}_{j1},\dots,\boldsymbol{\theta}_{jn_j} \mid P_j \overset{\text{iid}}{\sim} P_j \\ &P_1,\dots,P_d \mid q, h, p_0 \sim V - NIw(q, h_j, p_0). \end{aligned} \qquad (3)$$

In this work, the kernel $f(y \mid \boldsymbol{\theta})$ represents the density of a univariate normal distribution with parameter $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. We assume $q(m)$ to be the p.m.f. of a $1-$shifted Poisson distribution with parameter Λ and $h_j(s)$ is the density of a gamma distribution with shape parameter γ_i and rate equal to 1. Finally, $p_0(\boldsymbol{\tau})$ is the density of a conjugate normal inverse gamma prior with parameters μ_0, κ_0, ν_0 and σ_0^2.

3 Group-dependent clustering

The hierarchical model in (3) allows to define a group-dependent clustering based on the latent variables $\boldsymbol{\theta}_{ji}$. First, we introduce latent allocation variables c_{ji} such that $c_{ji} = m$ if $\boldsymbol{\theta}_{ji} = \tau_m$. Then, we denote $\mathcal{M}^{(a)}$ the set of couples (j,m) such that $\exists i$ for which $c_{ij} = m$ and we define the number of *allocated columns* as

$$M^{(a)} = \#\left\{m : \text{there exists one couple}(j,m) \in \mathcal{M}^{(a)}, j = 1, \ldots, d\right\}.$$

We denote $\mathcal{M}^{(na)}$ the complement of $\mathcal{M}^{(a)}$. Hence, for every pair (j,m), we define $n_{jm} = \#\{(j,i) : c_{ji} = m\}$. Note that

$$(j,m) \in \mathcal{M}^{(na)} \Rightarrow n_{jm} = 0$$

$$(j,m) \in \mathcal{M}^{(a)} \Rightarrow n_{jm} \geq 0.$$

Finally, let $c_1^*, \ldots, c_{M^{(a)}}^*$ be the allocated columns, that is, the indexes within $\{1, \ldots, M\}$ such that $(j, c_k^*) \in \mathcal{M}^{(a)}$.

We are now ready to define, for each group j, the clustering $\rho_j = \{A_{j1}, \ldots \ldots, A_{jM^{(a)}}\}$, where $A_{jk} = \{(j,i) : \ (j, c_{ki}^*) \in \mathcal{M}^{(a)}\}$ and $k = 1, ..., M^{(a)}$. In other words, A_{jk} is the set of data points of group j belonging to the k-th cluster. Note that, a distinctive feature of our setting, is that A_{jk} can be an empty set. Nevertheless, if $A_{jk} = \emptyset$ appears in ρ_j, it means that there is at least another group $\tilde{j}$ such that $A_{\tilde{j}k}$ is not empty.

We build upon the work Argiento & Iorio, 2019 and James *et al.* , 2009 to derive the joint distribution of the clustering $\rho_1, \ldots, \rho_d$, induced by the hierarchical mixture model (3). This turns out to be:

$$\pi(\rho_1,...,\rho_d,M^{(a)}) = \int_0^{\infty}\cdots\int_0^{\infty}\prod_{j=1}^{d}\frac{1}{\Gamma(n_j)}u_j^{n_j-1}\prod_{k=1}^{M^{(a)}}\kappa_{\gamma_j}(n_{jk},u_j)$$
$$\exp\left[-\Lambda\left(\prod_{j=1}^{d}\psi_{\gamma_j}(u_j)-1\right)\right] \tag{4}$$
$$\Lambda^{M^{(a)}-1}\left[\Lambda\prod_{j=1}^{d}\psi_{\gamma_j}(u_j)+M^{(a)}\right]du_1\dots du_d,$$

where $\psi_{\gamma_j}(u_j)=\frac{1}{(u_j+1)^{\gamma_j}}$ is the Laplace transform of a gamma distribution with shape γ_j and rate equal to 1, while $\kappa_{\gamma_j}(n_{jk},u_j)=\frac{\Gamma(\gamma_j+n_{jk})}{\Gamma(\gamma_j)}\frac{1}{(u_j+1)^{n_{jk}+\gamma_j}}$ is its relative cumulant function. The joint distribution in (4) enables us to build a Gibbs sampler for sampling from the full posterior distribution. We omit here the details for brevity. We will illustrate the performance of our model over a set of simulated and real data.

References

ARGIENTO, RAFFAELE, & IORIO, MARIA DE. 2019. Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *arXiv: Methodology*.

ARGIENTO, RAFFAELE, CREMASCHI, ANDREA, & VANNUCCI, MARINA. 2020. Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association*, **115**(529), 318–333.

CAMERLENGHI, FEDERICO, LIJOI, ANTONIO, ORBANZ, PETER, PRÜNSTER, IGOR, *et al.* . 2019. Distribution theory for hierarchical processes. *Annals of Statistics*, **47**(1), 67–92.

JAMES, LANCELOT F, LIJOI, ANTONIO, & PRÜNSTER, IGOR. 2009. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, **36**(1), 76–97.

REGAZZINI, EUGENIO, LIJOI, ANTONIO, & PRÜNSTER, IGOR. 2003. Distributional results for means of normalized random measures with independent increments. *Annals of Statistics*, 560–585.

TEH, YEE WHYE, JORDAN, MICHAEL I, BEAL, MATTHEW J, & BLEI, DAVID M. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, **101**(476), 1566–1581.

A MACHINE LEARNING APPROACH IN STOCK RISK MANAGEMENT

Salvatore Cuomo [1] , Federico Gatta[1] , Fabio Giampaolo[1] , Carmela Iorio[2]
and Francesco Piccialli[1]

[1] Department of Mathematics and Application 'R. Caccioppoli', University of Naples Federico II, Italy, (e-mail: salvatore.cuomo@unina.it, federico.gatta@unina.it, fabio.giampaolo@unina, francesco.piccialli@unina.it)

[3] Department of Industrial Engineering, University of Naples Federico II, Italy, (e-mail: carmela.iorio@unina.it)

ABSTRACT: In this paper, we propose a novel approach in stock clustering with the purpose of the construction of a portfolio optimization strategy. The idea is to exploit hierarchical Neural Network Principal Component Analysis and Adaptive LASSO in combination with the Arbitrage Pricing Theory in order to group stocks whose returns are affected by the same risk factors, and then eliminate such dependence through an appropriately constructed portfolio. We test our proposal on the Italian stock market.

KEYWORDS: neural network principal component analysis, stocks clustering, arbitrage pricing theory, pure alpha strategy

1 Introduction

In this work, we propose a novel technique in stock risk management through the construction of an appropriate pure alpha strategy. To do this, we exploit the *Arbitrage Pricing Theory* (APT) (Ross, 1976) that, given a market made up of M stocks $\Omega = \{1,...,M\}$, explain the stocks returns $X^{(j)}$, $j \in \Omega$ with a collection of standard random variables common to all stocks called *risk factors* F_i with $i = 1,...,n$. So, let's $\alpha^{(j)}$ be the intercept and $\varepsilon^{(j)}$ the error:

$$X^{(j)} = \alpha^{(j)} + \beta_1^{(j)} F_1 + ... + \beta_n^{(j)} F_n + \varepsilon^{(j)} \tag{1}$$

So, the task is to identify an appropriate set of risk factors. In literature, there are mainly two approaches: *macroeconomic*, which searches outside of the data; *statistical*, which extracts the risk factors from the data itself. We follow the statistical approach. A work of this type is that of Ladrón de Guevara Cortés *et al.*, 2019 that is the starting point for our model in that it uses the hierarchical Neural Network Principal Component Analysis (hNNPCA). As for time series clustering, we report the survey of Aghabozorgi *et al.*, 2015. In the second section, we show the data analysis techniques that we exploit. In the third section, we propose our methodology for clustering and investment. In the fourth section, we test our strategy on the Italian stock market.

2 Data Analysis Tools

2.1 Hierarchical Neural Network Principal Component Analysis

The hNNPCA is a technique of dimensionality reduction based on a neural network (NN) with 5 layers, such that both input and output are $X_t = [X_t^{(j)}]_{j \in \Omega}$. The central layer has dimension n, equal to the number of series to be extracted, known as *principal components* (PCs), and its neurons give us their value. The loss function is $E = \sum_{k=1}^{n} E_k$, where E_k is the Mean Square Error (MSE) calculated on the sub-NN obtained considering only the first k PCs.

2.2 Adaptive Least Absolute Shrinkage and Selection Operator

The *Adaptive Least Absolute Shrinkage and Selection Operator* (A-LASSO) is a feature selection technique that adjusts the LASSO estimator weighting the contribution of each coefficient, when computing the l_1 norm, with a weight that can be obtained from an Ordinary Least Squares (OLS) regression. Namely, given a linear model with K observations and n inputs: $X^{(j)} = \sum_{i=1}^{n} F_{i,t} \beta_i^{(j)} + \varepsilon_t^{(j)}$, the A-LASSO estimates the coefficients $\beta^{(j)}$ as the argmin of:

$$[\,\frac{1}{K}\sum_{t=1}^{K}(X_t^{(j)} - \sum_{i=1}^{n} F_{i,t}\beta_i^{(j)}) + \lambda \sum_{i=1}^{n} |\beta_i^{(j)} v_i|\,] \quad v_i = |\hat{\beta}_{OLS,i}|^{-\tau},\ \lambda > 0,\ \tau > 0 \quad (2)$$

The A-LASSO is exploited only for feature selection, so its final results is $\mathcal{A}_j = \{i \in \{1,\dots,n\}\ s.t.\ \beta_i^{(j)} \neq 0\}$. As for the regression, we exploit that of Fama and MacBeth (FMB) that is performed by dividing the train data into subsets and averaging the OLS coefficients obtained in the subsets.

3 The Methodology

3.1 Stocks Clustering

Firstly, we use the hNNPCA to obtain n PCs that, after the standardization process, are used as risk factors in equation 1. The underlying idea is that not all PCs affect the returns of all the stocks, so we apply the A-LASSO to perform feature selection. The hyperparameters λ and τ are set with grid-search searching to minimize the estimate of the MSE provided by the 3-fold nested cross-validation. In this stage we discard the combinations of hyperparameters that save less than 2 or more than 4 PCs, to prevent strong regularizations or complex models. So, for each stock j, we have a subset of PCs $\mathcal{A}_j$ that really affects j returns. After introducing the equivalence relationship between stocks $j \sim l \iff \mathcal{A}_j = \mathcal{A}_l$, the clusters are the equivalence classes of $\sim$.
Two strengths of our strategy are that we don't need to know in advance the number of clusters to create and we don't need to establish a similarity measure between the considered time series. These are, according to Aghabozorgi, difficult points in the traditional clustering algorithms. However, we have to set λ and τ, and not all the obtained clusters are usable in practice.

3.2 Pure Alpha Strategy

Now, we use the clustering to obtain an investment strategy. Fix a class $\tilde{\mathcal{A}}$, assume that $|\tilde{\mathcal{A}}| = n$ and consider a portfolio (equation 3) made up by $n+1$ stocks in $\tilde{\mathcal{A}}$ (without loss of generality $0,...,n+1$). The coefficients are estimated with FMB and the weights $\gamma^{(j)}$ indicate the exposition on the stocks.

$$X^{(Port)} = \sum_{j=0}^{n} \gamma^{(j)} X^{(j)} = \sum_{j=0}^{n} \gamma^{(j)} \alpha^{(j)} + \sum_{i \in \tilde{\mathcal{A}}} \left(\sum_{j=0}^{n} \gamma^{(j)} \beta_i^{(j)} \right) F_i + \sum_{j=0}^{n} \gamma^{(j)} \varepsilon^{(j)} \quad (3)$$

A *pure alpha strategy* is a portfolio (designed to reduce the riskiness) s.t. the total exposition on the F_i is nil. Furthermore, for the law of large numbers, we can neglect the contribution of the $\varepsilon^{(j)}$. So, to determine the weights, we impose the l_1 norm of Γ equal to 1 and we find that there are only two admissible vectors of weights. We chose the one with the higher expected return.
If we have more than $n+1$ stocks in $\tilde{\mathcal{A}}$, then we still consider portfolios made up by $n+1$ stocks and we chose the one that maximizes the expected return.

4 A Real Application in Italian Stock Market

Now, we propose a real application in the Italian stock market. Ω is made up of 30 stocks, whose time series are supplied by mercati.ilsole24ore.com. The data are from 2010-10-26 to 2020-08-31 (train set) and from 2020-09-01 to 2020-12-31 (test set). We extract 6 PCs from the train set and we obtain 16 clusters (only 3 usable in the investment methodology). Then, the results of the pure alpha strategy in the test set are compared with those of the Italian Index **FTSE MIB**, see figure 1.

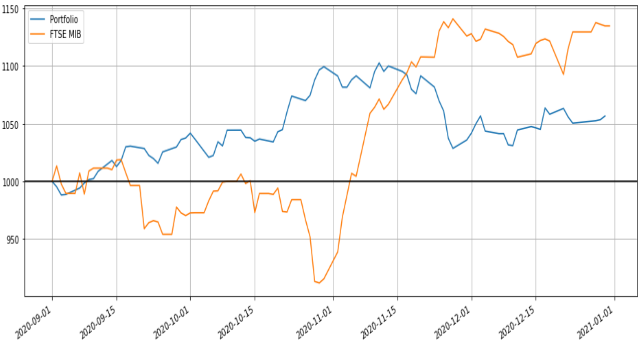


Figure 1. *Portfolio (blue) vs FTSE MIB (orange) in the period 01/09/20 - 31/12/20*

From the figure, we can see that the proposed investment methodology is quite good. In fact, it achieves a profit (even if it isn't as big as the FTSE MIB one), and it seems to be safer than the index, with quite constant growth and less downward peaks.

References

Aghabozorgi, Saeed, Shirkhorshidi, Ali Seyed, & Wah, Teh Ying. 2015. Time-series clustering–a decade review. *Information Systems*, **53**, 16–38.

Ladrón de Guevara Cortés, Rogelio, Torra Porras, Salvador, & Monte Moreno, Enric. 2019. Neural Networks Principal Component Analysis for estimating the generative multifactor model of returns under a statistical approach to the Arbitrage Pricing Theory. Evidence from the Mexican Stock Exchange. *Computación y Sistemas*, **23**(2), 281–298.

Ross, Stephen A. 1976. The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, **13**(3), 341 – 360.

PATHMOX SEGMENTATION TREES TO COMPARE LINEAR REGRESSION MODELS

Cristina Davino[1], Giuseppe Lamberti[2]

[1] Department of Economics and Statistics, University of Naples Federico II, (e-mail: cristina.davino@unina.it)

[2] Department of Business, Universitat Autonoma de Barcelona, (e-mail: giuseppe.lamberti@uab.cat)

ABSTRACT: The estimation of a dependency model for a group as a whole does not take into account possible heterogeneity, i.e., the presence of possible partitions characterised by different dependency structures. We propose a procedure that exploits the potential of segmentation trees to identify partitions in an initial set of data characterised by different linear regression patterns.

KEYWORDS: "pathmox approach", "linear regression", "heterogeneity", "F-Fisher".

1 Introduction

Segmentation trees have been attracting a great deal of attention as model comparison tools, with research mainly motivated by the fact that segmentation trees allow identification of partitions of data characterised by different dependency structures. Few algorithms have been proposed by the statistical community that combine model estimation and segmentation trees, outside the MOdel-based recursive partitioning (MOB) procedure proposed by Zelies *et al.* (2008). In a new approach we generalize the pathmox algorithm developed by Lamberti *et al.* (2016) to the context of linear regression models, using a model comparison test to identify the most significant partitions (i.e., subgroups) in data. Further developments of the proposed approach will involve extensions to other contexts such as quantile regression.

2 State-of-the-art

Analysis of a dependency model can be furthered by assessing whether a model and/or the impact of regressors on dependent variables differ if heterogeneity is observed. In other words, it may be interesting to assess differences between a global model estimated on the whole set of observations and models nbased on sub-groups identified on the basis of known categorical variables external to the model. These variables may identify partitions characterised by a dependency structure heterogeneity. The most popular approaches to comparing regression models rely on comparative statistical testing or on recursive methods. The comparison approach consists of comparing coefficients related to a model common to all the data (i.e., a restricted model representing a homogeneous situation) and another model that reflects the interactions between categorical and predictor variables (i.e., an unrestricted model corresponding to a heterogeneous situation). The comparison approach, which allows for analysis of one categorical variable at a time, is reflected in the F-tests developed by Chow (1960) and Lebart *et al.* (1979), based on an assumption of the normality of the residuals of the two models. Comparison is done by calculating restricted deviance (SSR_0) and unrestricted deviance (SSR_1). The latter will be lower if interaction between categorical and predictor variables is significant. Under the null hypothesis, if both types of deviance are equal, then the categorical variables produce no differences in model coefficients. This null hypothesis is tested by computing an F–statistic:

$$F = \frac{(SSR_0 - SSR_1)/(n-2p)}{SSR_1/p} \tag{1}$$

The recursive approach, based on multiple model comparisons, ranks variables that produce differences in the model coefficients. The outcome is a tree where each node represents a model. Partitions are obtained by comparing the effect of each categorical variable on the model coefficients and choosing the partitions that produce the biggest differences. This approach requires a criterion to quantify differences in the model coefficients. In case of the MOB procedure this criterion is based on a fluctuation test that measures coefficient instability (Zelies and Hornick, 2007) as caused by a categorical variable. High instability points to a significant effect of the variable. Tree partitions are defined according to the variables that produce the highest instability.

3 Pathmox in a nutshell

Pathmox (Lamberti *et al.*, 2016), developed to detect heterogeneity in models, is a recursive algorithm based on segmentation trees. While pathmox was introduced in the context of partial least square structural equation modelling, it can be generalized to other contexts when a suitable test for comparing models is available. The algorithm applies binary segmentation principles to produce a tree with different models in each node. It starts by fitting a global model to all the data (i.e., the tree root) and identifies models with the most significant differences in child nodes. The most different models are identified by minimizing the sum of the squares of the residuals of the models estimated in each child node. The available data are recursively partitioned according to categorical variables – not included in the model – that yield the most significant differences in the child nodes. Partitions are identified using a test that determines the degree of difference between two compared sub-models. Finally, pathmox avoids overfitting using stopping rules based on maximum depth, minimum node size and non-significance of the partitioning criterion. As the partitioning criterion we propose the hypothesis test as proposed by Lebart *et al.* (1979) and Chow (1960) to compare two linear regression models.

4 Employee satisfaction: a pathmox application

Using data referring to an organizational study of 2,000 employees in a Spanish financial institution, we applied the pathmox approach in an empirical analysis of the impact of work climate satisfaction on overall employee satisfaction. Overall satisfaction and specific work climate aspects (empowerment, company reputation, supervisor leadership, pay and work conditions) were scored on a 5-point Likert scale. The following categorical variables, reflecting specific employee characteristics, were considered as potential sources of heterogeneity: *age* (<31, 31-45, >45 years), *gender*, *marital status* (married, not married), *education* (secondary, graduate, post-graduate), *job grade* (low, intermediate, high) and *antiquity* in the organization (<2004, 2005-2009, 2009-2014, >2014).

Pathmox analysis results are reported in Figure 1 and Table 1. We set maximum depth to two levels, bounded the final number of segments to a maximum of four and set the minimum admissible node size to 10% of the total sample. The significance threshold for the partitioning algorithm was p=0.05. The pathmox algorithm identified *job grade* as the variable with the greatest power, distinguishing between low-intermediate grade and high grade employees (LM1 and LM2, respectively). LM1 (low-intermediate grade) was further differentiated according to *antiquity*. On the basis of job grade combined with antiquity, we could characterise partitions and assign labels to subgroups. Thus, LM2 can be defined as the group of managers, LM3 as senior employees and LM4 as junior employees. Finally, the global model coefficients were compared with the coefficients for the three models estimated for the sub-samples identified by the terminal nodes (Table 1), showing that, in terms of satisfaction, managers primarily valued empowerment followed by company reputation, senior employees valued empowerment, while junior employees mainly valued pay and leadership.

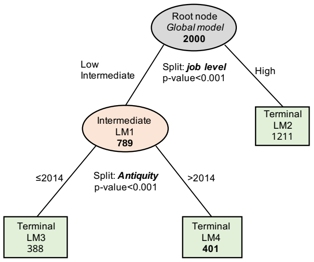


Figure 1. *Pathmox tree*

Table 1. *Coefficient comparison for global and terminal nodes.*

	LM β coefficients				
	Empowerment	Company reputation	Supervisor leadership	Pay	Work conditions
Global model	0.328	0.190	0.158	0.169	0.181
LM2: managers	0.267	0.209	0.116	0.118	0.191
LM4: senior	0.517	0.247	0.142	0.120	0.201
LM3: junior	0.271	0.052^{NS}	0.333	0.342	0.121

^{NS} indicates non-significance according to the t-test

Our results suggest that pathmox can be used to compare regression models, opening up a future research line in other contexts such as quantile regression. While the algorithm allows partitions to be identified where differences between model coefficients are greatest, it has the limitation that no overall significance criterion is considered once each partition is identified. This important aspect needs to be considered in a future version of the algorithm. Note that pathmox aims to identify the most significantly different sub-groups, unlike a classic decision tree where the objective is to obtain the best prediction based on splitting observations into sub-groups. Therefore, the only similar method is the MOB proposed by Zelies *et al.* (2008), which, however, uses a different criterion to identify the best partitions. A comparison of both approaches will be a natural next step in our research.

References

CHOW, G.C. 1960. Test of equality between sets of coefficients in two linear regressions. *Econometrica.*, **28**, 591–605.

LAMBERTI, G., ALUJA T., & SANCHEZ, G. 2016. The Pathmox approach for PLS path modeling. *Applied Stochastic Models in Business and Industry.*, **32**, 453–468.

LEBART, L., MORINEAU A., & FEENELON, J.P. 1979. *Traitement des donnees statistiques.* Paris: Dunod.

ZEILEIS, A., & HORNIK, K. 2007. Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica.*, **61**, 488–508.

ZEILEIS, A., HOTHORN T., & HORNIK, K. 2008. Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics.*, **17**, 492–514.

ANGULAR HALFSPACE DEPTH: CLASSIFICATION USING SPHERICAL BAGDISTANCES*

Houyem Demni[1], Davide Buttarazzi[1], Stanislav Nagy[2],
and Giovanni C Porzio[1]

[1] Department of Economics and Law, University of Cassino and Southern Lazio
(e-mail: houyem66@gmail.com, davidebuttarazzi@outlook.com, porzio@unicas.it)

[2] Department of Probability and Mathematical Statistics, Charles University
(e-mail: nagy@karlin.mff.cuni.cz)

ABSTRACT: Directional data lies on the surface of the unit sphere. Exploiting new results on the computation and the properties of the angular halfspace depth, we introduce the spherical version of the bagdistance, applicable to directional data. A bagdistance-based classification method for directional data is considered. The proposed method will be compared with other directional classifiers by means of a simulation study.

KEYWORDS: angular depth, bagdistance, directional data, supervised learning.

1 Introduction

Depth functions are nonparametric tools that assess how “centrally located”, or “inner” is a point with respect to (w.r.t.) a given probability distribution. They have been successfully adopted in supervised classification analysis. However, many depths suffer when evaluating points that lie in the tails of the distribution. This is because the depth functions are typically not robust at their lowest values, and also because they can easily assign constant zero depth to many points when evaluated w.r.t. datasets (the so-called outsider issue). An example of an important depth sharing all these shortcomings is the standard *halfspace depth* defined in Euclidean spaces $\Re^q$, $q \geq 1$.

Contrary to the depths, distance functions are much more powerful when dealing with points at the extremes of the distribution. Nevertheless, they generally suffer from robustness issues as well (unless some robustified versions

*The work of H. Demni and G.C. Porzio has been partially funded by the BiBiNet project (grant H35F21000430002) within the POR-Lazio FESR 2014-2020. The work of S. Nagy was supported by the grant 19-16097Y of the Czech Science Foundation, and by the PRIMUS/17/SCI/3 project of Charles University.

are adopted), and for a fruitful use of the distances in classification, certain assumptions on the data distribution typically need to be imposed (e.g., ellipticity of the underlying distribution in the case of the Mahalanobis distance).

For these reasons, and to introduce a supervised classification rule for Euclidean data, Hubert *et al.*, 2017 proposed to combine the information from these two approaches to obtain the so-called *bagdistance*, a function which joins the depth and the distance to obtain a measure of how close/inner is a point w.r.t. a given distribution. Bagdistances are robust, nonparametric, and able to manage information in the tails of the distribution.

In this work, we introduce the bagdistance for directional data. To do so, we use the angular halfspace depth, being the directional analogue of the standard halfspace depth from $\Re^q$. We also evaluate the performance of the bagdistance within the setting of supervised classification for directional data.

Our short paper is organized as follows. Section 2 provides some background on the bagdistance in the Euclidean case, while in Section 3, the spherical bagdistance and a directional classifier based on it are introduced.

2 The bagdistance for Euclidean data

Let Y be a random variable in $\Re^q$ with distribution P_Y, and let θ be its halfspace median (the point that maximizes the halfspace depth w.r.t. P_Y, or the barycentre of the set of such points if not a singleton). Denote by $B(Y) \subset \Re^q$ the smallest halfspace depth central region of P_Y (i.e., an upper level set of the halfspace depth of P_Y) that contains at least 50 % of the P_Y-probability mass. The bagdistance of x to Y is given by the ratio of the Euclidean distances of x to θ, and $c(x)$ to θ:

$$BD(x,P_Y) := \begin{cases} 0 & \text{if } c(x) = \theta, \\ \|x-\theta\| \,/\, \|c(x)-\theta\| & \text{otherwise,} \end{cases}$$

where $c(x)$ is the intersection of the boundary of the bag $B(Y)$ and the ray from the halfspace median θ passing through x.

3 The spherical bagdistance and a classification rule

Directional data can be viewed as realizations of a random variable X whose support is the unit hyper-sphere $S^{(q-1)} := \{x \in \Re^q\colon\ \|x\| = 1\}$. For directional data, the spherical bagdistance can be introduced in complete analogy with the bagdistance for Euclidean data.

We first define the directional variant of the halfspace depth. Let X be a directional random variable with distribution P_X. The *angular halfspace depth* ahD of a point $x \in S^{(q-1)}$ w.r.t. P_X can be defined considering the collection $\mathscr{H}_0$ of closed halfspaces in $\Re^q$ whose boundary contains the origin:

$$ahD(x, P_X) := \inf\{P_X(H) \colon H \in \mathscr{H}_0,\ \ x \in H\} \in [0,1].$$

Denote by $aB(X) \subset S^{(q-1)}$ the *angular bag* of X, defined as the smallest angular depth central region containing at least 50 % of the P_X-probability mass. Such a region always exists; its properties are detailed in the contribution of *P. Laketa* and *S. Nagy* in the present book of short papers. The *spherical bagdistance* from $x \in S^{(q-1)}$ to X is defined as the ratio of the arc distance between x and the angular halfspace median $\tilde{\theta}$ (a maximizer of the angular halfspace depth of X), and the arc distance between $c_{aB}(x)$ and $\tilde{\theta}$. Here, $c_{aB}(x)$ is the intersection between the boundary of the angular bag $aB(X)$ and the geodesic from $\tilde{\theta}$ to x. Altogether, we define

$$SBD(x, P_X) := \begin{cases} 0 & \text{if } c_{aB}(x) = \tilde{\theta}, \\ \arccos(x^\top \tilde{\theta}) / \arccos(c_{aB}(x)^\top \tilde{\theta}) & \text{otherwise.} \end{cases}$$

Similarly as the usual bagdistance in $\Re^q$, the spherical bagdistance can be exploited for supervised classification of directional objects. Formally, considering K directional distributions on $S^{(q-1)}$, a directional classifier is defined as the function $class : S^{(q-1)} \to \{1, \ldots, K\}$. Given a training set composed of K empirical distributions $\hat{P}_{X_i}$, $i = 1, ..., K$, the directional bagdistance classifier is then defined as the rule $class_{bag}$ such that:

$$class_{bag}(x) := u(SBD(x; \hat{P}_{X_1}), ..., SBD(x; \hat{P}_{X_i}), ..., SBD(x; \hat{P}_{X_K})),$$

where $u : \Re^K \to \{1, ..., i, ..., K\}$ is some discriminating function. That is, the classifier is a rule defined on a Euclidean space given by the bagdistances of the training set values w.r.t the directional distributions defined on a Riemannian manifold. For the choice of the discriminating function, we refer to the literature available for depth based classifiers, which includes the linear (LDA), quadratic (QDA) and k-NN classifiers (see e.g., Demni *et al.*, 2021).

In line with such a strategy, a simulation study with data generated according to a Kent distribution for each group has been performed. First results are promising: the spherical bagdistance classifier reaches the same level of correct classification as achieved by the empirical Bayes, at least under some circumstances. To exemplify, boxplots of the misclassification rates of the proposed classifier and of the empirical Bayes classifier under Kent are reported in Figure 1. The two Kent distributions have equal locations and ovalness, and different concentrations (the simulation setting described in Setup 2 in Demni & Porzio, 2021 has been adopted). The training set size is 400 (200 from each group), while the size of the testing set is 200; the number of replications is 100. Misclassification errors are essentially equivalent, with some preference to be given to the LDA and QDA solution. Performances under other simulation settings and comparison with other directional classifiers are under investigation.

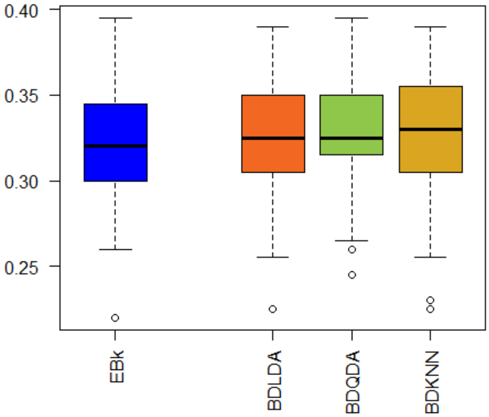


Figure 1: Misclassification rates of the empirical Bayes under Kent (EBk), and the spherical Bagdistance classifier (BD) when associated with the LDA, QDA, and *k*-NN classification rule. Data generated according to Kent distributions.

References

DEMNI, HOUYEM, & PORZIO, GIOVANNI C. 2021. Directional DD-classifiers under non-rotational symmetry. *IEEE Xplore, submitted.*

DEMNI, HOUYEM, MESSAOUD, AMOR, & PORZIO, GIOVANNI C. 2021. Distance-based directional depth classifiers: a robustness study. *Communications in Statistics – Simulation and Computation, in press.*

HUBERT, MIA, ROUSSEEUW, PETER, & SEGAERT, PIETER. 2017. Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*, **11**(3), 445–466.

NEURAL NETWORKS FOR HIGH CARDINALITY CATEGORICAL DATA

Agostino Di Ciaccio

Department of Statistics, University of Rome “La Sapienza”,
(e-mail: agostino.diciaccio@uniroma1.it)

ABSTRACT: If we want to apply neural networks to categorical data, we must necessarily adopt a coding strategy. This is a common problem for many multivariate techniques and several approaches have been suggested. In this paper, a method is proposed to analyze categorical variables with high cardinality. An application to simulated data illustrates the interest of the proposal.

KEYWORDS: encoding categorical data, neural networks, high cardinality attributes.

1 Introduction

Several machine learning algorithms cannot handle directly categorical variables and, in any case, categorical data can pose a serious problem if they have too many categories. Postal code is a good example of a categorical variable with high cardinality. This paper starts with some considerations on the currently used approaches, then an efficient encoding method is proposed for supervised neural networks when categorical variables with high cardinality need to be analyzed.

2 Approaches to quantify categorical features

Several methods have been proposed to encode categorical variables (a recent review is Hancock et al. 2020). From our point of view, they can be classified as:

1- Methods that do not use the target variable. In this category we find rather crude methods, such as the *Label Encoder* or the *Hashing Encoder*. The quantifications obtained are essentially arbitrary.

2- Methods that use only the target variable. The *Target Encoder* (TE) replaces the categorical variable with the conditional means of the target variable. This method often produces data leakage, to limit this inconvenience the *Leave one out Encoder* or the *Catboost Encoder* have been proposed.

3- Methods based on *One Hot Encoding* (OHE). In this approach a new binary variable is introduced for each category, indicating the presence or absence of that category. The eventual exclusion of one category is due to the multicollinearity problem (the dummy variable trap), but applying machine learning models, as the neural networks, it is necessary to include all the categories, otherwise we would never consider the omitted category.

3 Single and multiple quantifications by OHE

One Hot Encoding is the most used method. The coding in dummies does not depend directly on the target. Despite its great use, some drawbacks of OHE are well known: the tendency of dummy variables to cause overfitting; the introduction of many new orthogonal variables, which can slow down or affect learning; memory problems.

The encoding of categorical variables has been extensively studied in the approach based on Optimal Scaling (OS, Gifi 1990) where the *embedding* of the categories in a p-dimensional space was proposed. Given a categorical variable X which can assume the values $[a_1, a_2, ..., a_k]$, with k the number of categories, n the number of observations, then $G = [g_1, g_2, ..., g_k]$ is the indicator matrix with dimension $n \times k$. Let **c** a vector of k real values, the quantification of X is the vector:

$$\mathbf{x} = \mathbf{Gc} = \sum_{h=1}^{k} c_h \mathbf{g}_h \tag{1}$$

The values of **c** are the quantifications of the k categories and have to be estimated. The vector of the quantified data **x** is a linear combination of the indicator variables, which are an orthogonal base of R^k, then is defined in a subspace of R^k. To obtain ordered quantifications in the OS, the order indicator matrices, with non-negativity constraints on the coefficients, can be used (Gifi 1990).

In expression (1) we considered a single quantification for a categorical variable. There are several reasons that may lead to consider two or more quantifications of the same variable (Di Ciaccio 2020). Considering a regressive problem, in OS (MORALS, Young et al. 1976) it is possible to obtain a multiple quantification by means of copies of the variables (Gifi 1990). After choosing the number p of quantifications, we can extend (1) as:

$$\underset{n\times p}{\mathbf{X}} = \underset{n\times k}{\mathbf{G}}\ \underset{k\times p}{\mathbf{C}} = \sum_{h=1}^{k} \underset{n\times 1}{\mathbf{g}_h} \cdot \underset{1\times p}{\mathbf{c}_h} \tag{2}$$

In neural network applications, fixing a low p, equal to 2 or 3, is usually enough for a good quantification of categorical variables even with high cardinality.

To introduce quantification (2) in a neural network it is necessary to define, for each categorical variable, a distinct input and a dense layer with p neurons without bias and with linear activation function. In the next layer the outputs, coming from all the variables, must be concatenated. For example, given 3 input categorical variables, each with 100 categories, and one hidden layer containing 512 neurons, using this approach we must estimate (considering a regression problem and p=2) 4.697 weights. Given t=512, p=2, m=3, k_j=100 for each j, the Neural Network can be written:

$$\hat{\mathbf{y}} = \beta_0 + \sum_{s=1}^{t} \beta_s \phi\left(\sum_{j=1}^{m} \sum_{r=1}^{p} \mathbf{G}_j \mathbf{c}_j^r w_{jrs} + w_{0s} \right) \tag{3}$$

where $\phi(.)$ is the activation function of the hidden layer, $\mathbf{c}_j^r$ is the quantification of the j-th variable on the r-th dimension. Conversely, in the classical OHE encoding:

$$\hat{\mathbf{y}} = \beta_0 + \sum_{s=1}^{t} \beta_s \phi\left(\sum_{j=1}^{m} \sum_{r=1}^{k_j} \mathbf{G}_j w_{jrs} + w_{0s} \right) \tag{4}$$

obtaining 154.625 weights to estimate.

G$_j$ can be very big sparse matrices (sparsity equal to $1-1/k_j$), but we can avoid building such an inefficient coding estimating the dense matrix of quantifications **C**$_j$ of expression (3) without building the sparse matrix **G**$_j$.

In the first step, for a categorical variable *X*, the *k*-dimensional 'vocabulary' **V** of the categories have to be created and indexed. Then all the categories in the data will be substituted by the corresponding numerical index in the vocabulary, in a similar way to what the Label Encoder does. Call a_i the modality assumed by the categorical variable, and $\mathbf{v}[a_i]$ the index in the vocabulary corresponding to this modality. The *i*-th row of the $(n \times p)$ matrix of the quantified variable *X* can be expressed as:

$$\mathbf{x}_i = \mathbf{C}[\mathbf{v}[a_i]] \tag{5}$$

Each line of the quantification matrix **C** can be seen as the *p*-dimensional representation of one category. Inspired by Natural Language Processing, Guo & Berkhahn's (2016) *entity embedding* technique takes a similar approach. To obtain the estimate of **C** in a supervised neural network, the gradient descent and the backpropagation can be used, where the matrix **C** is initialized with random values taken from a standardized normal and subsequently updated through an iterative procedure to minimize the loss function, which in the case of regression is the classic Sum of Square Error. We call this technique LEE, Low Embedding Encoder, and to illustrate the proposed approach, a small simulation for a regression problem was build. Given three qualitative variables X_1, X_2, X_3 with 200 categories each (coded as the integers between 1 and 200), for each variable 20,000 observations were extracted randomly from a uniform distribution, then *Y* was computed by the rules:

$(X_1 > X_2 \ and \ X_3 < 100) \rightarrow Y \sim N(20,1.5)$

$(X_1 \leq X_2 \ and \ X_3 < 100) \rightarrow Y \sim N(10,1.5)$ $\quad$ $else \quad Y \sim N(1,1.5)$

There are only 3 expected values $E(Y \mid x_1, x_2, x_3)$, i.e. (1, 10, 20), so an optimal regressive model should predict these values. Note that the expected value of *Y* depends on the interaction of the three categorical variables and that the three conditional distributions of *Y* overlap in the tails. The dataset was then splitted as training-set (50%) and test-set (50%). Regression algorithms such as MORALS or Regression Tree cannot make a satisfactory prediction on this data unless introducing explicitly the interaction terms into the model, producing thousands of dummy variables. On the contrary, neural networks are able to autonomously detect the interactions, then a small neural network was chosen to predict the target *Y* in our simulation. The network includes an input layer, two hidden layers with 8 and 3 neurons (*elu* activation function), and 1 output neuron with linear activation function. With the LEE approach, each categorical variable is considered a separate input and one dense layer with 2 neurons (*p* = 2) and no bias, for each categorical variable, is added to the input. If we want to avoid sparse matrices, an *embedding* layer can be added, for each original categorical variable, using (5). It was also checked that the results obtained did not improve, on the test-set, by changing the size of the network or the number of iterations. Although the *Target Encoder* was applied also with a bigger neural network, with 32 neurons in each hidden layer, the result is very poor even on the training-set, as this encoding prevents interactions from being identified.

Table 1. *Comparison between three approaches*

	MSE - train	*MSE - test*	*n. parameters*
OHE	2.11	6.18	4839
LEE	2.55	4.82	1287
Target Encoder	61.47	61.48	1217

Figure 1. *OHE on the test-set*

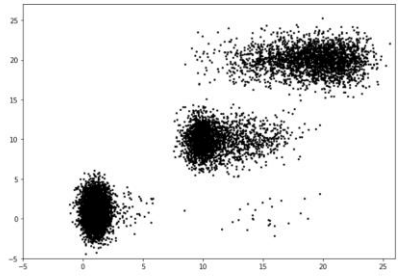
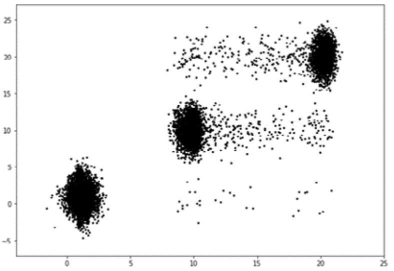


Figure 2. *LEE on the test-set*



4 Conclusions

The proposed method LEE allows to apply neural networks to categorical variables with high cardinality, reducing the number of parameters and memory resources. The results obtained show an increased predictive capacity of the neural network thanks to the more efficient architecture.

References

DI CIACCIO, A. 2020. Categorical Encoding for Machine Learning. *Book of short papers SIS2020*, A. Pollice et al. eds., ISBN 9788891910776, Pearson Italia.

GIFI, A. 1990. *Nonlinear Multivariate Analysis*. John Wiley & Sons, New York.

GUO, C., & BERKHAHN, F. 2016. Entity embeddings of categorical variables. *arXiv*:1604.06737.

HANCOCK, J.T., & KHOSHGOFTAAR, T.M. 2020. Survey on categorical data for neural networks. *Journal of Big Data*,**7**,28, https://doi.org/10.1186/s40537-020-00305-w

YOUNG, F.W., DE LEEUW, J., TAKANE, Y. 1976. Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. *Psychometrika*, v. **41**, n. 4.

ALI-MIKHAIL-HAQ COPULA TO DETECT LOW CORRELATIONS IN HIERARCHICAL CLUSTERING

F. Marta L. Di Lascio[1], Andrea Menapace[2], and Roberta Pappadà[3]

[1] Faculty of Economics and Management, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy, (e-mail: marta.dilascio@unibz.it)

[2] Faculty of Science and Technology, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy, (e-mail: andrea.menapace@unibz.it)

[3] Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, Italy, (e-mail: rpappada@units.it)

ABSTRACT: In this work we introduce a new dissimilarity measure based on the Ali-Mikhail-Haq copula, motivated by the empirical issue of detecting low correlations and discriminating variables with very similar rank correlation. This issue arises from the analysis of panel data concerning the district heating demand of the Italian city Bozen-Bolzano. In the hierarchical clustering framework, we empirically investigate the features of the proposed measure and compare it with a classical dissimilarity measure based on Kendall's rank correlation.

KEYWORDS: Ali-Mikhail-Haq copula; cluster analysis; dissimilarity measure; low correlation.

1 Introduction

Copula-based measures of association have been employed in clustering procedures in a variety of applied contexts, since they allow to describe complex multivariate dependence structures and address specific features of the joint distribution of random variables, such as asymmetries and tail dependence (Durante & Sempi, 2015). For instance, the copula approach made it possible to define pairwise dissimilarities in terms of concordance or tail dependence measures (see, e.g., Fuchs *et al.*, 2021, and the references therein).

While many contributions in this context have focused on detecting high association between extremely low/high values, in this paper we focus on modeling weak correlation and the ability to discriminate objects with low and very similar degree of dependence. This issue comes from the features of the district heating (DH hereafter) demand from residential users of the Italian city of Bozen-Bolzano. We thus propose a new dissimilarity measure based on the

Ali-Mikhail-Haq (AMH hereafter) copula, and empirically compare it with a classical dissimilarity measure based on Kendall's τ coefficient.

The contribution is organized as follows. First, we introduce the copula-based dissimilarity measures (Sect. 2). Second, we present the cluster analysis performed to compare the proposed AMH-based dissimilarity with the one based on Kendall's τ (Sect. 3). Finally, Sect. 4 summarizes the main findings.

2 Kendall's τ- and AMH-based dissimilarity

Here, we want to perform an agglomerative hierarchical clustering (AHC hereafter) of m continuous random variables $(X_1, \ldots, X_m)$ defined on the same probability space by taking into account their stochastic dependence. A typical dissimilarity measure used in the AHC algorithm can be defined in terms of Kendall's τ coefficient as follows

$$d^{\tau}_{jj'} = \sqrt{2(1-\tau_{jj'})} \in [0,2] \tag{1}$$

where $\tau_{jj'}$, $j, j' \in \{1, \ldots, m\}$, is computed from n observations of the pair $(X_j, X_{j'})$. From a different perspective, one can assume a specific copula function, motivated by its ability to capture some features of the joint behaviour observed from the data. Here we focus on the AMH copula function $C(u_1,u_2) = (u_1u_2)/(1-\theta(1-u_1)(1-u_2))$, where $\theta \in [-1,1]$. The AMH copula is very suitable for modeling low degree of association since the corresponding range for τ is $[-0.1817, 0.3333]$. Hence, we introduce a new dissimilarity measure

$$d^{\text{AMH}}_{jj'} = \sqrt{2(1-\theta_{jj'})} \in [0,2] \tag{2}$$

where $\theta_{jj'}$ is the dependence parameter of the AMH copula that can be estimated via one of the methods in the literature (see, e.g., Gunky *et al.*, 2007).

3 Application to district heating demand

We analyse time series data concerning the heat demand (in kWh) of $m = 41$ residential users connected to the DH of Bozen-Bolzano, which has been identified as a key technology for the development of sustainable cities. We consider $n = 150$ hourly observations in the period Jan 1–Jan 14, 2016. We first tackle serial dependence in the original time series by adopting a dynamic panel regression model (Wooldridge, 2002), that takes into account the relationships between DH demand and meteorological variables, such as temperature and solar radiation. Then, the residual time series are used to estimate the 41×41 dissimilarity matrices based on Eqs. (1) and (2) to use in the AHC algorithm. The crucial point is that all pairs of users have a quite low Kendall's τ (the minimum is -0.2, the highest value is 0.39). Thus, in principle, d^{AMH} should be able to better distinguish objects with low and very similar degree of association. On the basis of both the informativeness of the final clusters and the separation index by Akhanli & Hennig, 2020, we decide to adopt the complete linkage method and cut the dendrogram at $k = 3$ for both the dissimilarities.

Fig. 1 displays the mean daily pattern of each user (hourly heat demand over daily average heat demand (Menapace *et al.*, 2019)) by cluster, according to d^{τ} and d^{AMH}. As can be seen, a certain degree of internal homogeneity is obtained in both cases, denoting an overall good quality of the results. However, by using static features of the buildings, such as heating surface (in m^2), age class (in years), and energy class (in yearly kWh/m^2), we can highlight the diversity between the obtained partitions. The clusters based on d^{τ} are quite similar in terms of heating surface with median values in the range $(3656, 4076)$, and even though are better separated in terms of age class and energy class, they also present a source of variability. Indeed, the 75% of buildings in cluster 1 was built between 1961 and 1990, in cluster 2 almost the 70% of buildings is dated after 1981), while cluster 3 has a larger variability, and contains both recently-constructed and old energy-renovated buildings, with relatively low energy class (the third quartile is equal to 120). On the contrary, the d^{AMH} produces groups that are different in terms of heating surface (the medians are 3969, 5382, and 3102, respectively) and show within-homogeneity with respect to the energy and age class (e.g. buildings in cluster 3 are old, i.e. mostly dated before 1990, and non-efficient with first and third quartiles of energy class equal to 120 and 145, respectively).

4 Conclusions

We have introduced a new dissimilarity measure based on the Ali-Mikhail-Haq copula and empirically showed its ability to detect low correlations and discriminate among them. The application to district heating demand illustrates that the proposed measure seems to produce clusters that have a clear interpretation in terms of the relevant features, thus leading to a valuable tool to support the management and planning of a district heating system.

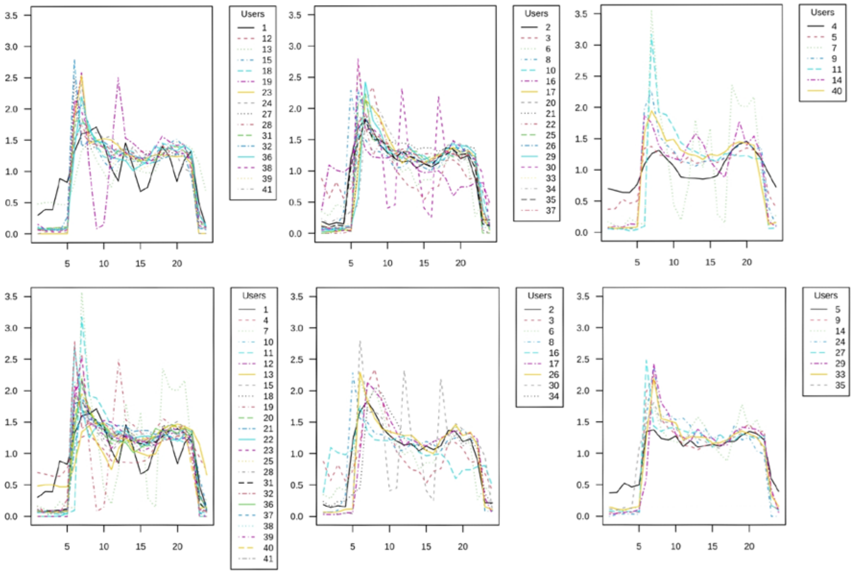


Figure 1. *Mean daily pattern (hours in x-axis) of DH users according to AHC based on d^{τ} and d^{AMH} (panels by rows) in cluster 1, 2, and 3 (panels by columns).*

References

Akhanli, S, & Hennig, C. 2020. Comparing clusterings and numbers of clusters by aggregation of calibrated clustering validity indexes. *Stat. Comput.*, **30**, 1523–1544.

Durante, F, & Sempi, C. 2015. *Principles of Copula Theory*. CRC Press, Boca Raton.

Fuchs, S, Di Lascio, F M L, & Durante, F. 2021. Dissimilarity functions for rank-invariant hierarchical clustering of continuous variables. *Comput. Stat. Data An.*, **159**, 107201.

Gunky, K, Silvapulle, M J, & Silvapulle, P. 2007. Comparison of semiparametric and parametric methods for estimating copulas. *Comput. Stat. Data An.*, **51**(6), 2836–2850.

Menapace, A, Righetti, M, Santopietro, S, Gargano, R, & Dalvit, G. 2019. Stochastic characterisation of the district heating load pattern of residential buildings. *Euroheat and Power*, **16**(3–4), 14–19.

Wooldridge, J. 2002. *Econometrics analysis of cross section and panel data*. Cambridge: MIT Press.

HIGHER EDUCATION AND EMPLOYABILITY: INSIGHTS FROM THE MANDATORY NOTICES OF THE MINISTRY OF LABOUR

Maria Veronica Dorgali[1], Silvia Bacci[1], Bruno Bertaccini[1] and Alessandra Petrucci[1]

[1] Departments of Statistics Informatics Applications "G.Parenti", University of Florence (e-mail: mariaveronica.dorgali@unifi.it, silvia.bacci@unifi.it, Bruno.bertaccini@unifi.it, alessandra.petrucci@unifi.it)

ABSTRACT: The Bologna Process has brought significant changes in the national education systems, increasing student mobility and expanding available options of education and training. Thus, an academic degree may no longer be sufficient to access the most prestigious and remunerative occupational positions. Relying on two sources of data, the Mandatory Notices of the Ministry of Labour datasets and the administrative database of University of Florence (UNIFI) students, this work aims to provide an overview of UNIFI graduates' employment and labour market participation. Preliminary results are provided.

KEYWORDS: bivariate random-effects probit model, higher education, logit model, occupational condition.

1 Introduction

In the twenty-first century the system of Higher Education (HE) in Italy has undergone profound, structural changes with a substantial increase in the number of higher education institutions (HEIs). The Bologna Process has brought significant changes in the education system, increasing student mobility and expanding available options of education and training. Thus, an academic degree may no longer be sufficient to access the most prestigious and remunerative occupational positions (Breen & Goldthorpe,1997). According to Rostan and Stan (2017) Italian graduates' employment conditions can be explained according to two main points. Firstly, even if Italy is one of the most industrialised country in Europe, its production system is characterized by small and medium size firms, poorer capacity for innovation and private and public sectors less developed than in other advanced economies. Moreover, R&D investments are insufficient and, in the last two decades, public sector lost its capacity of being the major employer of Italian graduates (ANVUR,2014). In addition, access to the liberal professions is limited by the high degree of entry regulation and the proportion of graduates employed in professional and managerial jobs has declined since 1990 (Ballarino et al., 2016). In few words, the national economy seems to lack the characteristics to valorise and reward qualified human capital (Rostan e Stan, 2017). Secondly, the expansion of HE in Italy is often not associated with the demand of skilled workers and can be explained by other factors, such as the increase of family income, the pressure of some social classes to obtain or maintain education advantages and the role of state and academy (Rostan e Stan, 2017). In this perspective, the growth of the education system has led to an oversupply of graduates, especially in some fields, worsening the employment and working conditions of degree holders (Rostan e Stan, 2017). As underlined by Assirelli et al. (2018), the 2015 unemployment rate among individuals aged 25 to 34 was higher than the corresponding value for upper secondary graduates.

In this contribution, we aim at studying in deep the topic at issue, relying on two main sources of data: the Mandatory Notices (MN) of the Italian Ministry of Labour and the administrative database of the University of Florence (UNIFI). In particular, we focus on detecting the determinants of two main variables of interest: (i) the probability of being employed and (ii) conditionally on being employed, the probability of having a permanent job.

2 Data

The analysis is based on the integration of the MN database and the UNIFI administrative archive.

The MN database is provided by the Ministry of Labour and collects information on the job contracts signed by graduates in the years after graduation, such as type of contracts (open-ended, fixed term, short term, permanent, etc.), number of working days per contract, contract effective date, graduate age and gender, economic sector. Self-employment jobs are not included in the MN database.

The UNIFI administrative archive allows us to integrate the MN dataset with information about graduates, such as enrolment date, graduation date, graduation mark, type of high school, high school graduation mark, description of the degree course, level of degree course (i.e., bachelor vs. master degree), and field of study.

The two datasets were merged using a probabilistic record linkage approach. The archive contains data on about 262,250 contracts signed by 46,931 UNIFI graduates from 1 January 2008 to 31 December 2016. All the information refers to UNIFI students that obtained their degree between 2008 and 2016. Overall, more than 60% of contracts were signed after graduation, the 37.17% within 3 years from graduation and almost the 29% more than 3 years after graduation.

Focusing on the contract signed after graduation and on those signed while studying (or during university) the most common contract among UNIFI students (bachelor and master level graduates and five-years masters) was the temporary one (59.13%); only the 10.35% of contracts were permanents. The 19.81% of contracts belongs to the category “Others” that includes “atypical” or “non standard” contracts. More in detail, permanent contracts were, respectively, the third (8.77%) and the fourth

(8.59%) most common type of contract among bachelor and master degree graduates, respectively.

3 Preliminary analyses

As preliminary analyses, we estimated two logistic regression models to detect the determinants of the probability to get the first job one year after graduation (Table 1) and the probability, one years after graduation, to obtain a permanent job contract (Table 2).

Table 1 Logistic regression results (Y=obtain the first job one year after graduation)

Variable	Bachelor graduates		Master graduates	
	Estimate	SE	Estimate	SE
Intercept	-1.8040	0.0730***	0.3320	0.0879***
Gender (Ref:" Female")				
Gender: Male	-0.0498	0.0388	-0.0193	0.0636
Age at first job (Ref=23-26)				
Age at first job: 20-23	3.2130	0.0754***	-2.4416	0.1428***
Age at first job: 26-30	4.2089	0.0831***	0.8063	0.0646***
Age at first job: 30+	3.9773	0.0990***	1.0651	0.0855***
Final grade (Ref:"106-110")				
Final grade: 75-95	-0.6303	0.0583***	0.0928	0.0944
Final grade: 96-100	-0.2899	0.0571***	0.0788	0.0914
Final grade: 101-105	-0.1794	0.0547**	0.0541	0.0830
Study area (Ref: Literature)				
Study area: Scientific	0.3496	0.0471***	0.2049	0.0751**
Study area: Social	0.5700	0.0479***	0.3174	0.0724***
Outside of prescribed time (Ref: "No")				
Outside of prescribed time: Yes	-1.6370	0.0562***	-0.7235	0.0728***
Honours (Ref: "No")				
Honours: Yes	0.1543	0.0700*	-0.0536	0.0854

Table 2 Logistic regression results (Y=obtain a permanent job contract one year after graduation)

Variable	Bachelor graduates		Master graduates	
	Estimate	SE	Estimate	SE
Intercept	3.1912	0.1173***	3.6304	0.2020***
Gender (Ref:" Female")				
Gender: Male	-0.0163	0.07437	-0.2781	0.1231*
Age at first job (Ref=23-26)				
Age at first job: 20-23	0.0491	0.1095	-0.1897	0.2109
Age at first job: 26-30	-0.2141	0.0800**	-0.0822	0.1387
Age at first job: 30+	-0.6456	0.1140***	-0.5268	0.1564***

Final grade (Ref:"106-110")				
Final grade: 75-95	-0.2175	0.1123	-0.5843	0.1862*
Final grade: 96-100	-0.1871	0.1102	-0.2413	0.1942
Final grade: 101-105	-0.0990	0.1070	-0.4297	0.1725**
Study area (Ref: Literature)				
Study area: Scientific	-0.0363	0.0899	-0.3276	0.1558*
Study area: Social	0.0968	0.0915	-0.1453	0.1543
Outside prescribed time (Ref: "No")				
Outside prescribed time: Yes	-0.1631	0.0916	-0.1619	0.1507
Honours (Ref: "No")				
Honours: Yes	-0.0692	0.1315	-0.0256	0.1934

Looking at the results of these preliminary analyses, it seems that age at first job, the final graduation mark, the study area, and being outside of the prescribed degree path play an important role in predicting professional achievements of bachelor's and master's graduates.

4 Further developments

The preliminary analyses above displayed represent a first step of our study. These analyses are static, because they refer to the job condition of graduates a year after the degree. To allow a dynamic analysis that takes into account the longitudinal structure of data, we intend to estimate a bivariate random-effect probit model. In particular, we will model, at any time occasion, the employment status (employed vs. unemployed) and the type of job contract (permament vs. temporary), given the employment status.

References

ANVUR 2014: Rapporto sullo stato del sistema universitario e della ricerca 2013. Rome: ANVUR.

ASSIRELLI, G., BARONE, C., & RECCHI, E. 2018. You Better Move On: Determinants and Labour Market Outcomes of Graduate Migration from Italy. *International Migration Review*., **53**,4-25.

BALLARINO, B., BARONE, C., & PENNICHELLA, N. 2016. Origini sociali e occupazione in Italia. *Rassegna Italiana di Sociologia*., **57**, 103–34.

BREEN, R., & GOLDTHORPE, J. H. 1997. Explaining educational differentials: Towards a formal rational action theory. *Rationality and Society*., **9**, 275–305.

ROSTAN, M., STAN., A 2017. Italian graduates' employability in times of economic crisis: overview, problems, and possible solutions. *Sociologica. SerieII*. doi: 10.4000/sociologico.1818.

An alternative to joint graphical lasso for learning multiple Gaussian graphical models

Lorenzo Focardi Olmi[1], Anna Gottard[1]

[1] Dipartimento di Statistica, Informatica, Applicazioni 'G. Parenti' (DiSIA), University of Florence, (e-mail: lorenzo.focardiolmi@unifi.it, anna.gottard@unifi.it)

Abstract: Gaussian graphical models are widely used to learn the conditional independence structure of a set of random variables. This is done through the nonzero elements of its precision matrix. In many practical situations, one needs to estimate multiple graphical models due to a group structure of the data. We propose a neighbourhood approach to jointly learn multiple Gaussian graphical models. Our method estimates the edge set of each graph through joint lasso regression, and a constrained maximum likelihood method is then used to obtain precision matrices. The estimation procedure can be refined with prior information about relations among groups.

Keywords: Gaussian graphical models, graphical lasso, joint lasso, multiple graphs

1 Introduction

Graphical models represent conditional independence relations among a set of random variables via a graph. The graph structure recovery of a concentration graph model is equivalent to find the zero elements of a precision matrix (Lauritzen, 1996).

Several recent proposals have focused on estimating Gaussian graphical models when data come from more than one distinct subpopulations. In particular, Guo *et al.,* (2011) suggested a hierarchical penalty that forces a similar sparsity pattern across classes with no shrinking non zero elements. Danaher *et al.,* (2014) proposed a direct extension of Glasso (Friedman *et al.,* 2008) using two different convex penalties to force precision matrices to be similar. Dondelinger & Mukherjee (2018) developed a lasso type penalty to handle observations divided into groups in a regression setting.

In this work, we propose a nodewise regression approach to jointly estimate multiple Gaussian graphical models using a penalty similar to the one proposed by Dondelinger & Mukherjee (2018) for inducing similarities in zero entries of regression coefficients. A full estimate of the precision matrices is then obtained via constrained maximum likelihood approach in each group.

2 Nodewise multiple graphical models

Meinshausen & Bühlmann (2006) firstly proposed the idea of neighbourhood selection based on penalized linear regressions. Their proposal consists in performing d lasso regression procedures, one for each variable as response, given the other $d-1$ variables in the graph. To extend this procedure for group structured data, consider $Y = (Y^{(1)},\ldots,Y^{(K)})'$ from K different groups, where $Y^{(k)}$ is a $n_k \times d$ matrix. Within each group, we assume observations to be independent and identically distributed as $Y^{(k)} \sim N_d(0,\Sigma_k)$.

To extend neighbourhood selection to multiple graphs, we propose to adopt a penalty term similar to the one used in the joint lasso by Dondelinger & Mukherjee (2018). Estimation is achieved minimizing

$$\widehat{\Theta}^i = \underset{(\theta^{i,1},\ldots,\theta^{i,K})}{\arg\min} \sum_{k=1}^{K} \left(\frac{1}{n_k}||Y_i^{(k)} - Y_{-i}^{(k)}\theta^{i,k}||_2^2 + \lambda||\theta^{i,k}||_1 + \gamma \sum_{k'>k} \tau_{k,k'}||\theta^{i,k} - \theta^{i,k'}||_1 \right), \tag{1}$$

where λ, γ and $\boldsymbol{\tau} = \{\tau_{k,k'} : k' > k\}$ are tuning parameters. The last term of Equation (1) allows exact equality between coefficient from different groups where $\boldsymbol{\tau}$ allows to weight differently each couple of groups. The vector $\boldsymbol{\tau}$ allows to attribute a specific shrinkage only on some pairs of parameters and is set to $\mathbf{1}$ in the rest of the paper.

Similarly to Meinshausen & Bühlmann (2006), we use Equation (1) nodewise. The neighborhood of node i for the kth group is then $\widehat{\mathrm{ne}}_i^{(k)} = \{j \in \{1,\ldots,d\} : \widehat{\theta}_j^{i,k} \neq 0\}$, while the selected edge set is given by $\widehat{E}^{(k)} = \{(i,j) : i \in \widehat{\mathrm{ne}}_j^{(k)} \wedge j \in \widehat{\mathrm{ne}}_i^{(k)}\}$. To obtain an estimate of the precision matrix for each group, we adopt a two-step approach. We first learn the edge set, and then we use constrained maximum likelihood method with given zero elements. Let $\mathcal{S}^+_{\widehat{E}^{(k)}} = \{\Omega : \Omega \succ 0 \wedge \omega_{ij} = 0, \forall (i,j) \notin \widehat{E}^{(k)}\}$ be the set of positive definite matrices with support defined by $\widehat{E}^{(k)}$. The precision matrix estimate is

$$\widehat{\Omega}^{(k)} = \underset{\Omega \in \mathcal{S}^+_{\widehat{E}^{(k)}}}{\arg\min} \left(\mathrm{tr}(S^{(k)}\Omega^{(k)}) - \log\det(\Omega^{(k)}) \right) \qquad \forall k \in \{1,\ldots,K\}.$$

This two-step procedure assures a positive definite estimate. However, using the same data for model selection and parameter estimation is known to lead

Table 1. *Monte Carlo summary of performance*

Method	EL	FL	FNR	FPR
Structure learning			0.2673	0.0823
Data Carving	3.0059	0.5516	0.3087	0.0767
Data Splitting	3.1110	0.5590	0.4080	0.0700

to not valid inference (Tian & Taylor, 2018). Thus, post-selection inference, such as data splitting or carving procedures, needs to be used.

Nodewise regression relies on the selection of tuning parameters λ and γ. We used a slightly modified version of StARS (Stability Approach to Regularization Selection) algorithm proposed by Liu *et al.,* (2010).

For a chosen b, $1 < b < n$, we draw N random subsamples $X_1,\dots,X_N$ from Y each of size b. Given a value of λ and γ, we apply nodewise joint estimation in each subsample. Let $\widehat{D}(\lambda,\gamma)$ be the maximum among groups of the average of instability for each edge across subsamples. We use a Bayesian optmization technique based on Gaussian Processes (Snoek *et al.,* 2012) to obtain optimal values of tuning parameters, minimizing the instability measure $|\widehat{D}(\lambda,\gamma)-\beta|$ with β to be set. The performance of the proposed procedure is illustrated in the next section.

3 Monte Carlo simulations

This Monte Carlo study reports a simple setting with only two groups ($K = 2$). We generate a random graph structure with $d = 15$ nodes and the corresponding precision matrix $\Omega^{(1)}$as described in Danaher *et al.,* (2014). To generate $\Omega^{(2)}$, we randomly change some entries of $\Omega^{(1)}$ adding edges, removing them or varying partial correlation coefficients.

We simulate 50 datasets of dimension $n = 150$ from $Y = (Y^{(1)},Y^{(2)})'$ where $Y^{(i)} \sim N(0,\Sigma_i)$, Σ_i being the inverse of $\Omega^{(i)}$ and $i = 1,2$. Then we use tuning parameter selection with $\beta = 0.1$ and $N = 30$ to optimize the nodewise selection algorithm in three different situations: structure estimation only, precision matrix estimation using data carving ($p = 0.9$), precision matrix estimation using data splitting ($p = 0.5$), where p is the proportion of data using to estimate structure. We evaluate the edge selection performances using false negative rate (FNR) and false positive rate (FPR), while the estimate of the precision matrices is compared using entropy loss (EL) and Frobenius loss (FL). Simulation results are summarized in Table 1.

To summarize, if one is mainly interested in structure learning, our proposal a slightly better performance, in comparison with the other existing procedures, not reported here, mostly because of fewer false negative errors. If the aim is to estimate the precision matrix through a two-step procedure, it seems that data carving is a better option than data splitting.

References

DANAHER, PATRICK, WANG, PEI, & WITTEN, DANIELA. 2014. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **76**(2), 373–397.

DONDELINGER, FRANK, & MUKHERJEE, SACH. 2018. The joint lasso: high-dimensional regression for group structured data. *Biostatistics (Oxford, England)*, **21**, 219–235.

FRIEDMAN, JEROME, HASTIE, TREVOR, & TIBSHIRANI, ROBERT. 2008. Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics*, **9**(3), 432–441.

GUO, JIAN, MICHAELIDIS, GEORGE, ZHU, JI, *et al.* . 2011. Joint estimation of multiple graphical models. *Biometrika*, **98**(1), 1–15.

LAURITZEN, STEFFEN L. 1996. *Graphical Models.* Oxford University Press.

LIU, HAN, ROEDER, KATHRYN, & WASSERMAN, LARRY. 2010. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *Pages 1432–1440 of:* LAFFERTY, J. D., WILLIAMS, C. K. I., SHAWE-TAYLOR, J., ZEMEL, R. S., & CULOTTA, A. (eds), *Advances in Neural Information Processing Systems 23.* Curran Associates, Inc.

MA, J., & MICHAILIDIS, GEORGE. 2016. Joint structural estimation of multiple graphical models. **17**(09), 1–44.

MEINSHAUSEN, NICOLAI, & BÜHLMANN, PETER. 2006. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, **34**(3), 1436–1462.

SNOEK, JASPER, LAROCHELLE, HUGO, & ADAMS, RYAN P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. *Page 2951–2959 of: Advances in Neural Information Processing Systems.* NIPS 2012, vol. 25. Curran Associates Inc.

TIAN, XIAOYIANG, & TAYLOR, JONATHAN. 2018. Selective inference with a randomized response. *The Annals of Statistics*, **46**(2), 679–710.

FUNCTIONAL CLUSTER ANALYSIS OF HDI EVOLUTION IN EUROPEAN COUNTRIES

Francesca Fortuna[1], Alessia Naccarato[1] and Silvia Terzi[1]

[1] Department of Economics, Roma Tre University,
(e-mail: francesca.fortuna@uniroma3.it,
alessia.naccarato@uniroma3.it, silvia.terzi@uniroma3.it)

ABSTRACT: The contribution aims to study the evolutionary aspects of a well-being indicator in European countries. To this end, an evolutionary indicator is proposed by considering the indicator as a function and integrating the information provided by the well-being curve with its temporal dynamic reflected by the first derivative. Then, functional cluster analysis is considered to derive groups of geographical areas that account not only for the indicator's level, but also for its evolution.

KEYWORDS: Human Development Index, FDA, functional clustering.

1 Introduction

Well-being indicators are commonly used to support decision making and to assess the performance of countries. However, well-being indicators are generally considered from a static point of view, disregarding their temporal dynamics. Our aim is to exploit the evolutionary aspect of a well-being indicator. To this end, temporal sequences of well-being indicators are analyzed from a functional point of view. Thus, indicators are considered as functions rather than scalar vectors. This is a novel perspective in well-being processing, which allows to introduce new analytical tools, such as derivatives. Since the latter quantify a function's behavior in an evolutionary perspective, we suggest to integrate the information provided by the well-being curve with the information concerning its first order derivative. Specifically, we focus on the problem of clustering well-being curves using the functional k-means algorithm under different distances in order to identify specific common patterns among the countries. The procedure is applied to a real data set regarding the annual time series of the Human Development Index (HDI) for 44 European countries. We compare the clusters obtained by functional k-means algorithm with the clusters derived in a non-functional environment via a k-means algorithm applied to raw data of an evolutionary integrated HDI, say $EHDI$. The latter is defined as $EHDI = HDI[1 + f'(x)]$ and integrates HDI with the information provided by its first derivative, $f'(x)$, in order to discount for a decreasing or increasing evolution of the HDI.

2 Functional distances

To identify common patterns among the HDI curves, the functional k-means algorithm (Tarpey & Kinateder, 2003) is considered using the following distances:

$$d_0\Big(f_i(t), f_j(t)\Big) = \int_T \Big(f_i(t) - f_j(t)\Big)^2 dt, \quad \forall i \neq j;\ i, j = 1, 2, ..., n; \qquad (1)$$

where $f_i(t) = \sum_{k=1}^{K} a_{ik}\phi_k(t)$, is expanded in terms of K cubic B-splines functions (Ramsay & Silverman, 2005);

$$d_{0+1}\Big(f_i(t), f_j(t)\Big) = \sqrt{\int_T \Big(f_i(t) - f_j(t)\Big)^2 dt + \int_T \Big(f_i'(t) - f_j'(t)\Big)^2 dt}, \qquad (2)$$

where $f_i'(t)$ denotes the smoothing estimate of the first derivative of $f_i(t)$. The distances in (1) and (2) are the norm and the semi-norm in the Hilbert space, respectively. The semi-norm d_{0+1} accounts both for the level of the well-being curve and for its evolutionary dynamic.

3 Application

The prosed method is applied to the annual time series of the HDI indexes from 2000 to 2019 for 44 European countries. Functional cluster analysis is applied to HDI data, converted into a sample of smooth functions using $K = 5$ cubic B-splines basis, chosen by cross validation (left-hand side of Figure 2). Distances in (1) and (2) are considered choosing three clusters, corresponding to high, medium and low human development countries. The clustering results are the same, except for France and Italy which, by means of d_{0+1}, are assigned to the high human development cluster rather than the medium one. The clustering algorithm is also applied to the smoothed version of $EHDI$ using d_0 as a distance. The resulting configuration is the same as that obtained with d_{0+1} on the functional HDI. The high development group is characterised by the countries of Western and Northern Europe: Austria, Belgium, Germany, Liechtenstein, Luxembourg, Netherlands, Switzerland, United Kingdom, Denmark, Finland, Iceland, Ireland, Norway and Sweden. This group also includes Slovenia, the only country in South-Eastern Europe. The medium development group includes mainly Southern and Eastern European countries, plus France and 3 Northern countries: Estonia, Latvia and Lithuania. The low human development group is characterised by the countries of Eastern and South-Eastern Europe: Armenia, Azerbaijan, Georgia, Ukraine, Albania, Bosnia and Herzegovina, Republic of Moldova, North Macedonia and Turkey. Table 1 displays the cluster sizes and the average silhouette value for each scenario. To provide an insight on the role of the first derivative of the curve, the results are compared with those obtained in a non-functional framework. Specifically, the k-means algorithm is applied on the raw *EHDI*. Figure 1 shows the centroids obtained with the k-means algorithm and different distances. We remark that the right-hand side of Figure 1 shows the sequences of raw *EHDI* across the years, not the smoothed functions. Comparing the clusters obtained in the functional and the non-functional contexts, only six countries are assigned differently. Specifically, the non-functional algorithm downgrades Bulgaria, Romania, Russian Federation and Serbia, classifying them as low development countries. Indeed, as we can see from the right-hand side of Figure 2, Bulgaria presents a first

Table 1. *Clustering Results from k-means algorithm with different distances.*

Cluster	d_0	d_{0+1}	Raw *EHDI*
High	15	17	16
Medium	20	18	15
Low	9	9	13
Mean Sil.	0.5	0.5	0.6

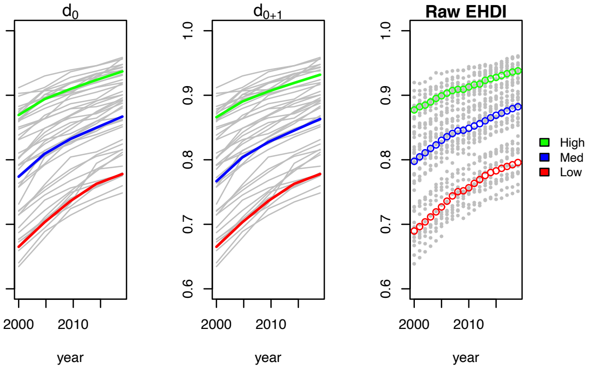


Figure 1. *Cluster centroids: k-means with different distances.*

derivative with a decreasing trend, but with high values especially in the first part of the domain. Romania has a first derivative with a fluctuating trend: strongly increasing until 2005, decreasing until 2015, increasing subsequently. Russian Federation and Serbia have flat first derivatives but with high values. Viceversa the non-functional algorithm upgrades France and Italy, including them in the high development cluster. However it is a partial upgrade, only with respect to the classification provided by d_0 (the d_{0+1} distance assigned both these countries to the high development cluster).

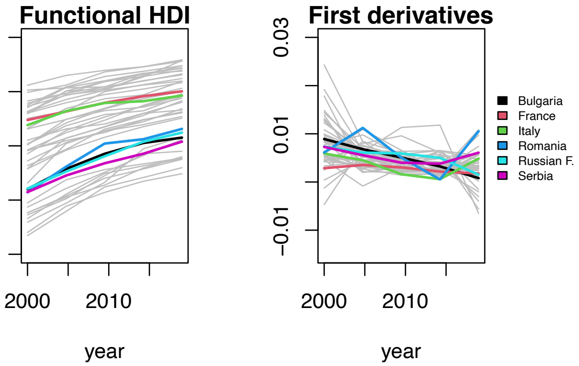


Figure 2. *Functional HDI and First derivatives of the European countries.*

4 Conclusions

FDA is a useful methodological framework for the analysis of well-being indicators as it allows to evaluate their evolution with additional tools. Specifically, the joint analysis of the level of well-being curves and their first derivatives can provide useful insight in countries' well-being improvement or worsening. In our application, the range of the first derivatives is very limited, thus the additional information concerning the indicator's trend has little effect on countries' classification.

References

RAMSAY, J. O., & SILVERMAN, B. W. 2005. *Functional Data Analysis*. New York: Springer.

TARPEY, T., & KINATEDER, K. K. J. 2003. Clustering functional data. *Journal of Classification.*, **20**, 93–114.

ESTIMATING BAYESIAN MIXTURES OF FINITE MIXTURES WITH TELESCOPING SAMPLING *

Sylvia Frühwirth-Schnatter[1], Bettina Grün[1] and Gertraud Malsiner-Walli[1]

[1] Institute for Statistics and Mathematics, WU (Vienna University of Economics and Business), (e-mail: Sylvia.Fruehwirth-Schnatter@wu.ac.at, Bettina.Gruen@wu.ac.at, Gertraud.Malsiner-Walli@wu.ac.at)

ABSTRACT: Finite mixtures result from convex combinations of arbitrary statistical models as components and thus allow to extend any statistical model. Specifying a prior on the number of components is natural in a Bayesian framework and results in a mixture of finite mixtures (MFM) model. Several sampling schemes for Bayesian estimation have been proposed, with most being only applicable to a specific component distribution or requiring extensive tuning. The recently proposed telescoping sampler extends the Markov chain Monte Carlo sampling scheme with data augmentation of the finite mixture model by sampling also from the posterior of the number of components. We will demonstrate the general applicability and performance of the telescoping sampler on mixture models with different component models.

KEYWORDS: Bayesian estimation, finite mixture model, Markov chain Monte Carlo sampling, transdimensional sampling.

1 Bayesian MFMs & Telescoping Sampling

Mixture models are a versatile model class which can be used for model-based clustering as well as density estimation. A finite mixture model is given by a convex combination of several distributions or models and hence any statistical model may be embedded within the mixture framework. In the following only mixture models with fixed component weights are considered, i.e., where the component sizes do not depend on any covariates, while parameteric distributions as well as regression models are covered for the components.

The application of finite mixture models in practice usually requires for estimation to fix the number of components a-priori and to then perform model selection to decide on a suitable number of components. In particular in Bayesian analysis, such a model selection step is complicated by the fact that

*The authors gratefully acknowledge support from the *Austrian Science Fund (FWF)*: P28740; and through the *WU Projects* grant scheme: IA-27001574.

the number of components in a finite mixture model does not necessarily correspond to the number of filled components given the observed data. As an alternative, Escobar & West (1995) consider Dirichlet process mixtures (DPMs) where the number of components is infinite and only inference on the number of filled components is performed. Malsiner-Walli *et al.* (2016) suggest sparse finite mixtures (SFMs), where the parameter for the prior on the weights is selected to imply that the number of components will be higher than the number of filled components, in this way allowing for posterior inference of the number of filled components.

Richardson & Green (1997) propose to use the specification of a mixture of finite mixtures (MFM) model where a prior on the number of components is included, to obtain posterior estimates for the number of components, the number of filled components and the parameter estimates. Richardson & Green (1997) also indicate the estimation of this model class using a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. Alternative approaches to perform Bayesian inference of the MFM model were considered by Stephens (2000) who suggests to use a Markov birth-death process and Miller & Harrison (2018) who re-use Chinese restaurant process (CRP) methods proposed for DPMs to sample the partitions.

Frühwirth-Schnatter *et al.* (2020) develop the *telescoping sampler* to perform inference for any kind of MFM where arbitrary component distributions or models as well as hierarchical priors may be included without complicating the sampling. They build on the data augmentation scheme suggested for finite mixtures by Diebolt & Robert (1994) and include a sampling step for the number of components. This implies that the telescoping sampler is straightforward to implement given that a MCMC sampling scheme for the components is available.

2 Empirical Demonstrations

Frühwirth-Schnatter *et al.* (2020) already present the application of the telescoping sampler on mixtures of univariate Gaussian distributions, which allows them to benchmark their sampler against RJMCMC and CRP sampling, on mixtures of multivariate Gaussian distributions and on latent class analysis models applied to multivariate categorical data. Following Frühwirth-Schnatter & Malsiner-Walli (2019), it is straightforward to investigate the use of the telescoping sampler also for mixtures of Poisson distributions, mixtures of generalized linear models and mixtures of skew normal and skew-t distributions and compare the performance to DPMs and SFMs. Section 2.1 presents

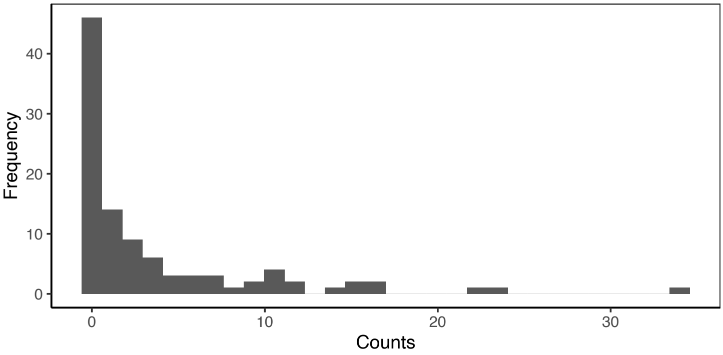


Figure 1. *Eye tracking data. Histogram of the observations.*

the results obtained with telescoping sampling for Poisson mixtures using the eye tracking data.

2.1 Poisson Mixtures: Eye Tracking Data

Figure 1 visualizes the count data on eye tracking anomalies in 101 schizophrenic patients studied among others by Frühwirth-Schnatter & Malsiner-Walli (2019). The overdispersion and the excess number of zeros present in the data set are clearly visible in the plot showing the frequency of counts. The MFM is fitted using the same hierarchical specification for the component means λ_k as used in Frühwirth-Schnatter & Malsiner-Walli (2019) when fitting SFMs and DPMs: $\lambda_k \sim \mathcal{G}(a_0, b_0)$ and $b_0 \sim \mathcal{G}(g_0, G_0)$, where the parameters of the gamma distribution are given by $a_0 = 0.1$, $g_0 = 0.5$ and $G_0 = g_0\bar{y}/a_0$ with $\bar{y}$ the mean of the observations. In addition the dynamic specification for the Dirichlet weights is used, i.e., the weights are a-priori drawn from a K-dimensional symmetric Dirichlet distribution $\text{Dir}_K(\alpha/K)$, with α having a hyperprior F-distribution $F(6,3)$. Four different priors on the number of components are considered: the discrete uniform prior on $\{1,2,\dots,150\}$, the shifted beta-negative-binomial priors $\text{BNB}(1,1,1)$ and $\text{BNB}(1,4,3)$ and the geometric prior $\text{Geo}(0.1)$. These priors vary in their prior mean, their regularization of additional components and the mass assigned to the tail.

The posterior distributions of the number of components K and the number of filled components K_+ obtained with telescoping sampling are summarized in Table 1. The influence of the prior on K is particularly noticeably for the posterior of K and a much less pronounced influence on the posterior of K_+ is discernible. Clearly results for all priors on K indicate that heterogeneity is

Table 1. *Posterior inference for K_+ and K. The posteriors are summarized by their modes, followed by the 1st, 2nd and 3rd quartiles.*

$p(K)$	$p(K_+\|\mathbf{y})$		$p(K\|\mathbf{y})$	
$\mathcal{U}(1,150)$	13	[12, 16, 21]	119	[50, 83, 118]
$\text{BNB}(1,1,1)$	10	[9, 12, 16]	11	[12, 21, 41]
$\text{Geo}(0.1)$	9	[9, 11, 15]	13	[12, 17, 25]
$\text{BNB}(1,4,3)$	6	[6, 8, 10]	7	[7, 9, 13]

present and a mixture with several components is needed to approximate the distribution of counts.

References

Diebolt, Jean, & Robert, Christian P. 1994. Estimation of Finite Mixture Distributions Through Bayesian Sampling. *Journal of the Royal Statistical Society B*, **56**(2), 363–375.

Escobar, Michael D., & West, Mike. 1995. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, **90**(430), 577–588.

Frühwirth-Schnatter, S., & Malsiner-Walli, G. 2019. From Here to Infinity: Sparse Finite Versus Dirichlet Process Mixtures in Model-Based Clustering. *Advances in Data Analysis and Classification*, **13**(1), 33–64.

Frühwirth-Schnatter, Sylvia, Malsiner-Walli, Gertraud, & Grün, Bettina. 2020. *Generalized Mixtures of Finite Mixtures and Telescoping Sampling*. arXiv:2005.09918 [stat.ME].

Malsiner-Walli, Gertraud, Frühwirth-Schnatter, Sylvia, & Grün, Bettina. 2016. Model-Based Clustering Based on Sparse Finite Gaussian Mixtures. *Statistics and Computing*, **26**(1), 303–324.

Miller, Jeffrey W, & Harrison, Matthew T. 2018. Mixture Models with a Prior on the Number of Components. *Journal of the American Statistical Association*, **113**(521), 340–356.

Richardson, Sylvia, & Green, Peter J. 1997. On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society B*, **59**(4), 731–792.

Stephens, Matthew. 2000. Bayesian Analysis of Mixture Models with an Unknown Number of Components – An Alternative to Reversible Jump Methods. *The Annals of Statistics*, **28**(1), 40–74.

A Bayesian framework for structural learning of mixed graphical models

Chiara Galimberti [1], Federico Castelletti [2] and Stefano Peluso [3]

[1] Department of Economics, Managment and Statistics, Università degli Studi di Milano-Bicocca, (e-mail: c.galimberti19@campus.unimib.it)

[2] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, (e-mail: federico.castelletti@unicatt.it)

[3] Department of Statistics and Quantitative Methods, Università degli Studi di Milano-Bicocca, (e-mail: stefano.peluso@unimib.it)

Abstract: Graphical models provide an effective tool to represent conditional independences among variables. While this class of models has been extensively studied in the Gaussian and categorical settings separately, literature which combines the two types of variables is narrow. However, mixed data are extremely diffuse in many applications where both continuous and categorical measurements are available. In this paper we propose a Bayesian framework for the analysis of mixed data. Specifically, we specifiy a likelihood function for *n* observations following a conditional Gaussian distribution, and assign suitable priors for the model parameters. Our end-result is a closed form espression for the marginal data distribution. The latter provides a primary input for the computation of the marginal likelihood under graph (independence) constraints and the development of an MCMC strategy for graph structural learning.

Keywords: conditional gaussian distribution, directed acyclic graph, graphical models, marginal likelihood, mixed variables

1 Introduction

Graphical models are particularly effective to represent conditional dependency structures in multivariate distributions (Lauritzen, 1996). In particular, inferring the unknown graph generating model from the data is possible using structural learning methodologies. In this contribution we focus on directed acyclic graphs (DAGs) where conditional dependencies between variables are represented through parent-child relationships.

Several works for structural learning of graphical models given continuous (Gaussian) or discrete/categorical data (Ising model) are available in the literature.

However, literature oriented to DAG structural learning given mixed data is extremely narrow. In the Bayesian framework, a unified approach which jointly models categorical and continuous data is also still lacking. The scope of this study is to develop a Bayesian methodology for DAG learning in the presence of mixed observations. Our ultimate goal is the development of an MCMC algorithm, along the lines of Castelletti *et al.*, 2018 and Castelletti & Peluso, 2021 for, respectively, the Gaussian and categorical case. In the next sections we illustrate some preliminary results relative to general Bayesian models for mixed variables together with some possible extensions to DAG-constrained models.

2 Model development

Our starting point is represented by the notion of Conditional Gaussian (CG) distribution introduced by Lauritzen & Wermuth, 1989. Let V be a finite set of nodes indexing a collection random variables $Z = (Z_1, \dots, Z_{|V|})^T$, which comprises both discrete and continuous quantities indexed by $\Delta \cup \Gamma = V$ respectively. The authors defined a general class of probability distributions of the form

$$f(z) = f(s,y) = \exp\left\{ g(s) + h(s)^T y - \frac{1}{2} y^T K(s) y \right\} \tag{1}$$

where s and y correspond to the level assumed by the categorical and continuous variables respectively. A probability distribution of the form (1) has CG-distribution if and only if $Z_\Gamma | Z_\Delta = s \sim \mathcal{N}_q(K(s)^{-1}h(s), K(s)^{-1})$ and the marginal distribution of the discrete variables is

$$\theta(s) = (2\pi)^{-\frac{q}{2}} |K(s)|^{-\frac{1}{2}} \exp\left\{ g(s) + \frac{1}{2} h(s)^T K(s)^{-1} h(s) \right\}, \tag{2}$$

for each level s assumed by Z_Δ. Moreover, if $K(s) = K$ the distribution is called *homogeneous*. An alternative representation of a CG-distribution, hereinafter adopted, is given in terms of moment-characteristics parameters $(\boldsymbol{\theta}, \xi, \Sigma)$.

Specifically, let $(X_1, \dots, X_p)$ be p categorical variables, $(Y_1, \dots, Y_p)$ q continuous variables. Let also I be the space of all possible configurations of the p categorical variables and $\boldsymbol{\theta} = \{\theta(s), s \in I\})$ where $\theta(s) = \Pr(X_1 = s_1, \dots X_p = s_p)$ is the probability to observe configuration $s = (s_1, \dots, s_p)$. Under the CG assumption we can write for each $s \in I$

$$Y_1(s), \dots, Y_q(s) \mid \boldsymbol{\mu}(s), \boldsymbol{\Omega} \sim \mathcal{N}_q(\boldsymbol{\mu}(s), \boldsymbol{\Omega}^{-1}). \tag{3}$$

We now consider a collection of n independent observations $\boldsymbol{x}_i = (x_{i,1},\ldots,x_{i,p})^T$, $\boldsymbol{y}_i = (y_{i,1},\ldots,y_{i,q})^T$, $i = 1,\ldots,n$. Categorical data $\{\boldsymbol{x}_i, i = 1,\ldots,n\}$, can be equivalently represented as a contingency table of counts $\boldsymbol{N}$ with elements $n(s) \in \boldsymbol{N}$ satisfying $\sum_{s\in I} n(s) = n$. Following Frydenberg & Lauritzen, 1989, the likelihood function can be written as

$$f(\boldsymbol{N},\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n\,|\,\boldsymbol{\theta},\{\boldsymbol{\mu}(s)\}_{s\in I},\boldsymbol{\Omega}) = \prod_{s\in I}\theta(s)^{n(s)}\prod_{s\in I}\prod_{i\in d(s)}\phi(\boldsymbol{y}_i\,|\,\boldsymbol{\mu}(s),\boldsymbol{\Omega}^{-1})$$

$$\propto \prod_{s\in I}\theta(s)^{n(s)}\prod_{s\in I}\prod_{i\in d(s)}|\boldsymbol{\Omega}|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i-\boldsymbol{\mu}(s))^T\boldsymbol{\Omega}(\boldsymbol{y}_i-\boldsymbol{\mu}(s))\right\}, \quad (4)$$

where $d(s)$ is the set of observations among $i = 1,\ldots,n$ with observed configuration s and ϕ is the Gaussian density. We then proceed by assigning the following prior distributions

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{A}), \quad \boldsymbol{\mu}(s)\mid\boldsymbol{\Omega} \sim \mathcal{N}_q(\boldsymbol{m}(s),(a_\mu\boldsymbol{\Omega})^{-1}), \quad \boldsymbol{\Omega} \sim \mathcal{W}_q(a_\Omega,\boldsymbol{U}), \quad (5)$$

where in particular $\mathcal{W}_q(a_\Omega,\boldsymbol{U})$ denotes a Wishart distribution having expectation $a_\Omega\boldsymbol{U}^{-1}$, $a_\Omega > q-1$ and $\boldsymbol{U}$ is a s.p.d. matrix. Under prior parameter independence, the posterior distribution is written after some calculations as

$$p(\boldsymbol{\theta},\{\boldsymbol{\mu}(s)\}_{s\in I},\boldsymbol{\Omega}\,|\,\boldsymbol{N},\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n) \propto \prod_{s\in I}\theta(s)^{a(s)+n(s)-1}$$

$$\cdot\prod_{s\in I}\left\{|\boldsymbol{\Omega}|^{\frac{1}{2}}|\exp\left\{-\frac{1}{2}(n(s)+a_\mu)(\boldsymbol{\mu}(s)-\bar{\boldsymbol{m}}(s))^T\boldsymbol{\Omega}(\boldsymbol{\mu}(s)-\bar{\boldsymbol{m}}(s))\right\}\right.$$

$$\left.\cdot|\boldsymbol{\Omega}|^{\frac{a_\Omega+n-q-1}{2}}\exp\left\{-\frac{1}{2}\text{tr}[(\boldsymbol{U}+\boldsymbol{S}+\boldsymbol{S}_0)\boldsymbol{\Omega}]\right\}\right., \quad (6)$$

with $\boldsymbol{S} = \sum_{s\in I}\text{SSD}(s)$,

$$\begin{aligned}\bar{\boldsymbol{m}}(s) &= \frac{a_\mu}{a_\mu+n(s)}\boldsymbol{m}(s)+\frac{n(s)}{a_\mu+n(s)}\bar{\boldsymbol{y}}(s),\\ \boldsymbol{S}_0 &= \sum_{s\in I}\frac{a_\mu n(s)}{a_m u+n(s)}(\boldsymbol{m}(s)-\bar{\boldsymbol{y}}(s))(\boldsymbol{m}(s)-\bar{\boldsymbol{y}}(s))^T,\end{aligned}$$

where $\text{SSD}(s) = \sum_{i\in d(s)}\boldsymbol{e}_i\boldsymbol{e}_i^T$, $\boldsymbol{e}_i = (\boldsymbol{y}_i-\bar{\boldsymbol{y}}(s))$ and $\bar{\boldsymbol{y}}(s)$ is the $(q,1)$ vector with sample means of $(Y_1,\ldots,Y_q)$ relative to observations $i \in d(s)$. It follows that

$$\begin{aligned}\boldsymbol{\theta}\,|\,\boldsymbol{N} &\sim \text{Dirichlet}(\boldsymbol{A}+\boldsymbol{N})\\ \boldsymbol{\mu}(s)\mid\boldsymbol{N},\boldsymbol{Y},\boldsymbol{\Omega} &\sim \boldsymbol{N}_q(\bar{\boldsymbol{m}}(s),[(a_\mu+n(s))\boldsymbol{\Omega})]^{-1})\\ \boldsymbol{\Omega}\,|\,\boldsymbol{Y} &\sim \mathcal{W}_q(a_\Omega+n,\boldsymbol{U}+\boldsymbol{S}+\boldsymbol{S}_0),\end{aligned} \quad (7)$$

where $\boldsymbol{Y}$ denotes the (n,q) data matrix, row-binding of the $\boldsymbol{y}_i$'s. Because of conjugacy, the marginal data distribution

$$m(\boldsymbol{Y},\boldsymbol{N}) = \int f(\boldsymbol{N},\boldsymbol{Y}\,|\,\boldsymbol{\theta},\{\boldsymbol{\mu}(s)\}_{s\in I},\boldsymbol{\Omega})p(\boldsymbol{\theta})\prod_{s\in I}p(\boldsymbol{\mu}(s))p(\boldsymbol{\Omega})\,d\boldsymbol{\theta}\prod_{s\in I}d\boldsymbol{\mu}(s)\,d\boldsymbol{\Omega},$$

can be computed as the ratio of prior/posterior normalizing constants. Care will be posed on score-equivalence (same marginal likelihood) for Markov equivalent DAGs; see also Peluso & Consonni, 2020.

3 Conclusion and further steps

We obtained a closed-form expression for the marginal likelihood of a complete (unconstrained) Bayesian model given mixed data. The subsequent step requires the computation of the marginal likelihood for a subset of (mixed) variables, e.g. $\{X_j | j \in C \subseteq \{1,\dots,p\}\} \cup \{Y_k | k \in D \subseteq \{1,\dots,q\}\}$. To this end we will adopt the procedure for prior parameter elicitation introduced by Geiger & Heckerman, 2002. The computation of the marginal likelihood of a given DAG will be at the basis of an MCMC algorithm for DAG structural learning.

References

CASTELLETTI, F., CONSONNI, G., DELLA VEDOVA, M., & PELUSO, S. 2018. Learning Markov equivalence classes of directed acyclic graphs: an objective Bayes approach. *Bayesian Analysis*, **13**(4), 1235–1260.

CASTELLETTI, FEDERICO, & PELUSO, STEFANO. 2021. Equivalence class selection of categorical graphical models. *arXiv preprint arXiv:2102.06437*.

DEGROOT, M. 2004. *Optimal statistical decisions*. John Wiley and Sons.

FRYDENBERG, M., & LAURITZEN, S. 1989. Decomposition of Maximum Likelihood in Mixed Graphical Interaction Models. *Biometrika*, **76**(3), 539–555.

GEIGER, D., & HECKERMAN, D. 2002. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, **30**(5), 1412–1440.

LAURITZEN, S. 1996. *Graphical models*. Oxford Press.

LAURITZEN, S., & WERMUTH, N. 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, **17**(1), 31–57.

PELUSO, STEFANO, & CONSONNI, GUIDO. 2020. Compatible priors for model selection of high-dimensional Gaussian DAGs. *Electronic Journal of Statistics*, **14**(2), 4110–4132.

MEASUREMENT ERROR MODELS ON SPATIAL NETWORK LATTICES: CAR CRASHES IN LEEDS

Andrea Gilardi[1], Riccardo Borgoni[1], Luca Presicce[1] and Jorge Mateu[2]

[1] Department of Economics, Management and Statistics, University of Milano - Bicocca, Milan, Italy (e-mail: andrea.gilardi@unimib.it)

[2] Department of Mathematics, Universitat Jaume I, Castellón, Spain

ABSTRACT: Road casualties represent the leading cause of death among young people worldwide, especially in poor and developing countries. This paper introduces a Bayesian hierarchical model to analyse car accidents on a network lattice that takes into account measurement error in spatial covariates. We exemplified the proposed approach analysing all car crashes that occurred in the road network of Leeds (UK) from 2011 to 2019. Our results show that omitting measurement error considerably worsens the fit of the model and attenuates the effects of spatial covariates.

KEYWORDS: CAR, Linear Networks, Network Lattices, Spatial Measurement Error

1 Introduction

As reported by World Health Organisation in 2018, car crashes are responsible for more than 1.35 million casualties each year, representing the leading cause of death among people aged 5-29 years, particularly those living in developing countries. In the last years, several authors developed sophisticated statistical models to analyse the spatial distribution of car crashes at the areal level (e.g. cities or census wards) and help the local authorities define safety measures.

Nevertheless, road casualties represent a classic example of events occurring on a linear network. This paper presents a Bayesian hierarchical model for car crashes developed on a network lattice that takes into account measurement error (ME) in spatial covariates. In particular, a Conditional Auto-Regressive (CAR) prior is introduced to adjust for ME in estimating road traffic volumes within the classical ME model paradigm. The Integrated Nested Laplace Approximation (INLA) framework is adopted for inference. This approach was found particularly convenient for large networks, as the one considered in this paper, while MCMC techniques may be challenging and time-consuming (Muff *et al.*, 2015).

2 Road network and car crashes

The statistical analysis introduced in Section 3 requires a specific data structure that was obtained after several preprocessing steps briefly described hereafter.

The *road network* was built using data extracted from Open Street Map (OSM), an online database that provides open-access geographic rich-attribute data worldwide. We downloaded the street segments that pertain to the most important* roads of Leeds and created a matrix of segments representing the elementary units of the statistical model.

A street network can also be seen as a graph object whose edges represent the road network segments and whose vertices are placed at junctions, intersections, and boundary points (Barthélemy, 2011). We took advantage of the graph representation to contract the street network removing redundant nodes, edges loops, duplicated roads, and several isolated clusters of segments that may create numerical problems (Gilardi *et al.*, 2020). Furthermore, we calculated the weighted edge betweenness centrality, a graph measure correlated with the spatial distribution of commercial activities, which is usually adopted to analyse congestion problems as a proxy for urban traffic (Barthélemy, 2011). Finally, we derived the edges' adjacency matrix, an essential ingredient for the CAR prior used below.

We analysed all car crashes involving personal injuries that occurred in the city of Leeds from 2011 to 2019 and became known to the Police Forces within thirty days from their occurrence. First, we downloaded the data from UK's official road traffic casualty database. Then, we excluded those car crashes that occurred farther than fifty metres from the closest road segment, and, finally, we projected the events to the nearest point of the network and counted the occurrences for each segment. The final sample included 15826 events distributed over 4253 segments covering approximately 1170 km.

3 Statistical methods

Let $y_i,\ i = 1,\dots,n$ represent the number of car crashes that occurred on the ith road segment. Following a classical hypothesis in the road safety literature, we assume that $y_i|\lambda_i \sim \text{Poisson}(e_i\lambda_i)$, where λ_i represents the car crashes rate and e_i is an exposure parameter equal to the geographical length of each segment.

*More precisely, we selected only those segments whose classification range from *Autostrada* (i.e. *Motorway*) to *Strada Comunale* (i.e. *Tertiary Road*).

In the first level of the hierarchy, we define a log-linear structure on λ_i, i.e.

$$\log(\lambda_i) = \beta_0 + \beta_z z_i + \beta_x x_i + \theta_i + \phi_i; \; i = 1, \ldots, n, \tag{1}$$

where β_0 denotes the intercept, z_i is an error free covariate representing the road-type of each segment, x_i is an unobservable error prone covariate representing the traffic volumes, while β_x and β_z are the corresponding coefficients. Finally, θ_i and ϕ_i denote spatially structured and unstructured random effects that are modelled using a reparametrisation and a network re-adaptation of Besag-York-Mollié (BYM) prior (Riebler *et al.*, 2016, Gilardi *et al.*, 2020).

The classical spatial ME model assumes that x_i can be observed only via a proxy, say w_i, such that

$$w_i = x_i + u_i + \varphi_i; \; i = 1, \ldots, n.$$

The terms u_i and φ_i represent the ME and denote, respectively, spatially structured and unstructured random effects that are also modelled using the BYM prior. In particular, parameter φ_i adds a spatial smoothing effect to the unobserved covariate x_i. In this paper, we assume that the edge betweenness centrality measure can approximate the unobservable traffic volumes.

At the second stage of the hierarchy, we specified an exposure model that relates x_i with the error-free predictor:

$$x_i = \alpha_0 + \alpha_z z_i + \varepsilon_i; \; i = 1, \ldots, n. \tag{2}$$

The parameter α_0 denotes the intercept, α_z is the coefficient of the error-free covariate, and ε_i is a normally distributed error component. Furthermore, we assigned independent $N(0, 10^3)$ priors to β_0, β_z, α_0, and α_z, i.e. the intercepts and the coefficients assigned to z_i in equations (1) and (2).

The third level completes the specification of the hierarchical model eliciting a $\mathrm{N}(0, 100)$ prior for β_x, i.e. the coefficient of the error-prone covariate, a $\mathrm{Gamma}(1, 5\text{e-}05)$ prior on the precision of u_i and φ_i, and Penalised Complexity priors for the parameters of BYM's re-adaptation (Simpson *et al.*, 2017).

4 Results and conclusions

We estimated the statistical model described in Section 3 using INLA methodology and compared the results with two simpler models: the first one completely ignores ME, while the second one adopts a classical ME without spatial smoothing effects. We found that omitting ME greatly attenuates the importance of traffic volumes, and excluding the spatial smoothing terms worsens

	No ME	ME	Spat. ME
β_x	0.01	1.064	2.95
β_0	-5.307	-9.90	-15.441
β_{primary}	0.61	0.56	0.40
$\beta_{\text{secondary}}$	0.57	0.68	1.05
DIC		33126	30466

Table 1. *Summary of DIC, posterior means of fixed effects, and error-prone covariate.*

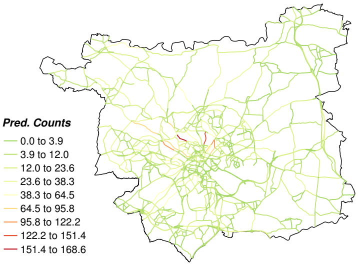


Figure 1. *Map displaying the posterior means of car crashes counts.*

the fit of the model. Motorways were found less prone to car crashes than the other road types, while the posterior distributions of fixed effects and common hyperparameters were found stable among the three models. We report in Table 1 a short summary of fixed effects' posterior means, while Figure 1 displays the posterior means of predicted counts. We can notice that it highlights a few road segments close to the city centre that would require a more detailed statistical analysis.

References

BARTHÉLEMY, MARC. 2011. Spatial networks. *Physics Reports*, **499**(1-3), 1–101.

GILARDI, ANDREA, MATEU, JORGE, BORGONI, RICCARDO, & LOVELACE, ROBIN. 2020. Multivariate hierarchical analysis of car crashes data considering a spatial network lattice. *arXiv preprint arXiv:2011.12595*.

MUFF, STEFANIE, RIEBLER, ANDREA, HELD, LEONHARD, RUE, HÅVARD, & SANER, PHILIPPE. 2015. Bayesian analysis of measurement error models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, 231–252.

RIEBLER, ANDREA, SØRBYE, SIGRUNN H, SIMPSON, DANIEL, & RUE, HÅVARD. 2016. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, **25**(4), 1145–1165.

SIMPSON, DANIEL, RUE, HÅVARD, RIEBLER, ANDREA, MARTINS, THIAGO G, & SØRBYE, SIGRUNN H. 2017. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 1–28.

THE L^p DATA DEPTH AND ITS APPLICATION TO MULTIVARIATE PROCESS CONTROL CHARTS

Carmela Iorio [1], Giuseppe Pandolfo[1], Michele Staiano[1] , Massimo Aria[2] and Roberta Siciliano[1]

[1] Department of Industrial Engineering, University of Naples Federico II, Italy, (e-mail: carmela.iorio@unina, giuseppe.pandolfo@unina.it, michele.staiano@unina.it, roberta@unina.it)

[2] Department of Economics and Statistics, University of Naples Federico II, Italy, (e-mail:massimo.aria@unina.it)

ABSTRACT: Control charts are used to identify non-random behaviours of a manufacturing process by monitoring changes in the distribution of the quality characteristics of the tested product. Process monitoring of related variables is usually referred to as a multivariate quality control problem. In many applications there is not enough information to justify the assumption of a specific form for the underlying process distribution. Thus, a non-parametric approach is a valid tool in a quality control process. Among possible non-parametric statistical techniques, data depth functions are gaining increasing interest in multivariate quality control. The aim of this work is to investigate the behaviour of a non-parametric approach based on the notion of the L^p depth in the statistical process control.

KEYWORDS: Non-parametric statistics, Q-charts, Data depth.

1 Introduction

Nowadays, industries collect a large amount of data on more than one variable. Hence in a quality control process there is more than one quality variable to be monitored simultaneously. A traditional control chart monitoring a single variable is not useful for detecting the overall quality of a process, as it is determined by the interaction of several related variables (Liu, 1995, Idris *et al.*, 2019). For this reason, multivariate analysis is becoming increasingly important within the statistical process control approaches (Woodall & Montgomery, 1999). Multivariate control charts are needed when dealing with more than one quality variable as overcome the drawback of obtaining incorrect control limits when dealing with related variables. As a matter of fact, the multivariate procedure takes into account the association between the components of a multivariate process.

Multivariate quality control studies was first conducted by Hotteling, 1947. For a more detailed description please refer to Montgomery, 2007.

Woodall, 2000 distinguishes the techniques of the control chart processing in Phase I (also called retrospective or preliminary phase) and in Phase II. Phase I uses charts with the purpose of defining whether a process is statistically under control when the first group are processed. In Phase II, the charts are used to check if the process is in control when future subgroups were being processed. In this last phase, it is assumed that the distribution of the process is known and most of the classical applications require the hypothesis that the process under consideration follows a multivariate normal distribution. However, in most industrial applications the distribution of a parametric multivariate control chart is difficult to estimate for processes with multiple quality characteristics. As such, all observations are considered as d-dimensional vector and therefore used to detect possible shifts in the d-dimensional distributions of the quality process. A statistical process control (SPC) procedure set up in a multivariate framework is more effective than a joint monitoring system consisting of a series of traditional univariate control charts (Crosier, 1988).

The most popular multivariate statistical process control charts are based on the Hotelling's T^2 statistics, that are a multivariate extension of Shewart's chart (or $\bar{X}$ control chart). Like the univariate counterparts, also the multivariate control charts can be distinguished into parametric and non-parametric types according to the distributive assumption underlying a control charts (e.g. normality) are verified or not. When the assumption of normality is not verified, the use of conventional (multivariate) control charts for process monitoring is questionable. Non-parametric control charts do not require distributive assumptions on process data and generally enjoy greater robustness, namely they are less sensitive to outliers than parametric control schemes. A survey of parametric multivariate SPC charts can be found in Bersimis *et al.*, 2007, while a review of non-parametric multivariate control charts can be found in Chakraborti & Graham, 2019.

Statistical depth functions are largely used in non-parametric statistics for the analysis of multivariate data. These are non-parametric functions that can provide a dimension reduction to high-dimensional problems. In this work, we will focus on L^p data depth to build a control chart. The L^p data depth have additional advantages over other existing depth based control charts already introduced in a multivariate SPC (i.e. they ensure an ease of computation even in high dimensions).

2 Our proposal

Depth-based methodology to construct control charts can be interpreted as a multivariate generalization of standard univariate. A depth function aims at providing the degree of centrality of a point x with respect to a distribution F in $\mathbb{R}^d$, denoted by $D(x,F)$. Hence, higher values of $D(x,F)$ correspond to deeper (more central) while smaller values indicate less central points (i.e. further away from the center with respect to F). Hence, a center-outward ranking of the data is provided. There are several notions of data depth function available in the literature. The halfspace, simplicial, Mahalanobis and L^p depths are some of the most popular ones. In this work, we adopt the notion of L^p depth introduced by Zuo, 2004 because of its ease of computation and (local and global) robustness properties. The depth is defined as follows:

$$L^p D(x,F) = \frac{1}{1+E(\|x-X\|_p)},$$

where $X \sim F$, $\|\cdot\|$ denotes the L^p-norm (when $p=2$ the Euclidean norm is derived) and $E(\cdot)$ is its expected value.

We conducted a Montecarlo simulation study to evaluate the performance of the Q-type control charts based on the L^p data depth in comparison with the Mahalanobis depth-based Q-type charts. The Q-type control chart is the multivariate analogue of the average univariate chart $\bar{X}$. We set $p=2$ for the computation of the L^p depth. The simulation study was designed as an analysis to evaluate the chart performances under multiple settings, defined with regard to the number of variables to be monitored (i.e., the dimension), the size of the reference sample, the size of the sub-group, and by considering different distributional settings (Normal, Skew-Normal and Cauchy). We considered both the in-control and out-of-control cases. Specifically, three out-of-control scenarios were evaluated including shift in the mean vector, change in the variance and a combination of both variance change and shift in the mean. We evaluated the performances in terms of average run length (ARL) and its standard deviation. ARL is defined as the expected number of samples required to get a first out-of-control signal, and it can be obtained by taking reciprocal of false alarm probability. Moreover, ARL is one of the performance measures used for comparing the control charts. Results obtained from both in-control and out-of-control cases indicate that Q-type charts based on L^2D perform better than those based on Mahalanobis regardless of the process distribution, the dimensionality and the size of both the reference and the sub-group samples.

3 Conclusion

In a statistical process control framework, we proposed to use control charts based on the L^p data depth. Our approach is fully non-parametric, meaning that the obtained charts are valid without parametric assumptions on the process distribution. In addition, these charts allow for the simultaneous detection of both the location change and the scale increase in a process. The performance of our proposal is investigated via a simulation study. The results show that the L^p depth based control charts are a promising alternative to the well-known Mahalanobis depth. Moreover, L^p depth is particularly appealing because of its computational ease even in high multidimensional spaces.

References

BERSIMIS, SOTIRIS, PSARAKIS, STELIOS, & PANARETOS, JOHN. 2007. Multivariate statistical process control charts: an overview. *Quality and Reliability engineering international*, **23**(5), 517–543.

CHAKRABORTI, S, & GRAHAM, MA. 2019. Nonparametric (distribution-free) control charts: An updated overview and some results. *Quality Engineering*, **31**(4), 523–544.

CROSIER, RONALD B. 1988. Multivariate generalizations of cumulative sum quality-control schemes. *Technometrics*, **30**(3), 291–303.

HOTTELING, H. 1947. Multivariate quality control, illustrated by the air testing of sample bombsights. *Techniques of statistical analysis*, 111–184.

IDRIS, SUWANDA, WACHIDAH, LISNUR, SOFIYAYANTI, TETI, & HARAHAP, ERWIN. 2019. The Control Chart of Data Depth Based on Influence Function of Variance Vector. *In: Journal of Physics: Conference Series*, vol. 1366. IOP Publishing.

LIU, REGINA Y. 1995. Control charts for multivariate processes. *Journal of the American Statistical Association*, **90**(432), 1380–1387.

MONTGOMERY, DOUGLAS C. 2007. *Introduction to statistical quality control*. John Wiley & Sons.

WOODALL, WILLIAM H. 2000. Controversies and contradictions in statistical process control. *Journal of Quality Technology*, **32**(4), 341–350.

WOODALL, WILLIAM H, & MONTGOMERY, DOUGLAS C. 1999. Research issues and ideas in statistical process control. *Journal of Quality Technology*, **31**(4), 376–386.

ZUO, YIJUN. 2004. Robustness of weighted L p–depth and L p–median. *Allgemeines Statistisches Archiv*, **88**(2), 215–234.

ANGULAR HALFSPACE DEPTH: CENTRAL REGIONS*

Petra Laketa[1]and Stanislav Nagy[1]

[1] Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
(e-mail: laketa@karlin.mff.cuni.cz, nagy@karlin.mff.cuni.cz)

ABSTRACT: The angular halfspace depth is an extension of the classical halfspace depth that is applicable to directional data. The upper level sets of this function serve as an analogue of the inter-quantile regions, and allow introduction of orders and rank statistics also for data living on the unit sphere. We explore the basic theoretical properties of these regions, and contrast them with the central regions defined in multivariate Euclidean spaces using the standard halfspace depth.

KEYWORDS: angular depth, central regions, directional data analysis, halfspace depth.

1 Angular halfspace depth

Statistical depth is a remarkable tool for ordering multivariate data. For a probability measure P in the Euclidean space $\mathbb{R}^d$, $d \geq 1$, the depth describes how much "centrally located" a point $x \in \mathbb{R}^d$ is with respect to P. The arguably most important depth in $\mathbb{R}^d$ is the *halfspace depth* that to each $x \in \mathbb{R}^d$ assigns

$$hD(x;P) = \inf\left\{P(H) : H \in \mathcal{H} \text{ and } x \in H\right\} \in [0,1], \tag{1}$$

for $\mathcal{H} = \left\{H_{y,v} \colon y \in \mathbb{R}^d, v \in \mathbb{R}^d \setminus \{0\}\right\}$ the set of all closed halfspaces $H_{y,v} = \left\{z \in \mathbb{R}^d \colon \langle z-y, v\rangle \geq 0\right\}$ in $\mathbb{R}^d$. Here we deal with directional data (Ley & Verdebout, 2017), meaning data generated from P whose support lies on the unit sphere $\mathbb{S}^{d-1} = \left\{x \in \mathbb{R}^d \colon \|x\| = 1\right\}$. For most such P, the depth (1) is trivially zero on $\mathbb{S}^{d-1}$, and is therefore of no use. In that situation, it is more natural to consider an angular variant of the halfspace depth introduced by Small, 1987. Let $\mathcal{H}_0 \subset \mathcal{H}$ be the collection of those halfspaces $H_{0,v} \in \mathcal{H}$ whose boundary contains the origin $0 \in \mathbb{R}^d$. The *angular halfspace depth* of $x \in \mathbb{S}^{d-1}$ with respect to a probability measure P on $\mathbb{S}^{d-1}$ is defined as

$$ahD(x;P) = \inf\left\{P(H) : H \in \mathcal{H}_0 \text{ and } x \in H\right\} \in [0,1]. \tag{2}$$

*This work was supported by the grant 19-16097Y of the Czech Science Foundation, and by the PRIMUS/17/SCI/3 project of Charles University. P. Laketa was supported by the OP RDE project "International mobility of research, technical and administrative staff at the Charles University" CZ.02.2.69/0.0/0.0/18_053/0016976.

The similarity of the depths (1) and (2) is obvious — one restricts to halfspaces from $\mathcal{H}_0$ when considering the angular depth. Many properties of the angular halfspace depth were explored by Liu & Singh, 1992. Here we first revise some of those known results, and then derive an array of new properties of the upper level sets of the function (2) for general probability measures on $\mathbb{S}^{d-1}$.

2 A hemisphere of constant depth

A peculiar property of the angular halfspace depth is that it has to be constant on a hemisphere of $\mathbb{S}^{d-1}$, for any P on $\mathbb{S}^{d-1}$. More precisely, Proposition 4.6 of Liu & Singh, 1992 says that there exists a hemisphere with a constant angular depth equal to $\alpha_0 = \inf_{x \in \mathbb{S}^{d-1}} ahD(x;P)$. That result is given without a proof and from the context it appears to be claimed for a closed hemisphere. In our first example we demonstrate that for general measures P one has to be cautious when formulating this statement.

Consider the probability measure P on the circle $\mathbb{S}^1$ in $\mathbb{R}^2$ (left panel of Figure 1) defined as a mixture of a uniform distribution on the upper halfcircle $\mathbb{S}^1_+ = \left\{(x_1,x_2) \in \mathbb{S}^1 : x_2 > 0\right\}$ and an atom at $a = (1,0)$ with equal weights $1/2$. For each $n = 1,2,\ldots$ consider a halfspace (beige region in Figure 1)

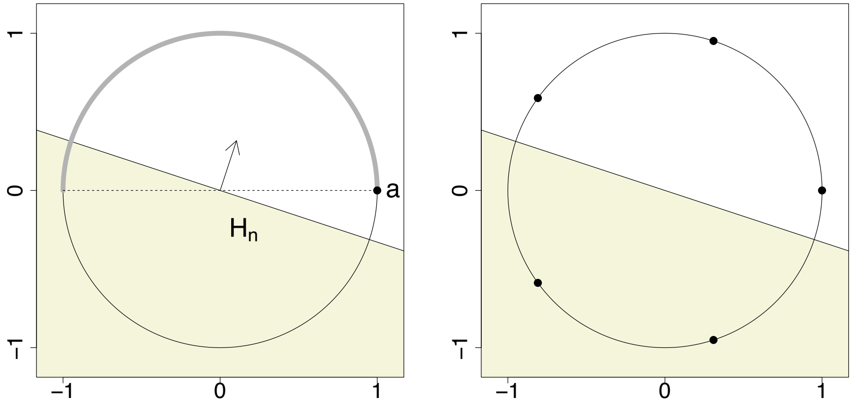


Figure 1. *Left: For P a mixture of uniform distribution on $\mathbb{S}^1_+$ (grey arc) and an atom at point a (black point) a closed hemisphere of constant depth does not exist. Right: A measure P with five atoms such that $ahD(\cdot;P)$ is constant on $\mathbb{S}^1$ and equal to $\alpha_0 = 2/5$.*

$$H_n = \left\{(x_1,x_2) \in \mathbb{R}^2 : x_1 \cos(\pi/2 - 1/n) + x_2 \sin(\pi/2 - 1/n) \leq 0\right\} \in \mathcal{H}_0$$

not containing a at an angle $\theta_n = -1/n$ with the x_1-axis. Since $\lim_{n\to\infty} \theta_n = 0$, surely $\lim_{n\to\infty} P(H_n) = 0$, meaning that $ahD(x;P) = 0$ for any point x in the lower halfcircle $\mathbb{S}^1_- = \left\{(x_1,x_2) \in \mathbb{S}^1 : x_2 < 0\right\}$. We obtain $\alpha_0 = 0$. On the other hand, $P(H) \geq 1/2$ for every $H \in \mathcal{H}_0$ that contains a, implying that $ahD(a;P) \geq 1/2 > \alpha_0$. Also, any $H \in \mathcal{H}_0$ that contains points from $\mathbb{S}^1_+$ is of positive P-mass. Overall we obtain that the angular depth is positive exactly in the set $\mathbb{S}^1_+ \cup \{a\}$, and there is no closed halfcircle of $\mathbb{S}^1$ of depth $\alpha_0 = 0$.

In our example there exists an open hemisphere $\mathbb{S}^1_-$ with constant depth α_0. That is not a coincidence — an analogous result for an open hemisphere of constant depth is possible to be proved* for any measure P on $\mathbb{S}^{d-1}$. It is however interesting to note that it may also happen that the depth (2) is constant on the whole sphere $\mathbb{S}^{d-1}$, and therefore equal to α_0 everywhere, see Figure 1.

3 Central regions of the angular halfspace depth

While for any P on $\mathbb{S}^{d-1}$ there always exists an open hemisphere $S_{min} \subset \mathbb{S}^{d-1}$ of minimal depth, its complement $\mathbb{S}^{d-1} \setminus S_{min}$ typically contains points of higher depth (2). The upper level sets of the angular halfspace depth therefore form a basis for generalizations of quantiles and inter-quantile regions to $\mathbb{S}^{d-1}$, in the same way as the level sets of the halfspace depth (1) do in $\mathbb{R}^d$. The *central region* of P at level $\alpha \geq 0$ is given by

$$D_\alpha = \left\{x \in \mathbb{S}^{d-1} : ahD(x;P) \geq \alpha\right\}. \tag{3}$$

The smallest non-empty region D_α presents a natural analogue of the median applicable to directional data. In analogy with the corresponding properties well established for the standard halfspace depth (1) in $\mathbb{R}^d$, it is possible to show that also regions (3) posses several attractive traits — they are closed and spherically convex sets in $\mathbb{S}^{d-1}$ that can be represented as intersections of closed spherical halfspaces (sets of the form $\mathcal{H}_0 \cap \mathbb{S}^{d-1}$). Formal proofs of the following statements will appear in our comprehensive treatment of the theory of the angular halfspace depth that is currently in preparation. In each of the statements P is a Borel probability measure on $\mathbb{S}^{d-1}$.

Upper semi-continuity. The mapping $\mathbb{S}^{d-1} \to [0,1] : x \mapsto ahD(x;P)$ is upper semi-continuous, i.e. for any $x \in \mathbb{S}^{d-1}$ and a sequence $\{x_n\}_{n=1}^{\infty} \subset \mathbb{S}^{d-1}$ that converges to x it holds $\limsup_{n\to\infty} ahD(x_n;P) \geq ahD(x;P)$. As a consequence, all depth regions (3) are closed sets.

*The proof of this claim is not difficult, but due to the limited available space we will present it elsewhere, together with all the other technical derivations outlined in the rest of this note.

Intersection of halfspaces. For any $\alpha > \alpha_0$ we can write

$$D_\alpha = \bigcap \left\{ \operatorname{int}(H) : H \in \mathcal{H}_0 \text{ and } P(\operatorname{int}(H)) > 1 - \alpha \right\} \cap \left(\mathbb{S}^{d-1} \setminus S_{min} \right),$$

where $\operatorname{int}(H)$ denotes the interior of H, which is an open halfspace. This result is weaker than the one for the usual halfspace depth (Proposition 6 of Rousseeuw & Ruts, 1999), where one can write a central region as an intersection of closed halfspaces from $\mathcal{H}$. As a consequence of our result we obtain that each D_α with $\alpha > \alpha_0$ is an intersection of a convex set in $\mathbb{R}^d$ and a hemisphere. The case $\alpha = \alpha_0$ gives trivially the whole sphere $D_{\alpha_0} = \mathbb{S}^{d-1}$.

4 Refinements for smooth measures

We say that a probability measure P on $\mathbb{S}^{d-1}$ is smooth if $P(\partial H) = 0$ for any boundary hyperplane ∂H of $H \in \mathcal{H}_0$. It is satisfied for any P that has a density with respect to the spherical Lebesgue measure on $\mathbb{S}^{d-1}$. For smooth P one obtains stronger results about the central regions of $ahD(x;P)$:

- the minimizing hemisphere S_{min} may be chosen to be closed, and satisfies the additional condition $P(S_{min}) = \alpha_0$;
- the depth $ahD(x;P)$ is continuous as a function of $x \in \mathbb{S}^{d-1}$;
- $D_\alpha = \bigcap \left\{ H \colon H \in \mathcal{H}_0 \text{ and } P(H) > 1 - \alpha \right\} \cap \left(\mathbb{S}^{d-1} \setminus S_{min} \right)$.

A useful application of these results is the construction of bagdistances for directional data presented by *H. Demni* in this book of short papers. In that contribution, bagdistances are used with success in a comprehensive simulation study of nonparametric classification of points in $\mathbb{S}^{d-1}$.

References

LEY, CHRISTOPHE, & VERDEBOUT, THOMAS. 2017. *Modern directional statistics*. Chapman & Hall/CRC Interdisciplinary Statistics Series. CRC Press, Boca Raton, FL.

LIU, REGINA Y., & SINGH, KESAR. 1992. Ordering directional data: concepts of data depth on circles and spheres. *Ann. Statist.*, **20**(3), 1468–1484.

ROUSSEEUW, PETER J., & RUTS, IDA. 1999. The depth function of a population distribution. *Metrika*, **49**(3), 213–244.

SMALL, CHRISTOPHER G. 1987. Measures of centrality for multivariate and directional distributions. *Canad. J. Statist.*, **15**(1), 31–39.

Clustering production indexes for construction with forecast distributions

Michele La Rocca [1], Francesco Giordano[1] and Cira Perna[1]

[1] Department of Economics and Statistics, University of Salerno, (e-mail: larocca@unisa.it, giordano@unisa.it, perna@unisa.it)

Abstract: In this paper we focus on a recent proposal for clustering nonlinear time series data in which dissimilarities are computed according to time series forecast distributions. The aim is to evaluate the impact of COVID-19 pandemic on the construction sector for a set of 21 European countries.

Keywords: Feedforward neural networks, bootstrap, nonlinear time series.

1 Introduction

In the last decades there has been a growing interest in time series clustering. Some recent approaches rely on the use of distance criteria which compare the forecast densities estimated by using a resampling method combined with a nonparametric kernel estimator (see Alonso *et al.*, 2006 and Vilar *et al.*, 2010). More recently, La Rocca *et al.*, 2021, have proposed a novel approach for clustering nonlinear autoregressive time series based on the use of a class of neural network models to approximate the original nonlinear process, combined with the pair bootstrap as a resampling device. The aim of this paper is to discuss the novel approach and to evaluate the impact of COVID-19 pandemic on the production index for construction, an important business cycle indicator, for a set of 21 European countries.

2 The clustering procedure in a nutshell

Let $\{Y_t, t \in \mathbb{Z}\}$ be a real valued stationary stochastic process modeled as a nonlinear autoregressive (NAR) model of the form $Y_t = g\left(\mathbf{x}_{t-1}\right) + \varepsilon_t$, where $g(\cdot)$ is an unknown (possibly) nonlinear regression function, $\mathbf{x}'_{t-1} = (Y_{t-1}, \ldots, Y_{t-p})$ and $\{\varepsilon_t\}$ are *iid* error terms, with $\mathbb{E}[\varepsilon_t] = 0$ and $\mathbb{E}[\varepsilon_t^2] > 0$. Let $\left(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(S)}\right)$ be S observed time series of length T generated from a DGP of the previous class, where $\mathbf{y}^{(i)} = \left(Y_1^{(i)}, \ldots, Y_T^{(i)}\right)$. The aim is to cluster time series based on their full forecast distribution at a specific future time $T+h$, with $h \geq 1$.

This approach accounts for the future dynamic behaviour of the time series, by using the L^r-norm distance $D_{r,ij} = \int \left| F^i_{T+h|T}(y) - F^j_{T+h|T}(y) \right|^r dy \quad r = 1,2,$ where $F^i_{T+h|T}(\cdot)$, $i = 1,\ldots,S$ is the forecast distribution function at a given future point $T+h$, of the series $\mathbf{y}^{(i)}$, conditioned on the information set available up to time T. Since the L^r-norm distance previously defined cannot be computed directly, La Rocca *et al.*, 2021 have proposed a strategy in which the unknown distributions are consistently estimated by using a feed forward neural network estimator and the pair bootstrap approach. In particular, given the forecast horizon h, the unknown function $g(\cdot)$ can be approximated by using the network

$$f_{mh}(\mathbf{x}_{t-h};\theta) = \sum_{k=1}^{m} c_k \psi\left(\mathbf{w}'_k \mathbf{x}_{t-h} + w_{k0}\right) + c_0 \qquad (1)$$

with $\theta = (c_0, c_1, \ldots, c_m, \mathbf{w}_1, \ldots, \mathbf{w}_m, w_{10}, \ldots, w_{m0})$, where m is the hidden layer size, $\mathbf{w}_k$ are the vectors of weights for the connections between input layer and hidden layer, c_k, $k = 1,\ldots,m$ are the weights of the link between the hidden layer and the output; w_{k0} and c_0 are the bias terms; $\psi(\cdot)$ is a proper chosen activation function and $\mathbf{x}'_{t-h} = (Y_{t-h},\ldots,Y_{t-h-p+1})$.

The general procedure is summarized in the following Algorithm.

Algorithm

```
1: Fix the forecast horizon $h \geq 1$. Let $\mathcal{X} = \{(Y_t, \mathbf{x}'_{t-h}), t = p+h,\ldots,T\}$
2: Fix the hidden layer size $m$ and the lag structure $p$ and estimate the weights of the network as $\hat{\theta}_h = \arg\min_\theta \frac{1}{T-p-h+1} \sum_{t=p+h}^{T} (Y_t - f_{mh}(\mathbf{x}_{t-h};\theta))^2$.
3: Compute the residuals from the estimated network defined as: $\hat{\varepsilon}_t = Y_t - f_{mh}\left(\mathbf{x}_{t-h};\hat{\theta}_h\right)$
4: Compute the centered residuals: $\tilde{\varepsilon}_t = \hat{\varepsilon}_t - \frac{1}{T-p-h+1}\sum_{t=p+h}^{T}\hat{\varepsilon}_t$.
5: Resample $\{(Y_t^*, \mathbf{x}'^*_{t-h}) = (Y_t^*, Y_{t-h}^*, \ldots, Y_{t-h-p+1}^*)$, $t = p+h,\ldots,T\}$, as an iid sample from the set of tuples $\mathcal{X}$.
6: Get the bootstrap estimate of the neural network weights:
       $\hat{\theta}_h^* = \arg\min_\theta \frac{1}{T-p-h+1}\sum_{t=p+h}^{T}\left(Y_t^* - f_{mh}\left(\mathbf{x}_{t-h}^*;\theta\right)\right)^2$.
7: Compute $\hat{Y}_{T+h}^* = f_{mh}\left(Y_T, Y_{T-1}, \ldots, Y_{T-p+1}; \hat{\theta}_h^*\right) + \varepsilon_{T+h}^*$ where $\varepsilon_{T+h}^*$ is a random sample from the centered residuals $\{\tilde{\varepsilon}_t\}$.
8: The bootstrap forecast distribution $F_{T+h|T}^*$ is given by the law of $\hat{Y}_{T+h}^*$ conditioned on $\mathcal{X}$.
```

Note that, as usual, the bootstrap distribution can be approximated by Monte Carlo simulations repeating B times steps 5-7, and then computing the empirical cumulative distribution function (ECDF) of $\hat{Y}^b_{T+h}$, $b = 1,2,\ldots,B$. As a resampling device, the pair bootstrap has been implemented, a suitable choice in the context of neural network models. Moreover, being the data generating process nonlinear, a direct multi-step forecasting approach is considered, where a separate neural network model is estimated for each forecasting horizon, and forecasts are computed only conditioning on the observed data.

3 An application to the European construction sector

The proposed procedure has been used to cluster the production index for construction (seasonally and calendar adjusted) for 21 European countries observed from January 2000 to December 2020 (base year 2015). The production index measures the activity in the building and construction industry, and it is considered a critical business cycle indicator. The dataset is available from the Eurostat website. The aim here is to identify the different group structure induced by the COVID-19 pandemic by using the forecast one-step ahead distribution for January 2020 (so excluding any observations from the COVID-19 pandemic), the forecast twelve-step ahead distribution for January 2021, the forecast one-step ahead distribution for January 2021 (we have trained all models up to December 2020).

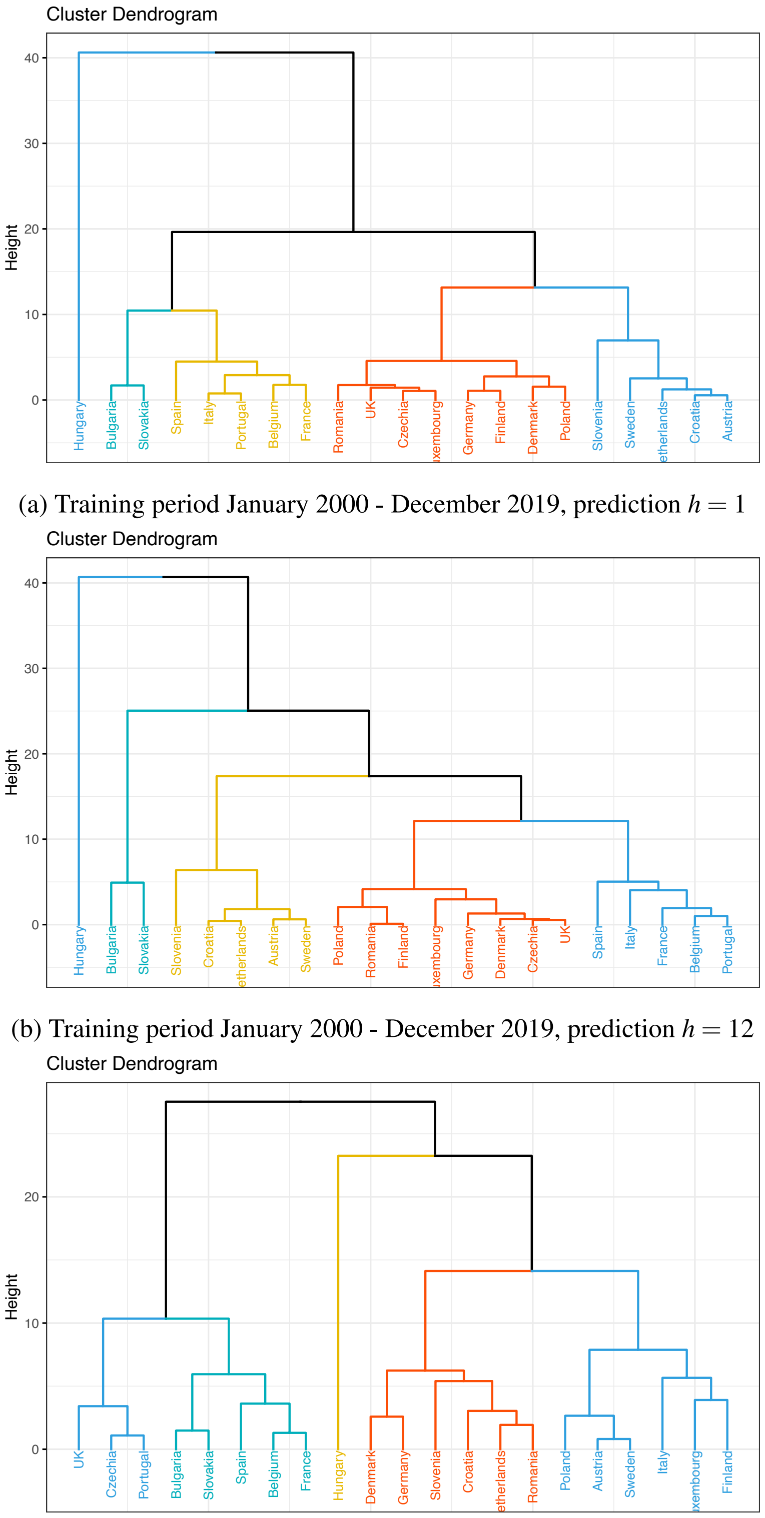
Apparently, the group structure would have been almost identical without the impact due to the COVID-19 pandemic, showing a somewhat stable economic evolution of all the countries considered in the application (see panels a and b). On the contrary, when based on models that include the year 2020 in the training period, where all countries experienced severe contractions in their economic activities, the dataset shows a pretty different group structure, indicating different routes and timelines for economic recovery (see panel c).

References

ALONSO, A.M., BERRENDERO, J.R., HERNÁNDEZ, A., & JUSTEL, A. 2006. Time series clustering based on forecast densities. *Computational Statistics & Data Analysis*, **51**(2), 762–776.

LA ROCCA, M., GIORDANO, F., & PERNA, C. 2021. Clustering nonlinear time series with neural network bootstrap forecast distributions. *Submitted.*

VILAR, J.A., ALONSO, A.M., & VILAR, J.M. 2010. Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics & Data Analysis*, **54**(11), 2850–2865.



(a) Training period January 2000 - December 2019, prediction $h = 1$

(b) Training period January 2000 - December 2019, prediction $h = 12$

(c) Training period January 2000 - December 2020, prediction $h = 1$

Figure 1: Construction indexes clustering based on h-step ahead forecast distributions and L^1-norm distance.

Clustering Longitudinal Data with Category Theory for Diabetic Kidney Disease

Maria Mannone[1][2], Veronica Distefano[1], Claudio Silvestri[1][3] and Irene Poli[1]

[1] European Centre for Living Technology, Ca' Foscari University of Venice, Italy, (e-mail: maria.mannone@unive.it, veronica.distefano@unive.it, claudio.silvestri@unive.it, irenpoli@unive.it)

[2] Department of Mathematics and Computer Sciences, University of Palermo, Italy

[3] Dipartimento di Scienze Ambientali, Informatica e Statistica, Ca' Foscari University of Venice, Italy

Abstract: In the framework of precision medicine, we investigate the similarity of diabetic kidney disease (DKD) patients through longitudinal data clusters. Starting with insights from category theory, we build patients' clusters according to the shapes of their trajectories, adopting the Fréchet distance. We group patients according to their behavior of the estimated glomerular filtration rate (eGFR), obtaining informative mean curves. Behavior pattern recognition can shed light on individualized treatments.

Keywords: longitudinal data clustering, category theory, Fréchet distance, precision medicine, DKD disease progress

1 Introduction

Precision medicine aims to find individualized therapeutic treatments according to patients' specific characteristics. To make accurate predictions it is crucial to retrieve information on the long-term reactions of patients to given treatments, investigating time trajectories of the disease progress (Karpati *et al.*, 2018). Each patient is identified by demographic and clinical variables at different time points. The similarity of behavior of patients across time can be accounted by *clusters of trajectories*. The final aim of this research is to build clusters of longitudinal data to identify the optimal treatment rule. Therefore, we intend to build patient clusters according to distances between patients at each time point, and distances of the same patient between time points. We can consider distance as a transformation, and distance variation as a transformation between transformations. Because the concept of transformations between transformations is the starting point of mathematical category theory[*] (Grandis, 2020), we can exploit its concepts for cluster analysis (Carlsson & Mémoli, 2013). We introduce comparisons between patient trajectories according to their shapes. Here, we compare patients' trajectories building clusters of shapes using the Fréchet distance (Genolini *et al.*, 2016), which takes into account shape variations. The novelty of our study is an integrative approach joining categories, clusters, and shape trajectories. For each patient we observe demographic and clinical variables, treatments, and response to the different treatments. This approach is derived for a real dataset concerning patients with diabetic kidney disease (DKD) from the DC-ren project.[†] Clustering is based on the response to the treatment, that is evaluated from the estimated glomerular filtration rate (eGFR). The identification of different evolution patterns can shed light on the best individualized drug combination. The paper is structured as follows: in Section 2 we introduce some theoretical concepts and the methodology we adopted, and in Section 3 we analyze the results of our study.

2 Methodology

Let us consider a dataset composed of n patients characterized by p observable variables at four time points t_0, t_1, t_2, t_3. Each patient is characterized as a triplet $(\mathbf{X}_i(t_k), \mathbf{D}(t_k), Y_i(t_k))$, where i is the individual (the patient); t_k is the time point $k = 0,1,2,3$; $\mathbf{X}_i(t_k)$ is a set of variables characterizing the individual; $Y_i(t_k)$ is the value of the response variable Y at t_k; $\mathbf{D}(t_k)$ stands for the given drug. We indicate the distance between patients i, i' with respect to the variable Y and time k as $d^Y_{i,i'}(t_k)$, and the distance between values observed at times $t_k, t_{k'}$ of the variable Y for the same individual i as $d^Y_i(t_k, t_{k'})$. We thus obtain an *enriched double category* with metrics in $\mathbb{R}$ (Grandis, 2020), having $x^Y_i(t_k)$, $k = 0,...3$, $i = 1,...,n$, as objects, and $d^Y_{i,i'}(t_k)$, $d^Y_i(t_k, t_{k'})$ as morphisms. The comparison of trajectories of different patients involves both of these distances. The time trajectory of the i-th patient is indicated as $path^Y_i$. Clustering is a functor $F_a : path^Y_i \rightarrow < path^{a,Y}_{\iota} >, \iota = 1,...,n' < n$, where n' is the number of significative curves, which groups similar trajectories in the same cluster.

[*] A category is constituted by objects (points) and morphisms (arrows). A *functor* maps objects and morphisms of a category into objects and morphisms of another category. *Natural transformations* map functors to functors.

[†] `https://dc-ren.eu/`. The project focuses on type 2 diabetes.

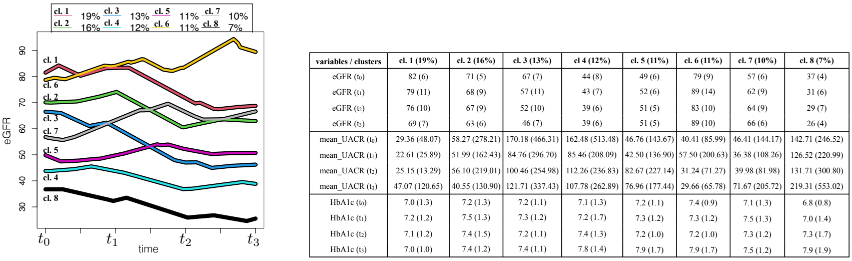
The $< path_{\iota}^{a,Y} >, \forall\iota$ is the representative curve of each cluster.[‡] Most of the existing research uses the Euclidean distance to compare trajectories. However, because we aim to compare trajectory shapes, we determine the Fréchet distance, and, according to this distance, we build patient clusters. The Fréchet distance is based on the comparison between pairs of points following the profiles of the curves they belong to. We analyzed the response to the treatment, measured according to the eGFR variable. We build patient clusters deriving the mean of eGFR trajectories. We then investigate the characteristics of patients, in relationship with demographic and clinical variables which characterize each patient. We evaluated the behavior of 241 DKD patients observed in a 4-year period, according to the DC-ren project. Computationally, we used an extension of the longitudinal k-means, *kmlShape* (Genolini *et al.*, 2016), with time scale 0.5.

3 Results

In this study, which involves clusters based on shapes of individual trajectories, without assuming a particular shape, we obtained a grouping of patients according to their similarity of eGFR behavior. Trajectories are evaluated upon the Fréchet distance between them, computed on the continuous variable eGFR. We obtain 8 patient clusters with similar eGFR shape of individual trajectories. In Figure 1a, we represent the obtained clusters. In the Table (Figure 1b), we show the behavior of clusters achieved with the Fréchet distance, with standard deviations at each time point. We find three main patients' subgroups: patients with decreasing eGFR (clusters 1, 3, 8); low decreasing eGFR (cl. 2, 4), and stable/increasing eGFR (cl. 5, 6, 7). To explain these behaviors, we consider some relevant clinical characteristics of patients, which are the ratio of urinary albumin and creatinine (mean UACR) and the glycated hemoglobin (HbA1c), shown in the Table (Figure 1b). Mean UACR is decreasing in cluster 6, while it is increasing in cluster 8. Decreasing or stable values of HbA1c, which characterize DKD patients (Karpati *et al.*, 2018), are shown by patients in cl. 6, while increasing values of HbA1c are shown by patients in cl. 8. We notice that there is a relationship between non-decreasing eGFR and stable HbA1c. Patients can receive different treatments, such as D_1, D_2, D_3, D_4. In cluster 3 most of the patients that change drug ($D_1 \rightarrow D_1 + D_2$) show a positive response to the treatment. However, patients in cluster 2 who change the drug

[‡] A different clustering method F_b gives us similar representative paths $< path_{\iota}^{b,Y} >$. Natural transformation $\alpha_{a,b} : F_a \rightarrow F_b$ maps clustering methods.

do not have a positive response. In cluster 7, all patients are keeping the same drug (D_1), and most of them show a positive response. Patients in clusters 6 and 7 display the best eGFR behavior; most of the patients in these clusters keep in fact a stable behavior and a positive response to the treatment. Patients in clusters 1 and 6 start from close eGFR values, but different UACR values; given $D_1 + D_2$, their response is different. Patients in clusters 1, 2, 5 change, receiving $D_1 + D_3$ and $D_1 + D_4$. These results will be considered in building a predictive system to envisage the best treatment for each individual.



variables / clusters	cl. 1 (19%)	cl. 2 (16%)	cl. 3 (13%)	cl 4 (12%)	cl. 5 (11%)	cl. 6 (11%)	cl. 7 (10%)	cl. 8 (7%)
eGFR (t_0)	82 (6)	71 (5)	67 (7)	44 (8)	49 (6)	79 (9)	57 (6)	37 (4)
eGFR (t_1)	79 (11)	68 (9)	57 (11)	43 (7)	52 (6)	89 (14)	62 (9)	31 (6)
eGFR (t_2)	76 (10)	67 (9)	52 (10)	39 (6)	51 (5)	83 (10)	64 (9)	29 (7)
eGFR (t_3)	69 (7)	63 (6)	46 (7)	39 (6)	51 (5)	89 (10)	66 (6)	26 (4)
mean_UACR (t_0)	29.36 (48.07)	58.27 (278.21)	170.18 (466.31)	162.48 (513.48)	46.76 (143.67)	40.41 (85.99)	46.41 (144.17)	142.71 (246.52)
mean_UACR (t_1)	22.61 (25.89)	51.99 (162.43)	84.76 (296.70)	85.46 (208.09)	42.50 (136.90)	57.50 (200.63)	36.38 (108.26)	126.52 (220.99)
mean_UACR (t_2)	25.15 (13.29)	56.10 (219.01)	100.46 (254.98)	112.26 (236.83)	82.67 (227.14)	31.24 (71.27)	39.98 (81.98)	131.71 (300.80)
mean_UACR (t_3)	47.07 (120.65)	40.55 (130.90)	121.71 (337.43)	107.78 (262.89)	76.96 (177.44)	29.66 (65.78)	71.67 (205.72)	219.31 (553.02)
HbA1c (t_0)	7.0 (1.3)	7.2 (1.3)	7.2 (1.1)	7.1 (1.3)	7.2 (1.1)	7.4 (0.9)	7.1 (1.3)	6.8 (0.8)
HbA1c (t_1)	7.2 (1.2)	7.5 (1.3)	7.3 (1.2)	7.2 (1.7)	7.3 (1.2)	7.3 (1.2)	7.5 (1.3)	7.0 (1.4)
HbA1c (t_2)	7.1 (1.2)	7.4 (1.5)	7.2 (1.1)	7.4 (1.3)	7.2 (1.0)	7.2 (1.0)	7.3 (1.2)	7.3 (1.7)
HbA1c (t_3)	7.0 (1.0)	7.4 (1.2)	7.4 (1.1)	7.8 (1.4)	7.9 (1.7)	7.9 (1.7)	7.5 (1.2)	7.9 (1.9)

(a)

(b)

Figure 1: Shape clusters (a) and Table of mean values (b).

Acknowledgements

This research activity is part of the project DC-ren that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 848011. We thank the researchers of the European Centre for Living Technology (ECLT) for helpful discussions and suggestions.

References

CARLSSON, G., & MÉMOLI, F. 2013. Classifying Clustering Schemes. *Foundations of Computational Mathematics*, **13**, 221–252.

GENOLINI, C., ECOCHARD, R., BENGHEZAL, M., DRISS, T., ANDRIEU, S., & SUBTIL, F. 2016. kmlShape: An Efficient Method to Cluster Longitudinal Data (Time-Series) According to Their Shapes. *PlosOne*, **11**(6).

GRANDIS, M. 2020. *Higher Category Theory*. Singapore: World Scientific.

KARPATI, T., LEVENTER-ROBERTS, M., FELDMAN, B., C., COHEN-STAVI., & RAZ I., BALICER, R. 2018. Patient clusters based on HbA1c trajectories: A step toward individualized medicine in type 2 diabetes. *PLos One*, **13**(11).

A Redundancy Analysis with Multivariate Random-Coefficients Linear Models

Laura Marcis [1], Maria Chiara Pagliarella [2] and Renato Salvatore [1]

[1] Department of Economics and Law, University of Cassino, (e-mail: laura.marcis@unicas.it, rsalvatore@unicas.it)

[2] Italian National Institute for Public Policy Analysis (e-mail: mc.pagliarella@inapp.org)

Abstract: Random-coefficients linear models can be considered as a particular case of linear mixed models, in which the random effects depend on the model fixed-effects design matrix. A Redundancy Analysis of estimates of the multivariate random-effects may be able to capture the leading contribution to this correlation. Starting from the standardized multivariate best linear predictors, we introduce the random effects reduced space by a weighted least-squares closed-form solution. The application shows the effect of the linear dependence of the random-effects in the space of the predictor variables.

Keywords: Redundancy analysis, linear mixed model, empirical best linear unbiased predictor, restricted maximum likelihood estimator.

1 Introduction

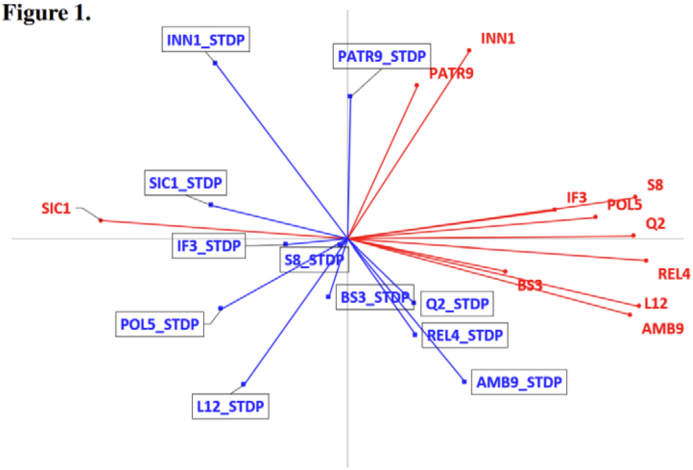
Random-coefficients linear regression models (*RCM*) represent a special case of linear mixed models (LMM, Demidenko, 2004), where the vector of regression coefficients for the subjects (e.g. repeated observations) is modeled in a second stage linear regression equation. In order to specify this type of models, it is convenient to define a two-stage hierarchical linear model, with a first stage that models within-subject observations, and as second stage we use a linear model for the random regression coefficients. Although in the basic linear mixed models the random effects are not correlated with the modeled response variables (unlike the fixed-effects with the random effects estimates), in the *RCM* this correlation depends on the fixed-effects design matrix of the regression model. Since this happens, one can be interested to know in which components of a multivariate model the random effects are related to the subspace spanned by the model covariates. In the same way, what components of the multivariate vector seem to be orthogonal to that subspace. *Redundancy*

Analysis (*RDA)* was originally introduced in order to capture the effect onto a reduced space of the linear dependence by a set of criterion variables from a set of predictors. A *RDA* of the predicted criterion variables by the best linear unbiased predictor may be quite representative (Marcis & Salvatore, 2020). This paper uses a *RDA* by a least-squares solution for an optimal fixed-effects estimate from the data provided by the random-coefficients linear model predictors of the criterion variables. The application study performs the method introduced on the official data by the Italian Equitable and Sustainable Well-being indicators.

2 Redundancy Analysis: model estimation and application study

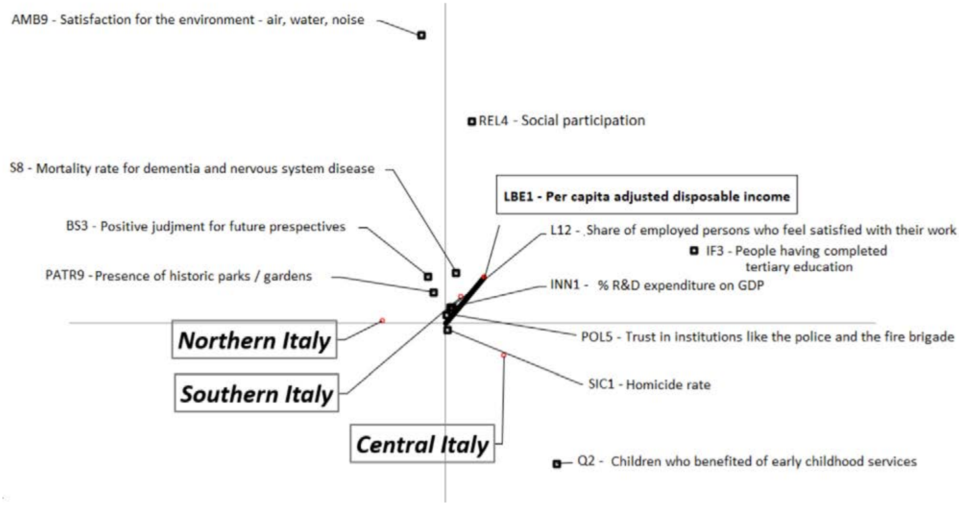
Given a q-variate random vector Y, consider the case when Y is partitioned in n subjects (groups), each of them with n_i individuals ($i = 1,...,m,;j = 1,...,n_i;s = 1,...,q$). We assume that the population model for the n subjects is $y_{i|q\times 1} = B'_{q\times p}x_{i|q\times 1} + A_{i|q\times r}z_{i|r\times 1}$, where B is the matrix of fixed regression coefficients. A_i is a matrix of q-variate r-dimensional vectors of random-effects, with $a_i = vec(A'_i) \sim N(0,\Sigma_a)$, $\Sigma_a = cov(vec(A'_i)) = \{\Sigma_{a,ss'}\}$, and $\Sigma_{a,ss'} = cov(vec(A'_{i,ss'}))$, where $s,s' = 1,...,q$, the $r\times r$ blocks of Σ_a . When $r = p$, the population model is a multivariate *RCM*, with $z_i = x_i$. Given a sample of $N = \Sigma_i\Sigma_j n_{ij}$ units (e.g. repeated measurements), then the model structure is $Y_{N\times q} = X^+_{N\times p}B_{p\times q} + Z^+_{N\times pm}A_{pm\times q} + E_{N\times q}$, with X^+ the matrix of data covariates, Z^+ the design matrix of random effects, and E the matrix of regression within-subject errors, $cov(vec(E)) = R$. Assuming both Y and X^+ as columnwise centered and standardized, we get in the general RCM setup $cov(a_{si},y_{si}) = DZ'_i = DX'_{si}$. If $F = \widetilde{Y}^{**}var(\widetilde{y})^{-\frac{1}{2}} - X^+B, \widetilde{Y}^{**} = \widetilde{Y} - E(\widetilde{Y})$, and $\widehat{\overline{\beta}} = (\overline{X}'\overline{\Sigma}^{-1}\overline{X})^{-1}\overline{X}'\overline{\Sigma}^{-1}\widetilde{y}^*$, $\beta = vec(B)$, $\Sigma = var(\widetilde{y}) = E\left\{(\widetilde{y}^*_{s'} - y^*_{s'})'(\widetilde{y}^*_s - y^*_s)\right\}$, the singular value decomposition of $\widehat{\widetilde{Y}}^{**} = \overline{Y} = X^+\widehat{\overline{\beta}}$ gives the common rescaled predictor's coordinates, $U_{\overline{Y}}\Lambda_{\overline{Y}}V'_{\overline{Y}}$, further noticing that $U^*_{\widetilde{Y}^{**}} = \widetilde{Y}V^{-1}\Lambda_{\overline{Y}}$ contains the row joint reduced coordinates in the space of $\widetilde{Y}^{**}$. In accordance with recent law reforms, the Equitable and Sustainable Well-being indicators (BES) - annually provided by the Italian Statistical Institute(ISTAT, 2017) - are designed to define the economic policies which largely act on some fundamental aspects of the quality of life. In order to highlight the result of the proposed *RDA,* we use 12 BES indicators relating to the years 2013-2016, collected at NUTS-2 level. In particular the variables are S8 (Age-standardised mortality rate), IF3 (People with tertiary education), L12 (Satisfaction with

Figure 1.



job), REL4 (Social participation), POL5 (Trust in institutions), SIC1 (Homicide rate), BS3 (Positive judgment for future perspectives), PATR9 (Presence of Parks/Gardens), AMB9 (Satisfaction for life), INN1 (Percentage of R&D expenditure), Q2 (Childhood services) and LBE1 (logarithm of per-capita adjusted disposable income). We use the latter as the predictor variable in the *RCM*, while the remaining 11 variables are dependent variables. The application uses the restricted maximum likelihood estimation, inside a SAS/IML code. To simplify the estimation process, we assume equicorrelation between the multivariate components of random effects. The linear mixed model with random coefficients highlights its analytical capabilities in the Figure 1. The plot features the standardized best predictors (STDP) and the original criterion variables in the space of the latter. As an example, while there is no correlation between INN1 (R&D expenditure) and AMB9 (satisfaction for the environment), the same best predictor variables register an inverse correlation. This evidence is supported by arguments, such as critical differences among Northern and Southern Italian Regions. Figure 2 shows the constrained RDA, in which the major contribution is given by the variables IF3, Q2, REL4, and AMB9. Interpreting the correlation between random effects and the LBE1 model covariate, this dependence is mainly explained indeed by the amount of the population that completed tertiary education (IF3). This means that most of the differences between Italian Regions reflect the dependence of the IF3 variable on the disposable income.

Figure 2.



References

DEMIDENKO, E. 2004. *Mixed Models: theory and applications*. Wiley and Sons.

MARCIS, L., & SALVATORE, R. 2020. Joint Redundancy Analysis by a Multivariate Linear Predictor. *Conference of the Italian Statistical Society*.

THE USE OF MULTIPLE IMPUTATION TECHNIQUES IN SOCIAL MEDIA DATA

Paolo Mariani [1], Andrea Marletta [1] and Matteo Locci [1]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, (e-mail: andrea.marletta@unimib.it, paolo.mariani@unimib.it, m.locci2@campus.unimib.it)

ABSTRACT: In the big data context, it is very frequent to manage the analysis of missing values. This is especially relevant in the field of statistical analysis, where this represents a thorny issue. This study proposes a strategy for data enrichment in presence of sparse matrices. The research objective consists in the evaluation of a possible distinction of behaviour among observations in sparse matrices with missing data. After selecting among the multiple imputation methods, an innovative technique will be presented to impute missing observations as a negative position or a neutral opinion. This method has been applied to a dataset measuring the interaction between users and social network pages for some Italian newspapers.

KEYWORDS: Social network data, Missing values, Multiple imputations

1 Introduction

The treatment of missing values is still a neglected phase in the field of quantitative analysis. In not statistical contexts, the most abused solution is the row elimination, that is to say the deletion of the observation with missing values. This operation could result misleading and the treatment of missing observations is more complex procedure. Firstly, it is necessary to conduct some preliminary analysis about the nature of this lack of information and to recognize the mechanism of the missing data. This relationship aims to evaluate the link between the observed value and the missing one. This lead to the well-known classification of Little and Rubin (Little, 1988) in MCAR (Missing Completely At Random) data, MAR (Missing at Random) data o NMAR (Not Missing At Random) data. Only after the identification of these mechanism, it is possible to find the best solution to solve the problem of missing values. If the complete case analysis has not been considered as a valid alternative, it is necessary to proceed with the imputation of the missing observation.

In this study, an innovative technique to discern the missing value from a behaviour for some individuals has been proposed using two steps. In the first step, the substitution of missing values is implemented using a threshold based on the number of expressed “Likes”. In particular, a missing value is considered as a “Dislike” only when a user has expressed a percentage of “Likes” that is higher than a selected threshold. Alternatively, if the percent of “Likes” is less than the threshold, a missing “Like” is imputed as a “Nothing”. The second step has been pursued using a multiple imputation technique known as MIMCA method (Multiple Imputation with Multiple Correspondence Analysis) (Audigier *et al.*, 2017). This procedure is applied to social media data from the official pages of 7 Italian newspapers.

2 The MIMCA approach

Multiple Imputation with Multiple Correspondence Analysis represents an available alternative as imputation technique for qualitative data. This approach allows to impute data sets with incomplete categorical variables. The principle of MI with MCA, as well as all the other multiple imputation techniques, consists in creating M different datasets to reflect the uncertainty on imputed values. In this context, each dataset is obtained with an algorithm called *iterative MCA*, which is useful to impute qualitative data. The iterative MCA algorithm consists in recoding the incomplete dataset as an incomplete disjunctive table Z, randomly imputing the missing values, estimating the principal components and loadings from the completed matrix and then, using these estimates to impute missing values according to the following reconstruction formula:

$$\hat{Z} = \hat{U}\hat{\Lambda}\hat{V}^T + M.$$

where $\hat{U}$, $\hat{\Lambda}$ and $\hat{V}$ are the left singular vectors, the diagonal matrix of singular values and the right singular vectors, respectively. The final version of matrix Z is obtained as $Z = W * Z + (\mathbb{I} - W) * \hat{Z}$, where $*$ is the Hadamard product and W is a matrix of weights where $w_{ij} = 1$ if z_{ij} is missing and $w_{ij} = 0$ otherwise. In this context, MCA is configured as a singular values decomposition applied on the triplet of matrices $(Z - M, \frac{1}{K}D_{\Sigma}^{-1}, R)$. The matrix Z represents the disjunctive table, M is a matrix whose rows are equal to the vectors of the means of each component of Z, D_{Σ} is a diagonal matrix with the proportions of individuals characterized by a specific category and R is the matrix of uniform weights assigned to individuals. After a first step of imputation, the procedure of iterative MCA is repeated many times until a convergence criterion is reached. In many cases, due to overfitting problems, a regularized version of this algorithm is used (Josse *et al.*, 2012).

This approach is part of the family of joint modelling MI method, which means that it is more computationally efficient than conditional models. In fact, this MI technique is based on Multiple Correspondence Analysis and then the number of parameters estimated is small. Another advantage of MI with MCA is the goodness of estimation even if the number of individuals is small. Finally, MI with MCA well represents less frequent categories in the step of imputation. This last is another property that derives from MCA.

3 Application on Italian newspaper social pages

The dataset used for this application is represented by users that expressed at least one “Like” in social media pages, websites, and forums concerning drugs and health. The research was conducted on 2,795 Italian subjects considering all interactions between people and brands and between products and services on Facebook. The selected category for Facebook pages is Italian newspapers. Each column of the dataset is a dummy variable that represents the presence or absence of a “Like.” The 7 Italian newspapers are: La Repubblica, Corriere della Sera, Il Fatto Quotidiano, Il Sole 24 Ore, La Gazzetta dello Sport, Il Messaggero, La Stampa.

Before performing the MIMCA approach, the entire dataset has been divided into training and validation set. In particular, a number of cells equivalent to 30% of the cells observed has been set to “missing value”. In order to create a validation set similar to the original data set, the proportion of each category (“Like”, “Dislike” and “Nothing”) has been maintained. The number of multiple data sets generated is equal to 100. The category to be imputed is selected by the majority rule. In other terms, among 100 imputations for each cell of the validation set, the category imputed at least 34 times is selected.

In order to evaluate the performances of MI with MCA, a confusion matrix has been created and summarized through an index of accuracy. This approach imputes more than 81% of the cells considered. Then, the performance of this technique is satisfactory.

Once the goodness of MIMCA has been proved, the process of data enrichment about cells without a category observed or imputed can be completed. In order to achieve this goal, $M = 100$ data sets have been imputed with MIMCA and, for each cell with a missing value, only those where a “Dislike” or a “Nothing” has been imputed are considered. Moreover, in order to minimize the simulation error due to the application of a bootstrap procedure, a threshold has been introduced. More specifically, considering only the data sets where a “Dislike” or a “Nothing” has been imputed, the imputation rule for each cell is the following:

- if the proportion of “Dislike” imputed is greater than or equal to 60%, then a “Dislike” is imputed;
- if the proportion of “Dislike” imputed is less than or equal to 40%, then a “Nothing” is imputed;
- if the proportion of “Dislike” is between 40% and 60%, then neither a “Dislike” nor a “Nothing” is imputed.

Table 1. *Distribution of “Like”, “Dislike” and “Nothing” after MI with MCA.*

	La Repubblica	**Corriere della Sera**	**Il Fatto Quotidiano**	**Il Sole 24 Ore**
“Like”	299	244	268	158
“Dislike”	24	8	57	47
“Nothing”	149	235	139	197
Missing Values	24	9	32	94
	La Gazzetta dello Sport	**Il Messaggero**	**La Stampa**	**Total**
“Like”	116	66	86	1237
“Dislike”	221	255	200	812
“Nothing”	155	169	170	1214
Missing Values	4	6	40	209

As can be noted from the table 1, few missing values are still present. In fact, in some cases the number of “Dislike” imputed in a specific cell is very similar to the number of “Nothing”. In particular, this behaviour is manifested when the proportion of “Dislike” (or “Nothing”) is between 40% and 60%. Even if there are some cases where missing values are not imputed, MI with MCA works well. In fact, the proportion of missing values is now equal to 6%.

References

AUDIGIER, V, HUSSON, F, & JOSSE, J. 2017. MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and computing*, **27**(2), 501–518.

JOSSE, J., CHAVENT, M., LIQUET, B., & HUSSON, F. 2012. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of classification*, **29**(1), 91–116.

LITTLE, R. 1988. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, **83**(404), 1198–1202.

PREDICTION OF GENE EXPRESSION FROM TRANSCRIPTION FACTORS AFFINITIES: AN APPLICATION OF BAYESIAN NON-LINEAR MODELLING

Federico Marotta[1], Paolo Provero[1] and Silvia Montagna[2,3]

[1] Dipartimento di Neuroscienze "Rita Levi Montalcini", Università degli Studi di Torino, Via Cherasco, 15, 10126, Torino, Italy (e-mail: federico.marotta@edu.unito.it, paolo.provero@unito.it)

[2] Dipartimento di Scienze Economico-sociali e Matematico-statistiche, Università degli Studi di Torino, Corso Unione Sovietica, 218/bis, 10134 Torino, Italy, (e-mail: silvia.montagna@unito.it)

[3] Collegio Carlo Alberto, Piazza Vincenzo Arbarello, 8, 10122 Torino, Italy

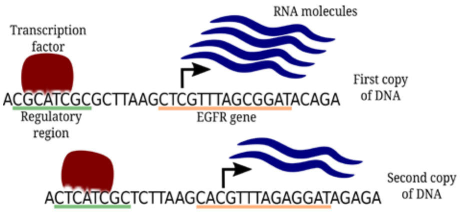
ABSTRACT: The prediction of gene expressions from DNA sequences is a relevant problem in biology. While most of the existing methods dedicated to this task use genotypes as predictors, here we propose a method based on transcription factor affinities, which have a clearer biological interpretation. This novelty, however, introduces new challenges for modelling, which we address leveraging on Bayesian non-linear modelling techniques.

KEYWORDS: Bayesian Methods, Gene Expressions, Non-linear Predictive Modelling.

1 Introduction

Scientists are often interested in predicting differences in the expression of a gene in different individuals solely from the DNA sequence of the individuals. The predicted expression can then be used in place of the real one when measuring the latter is too expensive, and the learnt relationship between DNA and expression can lead to a better understanding of how genes are regulated (Manor & Segal, 2013). The expression of a gene is the amount of RNA molecules it produces. Humans have two independent sets of DNA molecules, one coming from the father and one from the mother, therefore there are two copies of each gene. When measuring the expression, one simply sums the molecules produced by each copy.

When associating DNA to gene expressions, the first problem we face is how to encode the DNA (a 3-billion letter string from the alphabet $\{A, C, G, T\}$)



Person	log Expr. of EGFR	Affinity of TF 1 for copy 1 of region 1	…	Affinity of TF l for copy 2 of region r
Alice	3.5	8.4	…	1.1
Bob	4.1	7.7	…	0.6
Craig	3.3	9.4	…	0.5
Dave	3.8	10.2	…	0.8
Eve	3.4	8.1	…	1.2
Frank	4.2	9.5	…	0.5

Response (log $\mathbf{y}$) Predictors ($\mathbf{A}$)

Figure 1. *Left: Humans have two copies of DNA in each cell; the expression of a gene is the amount of RNA it produces. Transcription factors bind the DNA at the regulatory regions from where they activate or inhibit the expression of their target gene. Right: A plausible instance of our data set.*

into numbers to be used in a regression model. Most existing methods rely on genotypes (discrete variables taking values 0, 1, or 2 encoding single-letter differences in the DNA of different people), which do not allow for easy interpretation (*e.g.*, "If the DNA has an 'A' instead of a 'T', the expression of the gene will be higher"). Our first goal is to develop a more interpretable model.

Gene expression is mainly controlled by specialised proteins called transcription factors, which bind the DNA at particular locations (regulatory regions) by establishing weak chemical bonds. Different DNA sequences will have, therefore, different chemical *affinities* for the transcription factors. Since different individuals have different DNA sequences, it is possible to use the affinities for transcription factors as numerical (continuous) predictors in the predictive model of gene expression. Affinities have a far superior interpretation, exemplified by statements such as "If the affinity for this transcription factor is higher, the expression will be higher."

However, one needs to make an assumption about the relationship (*e.g.*, linear) between affinities and gene expression. de Boer *et al.* , 2020 models the logarithm of the expression as a linear function of the affinities. The model is developed for a type of yeast and achieves a good performance, but is still too simple for our application. Indeed, yeast has two important distinguishing features: 1) it is haploid, meaning that it has only one copy of DNA, whereas humans have two; and 2) its genes are regulated primarily by one regulatory region, whereas human genes typically have more than one.

In this paper, we set up a predictive model for the expression of the EGFR (Epidermal Growth Factor Receptor) gene, and explicitly address both limitations in de Boer *et al.* , 2020. Figure 1 provides a schematic of our application.

2 Methodology and results

Our dataset consists in the expression values of the EGFR gene for $n = 414$ individuals (from The GTEx Consortium, 2020), and in the affinity of each regulatory region for all transcription factors, for a total of $p = 358$ predictors.

We can take multiple regulatory regions into account (goal 2 above) via a straightforward modification of the model in de Boer *et al.* , 2020, which becomes: $\log(y) = \beta_0 + \sum_{g=1}^{r} \sum_{f=1}^{l} \mathrm{A}_{fg}\beta_{fg}$. Here y denotes the gene expression, $\{A_{fg}\}_{f=1,g=1}^{l,r}$ are the affinities, and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{rl})^\top$ is a vector of model parameters. Similarly to de Boer *et al.* , 2020, we sum over all transcriptions factors, indexed by f, but now also along the regulatory regions, g, of the gene.

Accommodating for both copies of DNA (goal 1 above) is more challenging. Biologically, we know that the effects of the two copies should be additive in the original scale of the expression, not in the log-transformed expression. At the same time, working with the expression in the original scale can be troublesome, for it is often not normally distributed. Therefore, we propose the following model for the expression of a single gene:

$$\log(\boldsymbol{y}) \sim \text{mvnormal}\left(\log\left(e^{\boldsymbol{A}^{(1)}\boldsymbol{\beta}} + e^{\boldsymbol{A}^{(2)}\boldsymbol{\beta}}\right), \sigma^2 \boldsymbol{I}\right). \tag{1}$$

Here $\boldsymbol{y}$ is an n-vector of expression values (one for each individual), $\boldsymbol{A}^{(i)}$, with $i \in \{1,2\}$, is the $n \times rl$ affinity matrix for copy i, where each column represents a transcription factor-regulatory region pair (r is the number of regions, l the number of transcription factors), and vector $\boldsymbol{\beta}$ ($lr \times 1$) encapsulates the coefficients of the affinities. By computing the exponential of $\boldsymbol{A}^{(i)}\boldsymbol{\beta}$, with $i \in \{1,2\}$, we obtain the effect of copy i on the expression in the original scale. We subsequently sum the two effects, and take the log of the sum to go back to the log-scale response. Importantly, the coefficient of a given transcription factor in a given regulatory region is the same for the two copies of DNA. We notice that for this reason our model does not fall in the class of generalised linear models (at least not obviously), as each coefficient β_j appears two times independently for two different predictors.

Model (1) is embedded in a Bayesian framework by placing a normal prior (with mean zero and variance τ) on all coefficients $\boldsymbol{\beta}$ independently, and a non-informative Jeffreys prior on σ^2. The Bayesian framework is chosen for two primary reasons: 1) Although in our specific application $p < n$, generally in genomics $p \gg n$ and regularisation is needed. Regularisation can be thought of as imposing a Bayesian prior on the underlying parameters. Specifically, the ridge regression estimator can be viewed as the Bayesian posterior mean

Table 1. *Results of the nested-cross validation. MSE is the mean squared error,* ρ *the correlation between true and predicted expression; averages and standard deviations of these quantities are computed across the 5-folds. Avg R^2 is the average of the squared correlations. Z is the Z-score computed via Stouffer's method, which combines the* ρ *of the five folds, and pval Z is the p-value of the Z-score.*

Gene	Avg MSE	Sd MSE	Avg ρ	Sd ρ	Avg R^2	Z	pval Z
EGFR	0.011	0.003	0.199	0.065	0.043	4.030	2.8e-5

estimator of $\boldsymbol{\beta}$ when imposing a Gaussian prior on $\boldsymbol{\beta}$. 2) Ability to encode knowledge via priors distributions. For example, we can exploit existing biological data about which transcription factors are bound to a region of DNA by giving the corresponding variables a less stringent regularisation.

To carry out an unbiased evaluation of the performance, we implemented a 5-fold cross-validation strategy. Table 1 summarises the results. While the average R^2 may seem small, we emphasise that low values are common in the prediction of gene expression and our model outperforms recently published genotype-based models (the R^2 achieved by Nagpal *et al.* , 2019 is only 0.005).

Thus, our method can model the underlying biological problem in a realistic way and provide meaningful results thanks to its interpretable predictors. In the future, it could be improved by considering interactions between transcription factors, which are also biologically important. Nevertheless, for the time being, we hope that non-linear models will find their way in the field of gene expression prediction, which currently is dominated by genotype-based linear models.

References

de Boer, Carl G., Vaishnav, Eeshit D., Sadeh, Ronen, *et al.* . 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.*, **38**(1), 56–65.

Manor, Ohad, & Segal, Eran. 2013. Robust Prediction of Expression Differences among Human Individuals Using Only Genotype Information. *PLoS Genet.*, **9**(3), e1003396.

Nagpal, Sini, Meng, Xiaoran, Epstein, Michael P., *et al.* . 2019. TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *Am. J. Hum. Genet.*, **105**(2), 258–266.

The GTEx Consortium. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**(6509), 1318–1330.

HIGH DIMENSIONAL MODEL-BASED CLUSTERING OF EUROPEAN GEOREFERENCED VEGETATION PLOTS

Francesca Martella [1], Fabio Attorre [2], Michele De Sanctis [2] and Giuliano Fanelli [2]

[1] Department of Statistical Sciences, Sapienza University of Rome, (e-mail: francesca.martella@uniroma1.it)

[2] Department of Environmental Biology, Sapienza University of Rome, Italy (e-mail: fabio.attorre@uniroma1.it, michele.desanctis@uniroma1.it, giuliano.fanelli@gmail.com)

ABSTRACT: An important challenge in complex vegetation systems is the classification of vegetation since it represents a useful tool for summarizing our knowledge of vegetation patterns and, consequently, for nature conservation, landscape mapping and land-use planning. It typically requires standard clustering methods that are capable of identifying groups of plots characterized by dominant and diagnostic species. When the data are high-dimensional, however, efficient clustering methods have to be considered. In this paper, we consider a robust model-based clustering, called Gaussian mixture models for high-dimensional data (HD-GMM) which takes into account for the specific subspace around which each cluster is located and, consequently, provides parsimonious modeling. Results are encouraging and deserve further discussion.

KEYWORDS: vegetation plots, high-dimensional data, finite mixture models

1 Introduction

Improving actions for nature conservation, landscape mapping and land-use planning is a key point in vegetation Science. The need for ecologists to develop appropriate management and conservation strategies has been widely recognized. The identification of homogeneous vegetation communities provides a useful way of summarizing our knowledge of vegetation in a certain area. Clustering represents an important tool to discover such communities and, in general, to draw insights from vegetation data. Summary of vegetation clustering methods can be found in several proposals that focus on this discipline (Sun et al., 1997). Attorre et al. (2020) propose a finite mixture model (FMM) for classifying georeferenced vegetation plots present in the Italian peninsula, including the two main islands (Sicily and Sardinia), but excluding the Alps and the Po plain, according to species composition and environmental variables. Previously, FMM has been applied to identify marine bioregions on the Western Australian continental margin (Woolley et al., 2013) and forest physiognomic types in Italy (Attorre et al., 2014). However, when we face with high-dimensional vegetation data, FMM, or more specifically, standard model-based clustering techniques, may show a disappointing behavior. This is mainly due to the fact that the number of parameters to be estimated usually depends on the dimension of the observed space and such approaches may therefore suffer from the so-called curse of dimensionality (Bellman, 1957). In this paper, we suggest the use of a robust model-based clustering, named Gaussian mixture models for high-dimensional data (HD-GMM) proposed by Bouveyron et al. (2007). We examine a database of 7955 georeferenced plots and 3181 plant species of evergreen forest vegetation, created in TURBOVEG by storing published and unpublished phytosociological plots collected over the last 30 years. These plant communities are scattered along the Mediterranean coastal area, whose main and distinctive ecological feature is the prolonged aridity in summer and rainfall mainly concentrated during winter and spring. Making use of HD-GMM, we assume that high-dimensional vetegation data live in subspaces with a dimensionality lower than the dimensionality of the original plant species space, limiting the number of parameters to estimate and, consequently, the computational time. Finally, as in the FMM framework, the plots are classified based on their plant species composition through a posteriori specific-plot probability and the clusters are defined to be homogeneous in that they include plots that show similar vegetation.

2 The model

Let $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ be the abundance data matrix, where the generic element y_{ij} represents the value of the measure of abundance for the j-th tree species, namely the percentage of biomass of a certain species with respect to the total biomass of vegetation, observed in the i-th plot of the study area, $(i = 1, \ldots, n,\ j = 1, \ldots, p)$. FMM assumes that each plot $\mathbf{y}_i$ is drawn from a mixture of K components in some unknown mixing proportions $\pi_1, \ldots, \pi_K$, with $\sum_{k=1}^{K} \pi_k = 1$. Each component identifies a cluster. When a (multivariate) Gaussian density is used to describe the component-specific distribution of observed plant species cover, the component is identified by a specific center, defined by the mean vector (as the observed values are on abundance scale, we may hypothesize that similar plots will be characterized by similar values of abundance of the same species), and a specific shape, summarized by the covariance matrix, which allows for varying dependence between cover values corresponding to different plant species for plots in that component. In other words, $\mathbf{y}_i$ has density function defined by:

$$f(\mathbf{y}_i|\Psi) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i \mid \mu_k, \Sigma_k), \tag{1}$$

where $\phi(\cdot)$ represents the cluster-specific p-variate Gaussian density with vector mean μ_k and covariance matrix Σ_k, for $k = 1, \ldots, K$, and $\Psi = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$ denotes the overall parameter vector. Unfortunately, FMM requires the estimation of a very large number of parameters (proportional to p^2) and therefore faces numerical problems in high-dimensional spaces. In this respect, HD-GMM assumed that high-dimensional data live around subspaces with a dimension lower than the considered species number, limiting to estimate the specific subspace and the cluster-specific intrinsic dimension. Formally, HD-GMM considers the following eigen-decomposition of the cluster-specific covariance matrix Σ_k:

$$\Sigma_k = \mathbf{D}_k^t \mathbf{A}_k \mathbf{D}_k \tag{2}$$

where $\mathbf{D}_k$ is a $(p \times p)$ orthogonal matrix having as columns the eigenvectors of Σ_k and $\mathbf{A}_k$ is a $(p \times p)$ diagonal matrix which contains the associated eigenvalues (sorted in decreasing order), $k = 1, \ldots, K$. It follows that, $\mathbf{A}_k$ represents the cluster-specific covariance matrix in the eigenspace of Σ_k. Moreover, it is assumed that $\mathbf{A}_k$ is reparametrized as a diagonal matrix having only $q_k + 1$ different eigenvalues:

$$\mathbf{A}_k = \mathrm{diag}(a_{k1}, \ldots, a_{kq_k}, b_k, \ldots, b_k), \tag{3}$$

with $a_{kj} > b_k$, $j = 1, \ldots, q_k$, $q_k \in \{1, \ldots, p-1\}$. In this way, the parameters a_{kj} describe the cluster-specific variance of the original data, while the unique parameter b_k models the variance of the noise which is isotropic and contained in a subspace, which is orthogonal to the subspace of the k-th cluster. The dimension q_k is unknown and represents the dimension of the cluster-specific subspace $\mathbb{E}_k$ which is spanned by the q_k first columns of $\mathbf{D}_k$, i.e. by the q_k first eigenvectors corresponding to the eigenvalues a_{kj}, with $\mu_k \in \mathbb{E}_k$. Notice that, if $q_k = p - 1$ for all $k = 1, \ldots, K$ then HD-GMM reduces to FMM. Following the classical parsimony strategy, a family of 28 parsimonious HD-GMMs is defined by constraining some (or all) parameters to vary within and between clusters. The more general HD-GMM is denoted by $[a_{kj} b_k \mathbf{D}_k q_k]$.

3 Conclusion

Thanks to the significant reduction of the number of parameters to be estimated, HD-GMM seems to be a promising approach when dealing with the analysis of high-dimensional complex vegetation systems data. This modelling may effectively highlight specific subspaces in the geographical patterns helping the interpretation of the clustering results.

References

ATTORRE, F., FRANCESCONI, F., DE SANCTIS, M., ALF, M., MARTELLA, F., VALENTI, R., VITALE, M. 2014. Classifying and Mapping Potential Distribution of Forest Types Using a Finite Mixture Model. *Vegetation Folia Geobotanica.*, **49**, 313–335.

ATTORRE, F., CAMBRIA, V.E., AGRILLO, E., ALESSI, N., ALFO, M., DE SANCTIS, M., MALATESTA, L., SITZIA, T., GUARINO, R., MARCEN, C., MASSIMI, M., SPADA, F., FANELLI, G. 2020. Finite Mixture Model-based classification of a complex vegetation system. *Vegetation Classification and Survey.*, **1**, 77–86.

BELLMAN, R. 1957. Dynamic Programming. *Princeton University Press.* Princeton.

BOUVEYRON, C., GIRARD, S., SCHMID, C. 2007. High-dimensional data clustering. *Computational Statistics and Data Analysis.*, **52(1)**, 502–519.

CATTELL, R. 1966. The scree test for the number of factors. *Multivariate Behavioral Research.*, **1(2)**, 145–276.

SCHWARZ, G. 1978. Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

SUN, D., HNATIUK, R.J., NELDNER, V.J. 1997. Review of vegetation classification and mapping systems undertaken by major forested land management agencies in australia.. *Australian Journal of Botany.*, **45(6)**, 929–948.

WOOLLEY, S.N.C., MCCALLUM, A.W., WILSON, R., OHARA, T.D., DUNSTAN, P.K., 2013. Fathom out: biogeographical subdivision across the Western Australian continental margin a multispecies modelling approach. *Diversity and Distributions.*, **19**, 1506–1517.

MULTIVARIATE OUTLIER DETECTION FOR HISTOGRAM-VALUED VARIABLES

Ana Martins [1], Paula Brito[2] , Sónia Dias [3] and Peter Filzmoser [4]

[1] Institute of Electronics and Informatics Engineering of Aveiro, Aveiro, Portugal (e-mail: a.r.martins@ua.pt)

[2] Faculdade de Economia, Universidade do Porto & LIAAD-INESC TEC, Porto, Portugal (e-mail: mpbrito@fep.up.pt)

[3] Instituto Politécnico de Viana do Castelo & LIAAD-INESC TEC, Portugal (e-mail: sdias@estg.ipvc.pt)

[4] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria (e-mail: peter.filzmoser@tuwien.ac.at)

ABSTRACT: A measure for outlier detection of multivariate histogram-valued variables based on the Mallows (SDO^2_M) distance is proposed. A case study with distributional data of repeated measurements of 10 patients' hematocrit and hemoglobin is presented. The $Q_3+3(Q_3-Q_1)$ criteria and, P_{95} and $P_{97.5}$ of a Chi-Square distribution with p-degrees of freedom (p number of variables) are used as cut-offs. Overall, the SDO^2_M along with the P_{95} cut-off are able to detect outliers in most analysed situations.

KEYWORDS: histogram-valued data, Mallows distance, outlier detection.

1 Introduction

Symbolic data were introduced to better describe and analyse data with intrinsic variability. Descriptive statistics (e.g., mean, median) and multivariate data analysis methods (e.g.,linear regression) for histogram-valued data analysis have been developed. Outlier analysis has first been addressed by Verde *et al.*, 2014. Following a different approach, we introduce a method for multivariate outlier analysis based on the Mallows distance. We define outlier as a data unit which is far apart from the center of the data cloud, here the barycenter. Results on a case study are presented.

2 Methods

Let $S=\{s_1,\dots,s_n\}$ be the set of entities under analysis, B the set of probability of frequency distributions over a set of sub-intervals $\{I_{i1},\dots,I_{iK_i}\}$ of an underlying domain $O \subseteq \mathbb{R}$, a histogram-valued variable is defined by a mapping $Y : S \rightarrow B$. Each realisation i of the histogram-valued variable, $Y(s_i)$, may be represented by the histogram

$$H_{Y(s_i)} = \left\{ \left[\underline{I}_{i1}, \bar{I}_{i1}\right[, p_{i1}; \ldots, \left[\underline{I}_{iK_i}, \bar{I}_{iK_i}\right], p_{iK_i} \right\}, \tag{1}$$

where $p_{i1} + \cdots + p_{iK_i} = 1$. Also, it is assumed that within each sub-interval $\left[\underline{I}_{i\ell}, \bar{I}_{i\ell}\right[$ the values of variable $Y(s_i)$ are uniformly distributed. Another representation of the histogram-valued variables is the quantile function,

$$\phi_i(t) = \begin{cases} \underline{I}_{i1} + \frac{t}{w_{i1}} r_{i1} & \text{if } 0 \leq t \leq w_{i1} \\ \underline{I}_{i2} + \frac{t - w_{i1}}{w_{i2} - w_{i1}} r_{i2} & \text{if } w_{i1} \leq t \leq w_{i2} \\ \vdots & \\ \underline{I}_{iK_i} + \frac{t - w_{iK_i - 1}}{1 - w_{iK_i - 1}} r_{iK_i} & \text{if } w_{iK_i - 1} \leq t \leq 1, \end{cases} \tag{2}$$

where $w_{ih} = \sum\limits_{\ell=1}^{h} p_{i\ell}$, $h = 1, \ldots, K_i$ and $r_{i\ell} = \bar{I}_{i\ell} - \underline{I}_{i\ell}$ for $\ell = \{1, \ldots, K_i\}$. The quantile functions are piecewise linear and even though the space of the quantile functions is only a semi-vector space, the arithmetic operations are simpler with this representation, which is preferred to represent histogram-valued data.

To identify multivariate outliers we propose a measure based on the Mallows distances to the multivariate means of quantile functions. The Mallows distance (d_M) is defined as $d_M\big(\phi_i(t), \phi_j(t)\big) = \sqrt{\int_0^1 \big(\phi_i(t) - \phi_j(t)\big)^2 dt}$, and the multivariate mean of quantile functions is the barycenter ($\bar{\phi}_b$), which is the solution of the minimisation problem: $Min \sum\limits_{i=1}^{n} \int_0^1 (\phi_i - \bar{\phi}_b)^2 dt$, leading to $\bar{\phi}(t) = \frac{1}{n} \sum\limits_{i=1}^{n} \phi_i(t)$, $t \in [0, 1]$. To easily implement this method, the approach by Hubert *et al.*, 2015 is adopted and, an outlyingness measure based on a one-dimensional projection of the observed data is computed. Thus, the Mallows outlyingness measure is

$$SDO_{M_i^2} = \sup_{||v||=1} \frac{d_M^2\left(\phi_i(t)v, \frac{1}{n} \sum\limits_{\substack{j=1 \\ j \neq i}}^{n} \phi_j(t)v\right)}{\frac{1}{n-1} \sum\limits_{i=1}^{n} d_M^2\left(\phi_i(t)v, \frac{1}{n} \sum\limits_{\substack{j=1 \\ j \neq i}}^{n} \phi_j(t)v\right)}, \tag{3}$$

where $v = (a_1, \ldots, a_p, b_1, \ldots, b_p)$ runs through a set of $2p$-dimensional vectors (for p variables) that project the histogram-valued data into a one-dimensional space, using the definition of linear combination proposed by Dias & Brito, 2015, that solves the problem of the semi-linearity of the space:

$$\phi_{\hat{W}_i}(t) = a_1 \phi_{X_{1i}}(t) - b_1 \phi_{X_{1i}}(1-t) + \cdots + a_p \phi_{X_{p_i}}(t) - b_p \phi_{X_{p_i}}(1-t), \qquad (4)$$

where $\phi_{X_{1i}}(t), \ldots, \phi_{X_{p_i}}(t)$ are the quantile functions of the observed histograms and $\phi_{X_{1i}}(1-t), \ldots, \phi_{X_{p_i}}(1-t)$ are the quantile functions of the corresponding symmetric histograms, $a_u, b_u \geq 0, u \in \{1, \ldots, p\}$, and $t \in [0,1]$.

To flag observations as outliers, possible alternative cut-offs for $SDO_{M_i^2}$ are Tukey's boxplot $Q_3 + 3(Q_3 - Q_1)$ criterion, and, by analogy with the classical case where the Mahalanobis distance to the mean is used (see Filzmoser *et al.*, 2005), the P_{95} or the $P_{97.5}$ of a Chi-Square distribution with p-degrees of freedom ($p =$ number of variables).

3 Case Study

An analysis of SDO_M^2 using distributional data of repeated measurements of the hematocrit (Y) and hemoglobin (X) values for 10 patients (Billard & Diday, 2006) was conducted. First, univariate outliers were considered, by perturbing the distribution of variable X for the first unit, in seven different ways (Out1 to Out7). SDO_M^2 was computed for the original hemogloblin distributions and for the seven situations where unit 1 is now an outlier. Then, bivariate outliers were considered, by perturbing the distributions of both variables for the first unit. SDO_M^2 was computed for all seven cross-situations between outliers of variables X and Y.

3.1 Results and Discussion

The ability of SDO_M^2 to detect outlier observations was studied, using all three cut-offs mentioned above. Overall, P_{95} is the cut-off that works the best to identify outliers in both cases. In the univariate case this cut-off fails to identify Out5 only and, in the bivariate case it fails to identify both Out5 and Out6 (Fig. 1). Note that in the bivariate case the outlier unit 1 for variable Y is fixed (only one perturbation considered). The $Q_3 + 3(Q_3 - Q_1)$ seems to be the worst, but this may reflect the fact that $n = 10$, which means that sample size is small to compute the quantiles. In fact, preliminary studies with larger n, suggest that this is the best cut-off (data not shown). In conclusion, SDO_M^2 with the P_{95} cut-off (and the Tukey's criterion) seem a promising approach to detect outliers in histogram-valued data.

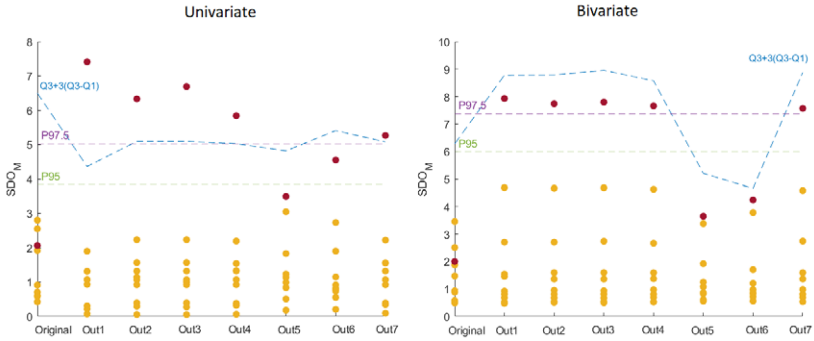


Figure 1. *SDO_M^2 measure for univariate and bivariate outliers and cut-offs P_{95}, $P_{97.5}$ and $Q_3 + 3(Q_3 - Q_1)$. Red dots represent the outlier observation (unit 1).*

References

Billard, L, & Diday, E. 2006. *Symbolic Data Analysis: Conceptual Statistics and Data Mining John Wiley.*

Brito, P. 2014. Symbolic data analysis: another look at the interaction of data mining and statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **4**(4), 281–295.

Dias, S., & Brito, P. 2015. Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining*, **8**(2), 75–113.

Filzmoser, P., Garrett, R.G., & Reimann, C. 2005. Multivariate outlier detection in exploration geochemistry. *Computers & Geosciences*, **31**(5), 579–587.

Hubert, M., Rousseeuw, P.J., & Segaert, P. 2015. Multivariate functional outlier detection. *Statistical Methods & Applications*, **24**(2), 177–202.

Verde, R., Irpino, A., & Rivoli, L. 2014. A box-plot and outliers detection proposal for histogram data: new tools for data stream analysis. *Pages 283–291 of: Analysis and Modeling of Complex Data in Behavioral and Social Sciences*. Springer.

A NONPARAMETRIC TEST FOR MODE SIGNIFICANCE

Giovanna Menardi[1], Federico Ferraccioli[1]

[1] Department of Statistical Sciences, University of Padova,
(e-mail: menardi@stat.unipd.it, ferraccioli@stat.unipd.it)

ABSTRACT: We propose a nonparametric test for the significance of a mode, with the aim of evaluating whether a region of relatively high observed density reflects the actual presence of a mode in the true distribution underlying a set of data. The method leverages on the correspondence between the the mathematical framework of Morse theory and the tools provided by gradient ascent approximation. This allows for building a sequence of asymptotically Normal realisations of the sample distribution of an estimated mode and the definition of a chi-squared test statistic.

KEYWORDS: Asymptotics, Gradient ascent, Hypothesis testing, Kernel estimator.

1 Introduction

Although often overlooked with respect to location measures as mean and median, inference on the modes of a distribution plays a central role in data analysis. In fact, modes represent informative summaries of a distribution, especially when data exhibit non-Gaussian features as, multimodality, skewness, or heavy tails. One question of interest typically arises when somewhat clumped data are observed, often at the tails of the empirical distribution, possibly inducing to wonder if they are real or just a spurious effect of sample variability. Similarly, in many applications where clustering is the final aim, one wishes to evaluate significance of detected groups. In astronomy, for example, a main goal is to establish if clusters of photon emissions are evidence of the presence of celestial energy sources or just express a strong background contamination. This problem has been often neglected by the inherent literature, mostly addressing related aims as the one of testing unimodality of a density function or the number of the modes (Chacón, 2020). Few contributions in the direction of interest are Duong *et al.*, 2008 and Genovese *et al.*, 2016.

In this work we propose a test to evaluate if a specific point is a true mode of the - unknown - probability density function underlying an observed set of data. We take advantage of formal definitions and theory underlying the modal concept of cluster (Chacón, 2015). The rationale we follow relies on the correspondence between the mathematical framework of Morse theory and the tools provided by gradient ascent algorithms. This allows us to define a sequence of realisations shown to be asymptotically normal, to approximate the sample distribution of an estimated mode.

2 Modes as critical points of the density

While intuitively clear, the problem of testing mode significance is firstly definitional. The concept of mode itself is, indeed, ambiguous, as for example the Uniform distribution can be regarded to as both unimodal or without modes. To overcome this problem and formalise our framework without any elusiveness, we shall restrict the analysis to smooth distributions, and exclude non-standard ones as, for example, functions with plateau. For our purpose, we resort to the framework provided by Morse Theory, a branch of differential topology which draws the relationship between the stationary points of a smooth real-valued functions on a manifold, and its global topology (Matsumoto, 2002).

Let $(\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ be a sample of realisations from a random variable X with unknown probability density $f : \mathbb{R}^d \to \mathbb{R}$, which we shall assume to be a Morse function, i.e. a function whose critical points are non-degenerate. We can define the autonomous system identified by the the gradient ∇f, to be $\frac{d\mathbf{x}(t)}{dt} = \nabla f(\mathbf{x}(t))$. Given an initial value $\mathbf{x} \in \mathbb{R}^d$, the integral curve $\nu_{\mathbf{x}} : \mathbb{R} \mapsto \mathbb{R}^d$ of the negative gradient $-\nabla f$ is the solution of the initial value problem

$$\nu_{\mathbf{x}}(0) = \mathbf{x} \qquad \nu'_{\mathbf{x}}(t) = -\nabla f(\nu_{\mathbf{x}}(t)), \tag{1}$$

namely, starting at a point $\mathbf{x}$, its integral curve moves it according to the gradient of f, to eventually reach, except for a set of null measure, the destination $\lim_{t\to\infty} \nu_{\mathbf{x}}(t)$. By the Morse theory, the set of destinations $\Theta = \{\theta \in \mathbb{R}^d : \theta = \lim_{t\to\infty} \nu_{\mathbf{x}}(t), \mathbf{x} \in \mathbb{R}^d\}$ is the set of distinct modes of f. Since integral curves never intersect except at critical points, Θ allows to identify a unique partition $\{\mathcal{D}_\theta\}_{\theta\in\Theta}$ of $\mathbb{R}^d$ in distinct regions $\mathcal{D}_\theta = \{\mathbf{x} : \lim_{t\to\infty} \nu_{\mathbf{x}}(t) = \theta\}$ which represent the "basins of attraction" of each mode θ and include all points whose integral curve having them as starting points has destination θ.

In the lack of information about the true modal structure of f, testing the significance of a mode recasts to defining the system of hypotheses

$$H_0 : \theta_0 \in \Theta \quad \text{vs} \quad H_1 : \theta_0 \notin \Theta, \tag{2}$$

for some $\theta_0 \in \mathbb{R}^d$. While apparently composite, the null hypothesis is fact a simple one, as the - yet unknown - partition of $\mathbb{R}^d$ in the set $\{\mathcal{D}(\theta)\}_{\theta\in\Theta}$ allows us to intend H_0 as "θ_0 is the mode of the domain $\mathcal{D}(\theta)$ where it belongs"; hence, under the null hypothesis, it holds: $\theta_0 = \arg\min_{x\in\mathcal{D}(\theta_0)} -f(x)$.

Gradient descent algorithms find iterative solutions to general optimisation problems with suitable smoothness properties. In the current framework, the problem can be faced via the discretisation of the integral curve (1)

$$\theta_{(t+1)} = \theta_{(t)} + \eta \nabla f(\theta_{(t)}), \tag{3}$$

where η is the step size, usually selected to guarantee convergence. In our framework, the target function f may be suitably replaced by a nonparametric kernel estimate $\hat{f} = \frac{1}{n}\sum_{i=1}^{n} K(\frac{\mathbf{x}-\mathbf{x}_i}{h})$, with bandwidth $h > 0$ and kernel K which we take to be a symmetric probability density; hence, the estimated gradient $\hat{\nabla} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \nabla K_h(\mathbf{x}-\mathbf{x}_i)$ is plugged into the (3).

Under regularity conditions, the asymptotic distribution of the kernel gradient estimator (Duong *et al.*, 2008) is shown to be $\nabla \hat{f}(x) \dot{\sim} N\left(\nabla f(x), \frac{1}{n}\Sigma_\nabla\right)$, with $\Sigma_\nabla = [h^{-d+2} R(\nabla K) f(\mathbf{x})]$ and $R(\nabla K)$ a constant depending on the kernel. In order to develop a test statistic for the (2), we adapt to the current framework the rationale of Liang & Su, 2019, which discuss moment-adjusted stochastic gradient descent for optimisation in statistical inference, so that the (3) gets:

$$\begin{aligned} \theta_{(t+1)} &= \theta_t + \eta \nabla \hat{f}(\theta_{(t)}) - \eta[\nabla \hat{f}(\theta_{(t)}) - \mathbb{E}(\nabla \hat{f}(\theta_{(t)}))] \\ &= \theta_{(t)} + \eta h^{\frac{d+2}{2}} R(\nabla K)^{-\frac{1}{2}} \frac{\nabla \hat{f}(\theta_{(t)})}{\sqrt{f(\theta_{(t)})}} - \frac{\eta}{\sqrt{n}} \varepsilon_{(t)} \end{aligned}$$

where the last step comes from a classic standardisation idea and $\varepsilon_{(t)} \sim N(0, I_d)$. Starting from $\theta_{(0)} = \hat{\theta}$, which under H_0 we expect to lie in $\mathcal{D}(\theta_0)$, we may then produce a random sequence $\theta_{(1)}, \ldots, \theta_{(T)}$ of sample modes by simply generating an artificial sample of $\varepsilon_{(t)} \sim N(\mathbf{0}, I_d)$ and applying the update mechanism (4), where f is replaced by $\hat{f}$. H_0 is afterwards rejected for large values of the asymptotically chi-squared distributed test statistic

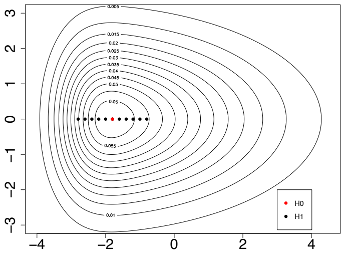
$$(\overline{\theta} - \theta_0)^\top \hat{\Sigma}_\theta^{-1} (\overline{\theta} - \theta_0) \dot{\sim} \chi^2_d,$$

where $\overline{\theta}$ and $\hat{\Sigma}_\theta$ are respectively the mean and covariance matrix of the sequence of sample modes $\theta_{(1)}, \ldots, \theta_{(T)}$.

3 Empirical study

A simulation study has been run to evaluate the behaviour of the proposed test with respect to the Type-I probability error and the power. The simple rule of thumb of selecting h as asymptotically optimal for Normal data has been used, and T has been set to 500.

Figure 1 displays the results - associated to the best value of η - obtained



$d(\theta,\theta_0)$	$\mathbb{P}$(reject H_0)	
	left	right
0	0.06	0.06
0.2	0.13	0.05
0.4	0.29	0.08
0.6	0.57	0.2
0.8	0.78	0.36
1.00	0.92	0.55

Figure 1. *Contour plot of a density and associated estimated Type I error probability for a nominal* $\alpha = 0.05$ *and power for increasing distance from* H_0.

by drawing 500 samples of size $n = 1000$ from a bivariate skew distribution. The test shows an overall good control of type I error, with a slight tendency to be anti-conservative. The power rightly increases as the true mode departs from θ_0, with higher values associated to the steepest side of the density, and lower ones to the most gentle side. This confirms that the testing problem is strictly related to the local curvature around the true mode, and hence to the eigenvalues of the Hessian. For brevity, further results are not reported here, broadly confirming the illustrated behaviour. More challenging settings, such as multimodal ones with overlapping modal regions, in general require a higher sample size to guarantee the control of Type I probability error.

Further discussion is needed to provide insights about the test in higher dimensional settings, along with the sensitivity to different choices of η and h.

References

CHACÓN, J. 2015. A population background for nonparametric density-based clustering. *Stat. Sc.*, **30**, 518–532.

CHACÓN, J. 2020. The modal age of statistics. *Int. Stat. Rev.*, **88**, 122–141.

DUONG, T., COWLING, A., KOCH, I., & WAND, M. 2008. Feature significance for multivariate kernel density estimation. *Comp.Stat.& Data An.*, **52**, 4225–4242.

GENOVESE, C., PERONE-PACIFICO, M., VERDINELLI, I., & WASSERMAN, L. 2016. Non-parametric inference for density modes. *J.Roy.Stat.Soc. B*, **78**, 99–126.

LIANG, T., & SU, W. 2019. Statistical inference for the population landscape via moment adjusted stochastic gradients. *J.Roy.Stat.Soc. B*, **81**, 431–456.

MATSUMOTO, Y. 2002. *An introduction to Morse theory*. Amer. Math. Soc.

VISUALIZING CLUSTER OF WORDS: A GRAPHICAL APPROACH TO GRAMMAR ACQUISITION

Massimo Mucciardi[1], Giovanni Pirrotta[2], Andrea Briglia[3] and Arnaud Sallaberry[4]

[1] Department of Cognitive Science, Education and Cultural Studies, University of Messina, (e-mail: massimo.mucciardi@unime.it)

[2] University of Messina, (e-mail: gpirrotta@unime.it)

[3] Université "Paul Valéry » Montpellier3 (e-mail: andrea.briglia@univ-montp3.fr)

[4] LIRMM, University of Montpellier, CNRS, & AMIS, Université "Paul Valéry » Montpellier3 (e-mail: arnaud.sallaberry@lirmm.fr)

ABSTRACT: Language has been traditionally considered as a qualitative phenomenon that mainly requires hermeneutical methodologies in order to be studied, yet in recent decades thanks to advances in data storage, processing and visualization - there has been a growing and fertile interest in analysing language by relying on statistics and quantitative methods. In light of these motivations, we think it is worthwhile to try to explore databases made up of transcripted infant children spoken language in order to verify whether and how underlying patterns and recurrent sequences of learning stages work during acquisition. So, we think that the Expectation Maximization clustering method combined with an innovative graphical visualization can be useful to evaluate the development of linguistic structures over time in a reliable way.

KEYWORDS: first language acquisition, EM clustering, graphical visualization, phonetic variation rate, POS Tags.

1 General Framework

First language acquisition can be studied and modelled by using statistical tools: experiments have shown how specific *innately biased statistical learning mechanisms* are activated during in vitro settings where children easily learn how to keep memory of the transitional probability between syllables to spot word' boundaries [6]. Statistical and computational methods have contributed to important advances in the understanding of language acquisition: corpus analysis is one of the most rigorous ways to account for pattern, regularities and learning stages in a sound and replicable procedure [2]. In a very abstract form, first language acquisition could be viewed as a mixture of deterministic and random processes. It is deterministic because rules and constraints applied to human cognition are partly known. It is partly a random process because the amount of variability between children and within a single child is largely acknowledged and represents at the same time what is interesting and what is difficult in modelling child language studies. Romberg and

Saffran [4] assert that in language acquisition, the term `statistical learning' is most closely associated with tracking sequential statistics in word segmentation or grammar learning tasks. Knowing these rules and constraints does not allow us to predict the outcome of a child beginning to be immersed in his/her native language. All we know is that around the age of 5/6, she/he will master his/her own language/s. We know approximately the learning stages, the date of his/her first word, and the rough order of consonant acquisition. Interesting theories have been developed about the patterns of errors (e.g. phonetic variation) that the child will most likely make, but it is to date vary hard to model language acquisition. The types of patterns tracked by a statistical learning mechanism could be quite simple, such as a frequency count, or more complex, such as conditional probability [4]. In other words, learning a language (here conceived as a statistical structure of the environment) is in some ways a process that bring a child to minimize long-term prediction error. Clustering text is an important phase in data analysis. The common task in text clustering is to handle text in a multidimensional space, and to partition corpora into groups, where each group contains sentences that are similar to each other according to some grammatical indicators. Considering the above, in this paper we propose a new statistical strategy to evaluate the development of child linguistic structures over time in a reliable way based on clustering and visualization of words. The clusters are sufficiently explanatory for understanding first language acquisition as well as seem efficient for clustering performance. The paper is organized as follows: section 2 describes the data structure and the model applied; section 3 briefly provides the analysis strategy, the principal elaborations and visual interface for clustering.

2 Data Structure and Model

CoLaJE is a database composed of seven children that have been videorecorded in vivo approximately one hour every month from their first year of life until they were five (see https://www.ortolang.fr/). In this exploratory research, statistical treatments have been tested only on two children (Adrien and Madeleine) because the transcriptions obtained from these corpora are the most complete. Code for the Human Analysis of Transcripts (CHAT) provides a standardized format for producing computerized transcripts of conversational interactions. By analyzing, cleaning, filtering and normalizing all the available original CHAT transcripts we aimed at producing two corpora composed of the overall amount of what the children said through the years. A total of 8214 and 7168 database annotated sentences containing more than 100 variables were collected[1]. Some useful measures have been calculated such as: child age in years (Time) and Sentence Phonetic Variation Rate (SPVR) [1]: the SPVR is obtained by comparing mod and pho in order to measure how the relation between varied and correct form evolves over time. In the single sentence i (with $i = 1....N$),

[1] Due to lack of space in this paper we present the results for the Adrein dataset only. All other calculations are available on request.

$$SPVR_i = \left(TNPV_i / CTWT_i\right) \cdot 100 \quad (1)$$

where TNPV is the Total Number of Phonetic Variations of the words - total number of the difference between what the child really says (called "pho") and what he should have said according to the adult norm called ("mod") - and CTWT is the Child Total Words Tokenized. Hence, SPVR can assume the value 0% when the child does not make any error and 100% when the child does not pronounce all the words contained in the sentence correctly. Then, we applied Part-Of-Speech Tagger (POS Tags), a software that reads text in a given language and assigns parts of speech to each word such as noun, verb, adjective. We used Stanza Core NLP engine [22] to tag all CHI words by using Universal Dependencies as a standard of reference for part-of-speech classification [7]. Considering the nature of the variables (count data), we use finite multivariate Poisson mixtures in the EM procedure. We recall that EM clustering is an iterative method relying on the assumption that the data is generated by a mixture of underlying probability distributions, where each component represents a separate group, or cluster. The method provides the optimal number of clusters in any empirical situation, by using a two step iterative algorithm [3]. According to this approach to estimate mixture parameters we computing the maximum likelihood estimate (MLE) with the EM algorithm. In the next paragraph we will see the results of the principal estimates

3 Principal Results

To extend previous research [1], we divide our database in nine strata considering 3 different age classes of the child (L=1.97-2.64; M= 2.71-3.39 H=3.46-4.33 expressed in years and months) and 3 classes of SPVR (L= 33; M=>33 and 66; H>66 expressed in percent). In total we get 9 strata (from LL to HH). By framing the analysis in this way, we turn EM clustering algorithm into a potentially interesting method that could provide a reliable way to observe linguistic structures development over time. In tables 1 we summarize the main results obtained from clustering through a overview on the most influential POS tags for each strata and its related clusters for the dataset examined. In addition, the means of the POS are calculated in each strata (data not shown). We can observe that VERB occupies an increasing important role in development: it is almost absent in Adrien (dataset 1) during the earlier ages strata, it develops sharply in median age strata while it is present in almost any sentence in the upper age strata. It is clear that VERB causes an increase in the SPVR, as their values are higher in higher error rate strata (more than 33 percent). We can also observe that the parts of speech such as PRON (pronoun), VERB, SCONJ (subordinating conjunction) - which could be considered as markers of longer sentences - increase their importance. For visualization of clusters, we propose an interactive and visual interface to better this analysis. It has been designed considering a list of requirements defined in regards of the data structures and variables extracted by the clustering technique and the tasks one should be able to perform on such data. These are the main features: 1) visualize the clusters by age and SPVR; 2) visualize the distribution of POS tags in the clusters;

3) visualize the different values characterizing the clusters (age, SPVR, number of POS tags, number of sentences) and the POS tags (number of occurrences in a cluster, percentage, mean, Fisher coefficient, p-value); 4) visualize the list of sentences of a cluster; 5) visualize the relative and absolute evolution of the number of POS tags when child grows up (see the following link for all the details http://advanse.lirmm.fr/EMClustering/). In conclusion, we would suggest that these preliminary results represent a fair attempt to visualize child language development through clusters of words grouped by several criteria (age, grammatical properties, correct pronunciation). We can cautiously say that in this first stage of research the EM algorithm can provide us some mild descriptions in the classification of POS tags.

Table 1. *EM clustering results by strata - Dataset 1 (Adrien) (# - clusters number in brackets - POS sorted for ANOVA post-hoc F-test (in bold) p <0.05) (First 10 POS)*

Ordered POS	LL (3)	LM (2)	LH (4)	ML (5)	MM (3)	MH (3)	HL (4)	HM (5)	HH (5)
POS1	**INTJ**	**VERB**	**PRON**	**CCONJ**	**ADP**	**PRON**	**PRON**	**NOUN**	**AUX**
POS2	**DET**	**PROPN**	**ADV**	**PRON**	**ADV**	**AUX**	**DET**	**DET**	**NOUN**
POS3	**ADP**	ADV	**DET**	**NOUN**	**DET**	**NOUN**	**VERB**	**PRON**	**VERB**
POS4	**NOUN**	NOUN	**VERB**	**AUX**	**SCONJ**	**DET**	**NOUN**	**ADJ**	**DET**
POS5	**SYM**	INTJ	**NOUN**	**VERB**	**CCONJ**	**ADP**	**SCONJ**	**AUX**	**PRON**
POS6	**ADV**	PRON	**INTJ**	**NUM**	**INTJ**	**ADV**	**ADP**	**VERB**	**NUM**
POS7	**PROPN**	DET	**PROPN**	**SYM**	**NOUN**	**PROPN**	**AUX**	**ADP**	**ADJ**
POS8	**PRON**	AUX	**AUX**	**ADV**	**ADJ**	**SCONJ**	**ADV**	**ADV**	**ADP**
POS9	**VERB**	NUM	**ADJ**	**DET**	**NUM**	**VERB**	**ADJ**	**SCONJ**	**ADV**
POS10	**X**	CCONJ	**SCONJ**	**PROPN**	**PROPN**	**INTJ**	**CCONJ**	**X**	**X**

References

[1] Briglia A., Mucciardi M., Sauvage J. 2020. Identify the speech code through statistics: a data driven approach. Proceedings SIS 2020 (Pearson Editions).

[2] Chater, N. Manning, C. D. 2006. Probabilistic models of language processing and acquisition Trends in Cognitive Sciences 10(7), 335-44.

[3] Dempster A.P., Laird N.M., Rubin D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B: Methodological 39: 1-38.

[4] Romberg, A.R, Saffran, J.R. 2020. Statistical learning and language acquisition. Wiley Inter discip Rev Cogn Sci. 1(6): 906-914.

[5] Qi, P., Zhang Y., Zhang Y., Bolton J., Manning C. J. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations.

[6] Saffran J. R., Aslin R. N., Newport E. L. 1996 Statistical learning by 8-Month-Old infants. Science, vol. 274,. 1926-1928.

[7]Universal Dependencies. 2021. Retrieved from https://universaldependencies.org/fr/ pos/index.html

ROBUSTNESS METHODS FOR MODELLING COUNT DATA WITH GENERAL DEPENDENCE STRUCTURES

Marta Nai Ruscone [1]and Dimitris Karlis[2]

[1] DIMA - Department of Mathematics, University of Genoa, (e-mail: marta.nairuscone@unige.it)

[2] Department of Statistics, Athens University of Economics and Business, (e-mail: karlis@aueb.it)

ABSTRACT: Bivariate Poisson models are appropriate for modelling paired count data. However, the bivariate Poisson model does not allow for a negative dependence structure. Therefore, it is necessary to consider alternatives. A natural way is to consider copulas to generate various bivariate discrete distributions. While such models exist in the literature, the issue of choosing a suitable copula has been overlooked so far. Different copulas lead to different structures and any copula misspecification can render the inference useless. In this work, we consider bivariate Poisson models generated with a copula and investigate its robustness under outliers contamination and model misspecification. Particular focus is on the robustness of copula related parameters. English Premier League data are used to demonstrate the effectiveness of our approach.

KEYWORDS: copula, dependence, outliers, robustness.

1 Introduction

Bivariate Poisson models are appropriate for modelling paired count data exhibiting correlation. Paired count data arise in a wide context including, for example, sports (e.g. the number of goals scored by each one of the two opponent teams in soccer). Several models are available that can incorporate different structures and marginal properties, see for example Karlis & Ntzoufras, 2003. See also the work in Nikoloulopoulos, 2013 for defining models with copulas. While several extensions and models have been proposed, up to our knowledge, issues of robustness have been overlooked. Following da Fonseca & Fieller, 2006, there are two kinds of achieved robustness that one should consider. The first one refers to contamination from outlier observations or, better, from observations that are unexpected under a certain model. The second one refers to model deviation, i.e. a researcher would like to fit the model with such a method that even if the model is not correct the method would protect from deriving inconsistent results.

In this work, we consider a copula based bivariate Poisson distribution. We apply a minimum distance estimation methodology using Hellinger distance. We investigate its robustness under outliers contamination and model misspecification. Particular focus is given on the robustness of copula related parameters that measure the association exhibited by paired count data. The effectiveness of this methodology is examined on data from English Premier League 2013-2014.

2 Copulas

Copula are functions that join multivariate distributions to their marginal distributions (Nelsen, 2007). They describe the dependence structure existing across marginal random variables. In this way we can consider bivariate distributions with dependency structures different from the linear one that characterizes the multivariate Gaussian distribution.

A bivariate copula $C : I^2 \to I$, with $I = [0,1]$, is the cumulative bivariate distribution function of the random variables (U,V) with uniform marginal distributions in $[0,1]$. It is define as:

$$C(u,v;\theta) = P(U \leq u, V \leq v; \theta), \quad 0 \leq u \leq 1 \quad 0 \leq v \leq 1 \tag{1}$$

where θ is a parameter measuring the dependence between U and V.

Let (Y_1,Y_2) be a bivariate random vector with marginal cdfs $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ and joint cdf $F_{Y_1,Y_2}(y_1,y_2;\theta)$. There always exists a copula function $C(\cdot,\cdot;\theta)$ such that

$$F_{Y_1,Y_2}(y_1,y_2;\theta) = C\big(F_{Y_1}(y_1), F_{Y_2}(y_2);\theta\big), \quad y_1,y_2 \in \mathrm{I\!R}. \tag{2}$$

This result states that each joint distribution can be expressed in terms of two separate but related issues, the marginal distributions and the dependence structures between them. The dependence structure is explained by the copula function $C(\cdot,\cdot;\theta)$.

When Y_1 and Y_2 are discrete random variables taking values on some lattice, Ω, the copula C is unique in $(y_1,y_2) \in \Omega$ but not elsewhere. Thus, in the discrete case the mapping from two marginals and a copula $\{F_1,F_2,C\}$ to a bivariate distribution $F(Y_1,Y_2)$ is not one-to-one. However, this is not-uniqueness is of no consequence as the region outside Ω is not of interest in the discrete case (Nelsen, 2007).

3 Bivariate count models based on copulas

For count data, a common starting point is to use the Poisson distribution for the marginals:

$$f(y;\mu_j) = \mu_j^y e^{-\mu_j}/y!, \qquad j=1,2 \quad y=0,1,\ldots \tag{3}$$

where $\mu_j > 0$. Models based on copulas in the case of bivariate counts offer the advantage of allowing easy generalization to several different models which is not easy in general. Take, for instance, the Frank copula:

$$C(u,v;\gamma) = -\gamma^{-1}\log\left[1+\frac{(\exp^{-\gamma u}-1)(\exp^{-\gamma v}-1)}{\exp(-\gamma)-1}\right], \quad \gamma \in R-\{0\},\ u,v \in [0,1]. \tag{4}$$

Then

$$F(y_1,y_2;\mu_1,\mu_2,\gamma) \equiv C(F(y_1;\mu_1),F(y_2;\mu_2);\gamma), \tag{5}$$

is a well defined distribution function with a dependence structure. It's probability mass function is

$$\begin{aligned} P(Y_1=y_1,Y_2=y_2;\mu_1,\mu_2,\gamma) \quad = \quad & F(y_1,y_2;\mu_1,\mu_2,\gamma) - F(y_1-1,y_2;\mu_1,\mu_2,\gamma) \\ & -F(y_1,y_2-1;\mu_1,\mu_2,\gamma) + F(y_1-1,y_2-1;\mu_1,\mu_2,\gamma) \end{aligned} \tag{6}$$

In the present work we focus on bivariate models. For a review of discrete valued models based on copulas see Nikoloulopoulos, 2013.

4 Minimum distance estimation

In discrete data, model robustness and efficiency can be achieved almost at the same time, by defining distances that downweight some observations Lindsay, 1994. The minimum distance estimators can be interpreted as weighted likelihood estimators, the weights are determined by some kind of distance between observed and expected frequencies. For example, consider Minimum Hellinger distance estimators based on minimizing

$$\sum_x \left(d(x)^{1/2} - m_\beta(x)^{1/2}\right)^2$$

where $d(x)$ is the observed relative frequency and $m_\beta(x)$ is the probability mass at x with the assumed model with parameters of interest β. It turns out that this quantity leads to estimating equations of the form

$$\sum_x \left(\frac{d(x)}{m_\beta(x)}\right)^{1/2} \frac{\partial m_\beta(x)}{\partial \beta} = 0$$

directly comparable to the ML estimating equations

$$\sum_x \frac{d(x)}{m_\beta(x)} \frac{\partial m_\beta(x)}{\partial \beta} = 0$$

which actually implies that we weight the observations differently (see Lindsay, 1994).

In this work we extend the approach for bivariate count models defined by copulas aiming at deriving robust estimators for both the marginal and the copula parameters. Now x implies a pair of observations. Also, in our case the parameters β to estimate are those of the marginal distribution plus the copula parameter(s).We have also developed an iterative algorithm that facilitates the estimation. In the bivariate case we are interested in the relative frequencies are still reasonable estimators of the underlying probabilities but we need larger sample sizes for that. As we move on higher dimensions, problems similar to that of the regression setting may occur.

5 Application

Bivariate count models are widely used for modelling the outcome of a football game. The two counts refer to the number of goals scored by each team. It seems natural to assume some dependence between the goals to represent the competitive nature of soccer. Our data refer to all scores from English Premier League 2013-2014 where a series of unexpectedly large scores have occurred. We apply a robust approach to estimate the parameters of the model to reduce the effect of the large scores.

References

DA FONSECA, V GRUNERT, & FIELLER, NRJ. 2006. Distortion in statistical inference: the distinction between data contamination and model deviation. *Metrika*, **63**(2), 169–190.

KARLIS, D., & NTZOUFRAS, I. 2003. Analysis of sports data by using bivariate Poisson models. *J. R. Stat. Soc.*, **52**(3), 381–393.

LINDSAY, B. G. 1994. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *Ann. Stat.*, **22**(2), 1081–1114.

NELSEN, R B. 2007. *An introduction to copulas*. New York: Springer.

NIKOLOULOPOULOS, A.K. 2013. Copula-based models for multivariate discrete response data. *Pages 231–249 of: Copulae in Mathematical and Quantitative Finance*.

Bayesian analysis of a water quality high-frequency time series through Markov Switching Autoregressive Models

Roberta Paroli [1], Luigi Spezia [2], Marc Stutter[3] and Andy Vinten[3]

[1] Dipartimento di Scienze Statistiche, Università Cattolica SC, Milano, (e-mail: roberta.paroli@unicatt.it)

[2] Biomathematics & Statistics Scotland, Aberdeen, (e-mail: luigi@bioss.ac.uk)

[3] The James Hutton Institute, Aberdeen, (e-mail: marc.stutter@hutton.ac.uk and andy.vinten@hutton.ac.uk)

Abstract: In order to provide simulation inputs for investigations on diffuse water pollution and support rural land management policy on soil and water management, a turbidity time series recorded in a Scottish stream for more than a year, along with two covariates, is considered. Turbidity time series have complex dynamics because they are non-linear, non-Normal, non-stationary, with a long memory, and present missing values. Given these issues the turbidity process is analysed by Markov switching autoregressive models under the Bayesian paradigm using novel evolutionary Monte Carlo algorithms. Hence, it is possible to efficiently fit the actual data, reconstruct the sequence of hidden states, restore the missing values, and classify the observations into a few regimes, providing new insight on turbidity dynamics.

Keywords: non-homogeneous hidden Markov chain; path sampling; population Markov chain MonteCarlo; water quality; Wemyss catchment.

1 Introduction and Data

Evidence of the effectiveness of diffuse pollution control measures is needed to support rural land management policy on soil and water management. For key pollutants (e.g., suspended sediment or particulate phosphorus), such evidence is difficult to obtain, because of the cost of sampling and chemical analysis of storm event driven changes in concentrations and loads in streams and rural drainage features. Some works have investigated the use of continuous automated turbidity as a proxy to estimate particulate phosphorus, fine sediment or hydrophobic pollutant loads using site specific calibrations of turbidity versus the pollutant of interest, with some success. The turbidity trace along with discharge and other data may contain hidden temporal patterns (Birkel et al.

(2012)), that can be used to understanding of sources and processes delivering them to surface waters.

In order to provide simulation inputs for further investigations on diffuse pollution, a time series of turbidity data recorded in the Wemyss catchment (Scotland) from 1st January 2011 (00:00) to 5th January 2012 (15:15) is analysed here. Measurements (in NTU) were taken every 15 minutes; thus, the length of the series is 35,486 points, with 470 missing values (1.38% of the total number of observations). Two time series of explanatory variables are also available, without missing values and recorded with the same time resolution of turbidity: stage height (in cm) and rainfall (in mm).

A few complex issues need to be taken into account when modelling turbidity time series: non-Normality, non-linearity, non-stationarity, and long memory. Non-Normality is observed when the data density is multimodal or asymmetric or kurtic and the data cannot be considered as realizations from a Gaussian process. Non-linearity is assumed when the whole series does not show the same statitical peculiarities over all the observations, but they can be classified into a few homogeneous groups. Non-linearity can also be assumed when the series exhibits asymmetries. Weak non-stationarity is caused by generating processes having time-varying means and autocovariances. Finally, when the series shows high autocorrelations at the higher lags, with a slow decay, the observations are realizations from a long-memory process.

Because of these issues the turbidity time series considered here was analysed by Markov switching autoregressive models (MSARMs). This class of models is a popular tool within the econometrics community to model complex time series. Although they are extremely powerful, MSARMs have been considered quite rarely in other disciplines. Among the few applications in environmental sciences see Spezia et al. (2004) and Paroli and Spezia (2008) for air pollutant concentrations; Birkel et al. (2012) for isotope signatures; Montbet and Ailliot (2017) for air temperatures; Ailliot and Montbet (2012) for wind time series.

2 Model and Inference

MSARMs are pairs of discrete-time stochastic processes, one observed and one latent, or hidden. The hidden process is a finite-state Markov chain, whereas the observed process, given the Markov chain, is conditionally autoregressive. The dynamics of the observed process is driven by the dynamics of the latent one, so that each observation depends on the contemporary state of the Markov chain. By this theoretical structure, MSARMs allow: *i*) modelling non-linear and non-Normal time series by assuming that different autoregressions, each one depending on a hidden state, alternate according to the Markovian regime switching; *ii*) modelling a long-memory process; *iii*) classifying the observations into a small number of homogeneous groups, labelled as the regimes of the Markov chain.

Covariates, i.e. stage height and rainfall, were also incorporated into the model through both the hidden Markov chain (the transition probabilities are time-varying and dependent on the two dynamic exogenous variables) and the observed process (the two state-dependent exogenous variables are added to the past observations). Thus, we have time-varying means and autocovariances, and hence, a non-stationary model. Finally, the slow decay of the autocorrelation function is due to both the non-linearity of the series and the automatic recording of the data at a high temporal frequency. Non-linear time series with structural changes produce realizations that appear to have long memory. Given that structural changes can be efficiently described by stochastic regime switching models, we adopted MSARMs to highlight the changes in the turbidity dynamics, classify the observations into a few states, and fit the long memory process of the turbidity dynamics.

Because of the multimodal posterior density an efficient simulation-based evolutionary Monte Carlo (EMC) method is developed to better traverse the posterior surface and, so, fit the actual data and classify temporal correlated observations into a few homogeneous groups. EMC is a Markov chain Monte Carlo method which processes a population of chains in parallel, exchanging information one another. An advanced EMC algorithm is proposed here for Bayesian inference and model choice. This original EMC algorithm and its application to MSARMs represent a further methodological contribution of the paper. We introduce novel random walk crossover operators and made the EMC algorithm more efficient by flattening the likelihood only, and not the posterior, as in common practice. Thus, the same algorithm can be run for both model choice and parameter estimation (including the fitted values, the hidden states, and the missing values).

3 Results

The Bayesian analysis was developed in two steps: model choice and parameter estimation. The choice of the best model among the many available which differed for the number of states of the hidden Markov chain (m) and the autoregressive order (p), was performed computing the logarithms of the marginal likelihoods by EMC via the power posteriors, for any $m = 1, \ldots, 4$

and $p = 0, \ldots, 6$. The best model was characterized by three hidden states ($m = 3$) and autoregressions of the fourth order ($p = 4$).

Given the dimensions of the model and the identifiability constraint, the whole set of parameters was estimated. They show that covariates in the observed process have a positive coefficient, that is the level of turbidity increases when stage height and/or rainfall increase. On the other hand, the covariates in the hidden process have a negative coefficient, that is the probabilities of state transitions decrease when stage height and/or rainfall increase, while the diagonal probabilities of the transition matrices increase.

The model fit was very satisfactory, as shown by the comparison of actual and fitted values. The model performance was assessed by the root mean square error and the mean absolute error, which are very low. They are 0.164 (2% of the range of the data) and 0.282 (3%), respectively. All observations were within the 99% credibility interval.

This methodology will be generalized and used in a further study based on turbidity observations recorded in several catchments. In fact, a hierarchical linear regression model will be developed in a longitudinal study by taking the different sequences of hidden states and state-dependent parameters from each model associated with any of the several catchments as explanatory variables for the analysis of particulate phosphorus.

References

BIRKEL C., SOULSBY C., TETZLAFF D. DUNN S., & L., SPEZIA. 2012. High-frequency storm event isotope sampling reveals time-variant transit time distributions and influence of diurnal cycles. *Hydrological Processes*, **26**, 308–316.

P., AILLIOT, & V., MONTBET. 2012. Markov-switching autoregressive models for wind time series. *Environ. Modell. Softw.*, **30**, 92–101.

R., PAROLI, & L., SPEZIA. 2008. Bayesian inference in non-homogeneous Markov mixture of periodic autoregressions with state-dependent exogenous variables. *Comput. Statist. Data Anal.*, **52**, 2311–2330.

SPEZIA L., PAROLI R., & DELLAPORTAS, P. 2004. Periodic Markov switching autoregressive models for Bayesian analysis and forecasting of air pollution. *Statistical Modeling*, **4**, 19–38.

V., MONTBET, & P., AILLIOT. 2017. Sparse vector Markov switching autoregressive models. Application to multivariate time series of temperature. *Comput. Statist. Data Anal.*, **108**, 40–51.

DETECTING THE EFFECT OF SECONDARY SCHOOL IN HIGHER EDUCATION UNIVERSITY CHOICES [*]

Mariano Porcu [1], Isabella Sulis[1] and Cristian Usala[1]

[1] Department of Political and Social Sciences, University of Cagliari, (e-mail: mrporcu@unica.it, isulis@unica.it, cristian.usala@unica.it)

ABSTRACT:
The paper investigates the relationship between students' university choices and their secondary school background. The main aim is to assess the effect that secondary schools have in advising university applications toward local or non local institutions, also on the light of the tertiary education supply in students' area of residence. For this sake, four typologies of students have been identified and a multilevel model has been adopted to jointly consider the secondary schools effect on the probability to belong to one specific category conditional upon students' subject of study, and the characteristics of their local areas. Moreover, we provide a robust definition of local and non local universities by defining multiple criteria for the definition of non local universities and taking into account the uncertainty in the definition of the catching areas.

KEYWORDS: higher education, mobility choices, multilevel model, secondary school, uncertainty

1 Introduction

In the last decade there has been an increasing interest in Italian students' mobility choices for university studies as phenomenon which mirrors the inequalities in socioeconomic conditions between origin and destination areas and contributes in widening the already sharp disparities existing in the country (see Ciriaci, 2014; D'Agostino *et al.*, 2019; Attanasio & Enea, 2019). Despite the similar contexts, the literature is characterized by a high level of heterogeneity in the definition of local or non local universities for students, and consequently on the classification of students as *mover* or *stayer*, and on methods to account for distances between origin and destination places, with stud-

[*]This paper has been supported from Italian Ministerial grant PRIN 2017 "From high school to job placement: micro-data life course analysis of university student mobility and its impact on the Italian North-South divide.", n. 2017HBTK5P - CUP B78D19000180001.

ies which mainly focus on the mobility between macro-geographic areas, with emphasis on South-North trajectory, and other which also investigate the mobility patterns within macro-geographical areas. Starting from this evidence, our contribution is twofold. First, we investigate how secondary schools background affects students' preferences towards local or non local universities. To our knowledge, this is the first attempt to use data on Italian students to shed light on the role that secondary school have in students' location decision process. Our second contribution is to provide a robust definition of local and non local university choices by using multiple criteria based on students' traveled distance, the supply of education services in their local area and the uncertainty in the assignment of the local catchment area to each university.

2 Data and Methods

Our analysis relies on the administrative data collected from the Italian National Student Archive (NSA)* and the open database of the Italian Ministry of University and Research (MUR). We consider all Italian high-school leavers enrolled in an Italian university between a.y. 2016/2017 and a.y. 2018/2019 in a bachelor's programme. We define our dataset according to two rules. First, we do not consider the students enrolled in programs accessible with a national entry test since their choices are likely to depend on their ranking position rather than their preferences. Secondly, since we have information on high schools only from the a.y. 2016/2017, we retain in our sample only the students that left their high-school after 2015. Thus, starting from a population of 815,614 pupils, our data consists of 700,024 students, cross classified in 5,887 secondary schools and 297 university-city pairs.

Students' university choices are classified depending on: (i) the tertiary education supply in their local area, (ii) the chosen subject of study and (iii) the minimum travel time needed to reach the nearest university. At this aim, we define travel time as the minimum distance by car between two cities obtained by combining the ISTAT matrices on Italian cities with the data available on Google Maps. Then, we define two thresholds: d_{univ}, given by the distance between students' city and the nearest university, and d_{field}, defined considering only universities providing programmes in students' field of study. To avoid

*Data drawn from the Italian "Anagrafe Nazionale della Formazione Superiore" has been processed according to the research project "From high school to the job market: analysis of the university careers and the university North-South mobility" carried out by the University of Palermo (head of the research program), the Italian "Ministero Università e Ricerca", and INVALSI.

arbitrary assumptions on these thresholds and to assess results' sensitivity to the deterministic choice of the cut points, we apply Rubin's rule (Rubin, 1987) to combine the results obtained by using different thresholds. In particular, we generate multiple cut points values by increasing d_{univ} and d_{field} by a random amount of time $\delta \in [30;90]$. Thus, from students perspective, we have four categories of university choices: local, forced non local, free non local, and telematic. Universities are classified as 'local' when hosted in city closer than d_{univ} minutes of travel from students' residence. Non local universities are considered as 'forced' if the chosen university is the nearest one providing a programme in students' field of study (i.e. located closer than d_{field}), and as 'free' when students exceed both thresholds. The last category refers to students enrolled in distance-learning telematic universities.

The effect of secondary school background on students' university choices is estimated by specifying two cross-classified Multinomial Logit models which consider (a) the cross-classification of students in secondary schools and university city pairs and (b) the cross-classification of students in secondary schools and disciplinary fields. To take into account of the several curricula offered by secondary schools, we define the first level of clustering as the interaction between the high schools and the type of curricula offered. Moreover, we account for students' choice determinants by controlling for students' gender, macro area of residence, diploma grade, year of enrollment, years of delay in finishing the high school and an indicator that takes value 1 if the student has attended a lyceum.

3 Results and Discussion

Table 1 reports the results related to the model (b) which considers the clustering of students according to secondary school-curricula combinations and students' field of study, with the δ parameter set equal to 60. The results concerning the variance of the random terms suggest a clear effect of schools in students' choices to attend local or non local universities when accounting for differences in their field of study. Indeed, the variability of the high school-curricula effect is relevant in all the categories. The posterior predictions regarding the school random terms provide evidences on the role that schools have in orienting students' choices towards local, non local and telematic universities. Moreover, the results concerning the control variables in the fix effect component show that students' educational background and socioeconomic characteristics affect the probability to make different choices in terms of selection of local and non local universities.

Table 1. *Cross-Classified Multinomial Logit*

	Forced Non Local	Free Non Local	Telematic
Constant	-4.079	-4.214	-1.678
	[-4.198;-3.935]	[-4.359;-4.005]	[-2.199;-1.185]
Controls	Yes	Yes	Yes
Random effect parameters:			
High School × Curriculum	5.379	2.008	1.637
	[5.157;5.611]	[1.928;2.096]	[1.522;1.758]
Field of study	2.694	0.659	10.31
	[1.320;5.263]	[0.319;1.337]	[4.070;23.962]
Observations	700068		

In conclusion, the approach proposed in this work allowed us to assess the effect that secondary schools have in advising university applications toward local or non local institutions by accounting for the choice of the disciplinary field. Further analyses are still in progress to take into account of the uncertainty related to a deterministic definition of δ parameter. At this aim, multiple values of δ have been generated to assess results' sensitivity to the choice of the cut points. This uncertainty in thresholds definitions is taken into account by using Rubin's rules to combine the results.

References

ATTANASIO, M., & ENEA, M. 2019. La mobilitá degli studenti universitari nell'ultimo decennio in Italia. *Pages 43–58 of:* DE SANTIS, G., PIRANI, E., & PORCU, M. (eds), *Rapporto sulla popolazione. L'istruzione in Italia*. Bologna: Il Mulino.

CIRIACI, D. 2014. Does University Quality Influence the Interregional Mobility of Students and Graduates? The Case of Italy. *Regional Studies*, **48**(10), 1592–1608.

D'AGOSTINO, A., GHELLINI, G., & LONGOBARDI, S. 2019. Out-migration of university enrolment: the mobility behaviour of Italian students. *International Journal of Manpower*, **40**, 56–72.

DATABASE MOBYSU.IT [MOBILITÀ DEGLI STUDI UNIVERSITARI IN ITALIA]. Research protocol MUR - Universities of Cagliari, Palermo, Siena, Torino, Sassari, Firenze, Cattolica and Napoli Federico II. Scientific Coordinator Massimo Attanasio (UNIPA). Data Source ANS-MUR/CINECA.

RUBIN, D. B. 1987. Multiple Imputation for Nonresponse in Surveys. *Wiley, New York*.

SEMI-CONSTRAINED MODEL-BASED CLUSTERING OF MIXED-TYPE DATA USING A COMPOSITE LIKELIHOOD APPROACH

Roberto Rocci[1] and Monia Ranalli[2]

[1] Department of Statistical Sciences, Sapienza University of Rome
(roberto.rocci@uniroma1.it)

[2] Department of Statistical Sciences, Sapienza University of Rome
(monia.ranalli@uniroma1.it)

ABSTRACT: We propose a class of semi-constrained models for clustering ordinal and continuous data. Ordinal variables are assumed to be a discretization of some latent continuous variables jointly distribuited with the observed continuous variables as a finite mixture of Gaussians. Parsimonious modeling is obtained by reparameterizing the covariance matrices in terms of factor analysis models semi-constrained across the components. Parameter estimation is carried out using a EM-type algorithm to maximize a composite log-likelihood. The proposal is evaluated through a simulation study and an application to real data.

KEYWORDS: mixture models, factor analyzers, composite likelihood, EM algorithm, mixed-type data

1 Introduction

Complex data structures are characterized by the presence of heterogeneity and a large number of features of mixed type, i.e. ordinal and continuous. To capture heterogeneity, clustering methods are used to find subgroups in the population. The literature has been mainly developed for continuous variables with methods distance-based (e.g. *k*-means, Ward) or model-based. In the latter, finite Gaussian mixture models are the most commonly used (Hennig *et al.*, 2015). In order to reduce the large number of parameters caused by the high dimensionality of the data, parsimonious modelling is needed like in factor analysis modelling. The challenge to model ordinal data is mainly due to the lack of metric properties. Ordinal variables can be modeled properly adopting the underlying variable approach (Jöreskog, 1990) where the ordinal variables are assumed to be generated by thresholding some latent continuous variables. This allow us to model the dependence between ordinal and continouous variables by modeling the dependence between the latent and the observed continuous variables.

Taking these aspects into account, i.e. heterogeneity, high dimensional and mixed type data, we propose a Gaussian mixture model with a factor decomposition on component-specific covariance matrices. Parameters may be constrained to be equal or unequal across mixture components (McNicholas & Murphy, 2010) obtaining different degrees of parsimony. The ordinal variables corresponds to some variates of the mixture that are partially observed through a discretization (see e.g. Ranalli & Rocci, 2017).

Inference could be carried out through the likelihood function. However, the presence of ordinal variables requires the computation of many high dimensional integrals. This makes the evaluation of the likelihood computationally demanding, or prohibitive, as the number of ordinal variables increases. To solve the problem, the likelihood is replaced with a surrogate function, that is the composite likelihood, defined as the product of m-dimensional marginals or conditional events (Lindsay, 1988). Under some regularity conditions the corresponding estimators are consistent, asymptotically unbiased and normally distributed (see Ranalli & Rocci, 2017, and references therein). In general they are less efficient than the maximum likelihood estimators, but much more efficient in terms of computational complexity. In the current work, the composite likelihood is based on the product of all possible sub-likelihoods composed of two ordinal and all continuous variables. The computation of parameter estimates is carried out through an EM-type algorithm.

A simulation study as well as a real data analysis is presented in the extended version of the paper.

2 Model

Let $\mathbf{y}^{\bar{O}} = [y_1, \dots, y_{\bar{O}}]$ and $\mathbf{x} = [x_{\bar{O}+1}, \dots, x_P]$ be $\bar{O}$ continuous variables and $O = P - \bar{O}$ ordinal variables, respectively. Each ordinal variable has the associated categories $c_i = 1, \dots, C_i$ with $i = \bar{O} + 1, \dots, P$. Following the underlying response variable approach, the observed ordinal variables $\mathbf{x}$ are considered as a discretization of some continuous latent variables $\mathbf{y}^O = [y_{\bar{O}+1}, \dots, y_P]$. The relationship between $\mathbf{x}$ and $\mathbf{y}^O$ is

$$\gamma^{(i)}_{c_i-1} \leq y_i < \gamma^{(i)}_{c_i} \Leftrightarrow x_i = c_i,$$

where $-\infty = \gamma^{(i)}_0 < \gamma^{(i)}_1 < \dots < \gamma^{(i)}_{C_i-1} < \gamma^{(i)}_{C_i} = +\infty$ are the non observable thresholds defining the C_i categories. In our proposal $\mathbf{y} = [\mathbf{y}^{\bar{O}}, \mathbf{y}^O]$ follows a finite mixture of factor analyzers (McNicholas & Murphy, 2010)

$$f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi(\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g)$$

where ϕ is the multivariate normal density, $\boldsymbol{\Lambda}_g$ is the $P \times K$ matrix of factor loadings, and $\boldsymbol{\Psi}_g$ is the diagonal matrix of uniqueness that could be assumed of the isotropic form $\psi_g \mathbf{I}$. Each term may be constrained to be equal or unequal across mixture components. The result of imposing, or not, such constraints generates the family of eight parsimonious Gaussian mixture models, described in Table 1, $\boldsymbol{\Lambda}_g$ type C and U, and introduced by McNicholas & Murphy (2010) in the context of continuous data. Each member of this family has a number of covariance parameters that grows linearly with the data dimensionality. By assuming a common covariance structure, even more parsimonious models are obtained. Some identifiability constraints are imposed on thresholds and factor loadings. They are not discussed here for sake of space.

Table 1: The covariance structure of parsimonious Gaussian mixture models with a constrained (C), semiconstrained (S) or unconstrained (U) factor loadings matrix.

Model ID	$\boldsymbol{\Lambda}_g$	$\boldsymbol{\Psi}_g$	Isotropic	$\boldsymbol{\Sigma}_g$
CCC	C	C	C	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi \mathbf{I}_P$
CCU	C	C	U	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$
CUC	C	U	C	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi_g \mathbf{I}_P$
CUU	C	U	U	$\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g$
SCC	S	C	C	$\boldsymbol{\Lambda}\mathbf{L}_g^2\boldsymbol{\Lambda}' + \psi \mathbf{I}_P$
SCU	S	C	U	$\boldsymbol{\Lambda}\mathbf{L}_g^2\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$
SUC	S	U	C	$\boldsymbol{\Lambda}\mathbf{L}_g^2\boldsymbol{\Lambda}' + \psi_g \mathbf{I}_P$
SUU	S	U	U	$\boldsymbol{\Lambda}\mathbf{L}_g^2\boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g$
UCC	U	C	C	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi \mathbf{I}_P$
UCU	U	C	U	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$
UUC	U	U	C	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \psi_g \mathbf{I}_P$
UUU	U	U	U	$\boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$

With respect to the proposal of McNicholas & Murphy (2010), we introduce four semi-constrained models to add some extra flexibility, with a certain degree of parsimony (see Table 1, $\boldsymbol{\Lambda}_g$ type S). The flexibility is achieved by assuming that the matrix of factor loadings can be written in the form $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}\mathbf{L}_g$, where $\mathbf{L}_g$ is a positive definite diagonal matrix of factor saliences. They can be considered as constrained cases between the first and the last four models of Table 1. The latent factors in each cluster are the same but with different variances recorded by the matrices $\mathbf{L}_g^2$. This is a particular form of factorial invariance firstly introduced by Cattell (1944) and then developed by several authors, e.g. in the context of three-way analysis (see Giordani *et al.*, 2020, and references therein). A nice feature of the semi-constrained models is that, under mild conditions, the factors are unique. In other terms, it is not possible to rotate the factors as in the classical factor analysis model.

Our proposal has been tested by a simulation study and an application on real data not shown here for sake of space. In the first, the effectiveness of the composite likelihood approach has been investigated under various settings, such as different numbers of observations, groups and latent factors, in terms of estimates precision and ability of recovering the true partition. In the second, the model has been used to find latent groups in a dataset taken from the survey on academic graduates' vocational integration carried out by ISTAT in 2015.

References

CATTELL, RAYMOND B. 1944. "Parallel proportional profiles" and other principles for determining the choice of factors by rotation. *Psychometrika*, **9**(4), 267–283.

GIORDANI, PAOLO, ROCCI, ROBERTO, & BOVE, GIUSEPPE. 2020. Factor Uniqueness of the Structural Parafac Model. *Psychometrika*, **85**(3), 555–574.

HENNIG, CHRISTIAN, MEILA, MARINA, MURTAGH, FIONN, & ROCCI, ROBERTO. 2015. *Handbook of cluster analysis*. CRC Press.

JÖRESKOG, KARL G. 1990. New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, **24**(4), 387–404.

LINDSAY, BRUCE. 1988. Composite likelihood methods. *Contemporary Mathematics*, **80**, 221–239.

MCNICHOLAS, PAUL D., & MURPHY, THOMAS BRENDAN. 2010. Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*, **26**(21), 2705–2712.

RANALLI, MONIA, & ROCCI, ROBERTO. 2017. Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis*, **110**(C), 87–102.

ANTIBODIES TO SARS-COV-2: AN EXPLORATORY ANALYSIS CARRIED OUT THROUGH THE BAYESIAN PROFILE REGRESSION

Annalina Sarra [1], Adelia Evangelista[1], Tonio Di Battista[1] and Damiana Pieragostino[2]

[1] Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University "G. d'Annunzio" of Chieti-Pescara, (e-mail: annalina.sarra@unich.it, adelia.evangelista@unich.it, tonio.dibattista@unich.it)

[2] Center for Advanced Studies and Technology (CAST), University "G. d'Annunzio" of Chieti-Pescara, (e-mail: damiana.pieragostino@unich.it)

ABSTRACT: In this paper we aim at characterizing the immune response to SARS-CoV-2 in a cohort study of individuals who received the first dose of mRna vaccines Pzifer or AstraZeneca. To examine some covariate-related effects on anti-S1 spike IgG levels we adopted a statistical technique known as Bayesian Profile Regression (BPR). The BPR explores the link between a response variable and a set of associated covariate data through cluster membership and supervises the clustering assignment in a unified fashion. In our study, this methodology allowed us to identify three clusters, differentiated according to the antibody titer of respondents, and draw the profile of subjects whose amount of antibodies produced is significantly higher.

KEYWORDS: Bayesian Profile Regression, SARS-CoV-2, anti-S1 spike IgG levels, cluster profile

1 Introduction

To impede the progress of the COVID-19 pandemic, the scientific world has raced to identify and understand the immune response to SARS-CoV-2 infection. Many efforts have been directed towards the development of the vaccines to curtail the novel coronavirus. Currently, among the EU authorized COVID-19 vaccines, with greater than 90% efficacy to reduce the symptomatic infection risk, there are the Pfizer/BioNTech and AstraZeneca. So far, post infection immunity to SARS-CoV-2 is still unclear and much work needs to be carried out to characterize the immune response to the virus. This knowledge is crucial to give insights into the disease pathogenics and into the usefulness of bridge therapies. In this study, we analysed anti-S1 spike IgG levels in a cohort of

89 individuals: 39 people have received one dose of the Pfizer vaccine and 50 one dose of AstraZeneca. In order to identify the main covariates associated with immunoglobulin antibodies, we followed an analytic strategy based on Bayesian Profile Regression (Molitor *et al.*, 2010) conceived as a non parametric dimension reduction technique, set in a Bayesian framework, for clustering responses and covariates simultaneously. The remainder of this paper proceeds as follows. In section 2, we provide details of the theoretical background of the Bayesian Profile Regression technique. Section 3 considers the available data whereas the main results of the statistical analysis are presented in Section 4.

2 Bayesian Profile Regression

The Bayesian Profile Regression (BPR), is a Bayesian dimension reduction and clustering technique to jointly modeling an outcome variable and a number of potentially correlated predictors (Molitor *et al.*, 2010). This technique, links non parametrically a response variable to covariate data through cluster membership, so that the outcome and the clusters mutually inform each other (Hastie *et al.*, 2013). To deal with these joint effects, the BPR approach adopts as unit of inference a profile, formed from a sequence of covariates values. In what follows, for each unit i, y_i denotes the outcome of interest while $\mathbf{X_i}$=$(x_{i_1},,\dots,x_{i_P})$ represents the covariate profile that consists of p covariates that we are interested in studying. Additionally, $\mathbf{w_i}$ are the fixed effects which are constrained to only have a global (i.e. non-cluster specific) effect on the response y_i and $\phi_c^p(x_{ip})$ indicate the probability that the p-th variable in cluster c is equal to x_{ip}. The model of interest here can be described by two key components: a covariate model which assigns individual profile to clusters and a response model which links cluster of profiles to an outcome of interest via a regression model. The full data are then jointly modelled leading to the following likelihood

$$f(\mathbf{x_i}, y_i|\theta_{zi}, w_i, \psi) = \sum_c \psi_c f(\mathbf{x_i}|z_i = c, \phi_c) f(y_i|z_i = c, \theta_c, \Lambda, \mathbf{w_i}) \qquad (1)$$

where z_i =c is the allocation variable that indicates the cluster to which a unit i belongs, $\mathbf{\Lambda}$ is a vector of global (i.e., non-cluster specific) parameters, finally, ψ_c are the mixture weights. The mixture weights corresponding to a maximum of C clusters, denoted as $\psi_c, c = 1,,\dots,C$, will be modeled according to a “stick-breaking” representation of a Dirichlet process prior. Owing to the complexity of the model, inference is facilitated by Markov Chain Monte Carlo (MCMC) methods. A detailed description of the BPR can be found in Molitor *et al.*, 2010.

3 Data

In this paper, we refer to a longitudinal research carried out by the Center of Advanced Studies and Technology (CAST) of University “G. d’Annunzio” of Chieti-Pescara (Italy). This study looked at antibody response of 89 individuals who received the first dose of mRna vaccines Pzifer or AstraZeneca. IgG antibodies to SARS-CoV-2 were measured by a fully automated solid phase DELFIA (time-resolved fluorescence) immunoassay in a few drops of blood collected by finger prick and let dry on filter paper card. Subjects involved in the analysis were re-called at 7, 10, 15 days after the first injection of vaccine for re-determination of IgG levels, recoded according to the quartiles. All participants were also surveyed regarding post-vaccination symptoms, including presence (coded with 1) or absence (coded with 0) of distinct symptom types, such as: fatigue, headache, chills, muscle pain, fever and joint pain. Age, recoded in two classes (20-40 and 40-65 years) and gender of vaccinated people have been also determinated (0=Male and 1=Female).

4 Main results

The BPR estimation, performed through the R package PreMium (Liverani *et al.*, 2015), has produced a partition of anti-S1 spike IgG levels after 21 days from injection, recoded using the median as cut-off, and some potential explanatory variables (IgG levels at previous at different times, type of vaccine, side effects after vaccination, age, gender) into 3 clusters. Each group is characterized by similar covariate profiles, as well as by the same amount of the antibodies. The posterior distribution of all clusters specific parameters are represented in Fig.1. The left panel of each figure displays the MCMC posterior draws of the anti-S1 spike IgG levels after 21 days for the identified clusters; conversely the right panel of each figure shows the posterior distributions of the probability that an explanatory variable appears with one of the discrete categories across the identified groups. In the typical profile of the cluster 3 (red boxplot in Fig.1), associated with the highest amount of antibodies produced, there is a prevalence of people aged 20-40 years, who have

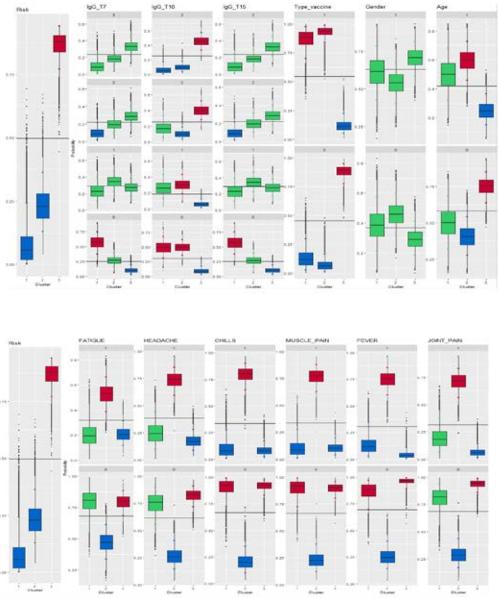


Figure 1. *Summary plot of the posterior distribution of parameter* ϕ_c*, for* $c = 1,2,3$

received the first dose of Pfizer vaccine. Furthermore, the majority of individuals belonging to this group has not experienced side effects while for them we observe a greater amount of anti-S1 spike IgG levels after 10 days from injection. Specular results characterize the first two clusters associated with a lower immunity response.

References

HASTIE, D.I., LIVERANI, S., AZIZI, L., RICHARDSON, S., & STÜCKER, I. 2013. A semi-parametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer. *BMC Medical Research Methodology.*, **13**, 129.

LIVERANI, S., HASTIE, D.I., PAPATHOMAS, M., & RICHARDSON, S. 2015. PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. *Spatial and Spatio-temporal Epidemiology.*

MOLITOR, J., PAPATHOMAS, M., JERRETT, M., & RICHARDSON, S. 2010. Bayesian profile regression with an application to the National survey of children's health. *Biostatistics.*, **11**, 484–498.

MODELING THREE-WAY RNA SEQUENCING DATA WITH MIXTURES OF MULTIVARIATE POISSON-LOGNORMAL DISTRIBUTIONS

Theresa Scharl[1] and Bettina Grün[2]

[1] Institute of Statistics, University of Natural Resources and Life Sciences, Vienna, (e-mail: Theresa.Scharl@boku.ac.at)

[2] Institute for Statistics and Mathematics, WU Vienna University of Economics and Business, (e-mail: Bettina.Gruen@wu.ac.at)

ABSTRACT: Mixtures of multivariate Poisson-lognormal distributions are used for modeling three-way RNA sequencing data. Taking the three-way structure into account, a range of specifications for the means and the variance-covariance matrices of the latent multivariate normal distribution emerge, leading to a more parsimonious and better interpretable clustering solution. We develop suitable specifications for an RNA sequencing dataset containing several biological units categorized by additional covariates. These include, for example, a regression setup for the means and time-series structure for the variance-covariance matrices. The models are fitted using maximum likelihood estimation with the expectation-maximization algorithm involving a variational E-step and their suitability investigated for a specific RNA sequencing dataset.

KEYWORDS: clustering, EM algorithm, multivariate Poisson-lognormal distribution, RNA sequencing

1 Modeling RNA Sequencing Data

RNA sequencing of time-course experiments leads to three-way count data where the dimensions are the genes, the time points and the biological units. Cluster analysis is used to group genes in dependence of their expression levels taking the development over time and across the biological units into account. Model-based clustering methods allow to embed the clustering problem within a statistical framework and the mixture models used may be adapted in a flexible way to the data structure and clustering aims by specifying suitable models for the components of the mixture.

The Poisson distribution is obvious to use for modeling count data. However, assuming independence between the time points and/or the biological units might be questionable. Silva *et al.* (2019) propose to use mixtures of multivariate Poisson-lognormal distributions to account for possible dependency structures via a latent multivariate normal distribution after transforming the data to a two-way format where the genes are in one dimension and time points and biological units are crossed out for the second dimension. Subedi & Browne (2020) also consider mixtures of multivariate Poisson-lognormal distributions for two-way data, but following Fraley & Raftery (2002) they propose parsimonious specifications of the variance-covariance matrix resulting from the decomposition into volume, shape and orientation. Taking the three-way structure into accout, Silva *et al.* (2018) also arrive at a more parsimonious parameterization of the variance-covariance matrix.

In all these contributions, maximum likelihood estimation of the finite mixture model for a fixed number of components is performed and a suitable model is selected based on information criteria such as BIC, AIC and ICL. The expectation-maximization (EM) algorithm is used for estimation with the cluster memberships as well as the latent multivariate normal observations are viewed as missing data. The EM algorithm is an iterative procedure where each iteration consists of an E- and an M-step. The expectation of the complete-data log-likelihood which results from combining the observed with the missing data conditional on current parameter estimates and the observed data is determined in the E-step. In the M-step the expected complete-data log-likelihood is maximized with respect to the parameters and new parameter estimates are obtained. In each iteration the log-likelihood is increased, ensuring that the algorithm converges to a fixed point if the log-likelihood is bounded. For mixtures of multivariate Poisson-lognormal distributions, the M-step is straighforward. However, the E-step is complicated. Silva *et al.* (2019) and Silva *et al.* (2018) use Bayesian Markov chain Monte Carlo methods to obtain an estimate for the expectation. Subedi & Browne (2020) propose to use a variational E-step.

2 The Mixture Model for Three-Way Data

Following Silva *et al.* (2018), a finite mixture model of multivariate Poisson-lognormal distributions implies the following data generating process for the observations $y_{i,jt}$, with i the gene index, j the biological unit index and t the time point index:

$$\begin{aligned} S_i &\sim \mathcal{M}(\boldsymbol{\eta}), \\ \Theta_{S_i}|S_i &\sim \mathcal{MN}(\boldsymbol{M}_{S_i}, \boldsymbol{U}_{S_i}, \boldsymbol{V}_{S_i}), \\ y_{i,jt}|S_i &\sim \mathcal{P}(b_j \exp(\Theta_{S_i,ij})), \end{aligned}$$

where S_i is the component membership of gene i, $\mathcal{M}(\boldsymbol{\eta})$ is the multinomial distribution with success probabilities vector $\boldsymbol{\eta}$, $\mathcal{MN}(\boldsymbol{M}_{S_i}, \boldsymbol{U}_{S_i}, \boldsymbol{V}_{S_i})$ is the matrix normal distribution which is equivalent to

$$\text{vec}(\Theta_{S_i})|S_i \sim \mathcal{N}(\text{vec}(\boldsymbol{M}_{S_i}), \boldsymbol{U}_{S_i} \otimes \boldsymbol{V}_{S_i}),$$

with vec() the vectorization operator and $\otimes$ the Kronecker product. A suitable constraint needs to be imposed on $\boldsymbol{U}_{S_i}$ and $\boldsymbol{V}_{S_i}$ to ensure identifiability. $\mathcal{P}(\lambda)$ is the univariate Poisson distribution with parameter λ given by the exponentiated ijth element of the latent normal variable Θ_{S_i} multiplied with a biological unit specific offset b_j.

Taking the specific three-way data structure into account the following model specifications might be considered:

(a) The mean matrix for each component $\boldsymbol{M}$ has dimension number of biological units times the number of time points. Assuming additive biological units and time point effects, this mean matrix would be given by:

$$\boldsymbol{M} = \boldsymbol{\alpha} \otimes \boldsymbol{\beta},$$

where $\boldsymbol{\alpha}$ are the mean biological effects and $\boldsymbol{\beta}$ are the mean time point effects. Additional interaction effects would indicate the need for a general $\boldsymbol{M}$.

(b) A more parsimonious specification of the mean vectors is possible if covariates are available to characterize the biological units using a regression model.

(c) The variance-covariance matrix $\boldsymbol{V}$ capturing time dependence could be specified in a more parsimonious way by assuming for example an underlying auto-regressive process, e.g., an AR(1) process.

(d) Assuming a correlation between the biological units might be questionable and the identity matrix could be specified for $\boldsymbol{U}$.

(e) Inspired by Fraley & Raftery (2002), different sets of parameters might either be assumed to be group-specific or the same across groups, thus allowing for a more parsimonious specification and easier interpretation of the fitted model.

3 Data

The available RNA sequencing data contains 4523 genes, 17 biological units and 4 time points for three biological replicates. The median value across the three biological replicates is used for analysis. Data pre-preprocessing reduces the number of genes by eliminating those which are not differentially expressed. The number of time points will also be reduced by using the first time point as baseline level. The biological units are characterized as wildtype or recombinant.

The mixture model can be fitted in R using the R package PLNmodels (Chiquet *et al.*, 2021a), available from the Comprehensive R Archive Network. The package implements the variant of the EM algorithm using a variational E-step, has been developed for modelling joint species abundances (Chiquet *et al.*, 2021b) and may be extended to cover the model specifications of interest for modeling three-way RNA sequencing data.

References

CHIQUET, JULIEN, MARIADASSOU, MAHENDRA, & GINDRAUD, FRANÇOIS. 2021a. *PLNmodels: Poisson Lognormal Models*. R package version 0.11.4.

CHIQUET, JULIEN, MARIADASSOU, MAHENDRA, & ROBIN, STÉPHANE. 2021b. The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances. *Frontiers in Ecology and Evolution*, **9**, 188.

FRALEY, CHRIS, & RAFTERY, ADRIAN E. 2002. Model-Based Clustering, Discriminant Analysis and Density Estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.

SILVA, ANJALI, ROTHSTEIN, STEVEN J., MCNICHOLAS, PAUL D., & SUBEDI, SANJEENA. 2018. *Finite Mixtures of Matrix Variate Poisson-Log Normal Distributions for Three-Way Count Data*. arXiv:1807.08380 [stat.ME].

SILVA, ANJALI, ROTHSTEIN, STEVEN J., MCNICHOLAS, PAUL D., & SUBEDI, SANJEENA. 2019. A Multivariate Poisson-Log Normal Mixture Model for Clustering Transcriptome Sequencing Data. *BMC Bioinformatics*, **20**(1), 394.

SILVA, H. ANJALI. 2018. *Bayesian Clustering Approaches for Discrete Data*. Ph.D. thesis, The University of Guelph.

SUBEDI, SANJEENA, & BROWNE, RYAN P. 2020. A Family of Parsimonious Mixtures of Multivariate Poisson-Lognormal Distributions for Clustering Multivariate Count Data. *Stat*, **9**(1), e310.

STACKING ENSEMBLE LEARNING WITH GAUSSIAN MIXTURES

Luca Scrucca [1]

[1] Dept. of Economics, University of Perugia (e-mail: luca.scrucca@unipg.it)

ABSTRACT: Stacking is an ensemble method which uses a meta-learning approach to learn how to best combine the predictions from two or more base statistical and machine learning algorithms. In this contribution we propose a stacking algorithm for classification using Gaussian mixtures as base learners.

KEYWORDS: Gaussian mixtures, classification, ensemble learning, stacking.

1 Introduction

Ensemble learning is a broad term referring to meta-learning methods that combine predictions provided by multiple learners or models to obtain a prediction that often is more accurate than any single prediction. Typically, ensemble learning is applied in supervised learning tasks, such as in regression and classification (Rokach, 2010). In this contribution we propose an algorithm for ensemble learning using Gaussian mixtures as base learners for classification tasks.

2 EDDA Gaussian mixture models for classification

Consider a training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ for which both the features vector $\boldsymbol{x}_i$ and the true class $y_i \in \{C_1, \ldots, C_K\}$ are known. Mixture-based classification models assume that the density within each class follows a Gaussian mixture distribution:

$$f(\boldsymbol{x}|C_k) = \sum_{g=1}^{G_k} \pi_{gk} \phi(\boldsymbol{x}; \boldsymbol{\mu}_{gk}, \boldsymbol{\Sigma}_{gk}), \qquad (1)$$

where G_k is the number of components within class k, π_{gk} are the mixing probabilities for class k, such that $\pi_{gk} > 0$ and $\sum_{g=1}^{G_k} \pi_{gk} = 1$, and $\boldsymbol{\mu}_{gk}$ and $\boldsymbol{\Sigma}_{gk}$ are, respectively, the mean vectors and the covariance matrices of component g within class k. Since this model is highly flexible, (i) parameter estimates are subject to high uncertainty unless a very large dataset is available, and (ii) it may easily lead to overfit. For these reasons a parsimonious mixture-based classification model, termed *Eigenvalue Decomposition Discriminant*

Analysis (EDDA) model, has been proposed (Bensmail & Celeux, 1996). This assumes that (i) the density for each class can be described by a single Gaussian component, i.e. $G_k = 1$ for all k in equation (1), and (ii) the class covariance structure is factorised as $\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{U}_k \boldsymbol{\Delta}_k \boldsymbol{U}_k^\top$.

The EDDA family contains 14 different models (see Scrucca *et al.*, 2016, Table 3), some of which are popular discriminant analysis models. For instance, if each class has the same covariance matrix, that is $\boldsymbol{\Sigma}_k = \lambda \boldsymbol{U} \boldsymbol{\Delta} \boldsymbol{U}^\top$ for all k, then EDDA is equivalent to the classical *Linear Discriminant Analysis* (LDA) model. If the class covariance matrices are unconstrained, that is $\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{U}_k \boldsymbol{\Delta}_k \boldsymbol{U}_k^\top$ for all k, then EDDA is equivalent to the *Quadratic Discriminant Analysis* (QDA) model. Finally, assuming the matrix of eigenvectors $\boldsymbol{U}$ is the identity matrix, features are conditional independent within each class and the so-called *Naïve-Bayes* models are obtained.

Classification of observation $\boldsymbol{x}$ can be obtained according to the MAP (*maximum a posteriori*) principle, that is by assigning an observation to the class with the largest posterior class probability computed via Bayes' theorem

$$\Pr(C_k|\boldsymbol{x}) = \frac{\tau_k f(\boldsymbol{x}|C_k)}{\sum_{g=1}^K \tau_g f(\boldsymbol{x}|C_g)},$$

where $f(\boldsymbol{x}|C_k)$ are the class-conditional densities, and $\tau_k = \Pr(C_k)$ are the prior class probabilities for each class C_k ($k = 1, \dots, K$).

Estimation of unknown parameters $(\tau_1, \dots, \tau_K, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K)$ for EDDA models can be obtained with a single M-step from the EM algorithm for Gaussian mixtures, with the conditional probabilities z_{ik} set to 1 if observation i belongs to class k and 0 otherwise.

3 Stacking EDDA for ensemble classification

In this section we propose a form of stacking, called *Super Learner* algorithm (Wolpert, 1992; Van der Laan *et al.*, 2007), which uses EDDA models as *base learners*. Let $\mathcal{M} = \{1, \dots, M\}$ be the set of EDDA models. The conditional probabilities of classifying an observation $\boldsymbol{x}_i$ to class C_k according to model $m \in \mathcal{M}$ estimated using training data $\mathcal{D}$ (the *level-zero* data) is indicated as $p_{ikm} = \Pr(C_k|\boldsymbol{x}_i; m, \mathcal{D})$, for $i = 1, \dots, n$ observations, $k = 1, \dots, K$ classes, and $m = 1, \dots, M$. Base learners can be used to generate cross-validation predictions, typically using V-fold cross-validation with $V = 10$, to get

$$\widehat{p}_{ikm}^{\text{CV}} = \widehat{\Pr}\left(C_k|\boldsymbol{x}_i; m, \mathcal{D}^{(-v(i))}\right),$$

where $v(i)$ indicates the fold containing the ith observation, and $\mathcal{D}^{(-v(i))}$ is the training set given by all the observations except those in the vth fold. The cross-validated predicted probabilities, along with the vector of original classes, is referred to as the *level-one* data.

The ensemble classifier or *metalearner* defines predicted classification probabilities as the convex linear combination of the base learners predictions:

$$\widehat{\Pr}(C_k|\boldsymbol{x}_i;\boldsymbol{\alpha}) = \widehat{p}_{ik} = \sum_{m=1}^{M} \alpha_m \widehat{p}_{ikm}^{\text{CV}},$$

where $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_M)$ are the ensemble weights, such that $\alpha_m \geq 0$ and $\sum_{m=1}^{M} \alpha_m = 1$. To completely specify the ensemble classifier the optimal combination of base learners is required. This can be achieved by minimizing the cross entropy loss function:

$$\mathcal{L}(\boldsymbol{\alpha}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log(\widehat{p}_{ik}), \tag{2}$$

where $y_{ik} = 1$ if the ith observation is from class k and 0 otherwise, and $\widehat{p}_{ik}$ is the estimated probability that the ith observation belongs to class k.

The optimization of the loss function in (2) is a constrained minimization problem which can be solved in different ways. One efficient approach is to remove the constraints on the ensemble weights by reformulating the problem as an unconstrained optimization using a different parameterization.

Let $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_M) \in \mathcal{S}^M := \left\{\boldsymbol{\alpha} \in [0,1]^M, \sum_{m=1}^{M} \alpha_m = 1\right\}$ be the $(M-1)$-dimensional unit simplex vector. Define the *Unit Simplex Transform* function which maps $\mathcal{S}^M \mapsto \Theta \in \mathbb{R}^{M-1}$ as

$$\theta_m = \text{logit}\left(\frac{\alpha_m}{1-\sum_{m'=0}^{m-1}\alpha_{m'}}\right) + \log(M-m) \qquad \text{for } m = 1,\dots,M-1,$$

where $\text{logit}(x) = \log(x/(1-x))$ and $\alpha_0 = 0$. Backward transformation can be obtained via the *Inverse Unit Simplex Transform* defined as

$$\begin{cases} \alpha_1 = z_1 \\ \alpha_m = \left(1 - \sum\limits_{m'=1}^{m-1} \alpha_{m'}\right) z_m & \text{for } m = 2,\dots,M-1 \\ \alpha_M = 1 - \sum\limits_{m=1}^{M-1} \alpha_m \end{cases}$$

where $z_m = \text{logit}^{-1}\{\theta_m - \log(M-m)\}$, and $\text{logit}^{-1}(x) = 1/(1+\exp(-x))$.

Thus, the unconstrained minimization of the cross entropy loss function in (2) can be pursued with respect to the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{m-1})$, and optimal stacking weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ are obtained via the inverse unit simplex transformation of the solution of such optimization. Numerical algorithms, such as the BFGS quasi-Newton method, typically require initialization of parameters. A natural choice is to consider $\alpha_m = 1/M$ for all $m = 1, \ldots, M$, which amounts to assign the same weight to all the models in the ensemble, and it is equivalent to set $\theta_m = 0$ for $m = 1, \ldots, M-1$. To improve exploration of the search space and to avoid getting trapped in local minima, a multiple restarts strategy can be implemented by generating uniformly distributed values on the M-simplex space, i.e. randomly drawn from a Dirichlet$(1, \ldots, 1)$ distribution.

4 Conclusion

In this contribution we have proposed an ensemble approach to classification based on stacking with Gaussian EDDA mixtures as base learners. The proposal has been applied to both simulated and real datasets (not included here due to space constraints), demonstrating that it is able to improve the overall classification accuracy compared to the best single model among the base learners.

References

BENSMAIL, H., & CELEUX, G. 1996. Regularized Gaussian Discriminant Analysis through Eigenvalue Decomposition. *Journal of the American Statistical Association*, **91**, 1743–1748.

ROKACH, LIOR. 2010. *Pattern Classification Using Ensemble Methods*. World Scientific.

SCRUCCA, LUCA, FOP, MICHAEL, MURPHY, T. BRENDAN, & RAFTERY, ADRIAN E. 2016. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, **8**(1), 205–233.

VAN DER LAAN, MARK J, POLLEY, ERIC C, & HUBBARD, ALAN E. 2007. Super learner. *Statistical Applications in Genetics and Molecular Biology*, **6**(1).

WOLPERT, DAVID H. 1992. Stacked Generalization. *Neural Networks*, **5**(2), 241–259.

A ROBUST QUANTILE APPROACH TO ORDINAL TREES

Rosaria Simone[1], Cristina Davino[2], Domenico Vistocco[1] and Gerhard Tutz[3]

[1] Department of Political Science, University of Naples Federico II, Italy, (e-mail: rosaria.simone@unina.it, domenico.vistocco@unina.it)

[2] Department of Economics and Statistics, University of Naples Federico II, Italy, (e-mail: cristina.davino@unina.it)

[3] Ludwig–Maximilians–Universität München, Germany, (e-mail: tutz@uni-muenchen.de)

ABSTRACT: We propose a quantile tree making use of the one-way quantile ANOVA to check whether two groups of observations of an ordinal response variable differ significantly in a group of quantiles. Specifically, at each node, a quantile ANOVA checks, for each of the available covariates, if the implied split induces significant differences in (at least one of) the selected quantiles. If several splits are significant, the final split will be that with the highest number of significant differences in quantiles, and among them, the one with the strongest overall effects. Since at each step, a multiple testing is applied, the selection of the split is based on the adjusted p–values with the Hochberg correction. An application to the profiling of voting probabilities is used to show the potentiality of the quantile tree for ordinal responses.

KEYWORDS: non-parametric trees, ordinal responses, quantile ANOVA.

1 Introduction and motivation

Decision trees (Breiman *et al.*, 1984) are supervised learning methods aiming at modeling and predicting the value of a response variable based on several explanatory variables. Since they mostly employ an ordinary least squares loss (OLS) function as splitting criterion, the corresponding decision rule is sensitive to outliers and/or skewness in the distribution of the response variable. Moreover, in compliance with classical OLS interpretative rules, results are to be read in light of the effect the predictors exert on the conditional mean of the response. Breiman *et al.*, 1984 extended the regression tree to the median tree through the use of least absolute deviations (LAD) as splitting criterion. Quantile regression trees represent the natural evolution to inspect the conditional quantiles of the response. The proposals in literature differ in the splitting criterion, the used quantiles (a fixed quantile for the whole tree vs different quantiles at the various splits), the type of approach (descriptive vs inferential). In case of a categorical response, the modal value of the terminal node is commonly used to assign the predictive value. However, modal values might not be unique and, in case of an ordinal response, these values perform poorly, since the modal value does not consider the ordering of the categories.

This paper addresses the case of ordinal dependent variables through a robust quantile tree. Given that quantiles can be always defined for ordinal rating data and do not need any scoring rule for categories, we introduce a tree methodology to study the effects of covariates on an ordinal response, exploiting quantile ANOVA (Wilcox, 2017), which is an effective approach to analyze the quantiles of an ordinal distribution. We implemented the recursive partitioning algorithm to detect significant differences in possibly many quantiles, given splitting covariates, at each partitioning level. Our approach is based on inference, i.e. it assesses whether the subgroups are significantly different from each other.

2 A quantile–based classification tree

This section briefly describes the proposed approach to grow a tree through a sequence of splits best discriminating the response variable in terms of a selected grid of quantiles. We refer to an ordinal response variable, even if the generalization to the continuous case is fairly straightforward. Several steps are needed when growing a tree, namely the splitting criterion, the classification rule, the stopping rule, the accuracy measure, among others. Due to the limited space, we discuss here only the splitting criterion, being the originality of the proposals.

Let R be a rating response collected on a support with m ordered categories, labelled using the first m natural numbers, without loss of generality. The splitting criterion enables to identify subgroups (child nodes) significantly different with respect to the quantiles, i.e. the selected location measure. Moreover, the goodness of fit of the decision rule is assessed using a measure that takes into account solely the ordinal nature of the dependent variable. Let $S(\mathbf{q})$ denote the set of quantiles of interest, $S(\mathbf{q}) = \{q_{\tau_{r_1}}, \ldots, q_{\tau_{r_h}}\}$, where q_τ is the quantile of order τ. At each node k, a quantile ANOVA (Wilcox, 2017) is carried out for each of the available covariate to check if the implied binary split induces significant differences in (at least one of) the selected quantiles $S(\mathbf{q})$. At a given node k, and for each candidate binary splitting variable D, whose levels are coded as 0 and 1, let $q_\tau^{(l)}$ and $q_\tau^{(r)}$ the two quantiles of order τ of the conditional response distributions $(R \mid D = 0)$ and

$(R \mid D = 1)$ associated to the left and right descendants, respectively. The procedure will test the null: $H_0 : q_\tau^{(l)} = q_\tau^{(r)}, \forall q_\tau \in S(\mathbf{q})$, against the alternative H_1 : at least one $q_\tau \in S(\mathbf{q}), q_\tau^{(l)} \neq q_\tau^{(r)}$. The chosen split will select the candidate split so that it is one of those with the highest number of significant differences in quantiles, and with the lowest p-value among the competitor splitting variables with the same number of significant differences. The p-vales in this step are adjusted with the Benjamini-Hochberg correction for multiple testing (Benjamini & Hochberg, 1995).

3 An application to German vote data

The performances of the proposed quantile tree for ordinal rating data are illustrated through an application to response profiles for the probability to vote for competing German parties. Data are taken from the GESIS ALLBUS German Social survey (GESIS ALLBUS Leibniz Institute for the Social Sciences, 2012). On a rating scale ranging from 1 = "*very unlikely*", to 10 = "*very likely*", respondents were asked to rate "*How likely it is that you would ever vote for this German party?*". Here we refer to interviewees collected in 2008, and we shall focus on probability to vote for Social Democratic Party (SPD). In the assessment of the electorate belief and behavior, the collection of ratings on the probability to vote for each of the candidate running in an upcoming electoral competition is a much more valuable source of information than the one based on classical voting intentions collected on nominal scales, as it allows to design targeted campaign and to locally understand and predict the electorate behavior. The splitting variables are related to the presence or absence of a series of personal or political-related characteristics of the interviews: participation of the respondent in the previous federal election (*votelast*), supporter of a particular political party (*supportpp*), marital status (*marital*), signing a petition (*petition*), gender, religion (*norel*), catholic, past use of a vote for protest a party (demo), past refuse to vote in some election out of a protest (*refusevote*). We use the following setting for growing up the tree: a maximum depth of 4, a nominal level $\alpha = 0.05$ for the testing procedure of the splitting phase, a minimum samples sizes of 250 at a node for a split to be attempted, and a minimum sample size of 50 required to children of a candidate split to be admissible. The final tree obtained using the grid of quantiles $S(\mathbf{q}) = \{q_{0.1}, q_{0.25}, q_{0.5}, q_{0.75}, q_{0.9}\}$ is shown in Figure 1: it includes nine terminal nodes and seven splitting variables out of the nine candidates (the splitting variables that determine the major number of effects are `demo` and petition). It is worth of notice that the extreme quantiles 0.1 and 0.9 are never chosen as the best quantiles and differences at the first decile are never significant. By following the different paths of the tree from the root to the terminal nodes it is possible to identify different profiles of respondents. For sake of space, results related to the distribution of the dependent variable in the terminal nodes are not shown but a further deepening of the analysis can be achieved by exploring the homogeneity of each profile of respondents.

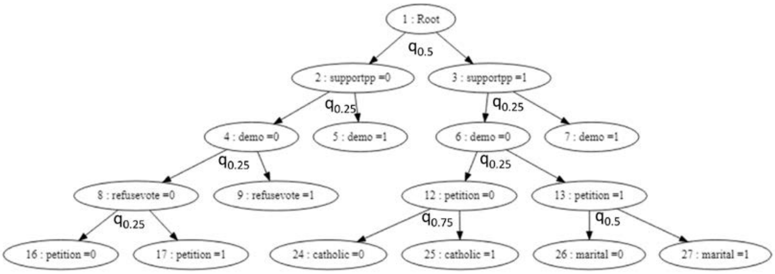


Figure 1. *Quantile tree for rating probabilities for SPD party*

References

BENJAMINI, Y., & HOCHBERG, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B.*, **57**, 289–300.

BREIMAN, L., J.H., FRIEDMAN, OLSHEN, R.A., & C.J., STONE. 1984. *Classification and Regression Trees.* Boca Raton: Chapman & Hall CRC.

GESIS ALLBUS LEIBNIZ INSTITUTE FOR THE SOCIAL SCIENCES, GE. 2012. German General Social Survey (ALLBUS) - Cumulation 1980-2014. Data file version 1.0.0. *GESIS Data Archive.*

WILCOX, R.R. 2017. *Introduction to Robust Estimation and Hypothesis Testing. 4th edition.* Amsterdam, The Netherlands: Elsevier.

THE DETECTION OF SPAM BEHAVIOUR IN REVIEW BOMB

Venera Tomaselli [1], Giulio Giacomo Cantone[2] and Valeria Mazzeo[2]

[1] Department of Political and Social Sciences, University of Catania (e-mail: venera.tomaselli@unict.it)

[2] Department of Physics and Astronomy 'E. Majorana', University of Catania, (e-mail: giulio.cantone@phd.unict.it; valeria.mazzeo@phd.unict.it)

ABSTRACT: In recent years, a new phenomenon called 'Review Bomb' has affected online rating systems. It occurs when a massive amount of accounts reviews a single product, usually negatively to make its reputation slump.
This study analyses the differences among legitimate users and 'review bombers', using common classifiers and techniques from spam detection to identify suspicious reviews, by looking at both content and user's features.

KEYWORDS: review bomb, online ratings, cold start, machine learning.

1 Introduction

Often, before purchasing a product or service, consumers ask for the opinion of their peers who already purchased it. This is commonly referred to as *word-of-mouth* (WOM). A positive opinion among WOM networks is regarded by marketing experts as a valuable and powerful source of reputation for brands. Online rating platforms, or 'review aggregators', are a case of technological innovation for electronic word-of-mouth (eWOM): by browsing a review aggregator, a consumer can read opinions of people who already purchased items (i.e., *evaluands*, such as products, services, place to visits, etc).

Aggregators take this name from the service of recommendation (i.e., a recommender system) they offer. They ask their registered users for submitting a numerical score in a constrained multipoint scale, and then summarise the scores into ratings and rankings (Tomaselli & Cantone, 2020). Scores collected in experimental settings respect methodological assumptions or normality (i.e., independence of observations) but scores collected in online (open) platforms are subject to two biases:

- Purchasing bias, people review what they purchase but they purchase what is already reviewed or, at least, already popular (a case of 'Matthew

Effect');

- Under-reporting bias, people review when they are extremely satisfied or unsatisfied.

The consequence of these biases is a J-shaped distribution of scores in online ratings (Hu *et al.*, 2017; Smironva *et al.*, 2020). These biases make easier to fraud the network of eWOM by injection of fake reviews submitted by the so-called 'sock puppet' accounts. Experimental results confirm that positive fake reviews have an impact on the success of online business (van de Rijt *et al.*, 2014). A consensus on the impact of negative fake reviews has not been reached yet.

Some recommender systems have information if the reviewer purchased the item (e.g., Amazon) but recommender systems generally do not know how much the user is experienced about the item (e.g., how much time spent interacting with that). This issue is related to the fake reviews: one could ask an uninterested friend with an account in the system to rig a review of a item. Should a case like this be considered fake? To overcome such issues, researchers have adopted the broader perspective of 'spam reviews' attack (Hussain *et al.*, 2019). Spam is not necessarily fake but it is an excess of information which is undesired or harmful for the purposes of the system. According to Aggarwal, 2016, a good spam attack, hard to detect, is deployed slowly in the time, so that the sock puppet mimicries the behaviour of a regular user.

Recently, another type of review spam attack has emerged, known as 'Review Bomb', occurring when a massive amount of accounts reviews, usually negatively, attack a single product to make its reputation slump (Tomaselli *et al.*, 2021). During a 'Review Bomb', is often unclear how many accounts are sock puppets and how many accounts are people ideologically driven to review the specific item, but most of them involved lack a history of previous reviews/ratings in the system (*cold-start* problem).

2 Dataset

The dataset includes $N = 59k$ English reviews on the video game *The Last of Us Part II* (TLOU2). TLOU2 was 'review bombed' since its publication date (June 19th, 2020) for ideological reasons (Tomaselli *et al.*, 2021). These reviews were written by registered users on the online platform `metacritic.com`.

From each review, the following metadata are extracted: *i*) username; *ii*) the date the current review was written; *iii*) text of the review; *iv*) score, in a scale

[1:10]; *v*) number of upvotes (i.e., likes) assigned to the review from users, *vi*) number of downvotes (i.e., dislikes) assigned to the review from users; *vii*) number of past ratings that a user provided on Metacritic; *viii*) number of past reviews that a user wrote on metacritic.com. Once collected data, the labelling procedure, consisting of assigning a binary class label whether the review was legitimate (0) or related to the bombing phenomenon (1), is performed.

3 Methods

In the present paper, we propose a methodology for analysing data from a real dataset of TLOU2 reviews, focusing on the online review bomb phenomenon. The data pre-processing stage (data cleaning and handling of missing values) consists of reducing noise words by removing all parts of text which are not relevant for the scope, i.e., punctuation, symbols, and stopwords. Simple Bag-Of-Words and weighted strategy such as Term Frequency-Inverse Document Frequency (TF-IDF) measures are applied to determine term's representativeness. In terms of review's content, some statistical features (e.g., number of punctuation marks, number of unique words, words per sentences) are also extracted.

Techniques for detecting spammer activities on online social networks (Abkenar *et al.*, 2020) and online review platforms (Liu *et al.*, 2017; Harris, 2018) allow to identify accounts involved in review bombing within this dataset. Extra engineered features, therefore, are created to better discriminate not legitimate reviews from legitimate one by looking at users' features, such as username length, username starting with/containing numbers among others.

To reduce the dimensionality of the data and improve the results of the analysis, the most relevant features are selected to enter the model. Popular statistical tests, such as Pearson's test and Chi-squared, are used for this purpose, since they can handle numerical and categorical variables, respectively.

Once got the most important features, these ones are then passed into the classification algorithms to produce a range of models to predict not legitimate reviews. A k-Fold Cross Validation technique is considered to compare different machine learning algorithms ((e.g., Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine); Nematzadeh *et al.*, 2015), generally used in spam (Al-Zoubi *et al.*, 2021) and fake news/reviews detection. Finally, model performance is evaluated by scoring the outcomes from a test set, using precision, accuracy, recall, and F_1 score (Zheng *et al.*, 2015) metrics.

References

ABKENAR, S. B., KASHANI, M. H., AKBARI, M., & MAHDIPOUR, E. 2020. Twitter Spam Detection: A Systematic Review. *ArXiv*, **abs/2011.14754**.

AGGARWAL, C. C. 2016. *Recommender Systems*. Springer-Verlag.

AL-ZOUBI, AM, ALQATAWNA, J., FARIS, H., & HASSONAH, MA. 2021. Spam profiles detection on social networks using computational intelligence methods: The effect of the lingual context. *Journal of Information Science*, **47**(1), 58–81.

HARRIS, C. G. 2018. Decomposing TripAdvisor: Detecting Potentially Fraudulent Hotel Reviews in the Era of Big Data. *Pages 243–251 of: 2018 IEEE International Conference on Big Knowledge (ICBK)*.

HU, N., PAVLOU, P., & ZHANG, J. 2017. On Self-Selection Biases in Online Product Reviews. *Management Information Systems Quarterly*, **41**(2), 449–471.

HUSSAIN, N., TURAB MIRZA, H., RASOOL, G., HUSSAIN, I., & KALEEM, M. 2019. Spam Review Detection Techniques: A Systematic Literature Review. *Applied Sciences*, **9**(5), 987.

LIU, P., XU, Z., AI, J., & WANG, F. 2017. Identifying Indicators of Fake Reviews Based on Spammer's Behavior Features. *Pages 396–403 of: 2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)*.

NEMATZADEH, Z., IBRAHIM, R., & SELAMAT, A. 2015. Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques. *2015 10th Asian Control Conference (ASCC)*, 1–6.

SMIRONVA, E., KIATKAWSIN, K., LEE, S. K., KIM, J., & LEE, C.-H. 2020. Self-selection and non-response biases in customers' hotel ratings – a comparison of online and offline ratings. *Current Issues in Tourism*, **23**(10), 1191–1204.

TOMASELLI, V., & CANTONE, G. G. 2020. Evaluating Rank-Coherence of Crowd Rating in Customer Satisfaction. *Social Indicators Research*.

TOMASELLI, V., CANTONE, G. G., & MAZZEO, V. 2021. The polarising effect of Review Bomb. *ArXiv*, **abs/2104.01140**.

VAN DE RIJT, A., KANG, S. M., RESTIVO, M., & PATIL, A. 2014. Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences*, **111**(19), 6934–6939.

ZHENG, X., ZENG, Z., CHEN, Z., YU, Y., & RONG, C. 2015. Detecting spammers on social Networks. *Neurocomputing*, **42**(02).

CLUSTERING MODELS FOR THREE-WAY DATA

Donatella Vicari[1] and Paolo Giordani[1]

[1] Department of Statistical Sciences, Sapienza University of Rome
(e-mail: donatella.vicari@uniroma1.it, paolo.giordani@uniroma1.it)

ABSTRACT: A novel clustering model for three-way data concerning a set of objects on which variables are measured by different subjects is proposed. The main aim of the model is to summarize the objects through a limited number of clusters. In order to exploit the three-way structure of the data, such clusters are assumed to be common to all subjects and variables and subjects are summarized through the PARAFAC model.

KEYWORDS: *K*-Means, PARAFAC, Variable weighting.

1 Introduction

Nowadays, it is very frequent to analyze data corresponding to variables measured on some objects by a set of subjects. Such three-way data can be stored in a (three-way) array or tensor. It can be interesting to discover clusters of homogeneous objects with respect to the variables measured by the subjects. However, classical (two-way) clustering techniques are usually inadequate to handle three-way data. To this purpose, several three-way extensions have been developed following the model-based (see, e.g., Viroli, 2011) or least-squares approaches. Here, we propose a new clustering model for three-way data according to the least-squares approach. It can be seen as a three-way extension of the well-known *k*-Means algorithm (MacQueen, 1967) where, in particular, the three-way nature of the model is exploited by considering the so-called PARAFAC model, independently developed by Carroll & Chang (1970) and Harshman (1970). In the PARAFAC model, data are summarized by a limited number of components. As such, the PARAFAC model represents a three-way extension of classical Principal Component Analysis.

The paper is organized as follows. In the next section, we briefly review alternative clustering models for three-way data. Section 3 deals with the proposed model. Some final comments are made in Section 4.

2 Related works

The clustering problem of three-way data has received a great deal of attention over the last few years. We can roughly distinguish two main classes of techniques aiming at partitioning the entities referring to a single way or to more ways simultaneously. A common case is referred to as bi-clustering or co-clustering when two ways, usually objects and variables, are clustered (for a comprehensive review see Madeira & Oliveira, 2004). In this paper, we are going to focus on the first class of models, seeking, without loss of generality, a partition of objects.

Wilderjans & Ceulemans (2013) introduced the so-called Clusterwise PARAFAC, where objects are assigned to a limited number of clusters and, simultaneously, a *standard* PARAFAC model is applied within each cluster. In other words, within each cluster, objects, variables and subjects are summarized through a limited number of components. Therefore, the main idea of the Clusterwise PARAFAC model is that objects assigned to the same clusters have the same component structure, whereas objects belonging to different clusters have different underlying components.

A different approach is followed by Rocci & Vichi (2005). First of all, the PARAFAC model is replaced by the Tucker3 model (Tucker, 1966). Tucker3 is more general than PARAFAC. In fact, the *standard* Tucker3 model allows for different numbers of components for objects, variables and subjects. Unfortunately, the Tucker3 solution suffers from rotational indeterminacy. On the contrary, the PARAFAC solution is unique (up to scaling and permuting the components) under mild conditions. Rocci & Vichi (2005) suggested to summarize only variables and subjects through components, whilst objects are partitioned into a reduced number of clusters. It follows that objects are analyzed asymmetrically with respect to variables and subjects. Specifically, objects are assigned to clusters following a *K*-Means-type procedure where the cluster prototypes lie onto the low-dimensional space spanned by the components for the variables and the subjects. The partition of the objects and the dimensionality reduction of both variables and subjects is performed simultaneously in such a way that the components explain the between-cluster variability. In this respect, Rocci & Vichi (2005) is actually a generalization of the Reduced *K*-Means method for standard two-way data (De Soete & Carroll, 1994).

In the next section, we present a new clustering model for three-way data, which takes inspiration from the previously mentioned proposals. Namely, consistently with Rocci & Vichi (2005), objects play an asymmetric role with respect to variables and subjects, and consistently with Wilderjans & Ceulemans (2013), the PARAFAC model is used for its simplicity and the uniqueness property. As we shall see, it can be interpreted as a K-Means type clustering model for three-way data.

3 The clustering model

Let us suppose J variables are measured on N objects by H subjects. Such data are stored in the three-way array $\underline{\mathbf{X}}$ of order ($N \times J \times H$), whose generic element is x_{njh}, expressing the measurement of object n ($n = 1, \dots, N$) with respect to variable j ($j = 1, \dots, J$) made by subject h ($h = 1, \dots, H$). The array $\underline{\mathbf{X}}$ can be seen as a collection of matrices, one for every subject. Therefore, matrix $\mathbf{X}_h$ ($h = 1, \dots, H$) of size ($N \times J$), usually referred to as slice, contains all measurements from subject h.

The most general model can be fully specified as follows:

$$\mathbf{X}_h = \mathbf{U}_h\mathbf{Y}_h + \mathbf{E}_h,\ h = 1, \ldots, H, \tag{1}$$

where $\mathbf{E}_h$ is the error term for subject h and $\mathbf{U}_h$ is the membership matrix of order ($N \times K$) for objects into clusters, being K the number of clusters. Matrix $\mathbf{U}_h$ is binary with only one entry equal to 1 per row and identifies a partition of the N objects into K disjoint clusters for subject h ($h = 1, \ldots, H$). Matrix $\mathbf{Y}_h$ ($h = 1, \ldots, H$) of order ($K \times J$) is the subject-specific prototype matrix. Thus, the model assumes a *different* partition among the slices referring to the subjects. In other words, separate partitions are sought by means of a K-means-type model for every subject.

In order to exploit the three-way structure of the data, i.e., to properly take into account that the same variables are observed on the same objects by the subjects, constrained versions of model (1) can be derived. For instance, we may assume that $\mathbf{U}_h = \mathbf{U}$, $h = 1, \ldots, H$. Then, we get

$$\mathbf{X}_h = \mathbf{U}\mathbf{Y}_h + \mathbf{E}_h,\ h = 1, \ldots, H. \tag{2}$$

Matrix $\mathbf{U}$ is the allocation matrix, fulfilling the same constraints as for $\mathbf{U}_h$. Therefore, model (2) identifies a *common* partition across subjects. As in model (1), different prototype matrices $\mathbf{Y}_h$ are assumed allowing for possible differences among subjects. Model (2) is therefore a K-means-type model with a consensus partition specified by $\mathbf{U}$.

Model (2) can be further extended by considering the PARAFAC model. Specifically, setting $\mathbf{Y}_h = \mathbf{D}_h\mathbf{B}$, model (2) can be rewritten as

$$\mathbf{X}_h = \mathbf{U}\mathbf{D}_h\mathbf{B} + \mathbf{E}_h,\ h = 1, \ldots, H, \tag{3}$$

where $\mathbf{D}_h$ ($h = 1, \ldots, H$) is the diagonal matrix of order ($K \times K$) with diagonal elements giving subject-specific weights for the K clusters. Matrix $\mathbf{B}$ of order ($K \times J$) measures the relevance of the variables for the K clusters. The three-way structure of the data is captured by the matrices $\mathbf{D}_h$ ($h = 1, \ldots, H$) and $\mathbf{B}$. In fact, since the same matrix $\mathbf{B}$ is assumed across subjects, the underlying idea of model (3) is that the slices are described by the same matrices $\mathbf{U}$ and $\mathbf{B}$, but in different proportions because $\mathbf{B}$ is weighted differently through the subject-specific matrices $\mathbf{D}_h$ ($h = 1, \ldots, H$).

The proposed model is a PARAFAC model with binary constraints on $\mathbf{U}$. It is a special case of the so-called NMFA/GENNCLUS model, mentioned by Carroll & Chaturvedi (1995). The solution is unique up to scaling and cluster labeling, as it holds for PARAFAC. Such a solution can be found according to the least-squares approach by minimizing the loss function

$$\textstyle\sum_h \| \mathbf{E}_h \|^2, \tag{4}$$

with respect to $\mathbf{U}$, $\mathbf{Y}$ and $\mathbf{D}_h$ ($h = 1, \ldots, H$), being $\| \cdot \|$ the Frobenius norm of matrices. For this purpose, an Alternating Least-Squares algorithm has been implemented.

Model (3) can be extended along various directions. For instance, it might be fruitful to introduce subject-specific weights for the variables tuning their importance in the clustering process. Such weights might be objectively estimated by minimizing loss function (4).

4 Concluding remarks

The paper introduced a novel *K*-Means type clustering model for three-way data involving the PARAFAC decomposition. The effectiveness of the proposal will be illustrated with simulated and real applications and its possible extensions will be presented during the meeting.

References

CARROLL, J. D., & CHANG, J. J. 1970. Analysis of individual differences in multidimensional scaling via an *n*-way generalization of Eckart–Young decomposition. *Psychometrika*, **35**, 283-319.

CARROLL, J. D., & CHATURVEDI, A. 1995. A general approach to clustering and multidimensional scaling of two-way, three-way or higher-way data. *Geometrical Representations of perceptual phenomena*. Mahwah, NJ: Lawrence Erlbaum, 295-318.

DE SOETE, G., & CARROLL, J. D. 1994. *k*-means clustering in a low-dimensional Euclidean space. *New approaches in classification and data analysis*. Heidelberg: Springer Verlag, 212-219.

HARSHMAN, R. A. 1970. Foundations of the PARAFAC procedure: Models and conditions for an ‘explanatory’ multi-modal factor analysis. *UCLA Working Papers in Phonetics*, **16**, 1-84.

MACQUEEN, J. B. 1967. Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, **1**, 281-297.

MADEIRA, S. C. & OLIVEIRA, A. L. 2004. Biclustering algorithms for biological data analysis: A survey. *IEEE Transactions in Computational Biology and Bioinformatics*, **1**, 24-45.

ROCCI, R. & VICHI, M. 2005. Three-mode component analysis with crisp or fuzzy partition of units. *Psychometrika*, **70**, 715-736.

TUCKER, L. R 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279-311.

VIROLI, C. 2011. Model based clustering for three-way data structures. *Bayesian Analysis*, **6**, 573-602.

WILDERJANS, T. F., & CEULEMANS, E. 2013. Clusterwise Parafac to identify heterogeneity in three-way data. *Chemometrics and Intelligent Laboratory Systems, **129***, 87-97.

USING EYE-TRACKING DATA TO CREATE A WEIGHTED DICTIONARY FOR SENTIMENT ANALYSIS: THE EYE DICTIONARY

Gianpaolo Zammarchi[1], Jaromír Antoch[2]

[1] Department of Economics and Business Sciences, University of Cagliari,
(e-mail: gp.zammarchi@unica.it)

[2] Department of Mathematics and Physics, Charles University,
(e-mail: antoch@karlin.mff.cuni.cz)

ABSTRACT: Extracting information from written texts is of paramount importance to many entities (e.g. businesses, public organizations, individuals), but the exponential growth of available data has made this task beyond any single human being or business. Sentiment analysis is a tool to automatically transform the information extracted into knowledge. One of the main challenges is to assess if a text is positive or negative, which can be tackled using a dictionary where each word has a positive or negative associated value and then combining single-words values to express an overall text sentiment. In order to use such lexicon-based approach, we need an existing dictionary or to build a new one. In this work we present a new dictionary for sentiment analysis developed using eye-tracking data to determine the relevance of words and we assess its performances against other existing dictionaries.

KEYWORDS: eye-tracking, sentiment analysis, lexicon, dictionary.

1 Introduction

Sentiment analysis is aimed at classifying texts into sentiments with a polarity (positive or negative) using different approaches. The lexicon-based approach is based on a dictionary, i.e. a base tool where hundreds or thousands of words are associated with a polarity (negative/positive). In order to classify the polarity of a text, each word is searched in the dictionary. If the word is present, the value assigned to that word will contribute to the overall text sentiment (along with the other words present both in the text and in the dictionary). To obtain a single value representative of the whole text a summarizing function (e.g. average or sum) is applied. An important challenge in sentiment analysis is the definition of weights to attribute to words, i.e. to have instruments to define which words should be assigned greater importance. In this sense, the eye tracking technology, which allows to measure the exact position of the eyes during the visualization of texts, images or other visual stimuli, can be of help to understand which words might be able to gain more attention from a reader and are thus potentially more relevant.

Aim of the present method is to develop a new dictionary for sentiment analysis using eye-tracking data as weights to attribute a different relevance to the words in a text, based on the attention they might receive.

2 Materials and methods

2.1 Development of the Eye-dictionary

To develop a dictionary based on eye tracking data, we focus on two main aspects: weights and polarities. Weights have been computed based on the ProvoCorpus, a large corpus including eye tracking data for 55 paragraphs taken from various sources (e.g. news articles, science magazines and public domain works of fiction). Each paragraph was read by an average of 40 participants. Across all texts, eye tracking data in the form of dwell time for each word (i.e. total reading time calculated as the summation of the duration across all fixations on a given word) are available for a total of 2,689 words (1,191 of which are unique). For each word w included in the corpus of eye tracking data, the average dwell time based on the total number of occurrences of the word in the corpus is calculated as in Eq. (1)

$$\frac{1}{n}\sum_{i=1}^{n} d_i^w \tag{1}$$

where n is the number of occurrences of a word w in the dataset and d^w is the dwell time for the word w. The average global dwell time for any word in the dataset is computed as in Eq. (2)

$$\frac{1}{m}\sum_{i=1}^{m} d_i \tag{2}$$

where m is the number of all occurrences of all words observed in the dataset and d_i is the dwell time for the occurrence i of a word in the dataset. Each weight v for each word w is then calculated as the ratio in Eq. (3)

$$v^w = \frac{\frac{1}{n}\sum_{i=1}^{n} d_i^w}{\frac{1}{m}\sum_{i=1}^{m} d_i} \tag{3}$$

and these values have been normalized using the min-max normalization. Polarities are computed using a large dataset of movie reviews including 50,000 texts, labeled as positive and negative reviews (Maas et al., 2011). To assess if a word has a positive or negative polarity, we compute a probability in the form of Eq. (4):

$$P(w_{pos}) = \frac{N_{w_{pos}}}{N_w} \qquad P(w_{neg}) = \frac{N_{w_{neg}}}{N_w} \tag{4}$$

where $P(w_{pos})$ is the probability that the word w is positive, $N_{w_{pos}}$ is the number of occurrences of the word w in positive labeled texts and N_w is the number of occurrences of the word w. The same computation is made for negatives. Given the probabilities in Eq. (4) we assign a polarity p to each word w as in Eq. (5)

$$p^w = \begin{cases} 1 & if\ P(w_{pos}) > P(w_{neg}) \\ 0 & if\ P(w_{pos}) = P(w_{neg}) \\ -1 & otherwise \end{cases} \quad (5)$$

Therefore, we assign the word w a positive (+1) or negative value (-1) in case $P(w_{pos})$ is greater or lower than 0.5, respectively. If the probability is exactly 0.5 the word w is assigned 0 (neutral). For each word, a final value s is then computed as the product of weights and polarities as in Eq. (6)

$$s^w = v^w \cdot p^w \quad (6)$$

2.2 Assessment of the performance of the Eye dictionary and comparison with existing dictionaries

The performance of the dictionary based on eye tracking data in the classification of sentiment polarity of texts has been assessed using two independent collections of labeled texts: 1,000 consumer reviews from Amazon (McAuley et al., 2013) and 1,000 consumer reviews from Yelp (Yelp dataset). For these texts, the performance of the Eye dictionary in the classification of sentiment polarity is compared with four existing dictionaries: Loughran-McDonald (2,702 words), SentiWordNet 3.0 (20,093 words), SO-CAL Google (3,290 words) and Hu Liu (6,874 words) extracted from the Lexicon package in R (Rinker, 2018). For each text, a polarity value is calculated as the algebraic sum of signed values assigned to each word by a dictionary. Finally, the number of texts correctly classified using the different dictionaries is compared.

3 Results

A total of 1,185 words for which weights and polarities were computed are included in the Eye dictionary (619 positive, 466 negative and 100 neutral). Table 1 shows the performance of the Eye dictionary and four other dictionaries in terms of precision, recall, F1-score and accuracy for the Yelp dataset (similar results were obtained using the Amazon dataset).

The Eye dictionary showed the best precision for positive texts, best recall for negative texts and the second-best accuracy after the Hu Liu dictionary. The Eye dictionary was able to correctly classify a higher number of texts compared to two of the four dictionaries (Loughran and Socal Google) in the Amazon dataset and three of the four dictionaries (Loughran, Sentiword and Socal Google) in the second dataset. Hu Liu was the only dictionary to show a better performance in both datasets.

Overall, all dictionaries only showed a modest performance in this preliminary analysis, which could be improved with the application of rules for handling cases such as presence of negations, amplifiers and downtoners. Notably, the Eye dictionary was able to achieve a performance similar or better compared to most of the other dictionaries even if it includes a much lower number of words.

Table 1. Comparison between Eye dictionary and four other dictionaries

	Eye dictionary		Loughran-McDonald		SentiWord Net		SO-CAL Google		Hu Liu	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Precision	0.60	0.55	0.38	0.30	0.54	0.56	0.48	0.42	0.58	0.68
Recall	0.39	0.74	0.46	0.23	0.63	0.46	0.74	0.19	0.81	0.41
F1-score	0.47	0.63	0.41	0.26	0.58	0.51	0.58	0.27	0.67	0.51
Accuracy	0.56		0.35		0.55		0.46		0.61	

4 Conclusions

In this work we present a new sentiment analysis dictionary built by leveraging eye tracking data to assign weights to words based on their ability to gain attention from a reader. To this aim, dwell time is used as a measure of relevance of a word. Future developments include the expansion of the number of words included in the dictionary as well as evaluation of its performance in the classification of text using rules to handle cases in which classification is particularly challenging, such as sentences including negations, amplifiers and downtoners.

References

KOTZIAS, D., DENIL, M., DE FREITAS, N., & SMYTH, P. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, USA, 597–606.

MAAS, A., DALY, R.E., PHAM, P.T., HUANG, D., NG, A.Y., & POTTS, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11).* ACL, USA, 142–150.

MCAULEY, J.J., & LESKOVEC, J. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. *RecSys '13: Proceedings of the 7th ACM conference on Recommender systems*, 165–172.

RINKER, T.W. 2018. lexicon: Lexicon Data version 1.2.1. http://github.com/trinker/lexicon

The book collects the short papers presented at the 13th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS). The meeting has been organized by the Department of Statistics, Computer Science and Applications of the University of Florence, under the auspices of the Italian Statistical Society and the International Federation of Classification Societies (IFCS). CLADAG is a member of the IFCS, a federation of national, regional, and linguistically-based classification societies. It is a non-profit, non-political scientific organization, whose aims are to further classification research.

Giovanni C. Porzio PhD, is Professor of Statistics in the Department of Economics and Law at the University of Cassino and Southern Lazio. His research interests include directional statistics, statistical learning, nonparametric multivariate analysis and data depth, graphical methods and data visualization.

Carla Rampichini PhD, is full professor of Statistics and head of the Department of Statistics, Computer Science and Applications 'G. Parenti' of the University of Florence. Her research interests relate to random effects models for multilevel analysis, multivariate analysis and evaluation of educational systems.

Chiara Bocci PhD, is a Researcher in Statistics at the Department of Statistics, Computer Science and Applications "G. Parenti" of the University of Florence. Her current research interests include statistical analysis of spatially referenced data, small area estimation methods, and statistical models for skewed variables.

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 978-88-5518-340-6 (PDF)
ISBN 978-88-5518-341-3 (XML)
DOI 10.36253/978-88-5518-340-6
www.fupress.com