CLADAG 2023

BOOK OF ABSTRACTS AND SHORT PAPERS

14th Scientific Meeting of the Classification and Data Analysis Group
Salerno, September 11-13, 2023

edited by

Pietro Coretto
Giuseppe Giordano
Michele La Rocca
Maria Lucia Parrella
Carla Rampichini

Pearson

SCIENTIFIC PROGRAM COMMITTEE

Carla Rampichini (chair, University of Florence - Italy)

Claudio Agostinelli (University of Trento - Italy)
Michela Battauz (University of Udine - Italy)
Antonio Canale (University of Padua - Italy)
Carlo Cavicchia (Erasmus University Rotterdam - Netherlands)
Claudio Conversano (University of Cagliari - Italy)
Eustasio del Barrio (University of Valladolid - Spain)
Roberto Di Mari (University of Catania - Italy)
Stefania Fensore (University of "G. d'Annunzio" - Italy)
Nial Friel (University College Dublin - Ireland)
Maria Giovanna Ranalli (University of Perugia - Italy)
Leonardo Grilli (University of Firenze - Italy)
Luigi Grossi (University of Padua - Italy)
Christian Hennig (University of Bologna - Italy)
Mia Hubert (KU Leuven - Belgium)
Alfonso Iodice D'Enza (University of Naples "Federico II" - Italy)
Julien Jacques (University of Lyon - France)
José Joaquim Dias Curto (ISCTE-Instituto Universitário de Lisboa- Portugal)
Michele La Rocca (University of Salerno - Italy)
Silvia Montagna (University of Turin - Italy)
Barbara Pawelek (University of Cracow - Poland)
Fulvia Pennoni (University of Milano-Bicocca - Italy)
Mario Rosario Guarracino (University of Cassino - Italy)
Katrijn Van Deun (University of Tilburg - Netherlands)
Simone Vantini (Politecnico di Milano - Italy)
Donatella Vicari (Sapienza University of Rome - Italy)
Helga Wagner (Johannes Kepler University Linz - Austria)
Hiroshi Yadohisa (Doshisha University - Japan)

LOCAL PROGRAM COMMITTEE

Michele La Rocca (chair, University of Salerno - Italy)

Pietro Coretto (University of Salerno - Italy)
Giuseppe Giordano (University of Salerno - Italy)
Paolo Rocca Comite Mascambruno (University of Salerno - Italy)
Marcella Niglio (University of Salerno - Italy)
Maria Lucia Parrella (University of Salerno - Italy)
Marialuisa Restaino (University of Salerno - Italy)
Domenico Vistocco (University of Naples "Federico II" - Italy)
Maria Prosperina Vitale (University of Salerno - Italy)

CLADAG 2023 BOOK OF ABSTRACTS AND SHORT PAPERS:
14th Scientific Meeting of the Classification and Data Analysis Group, Salerno, September 11-13, 2023
edited by Carla Rampichini, Michele La Rocca, Pietro Coretto, Giuseppe Giordano, Maria Lucia Parrella

Front cover: Genome sequence map, chromosome architecture and genetic sequencing chart abstract data,
© Tartila / Shutterstock

Published by Pearson Education Resources, Italia
www.pearson.it
ISBN: 9788891935632

INDEX

Preface

This book collects the abstracts and short papers presented at CLADAG 2023, the 14th 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 Economics and Statistics of the University of Salerno, under the auspices of the University of Salerno, 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 2023 conceived the Keynote Sessions to provide a fresh perspective on the state of the art of knowledge and research in the field. The scientific program of CLADAG 2023 is particularly rich. All in all, it comprises 5 Keynote Lectures, 31 Invited Sessions promoted by the members of the Scientific Program Committee, and 27 Contributed Sessions. 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. Special thanks are finally due to the members of the Local Organizing Committee and all the people who collaborated for CLADAG 2023. Last but not least, we thank all the authors and participants, without whom the conference would not have been possible.

Pietro Coretto
Giuseppe Giordano
Michele La Rocca
Maria Lucia Parrella
Carla Rampichini

Salerno, September 2023

Plenary Session

DISCRETE LATENT VARIABLE MODELS: RECENT ADVANCES AND PERSPECTIVES

Francesco Bartolucci[1], Michael Greenacre[2],
Silvia Pandolfi[1] and Fulvia Pennoni[3]

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

[2] Department of Economics, Universitat Pompeu Fabra, ES (e-mail: michael.greenacre@upf.edu)

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

ABSTRACT: After a review of the class of discrete latent variable models in terms of formulation and estimation methods, recent advances and perspectives regarding these models are illustrated. We consider in detail the stochastic block model for social networks and models for spatio-temporal data. Among these developments, we discuss, in particular, the analysis of longitudinal compositional data about expenditures of the Spanish regions over several decades.

KEYWORDS: Compositional data, data augmentation, expectation-maximization algorithm, spatio-temporal modeling, variational inference.

1 Introduction

In general terms, latent variable models include variables not directly observable to describe the relation between observable variables. Among these models, those based on the assumption that the latent variables follow a discrete distribution, namely discrete latent variable (DLV) models, are nowadays commonly used (for a recent review, see Bartolucci *et al.*, 2022). With respect to models based on continuous latent variables, DLV models present some advantages, such as the flexibility and capability of clustering units in different latent groups, also named components, classes, or states. Obviously, there are also issues that may complicate the use of DLV models such as the selection of the number of support points of the discrete distribution of the latent variables and the multimodality of the likelihood function.

The first aim of this work is to provide a critical review of DLV models in terms of formulation and estimation methods. Regarding the first aspect, we describe recent proposals that can be used to deal with complex data struc-

tures such as social networks and spatio-temporal data. In particular, for the analysis of social networks we consider the stochastic block model and its extended versions that may be used in a longitudinal context where individuals are repeatedly observed in terms of social behavior. For the analysis of spatio-temporal data, we illustrate models based on latent variables which are specific to each site and time of observation. We also consider recent formulations which may be used to make causal inference on a certain policy or treatment and that conceive potential versions of the latent variables to properly define causal effects (Lanza *et al.*, 2013).

Regarding estimation, we show that both frequentist and Bayesian inferential approaches rely either on methods that directly assign the units to the different components or methods in which this explicit assignment is avoided. Among the methods of the first type, it is worth recalling those based on the maximization, with respect to the model parameters and the assignment of units to the components, of the so-called classification likelihood and the corresponding Bayesian methods based on Markov chain Monte Carlo (MCMC) algorithms (Gelman *et al.*, 2011) with data augmentation, where the latent variables are considered on the same footing as the model parameters. Estimation methods of the second type are instead based on popular algorithms such as the expectation-maximization (EM Dempster *et al.*, 1977) applied to find the maximum likelihood estimate of the parameters and corresponding MCMC algorithms for Bayesian inference. We also describe variational methods (see, among others, Daudin *et al.*, 2008), used for complex contexts, and in general we pay attention to the problem of scalability (Bartolucci *et al.*, 2018).

The second aim of the present work is to illustrate a new possible application of the DLV models to the analysis of temporal and spatio-temporal compositional data, as is briefly described in the following section.

2 Analysis of spatio-temporal compositional data

This development is motivated by the availability of a recent dataset about the composition of the annual investments in different sectors of the Spanish economy, for a long period that goes from 1964 to 2020 (García *et al.*, 2023). In the present work we concentrate mainly on the simpler problem of the national data on the temporal scale, mentioning later how to broaden this to the more detailed spatio-temporal scale across the different autonomous regions of Spain. The data are thus collected in the $m \times 1$ vectors $\mathbf{y}_t$, $t = 1, \ldots, T$, where T is the number of time occasions and m the number of sectors. For the spatio-temporal framework, the data would be in vectors $\mathbf{y}_{it}$, $i = 1, \ldots, n$, $t = 1, \ldots, T$,

where n is the number of regions. The changing total amount invested across the years is, of course, important to analyze, but here it is the changing composition of the investments that is of interest, namely the amounts invested each year relative to their respective totals. Hence, compositional data are such that the sum of the elements of each compositional response vector is fixed at 1 or 100% (see Greenacre, 2021, for a recent review). This has crucial implications in terms of data analysis. Two approaches are presented here: first, an exploratory approach where the logratio transformation is used (Greenacre, 2018); and second, where the data are assumed to follow the Dirichlet distribution on the unit interval. For the logratio approach the simplest transformation is the so-called additive logratio transformation, where all compositional parts are expressed as a ratio with a fixed part, and then log-transformed. These transformed data can then be analyzed using existing approaches for multivariate interval-scale data, assuming multivariate normal distribution.

For the regional data at hand we formulate different models. The starting one is of hidden Markov type and does not account for the spatial dependence between the regions. It only accounts for temporal dependence. For every region, this model assumes that each time-specific vector of response variables $\mathbf{y}_{it}$, corresponding to parts of the composition, follows a Dirichlet distribution with parameters that depend on an underlying discrete latent variable. In symbols, we have

$$\mathbf{Y}_{it}|U_{it} = u \sim \text{Dir}(\boldsymbol{\alpha}_u),$$

where U_{it} is the underlying latent variable having support $\{1,\ldots,k\}$ and $\boldsymbol{\alpha}_u$ is the state-specific vector of parameters.

Moreover, each sequence of latent variables $U_{i1},\ldots,U_{iT}$ follows a Markov chain with initial probabilities and transition probabilities that, without covariates, are denoted by $\lambda_u = p(U_{i1} = u)$ and $\pi_{u|\bar{u}} = p(U_{it} = u|U_{i,t-1} = \bar{u})$, $t = 2,\ldots,T$. With unit-specific covariates, these probabilities are formulated by suitable logit parametrizations based on regression coefficients to account for the effect of such covariates. This formulation is based on the usual assumption that the response variables are conditionally independent given the latent variables. Regarding the parametrization of the Dirichlet distribution, we follow an approach that separates the effects of the latent states on the expected value and on the variance (see also Maier, 2014).

We also consider a spatio-temporal model where, following recent approaches (e.g., Bartolucci & Farcomeni, 2022), the latent state of a region in a certain year may depend, not only on the previous state, but also on the state of the neighbor regions. More precisely, each latent variable U_{it} is modeled conditionally on $U_{i,t-1}$ and U_{jt}, $j \in \mathcal{N}_i$, where $\mathcal{N}_i$ is the set of neighbors of region

i. Even in this case, multinomial logit parametrizations are adopted to include the effect of possible covariates. Again, we rely on the assumption of conditional independence between the response vectors given the latent variables that has an interesting interpretation and simplifies the estimation process.

References

BARTOLUCCI, F., & FARCOMENI, A. 2022. A hidden Markov space–time model for mapping the dynamics of global access to food. *Journal of the Royal Statistical Society, Series A*, **185**, 246–266.

BARTOLUCCI, F., BACCI, S., & MIRA, A. 2018. On the role of latent variable models in the era of big data. *Statistics & Probability Letters*, **136**, 165–169.

BARTOLUCCI, F., PANDOLFI, S., & PENNONI, F. 2022. Discrete latent variable models. *Annual Review of Statistics and Its Application*, **9**, 425–452.

DAUDIN, J.-J., PICARD, F., & ROBIN, S. 2008. A mixture model for random graphs. *Statistics and Computing*, **18**, 173–183.

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*, **39**, 1–22.

GARCÍA, F. P., IVARS, M. M., RADOSELOVICS, J. F. G., CANDAU, E. B., & DOMÍNGUEZ, J. C. R. 2023. El stock de capital en España y sus comunidades autónomas Análisis de los cambios en la composición de la inversión y las dotaciones de capital entre 1995 y 2022. *Documentos de Trabajo, Fundación BBVA*.

GELMAN, A., JONES, A., & MENG, X. L. 2011. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: CRC Press.

GREENACRE, M. 2018. *Compositional Data Analysis in Practice*. Boca Raton, FL: CRC Press.

GREENACRE, M. 2021. Compositional data analysis. *Annual Review of Statistics and its Application*, **8**, 271–299.

LANZA, S. T., COFFMAN, D. L., & XU, S. 2013. Causal inference in latent class analysis. *Structural Equation Modeling*, **20**, 361–383.

MAIER, M. 2014. DirichletReg: Dirichlet regression for compositional data in R. *Institute for Statistics and Mathematics, Report 125*.

SELECTIVE INFERENCE AFTER VARIABLE SELECTION BY THE RANDOMIZED GROUP LASSO METHOD

Gerda Claeskens[1], Sarah Pirenne[1], Snigdha Panigrahi[2], and Yiling Huang[2]

[1] ORStat and Leuven Statistics Research Center, KU Leuven, (e-mail: gerda.claeskens@kuleuven.be; sarah.pirenne@kuleuven.be)

[2] Department of Statistics, University of Michigan, (e-mail: psnigdha@umich.edu; yilingh@umich.edu)

ABSTRACT: The randomized group lasso method performs a selection of groups of variables in a model and returns estimates of the coefficients in the selected model. The lasso estimator is a special case when all groups have size one. Typically, one is interested in inference for only those model coefficients that appear in the selected model and non-selected coefficients are further ignored. In selective inference one obtains valid confidence intervals and P-values for the model coefficients after selection, when conditioning on the event of the selection. We consider this problem in the framework of a general class of loss functions and distributions, including the generalized linear models, but also quasi-likelihood models that can deal with overdispersed data, for example. Our method allows the models to contain both categorical or grouped covariates as well as continuous covariates. We use an additional randomization during the group lasso estimation stage, which allows us to define a post-selection likelihood. We show that this likelihood function can be used for selective inference when conditioning on the event of the selection. An additional bonus is the selective point estimator obtained from this likelihood, which accounts by construction for the selection of the variables by the group lasso method. The confidence intervals for the regression coefficients in the selected model can be constructed in the familiar Wald-type way and we show that they have bounded lengths. We illustrate the selective inference method for grouped lasso on data from the national health and nutrition examination survey.

KEYWORDS: group lasso estimation, likelihood estimation, post-selection inference, selective inference

TO GET THE BEST, TAME THE BEAST: ROBUST ML ESTIMATION FOR MIXTURE MODELS

Francesca Greselin [1]

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

ABSTRACT: This paper presents a brief review of constrained maximization of the likelihood, in combination with data-driven trimming, as a powerful technique for achieving robust classification and clustering in mixture models. Trimming is the simpler way to achieve robustness, being also highly intuitive. Originally developed for Gaussian components, this methodology has been successfully extended to various scenarios, including parsimonious mixtures, mixtures of factor analyzers, mixtures of regression and cluster-weighted mixtures, as well as to mixtures of skew and functional data. By effectively taming the complexities associated with parameter estimation, this approach yields an estimator which exists and is strongly consistent to the corresponding solution of the population optimum under widely general conditions.

KEYWORDS: Model-based classification, robustness, trimming, constrained estimation, outliers.

1 Introduction

Mixture models offer a highly flexible approach for statistical modeling of diverse random phenomena, especially when we posit that the observations arise from unobserved groups within the population. However, estimating Gaussian (and related) mixture models using the Maximum Likelihood (ML) approach introduces two significant challenges: *i)* the unboundedness of the likelihood function, that sets the ML as a mathematically ill-posed problem, and *ii)* the presence of contaminating data (background noise, pointwise contamination, unexpected minority patterns, etc.) that could severely affect the model fitting.

To *tame* the likelihood, researchers have adopted two essential techniques for parameter estimation in mixture models: eigenvalue constraints and trimming. The foundation of this methodology can be traced back to García-Escudero *et al.*, 2008, which has since evolved into a paradigm for robust model-based classification and clustering. The approach is well-regarded for its desirable theoretical properties and the availability of feasibile EM algorithms for its implementation. Constrained estimation prevents convergence

towards degenerate solutions, and mitigates the occurrence of non-interesting (spurious) local maximizers associated with complex likelihood surfaces. From Hathaway's seminal paper (Hathaway, 1985), these approaches, when coupled with impartial trimming, find applications in the context of robust statistical methods. Impartial trimming consists in excluding a small percentage of the less plausible observations, during the EM iterations, from contributing to model estimation. So doing, it protects the inferential results from the harmful effects of outliers.

Within the realm of this research stream, notable contributions include robust mixtures of factor analyzers (García-Escudero *et al.*, 2017) and the robust cluster-weighted model (García-Escudero *et al.*, 2016b). The introduction of their fuzzy versions (García-Escudero *et al.*, 2018b) revealed an intriguing interplay between the fuzzifier parameter and the scale. Additionally, advancements have been made in the treatment of skew components (García-Escudero *et al.*, 2016a). In the context of the semisupervised setting, when label noise interferes with the learning process, and whenever variable selection could be beneficial, the development of a specific robust approach is needed (Cappozzo *et al.*, 2020b, Cappozzo *et al.*, 2021).

However, in all such models, hyperparameter tuning remains an essential part of the inferential process, and ongoing research is focused on determining optimal settings for critical parameters such as the percentage of trimming, the number of components in the mixture, and the value for the eigenvalue constraint (Riani *et al.*, 2019). To support practitioners in this delicate task, researchers have proposed graphical and computational tools based on the combination of two exploratory steps (Cappozzo *et al.*, 2023).

Indeed, the literature presents several alternative methodologies for achieving robust model-based classification. Some of these approaches involve substituting Gaussian components with other elliptical distributions that possess heavier tails, such as the Student t (e.g., Greselin & Ingrassia, 2010) or the contaminated normal distribution (Punzo & McNicholas, 2016). These approaches withstand the presence of mild outliers. One key concept used to assess the robustness of an estimate in the presence of outliers is the breakdown point (Hampel, 1971). Its finite sample version is the maximum fraction of outliers which a given sample may contain without spoiling the estimate completely (Donoho, 1982). Among the models with good breakdown properties, we may mention a method for cluster detection and clustering with random start forward searches (Atkinson *et al.*, 2018), the optimally tuned robust improper maximum likelihood estimator, which uses an improper constant density for modeling outliers and noise (Coretto & Hennig, 2017), and

the weighted likelihood approach, aimed at downweighting outliers (Greco & Agostinelli, 2020). Here the weights are based on Pearson residuals stemming from robust Mahalanobis-type distances.

To conclude, the reviewed models deliver more reliable and stable estimation in the presence of outliers and noisy data, significantly enhancing model performance and facilitating more accurate statistical inference. The field of robust model-based clustering and classification is continually progressing, built upon solid results and theoretical foundations, while still presenting numerous intriguing challenges that await exploration. Exciting possibilities lie ahead, and the best is yet to come for researchers in this evolving domain.

References

ATKINSON, A.C., RIANI, M., & CERIOLI, A. 2018. Cluster detection and clustering with random start forward searches. *Journal of Applied Statistics*, **45**(5), 777 – 798.

CAPPOZZO, A., GRESELIN, F., & MURPHY, T.B. 2020a. A robust approach to model-based classification based on trimming and constraints: Semi-supervised learning in presence of outliers and label noise. *Advances in Data Analysis and Classification*, **14**(2), 327–354.

CAPPOZZO, A., GRESELIN, F., & MURPHY, T.B. 2020b. Anomaly and Novelty detection for robust semi-supervised learning. *Statistics and Computing*, **30**(5), 1545–1571.

CAPPOZZO, A., GRESELIN, F., & MURPHY, T.B. 2021. Robust variable selection for model-based learning in presence of adulteration. *Computational Statistics & Data Analysis*, **158**, 107–186.

CAPPOZZO, A., GARCÍA-ESCUDERO, L.A., GRESELIN, F., & MAYO-ISCAR, A. 2023. Graphical and computational tools to guide parameter choice for the cluster weighted robust model. *Journal of Computational and Graphical Statistics*, 1–20.

CORETTO, P., & HENNIG, C. 2017. Consistency, breakdown robustness, and algorithms for robust improper maximum likelihood clustering. *Journal of Machine Learning Research*, **18**(142), 1–39.

DONOHO, D.L. 1982. Breakdown properties of multivariate location estimators. *Ph.D. qualifying paper, Harvard University*.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., GRESELIN, F., INGRASSIA, S., & MAYO-ÍSCAR, A. 2017. Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *Statistics and Computing*, **27**(2), 377–402.

García-Escudero, L.A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. 2008. A general trimming approach to robust cluster analysis. *Annals of Statistics*, **36**(3), 1324–1345.

García-Escudero, L.A., Greselin, F, Mc Lachlan, G., & Mayo-Iscar, A. 2016a. Robust estimation of mixtures of Skew Normal Distributions. *In: Proceedings of the 48th Scientific Meeting of the Italian Statistical Society-Salerno (Italy), June 8-10, 2016*.

García-Escudero, L.A., Gordaliza, A., Greselin, F., Ingrassia, S., & Mayo-Iscar, A. 2016b. The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. *Computational Statistics & Data Analysis*, **99**, 131–147.

García-Escudero, L.A., Gordaliza, A., Greselin, F., Ingrassia, S., & Mayo-Iscar, A. 2018a. Eigenvalues and constraints in mixture modeling: geometric and computational issues. *Advances in Data Analysis and Classification*, **12**(2), 203–233.

García-Escudero, L.A., Greselin, F., & Mayo-Iscar, A. 2018b. Robust, fuzzy, and parsimonious clustering, based on mixtures of factor analyzers. *International Journal of Approximate Reasoning*, **94**, 60–75.

Greco, L., & Agostinelli, C. 2020. Weighted likelihood mixture modeling and model-based clustering. *Statistics and Computing*, **30**(2), 255–277.

Greselin, F., & Ingrassia, S. 2010. Constrained monotone EM algorithms for mixtures of multivariate *t* distributions. *Statistics and Computing*, **20**(1), 9–22.

Greselin, F., & Ingrassia, S. 2015. Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers. *Statistics and Computing*, **25**(2), 215–226.

Hampel, F.R. 1971. A general qualitative definition of robustness. *The annals of mathematical statistics*, **42**(6), 1887–1896.

Hathaway, R.J. 1985. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**(2), 795–800.

Punzo, A., & McNicholas, P.D. 2016. Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**(6), 1506–1537.

Riani, M., Atkinson, A.C., Cerioli, A., & Corbellini, A. 2019. Efficient robust methods via monitoring for clustering and multivariate data analysis. *Pattern Recognition*, **88**, 246 – 260.

RAGE AGAINST THE MEAN - AN INTRODUCTION TO DISTRIBUTIONAL REGRESSION

Thomas Kneib[1]

[1] Georg-August-University Göttingen, Chair of Statistics (e-mail: tkneib@uni-goettingen.de)

ABSTRACT: Distributional regression models that overcome the traditional focus on relating the conditional mean of the response to explanatory variables and instead target either the complete conditional response distribution or more general features thereof have seen increasing interest in the past decade. In this presentation, we will focus on generalized additive models for location, scale and shape as a flexible and versatile tool for distributional regression. We will introduce the underlying methodology and illustrate its application in different case studies. Furthermore, we will briefly review competing distributional regression approaches such as conditional transformation models or quantile and expectile regression.

KEYWORDS: conditional transformation models, generalized additive models for location, scale and shape, quantile regression, semiparametric regression.

ON GRAPH LIMITS AS MODELS FOR INTERACTION DATA

Sofia Charlotta Olhede[1]

[1] Institute of Mathematics, EPFL, Switzerland (e-mail: sofia.olhede@epfl.ch)

ABSTRACT: Network data has become a staple in many different applications, ranging from ecology, to neuroscience and systems biology. Its inference will of course depend on the application where we collect the network data, but I will discuss some general principles based on probabilistic symmetries such as permutation invariance. Just like other probabilistic invariances, the distributional invariance to permuting indices of a matrix of interactions implies a representation theorem (the Aldous-Hoover theorem). This representation is in terms of a graph limit function, or graphon. I will discuss the representation, how to make inferences based on this representation, what to do if distributional permutation invariance does not hold, and what to do if we have additional information such as time stamp of interactions, multiple interactions or additional covariate data.

KEYWORDS: network data, stochastic blockmodel, graph limit

Invited Papers

ENSEMBLE METHOD FOR TEXT CLASSIFICATION IN MEDICINE WITH MULTIPLE RARE CLASSES

Alessandro Albano [1], Mariangela Sciandra [1] and Antonella Plaia [1]

[1] Department of Economics, Business and Statistics, University of Palermo, (e-mail: alessandro.albano(mariangela.sciandra,antonella.plaia)@unipa.it)

ABSTRACT: The paper presents an ensemble method for text classification in the presence of multiple rare classes in the context of medical record data. Specifically, our study aims to classify clinical notes into multiple disease categories, including rare diseases. The Ensemble method involves combining the predictions of multiple machine learning models to predict the patient's diagnosis more accurately. We used three different machine learning algorithms, namely Support Vector Machine, Random Forest, and Naive Bayes, to generate three distinct models and combine their predictions through an ensemble method. The results demonstrate that the ensemble method improves the classification performance compared to individual models. We evaluated this approach on a dataset of 50,000 clinical notes with multiple rare classes.

KEYWORDS: text classification, ensemble method, machine learning, clinical coding.

1 Introduction

In the field of medicine, text classification is a crucial task for organizing and managing large volumes of medical documents. Proper classification of medical texts can aid in decision-making processes, clinical research, and the development of new treatments. Clinical coding is the task of transforming medical information in a patient's health records into structured codes, and machine learning algorithms have been widely used to classify medical documents automatically. Nonetheless, the accuracy of machine learning methods can be boosted by assembling various methods by combining their outputs. In this paper, we explore ensemble methods for text classification in medicine, specifically dealing with multiple rare classes.

In this paper, we propose using an ensemble method for Clinical coding, i.e., transforming medical records, usually presented as free texts written by clinicians, into structured codes in a classification system like the International

Classification of Diseases (ICD-9) code, involving 18 different labels. Our approach involves fitting multiple machine learning algorithms and combining their predictions to produce a final prediction. Specifically, we use Support Vector Machine, Random Forest, and Naive Bayes and combine their predictions with improving the accuracy of our classification results. Our study adds to the expanding research on clinical natural language processing (NLP), focusing on the specific problem of text classification in the context of medical records with multiple rare classes (imbalanced labels). The literature contains important contributions, such as the work of Alsentzer *et al.* (2019), demonstrating NLP applications in medical research and clinical practice, or the study by Harrison & Sidey-Gibbons (2021) that highlights the potential of NLP models to improve medical NLP tasks. Finally, Wu *et al.* (2022) provides a comprehensive survey of clinical NLP research and applications, specifically focusing on text classification.
The paper is organized as follows. In the next section, we present the experimental setup we used in our study, including the dataset, the machine learning algorithms, and experimental results. Finally, we conclude the paper and discuss future directions for research.

2 Experimental Setup

2.1 Data

In this study, we used the MIMIC-III (Medical Information Mart for Intensive Care III) dataset, a publicly available dataset of de-identified electronic health records of patients admitted to the intensive care unit (ICU) at Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset includes clinical notes such as discharge summaries, progress notes, and nursing notes. The MIMIC-III dataset is widely used in the research community for various tasks, such as predicting patient outcomes, identifying risk factors, and natural language processing.
The clinical notes from patients with different diagnoses, including rare ones (18 total different ones), were preprocessed to remove any personally identifiable information and to extract the relevant text for each diagnosis. Each note was then labelled with its corresponding diagnosis obtaining 50,000 records.

2.2 Machine learning methods

Our study used three machine learning algorithms: i) Support Vector Machine (SVM), a supervised learning algorithm that looks for the optimal hyperplane

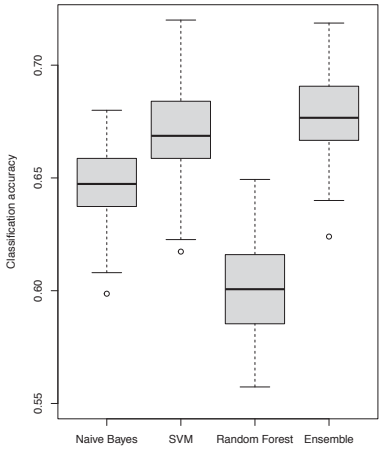
that separates the data into different classes. In our study, we used the linear kernel function to train the SVM model; ii) Random Forest (RF), an ensemble learning algorithm that combines multiple decision trees to improve the accuracy of predictions. RF works by randomly selecting a subset of variables and a subset of data samples to build multiple decision trees. In our study, we used 100 decision trees to build the RF model; iii) Naive Bayes (NB), which is a probabilistic machine learning algorithm that calculates the conditional probability of each variable given a class label and then uses Bayes' theorem to calculate the probability of each class given the variables. In our study, we used the Multinomial Naive Bayes variant to build the NB model.
We then used the Ensemble method to combine the predictions of the three machine learning models and produce a final prediction. Specifically, we used the majority voting method to combine the SVM, RF, and NB model predictions. The majority voting method works by selecting the class label predicted by most of the three models. In other words, if two or more models predict the same class label, that label is selected as the final prediction. If there is no majority, the class label predicted by the model with the iteration-specific highest accuracy is selected as the final prediction.

2.3 Results

The experiments' results (Fig.1a) indicate that the ensemble method achieved better results than individual models in predicting diseases from clinical notes. The median accuracy of the ensemble method was 67.7%, which is higher than the accuracy of individual models such as Naive Bayes (64.7%), SVM (66.9%), and Random Forest (60%), indicating that the ensemble method is more consistent in its predictions.
The results also show that the accuracy of the ensemble method was relatively stable across all quantiles of the accuracy distribution. The ensemble method was able to leverage the strengths of each model and compensate for its weaknesses.

3 Conclusion

In conclusion, our proposed ensemble method for text classification in medicine with multiple rare classes shows promising results for identifying and predicting various diseases from clinical notes. Our approach combines three machine learning algorithms (SVM, RF, and NB) to improve the accuracy of individual models. The results demonstrate that the proposed ensemble method is a



(a) Boxplot of accuracy scores.

	Dev. Std.	Median	Mean
NB	**0.017**	0.647	0.647
SVM	0.020	0.668	0.669
RF	0.020	0.600	0.600
EM	0.019	**0.676**	**0.677**

(b) Summary of accuracy scores.

promising approach for clinical coding, also when dealing with multiple rare classes or imbalanced datasets. Further research can explore the performance of the proposed ensemble method on larger datasets with a broader range of diseases, as well as the potential of incorporating other machine learning algorithms and techniques such as deep learning and active learning. In addition, exploring ways to reduce the computational complexity of the ensemble method without sacrificing performance is also an exciting avenue for future research.

References

ALSENTZER, EMILY, MURPHY, JOHN R, BOAG, WILLIE, WENG, WEI-HUNG, JIN, DI, NAUMANN, TRISTAN, & MCDERMOTT, MATTHEW. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

HARRISON, CONRAD J, & SIDEY-GIBBONS, CHRIS J. 2021. Machine learning in medicine: a practical introduction to natural language processing. *BMC medical research methodology*, **21**(1), 1–11.

WU, HONGHAN, WANG, MINHONG, WU, JINGE, FRANCIS, FARAH, CHANG, YUN-HSUAN, SHAVICK, ALEX, DONG, HANG, POON, MICHAEL TC, FITZPATRICK, NATALIE, LEVINE, ADAM P, *et al.* 2022. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *NPJ digital medicine*, **5**(1), 186.

DISTANCE-BASED AGGREGATION AND CONSENSUS FOR PREFERENCE-APPROVALS

Alessandro Albano [1], Mariangela Sciandra [1] and Antonella Plaia [1]

[1] Department of Economics, Business and Statistics, University of Palermo, (e-mail: alessandro.albano@unipa.it,mariangela.sciandra,antonella.plaia)

ABSTRACT: This paper proposes a distance-based aggregation and consensus method for preference-approvals, a type of preference data where individuals provide a list of approved alternatives in addition to a strict ranking. The proposed method aims to synthesize individual preference-approvals into a unified consensus representing the group's collective view. The consensus is the preference-approval, which minimizes the average distance with the whole set of voters. The proposed method has potential applications in group decision-making, recommendation systems, and social choice theory.

KEYWORDS: preference-approvals, preference aggregation, group decision-making, consensus

1 Introduction

In recent years, preference aggregation has received much attention due to its various applications in group decision-making, recommendation systems, and social choice theory. One type of preference data that has gained increasing popularity is preference-approvals, where individuals provide a list of approved alternatives in addition to a ranking (Brams & Sanver, 2009). In this paper, we propose a distance-based aggregation and consensus method for preference-approvals, which aims to synthesize individual preference-approval into a unified consensus representing the group's collective view. The proposed method finds the consensus as the preference-approval that minimizes the average distance with the whole set of voters. We employ a family of distances to evaluate the disagreement between preference-approvals and then use this to formulate an optimization problem to find the consensus preference-approval. This paper presents the notation and framework necessary to understand the proposed method describing the aggregation procedure. This method could advance preference aggregation and aid in practical decision-making scenarios.

2 Notation

Suppose a set of voters $V = \{v_1, \ldots, v_n\}$, with $n \geq 2$, are asked to order m different alternatives. The ranking π is a mapping function from the set of alterna-

tives $X = \{x_1, \ldots, x_m\}$ to the set of ranks $\pi = \{P_\pi(x_1), \ldots, P_\pi(x_i), \ldots, P_\pi(x_m)\}$, where $P_\pi : X \longrightarrow \{1, \ldots, m\}$ assigns the rank of each alternative.
In the framework of preference-approval modelling, each preference ranking, π, is paired with an approval vector, A. For any given set X of alternatives, we define approvals by partitioning X into the set of approved alternatives G and the set of rejected alternatives U. We represent a voter's preference-approval profile by a top-down order of alternatives with a horizontal bar: alternatives above the bar are approved, and those below are rejected.

$$\begin{array}{c} x_3 \\ \hline x_2 \\ x_1 \\ x_4 \end{array}$$

The preference-approval above is codified as follows:

$$\pi_1 = (2,3,1,4) \quad A_1 = (0,0,1,0).$$

To evaluate the disagreement between preference-approvals, Erdamar *et al.* (2014) introduced a family of distances. Specifically, given a parameter $\lambda \in [0,1]$, they define a distance for preference-approvals, denoted by d_λ, as a mapping from pairs of preference-approval profiles to the interval $[0,1]$.

$$d_\lambda\big((\pi_1, A_1), (\pi_2, A_2)\big) = \lambda d_K(\pi_1, \pi_2) + (1-\lambda) d_H(A_1, A_2) \tag{1}$$

where (π_1, A_1) and (π_2, A_2) are two preference-approval profiles for the same set of alternatives X of size m, d_K and d_H are respectively the Kemeny and Hamming distance. In a recent study, Albano *et al.* (2022) presented a generalized version of d_λ, denoted as D^r_λ. This extended distance measure incorporates a power-weighted mean as an aggregation function and accounts for discordance between pairs of items in the preference-approval profiles.

3 Aggregation procedure

Given a $n \times 2m$ matrix Π, whose l-th row represents the preference-approval associated with the l-th judge, the consensus preference-approval $(\hat{\pi}, \hat{A})$ is found by minimizing the average distance function d_λ for fixed λ:

$$(\hat{\pi}, \hat{A})_\lambda = \underset{(\pi, A) \in P^m}{\arg\min} \sum_{l=i}^{n} d_\lambda\big((\pi^{(l)}, A^{(l)}), (\pi, A)\big), \tag{2}$$

where P^m is the universe of all preference-approvals with m objects.
By construction, the minimization of d_λ entails the simultaneous minimization

of both rank and approval distances. Therefore, the problem is reduced to finding $\hat{\pi}$ and $\hat{A}$ such that:

$$(\hat{\pi} = \underset{\pi \in S^m}{\arg\min} \sum_{l=i}^{n} d_K(\pi^{(l)}, \pi),\ \hat{A} = \underset{A \in \{0,1\}^m}{\arg\min} \sum_{l=i}^{n} d_H(A^{(l)}, A)). \tag{3}$$

where S^m is the universe of the permutations (with ties) of m elements, and $d_H(A^{(l)}, A)$ and $d_K(\pi^{(l)}, \pi)$ are respectively the Hamming and the Kemeny distance between the preference and the approval part of the l-th row and the candidate consensus.

To find the Kemeny optimal ranking $\hat{\pi}$, we rely on the work of D'Ambrosio *et al.* (2015), who provided two accurate algorithms, called QUICK and FAST, for identifying the median ranking following the Kemeny approach. To find the approval consensus, $\hat{A}$, we compute the median approval vector by calculating the element-wise median of the binary approval matrices for all judges. In other words, we calculate the median of each column of the binary approval matrix, resulting in a final approval vector representing the consensus among the judges.

4 Case study

This section presents a case study, using data from the Eurobarometer*, website to demonstrate the effectiveness of the proposed method. The data consists of 27 rows (one per EU member country) and 9 columns representing alternatives concerning social values such as x_1: Equality between women and men, x_2: Fight against discrimination, x_3: Tolerance and respect for diversity, x_4: Solidarity among EU States, x_5: Solidarity between the EU and poor countries, x_6: Protection of human rights, x_7: Freedom of religion, x_8: Freedom of movement, and x_9: Freedom of speech. To obtain preference-approvals, alternatives are ranked in order of popularity for each country, and those that received more votes than the national average were considered acceptable. We used a hierarchical clustering procedure based on d_λ (with $\lambda = 0.75$) and found that the EU countries can be separated into two large clusters. Cluster 1 mainly comprises Western European countries (Austria, Belgium, Denmark, France, Italy, Luxembourg, Malta, Netherlands, Portugal, Spain, and Sweden). In contrast, Cluster 2 is manly composed of Eastern European countries (Bulgaria, Croatia, Cyprus, Czech Rep., Estonia, Finland, Germany, Greece, Hungary, Ireland, Latvia, Lithuania, Poland, Romania, Slovakia, and Slovenia). The consensus procedure has been applied to aggregate preference-approvals within

*`https://europa.eu/eurobarometer/surveys/detail/2612`.

each cluster and facilitates the interpretation. The two consensus preference-approvals are:

Cluster 1	Cluster 2
$x_1\, x_9$	x_6
x_6	x_9
———	x_8
x_4	x_4
x_2	———
x_3	$x_1\, x_3$
x_8	x_2
x_5	x_5
x_7	x_7

The two consensus clusters show different levels of agreement on certain alternatives. For instance, Cluster 1 consensus shows a higher preference for equality between women and men. In contrast, Cluster 2 consensus shows a higher preference for the solidarity between EU Member States and freedom of movement. Overall, the two consensus preference-approvals provide a more detailed and nuanced picture of how the EU countries express their views on the nine alternatives proposed.

5 Conclusions

In conclusion, this paper proposes a distance-based approach for aggregating and reaching a consensus on preference-approvals, providing a solution for extracting a common preference from a group with diverse preferences. The approach offers a framework for achieving consensus among individuals with diverse preferences and can help improve decision-making processes' effectiveness and efficiency. Moreover, this algorithm could be used within preference learning algorithms to make predictions. In future work, we aim to extend this approach to the generalized distance function presented by Albano *et al.* (2022), thus providing an algorithmic solution to achieving consensus through the extended preference-approval distance.

References

ALBANO, A., GARCÍA-LAPRESTA, JOSÉ LUIS., PLAIA, A., & SCIANDRA, M. 2022. A family of distances between preference-approvals. *Annals of Operations Research*, 1–29.

BRAMS, STEVEN J., & SANVER, M. REMZI. 2009. *Voting Systems that Combine Approval and Preference*. Berlin, Heidelberg: Springer Berlin Heidelberg. Pages 215–237.

D'AMBROSIO, ANTONIO, AMODIO, SONIA, & IORIO, CARMELA. 2015. Two algorithms for finding optimal solutions of the Kemeny rank aggregation problem for full rankings. *Electronic Journal of Applied Statistical Analysis*, **8**(2), 198–213.

ERDAMAR, BORA, GARCÍA-LAPRESTA, JOSÉ LUIS, PÉREZ-ROMÁN, DAVID, & SANVER, M REMZI. 2014. Measuring consensus in a preference-approval context. *Information Fusion*, **17**, 14–21.

FLEXIBLE EMPLOYMENT, A MACHINE LEARNING APPROACH

Marco Alfó [1], Dimitris Pavlopoulos[2] and Roberta Varriale[1]

[1] Department of Statistical Science, Sapienza University of Rome, (e-mail: marco.alfo@uniroma1.it, roberta.varriale@uniroma1.it)

[2] Department of Sociology, Vrije Universiteit Amsterdam, (e-mail: d.pavlopoulos@vu.nl)

ABSTRACT: Flexible employment is an important topic in scientific and political debate in Europe. The present work describes the use of machine learning techniques to predict the contract type, by using information coming from both survey and administrative data. Information on contract type come from linked data from the Labour Force Survey and the Employment Register of the Netherlands for the period 2007-2015.

KEYWORDS: flexible employment; machine learning; multi-source data

1 Introduction

Flexible employment is an important topic in scientific and political debate in Europe. In Eurozone countries, OECD statistics (https://stats.oecd.org) show that the incidence of temporary employment was 11.8 percent in 2021, while the probability of getting a job on a temporary contract increased by 36 percent between 2013 and 2019 (Latner, 2022). The Netherlands also saw a sharp increase in the incidence of temporary employment in 2021: from 13.7 percent in 2000 to 27.4 percent. In Italy, the increase was a bit lower, from 10.1 percent in 2000 to 16.6 percent in 2021. The role of temporary contracts can be assessed from a life course perspective: one of the main questions the research seeks to answer is whether temporary work is always a stepping stone to permanent employment or it should be rather considered as a trap of precarious jobs (Latner & Saks, 2022).

Data to study flexible employment dynamics may come from different sources, such as survey, administrative and statistical register data. Measurements from different sources may not agree for different reasons, including the presence of measurement error or misalignment in the definitions or between occasions of measurement. For example, there could be temporal misalign-

ment of the sources, structural lack of administrative information on irregular work, misalignment of employment definition in the available sources.

Findings on mobility from temporary to permanent employment can be severely biased due to measurement error, usually present in the information used for analysis, coming from both survey and statistical register data (see, for example, Pavlopoulos & Vermunt, 2015; and Pankowska *et al.*, 2021). A possible approach to deal with measurement error when multiple data sources are available is based on the use of latent variable models. In particular, latent variable models can be used to predict the true target value (here, the current type of contract) given the observed measurements in the data sources when all these data sources contain information closely related to the target variable, but none can be assumed to be error free (Filipponi *et al.*, 2021). An alternative approach to deal with data coming from multiple possibly discordant sources is based on Machine Learning (ML) tools for supervised classification (Varriale & Alfo', 2023). ML tools may be used to predict the individual target variable, and to extract important information from the data to learn more about the phenomenon in the form of a selection of possibly important predictors of the response.

The present work describes the use of some ML techniques, including decision trees and random forests, to predict the individual contract type. We use linked data drawn from the Labour Force Survey (LFS) and the Employment Register of the Netherlands for the period 2007-2015. The aim of this paper is to show how ML techniques can be used with longitudinal data to extract important information for the purpose of estimating the probability of a temporary employment contract in the life course, and to learn more about the phenomenon.

2 The context

The data sources providing information on types of employment contracts are the Labour Force Survey (LFS) administered by Statistics Netherlands and the Employment Register of the Netherlands (ER). LFS represents the main source of information on the labour market for official statistics. It produces information on employment and the main aggregates of the job offer - profession, sector of economic activity, hours worked, type and duration of contracts, training. LFS is harmonized at the European level as established by the EU Regulation 2019/1700 of the European Parliament and the Council. In the Netherlands, the LFS has a rotating trimonthly scheme and it is representative for the Dutch population aged 15 or more. Since 1999, respondents are inter-

viewed at 5 consecutive panel waves. The collected information refers to the moment of the interview, and the interviews are carried out during every week of the trimester. Table 1 show the LFS rotating scheme for two years.

Table 1. *LFS rotating scheme for two years.*

	$year_1$				$year_2$			
Sample	q_1	q_2	q_3	q_4	q_1	q_2	q_3	q_4
1	x	x	x	x	x	.	.	.
2	.	x	x	x	x	x	.	.
3	.	.	x	x	x	x	x	.
4	.	.	.	x	x	x	x	x
5	.	.	.	.	x	x	x	x
6	.	.	.	.	.	x	x	x
7	.	.	.	.	.	.	x	x
8	.	.	.	.	.	.	.	x

The ER is a register administered by the Institute for Employee Insurance (UWV), containing information on labour market and income for all insured workers in the Netherlands. The ER is constructed by collecting and matching information from various sources, i.e. the Tax Office, the Population Register and information drawn from temporary work agencies' registries (Bakker *et al.*, 2014). The submission of tax-reporting statements is compulsory for employers. However, while ER dataset contains monthly information, employers typically submit the relevant information only few times per year. This may, at least potentially, produce some errors, in particular for the information regarding the period between two consecutive submissions. Additional sources of measurement error in ER may result from administrative delays, wrong registration, and erroneous administrative procedures.

3 A machine learning approach

A ML approach for supervised classification is applied to predict individual contract type as a function of individual features and time. Categories of contract type can be classified as "permanent"/"non-permanent." The latter category can be divided into "fixed-term", "temporary or on-call", and "other". The response is contract type, and models considering both response at 2 and 4 categories as target variable are considered and estimated.

Let y_{ijt} denotes the binary indicator for contract type j at occasion t for individual i ($i = 1,\ldots,n$, t=1,...,T, $j = 1,\ldots,m$). In this work $T = 32$, each time t corresponding to a specific quarter of the year. We use multiple strategies of analysis. The first strategy involves using y_{ijt} as the target variable and all the information available in previous times as covariates. Therefore, we want to model the conditional expectation $E(y_{ijt}|x_{it}, x_{it-1}, \ldots, x_{i1})$. The second strategy uses as covariate also the information on the target variable y at time $t-1$, in order to take into account the longitudinal structure of the data by defining a formal for the conditional expectation $E(y_{ijt}|y_{ijt-1}, x_{it}, x_{it-1}, \ldots, x_{i1})$. In particular, we are assuming that the evolution of the contract type is governed by a first order Markov chain with transition matrix that do not depend on time. Last, we will consider a first order non homogeneous Markov chain where transition probabilities may depend on individual features and or time. ML techniques are applied using the R software.

The aim is to show how ML can be used with longitudinal data, both for prediction and to extract the relevant information.

References

BAKKER, B.F., VAN ROOIJEN, J., & VAN TOOR, L. 2014. The system of social statistical datasets of statistics netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, **30**(4), 411–424.

FILIPPONI, D., GUARNERA, U., & R., VARRIALE. 2021. Latent Mixed Markov Models for the Production of Population Census Data on Employment. *Book of short papers SIS 2021*, 112–117.

LATNER, J.P. 2022. Temporary employment in Europe: stagnating rates and rising risks. *European Societies*, **24**(4), 383–408.

LATNER, J.P., & SAKS, N. 2022. The wage and career consequences of temporary employment in Europe: Analysing the theories and synthesizing the evidence. *Journal of European Social Policy*, **32**(5), 514–530.

PANKOWSKA, P., BAKKER, B., OBERSKI, D., & PAVLOPOULOS, D. 2021. Dependent interviewing: a remedy or a curse for measurement error in surveys? *Survey Research Methods*, **15**(2), 135–146.

PAVLOPOULOS, D., & VERMUNT, J.K. 2015. Measuring Temporary Employment. Do Survey or Register Data Tell the Truth. *Survey Methodology*, **41**(1), 197–214.

VARRIALE, R., & ALFO', M. 2023. Multi-source statistics on employment status in Italy, a machine learning approach. *METRON*.

CLINICALLY USEFUL MEASURES IN SURVIVAL ANALYSIS: THE RESTRICTED MEAN SURVIVAL TIME AS AN ALTERNATIVE TO THE HAZARD RATIO

Federico Ambrogi [1] [2], Matteo Di Maso[1]

[1] Department of Clinical scinces and Community Health, University of Milan, (e-mail: federico.ambrogi@unimi.it, matteo.dimaso@unimi.it)

[2] Laboratory of Data Management and Analysis, IRCCS Policlinico San Donato, (e-mail: Federico.Ambrogi@grupposandonato.it)

ABSTRACT: Hazard ratios are ubiquitously used in time to event applications to quantify treatment effects. Although hazard ratios are invaluable for hypothesis testing, other adjusted measures of association, both relative and absolute, may be used to fully appreciate studies results, especially when the assumption of proportional hazards does not hold. In the following we will show the use of restricted mean survival time, a measure of association that received a lot of attention in the last years, estimated through the follow-up time. Direct regression models on RMST and Machine Learning approaches are available. Examples will be used to illustrate the different approaches.

KEYWORDS: restricted mean survival time, machine learning, direct regression.

1 Introduction

Restricted mean survival time (RMST) differences between groups have been advocated as useful measures of association in time to event studies. In fact, while the ubiquitously used hazard ratios are invaluable for hypothesis testing, measures of association based on RMST, both relative and absolute, may have a more plain clinical interpretation and help to fully elucidate study results.

Many recent contributions focused on estimates of the difference in RMST through follow-up times, instead of using a single time horizon. The resulting curve can be used to quantify the association in time units. Moreover regression models have been developed to directly regress RMST on covariate patterns. These methods are based either of IPCW or on pseudo-values (PV). In particular, the method based on PV is easily implementable with available software and makes possible to adopt Machine Learning methods, such as the Deep Neural Network (DNN) proposed by Zhao, 2021.

We investigated the ability of DNN to account for complex covariate patterns, such as interactions, using literature data as done in Ambrogi *et al.*, 2022.

2 Methods

In survival analysis the time T elapsed from an initial event to the possible occurrence of a terminating event is analysed. Generally, only a right-censored version of the random variable T is observe. Therefore, instead of the mean value of T the τ-restricted mean survival time (RMST) is used:

$$RMST(\tau) = \int_0^\tau S(t)dt \tag{1}$$

where $S(t) = P(T > t) = \exp(-\int_0^t \lambda(u)du)$ is the survival function and $\lambda(t)$ is the hazard function. The $RMST(\tau)$ represents the expected lifetime over a time horizon equal to τ. The $RMST(\tau)$ can be estimated non-parametrically based on the Kaplan-Meier estimator or model-based.

Direct regression of RMST as a function of covariate values was studied by Tian *et al.*, 2014, based on inverse probability of censoring weighting, and by Andersen *et al.*, 2004 based on pseudo-values.

A joint model for several τ-values, $\tau_1, \ldots, \tau_j, \ldots, \tau_M$, including an interaction term between the treatment and a function, $f(\cdot)$, of time, to model a time-varying treatment effect is

$$g(RMST(\tau|Z)) = h(\tau) + \beta Z + \gamma Z f(\tau). \tag{2}$$

Commonly used link functions are the **log**, the **logit** or the identity function. Estimation based on pseudo-values is discussed in Ambrogi *et al.*, 2022, while estimation based on IPCW is presented in Zhong & Schaubel, 2022.

Recently a deep neural network (DNN) model was presented for RMST prediction by Zhao, 2021 called DnnRMST. The DNN is based on pseudo-values estimated at multiple times during the follow-up and optimized using MSE. The DNN consists of an input layer, some hidden layers and a multiple output layer with M nodes, for the pseudo-values at the different times. The DNN can be implemented using the Keras library in R (Allaire & Chollet, 2022). Hyper-parameters can be selected using a random grid search over the number of nodes, dropout regularization, ridge regularization and learning rate.

3 Results

Data of a double blind randomised clinical trial studying the effect of prednisone versus placebo on survival in patients with liver cirrhosis, already used for RMST estimation in Andersen *et al.*, 2004, were used to illustrate the methods. The CSL1 trial showed an interaction effect between treatment and presence of ascites, as illustrated in figure 1. Top panels show patients without ascites, while bottom panels show patients with ascites. Left panels show the KM survival curves for treated and control groups. The central figure panels show the non parametric estimate of RMST for treated and control groups. Right panels show the difference between RMST curves for treated vs control groups estimated non-parametrically (solid line), with the direct model with pseudo-values (dotted) and with DnnRMST (red). It is possible to see that, even if the interaction is captured by the DNN, the estimates are not in lines with those of the non-parametric estimators.

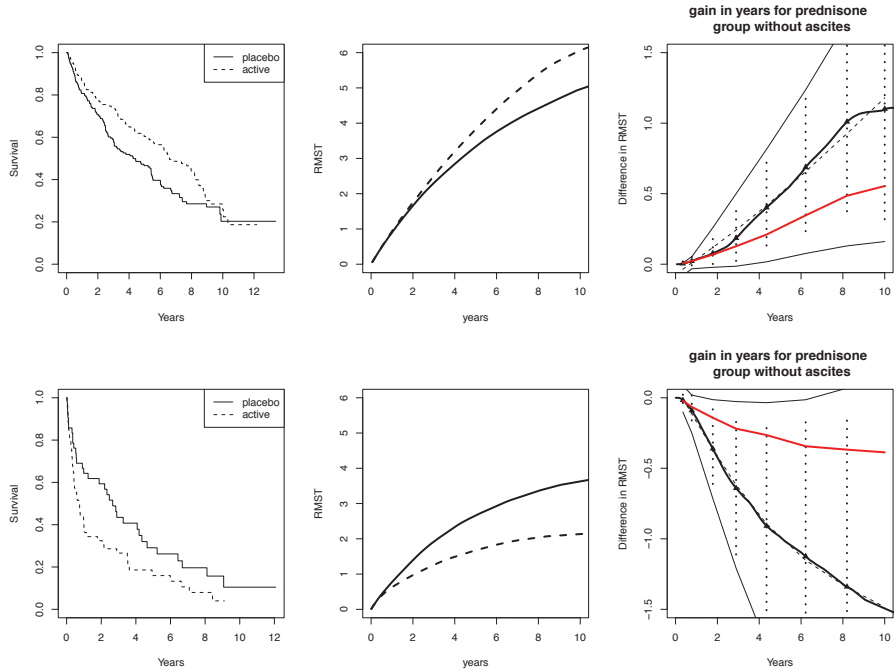


Figure 1. *Comparison of the nonparametric estimate of RMST with the one obtained using direct regression models and DnnRMST.*

4 Discussion

RMST has received a lot of attention in recent years. A possibility introduced for the first time by Royston & Parmar, 2011, is to estimate the difference of RMST curve through the time, to appreciate how the treatment comparison is evolving through time. Different regression methods have been proposed to estimate RMST as a function of time and Machine Learning techniques are also available. One interesting aspect is that of sample size. In fact, ML is in principle able to learn directly from data at the cost of hyper-parameters optimization. However, learning is data expensive and evaluating at which sample size the ML models are able to correctly reproduce complex data pattern is an open research question.

References

ALLAIRE, JJ, & CHOLLET, FRANÇOIS. 2022. *keras: R Interface to 'Keras'*. R package version 2.11.0.

AMBROGI, F., IACOBELLI, S., & ANDERSEN, P.K. 2022. Analyzing differences between restricted mean survival time curves using pseudo-values. *BMC Med Res Methodol*, **22**.

ANDERSEN, P. K., HANSEN, M. G., & KLEIN, J. P. 2004. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Anal*, **10**(4), 335–350.

ROYSTON, P., & PARMAR, M. K. 2011. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Stat Med*, **30(19)**, 2409–2421.

TIAN, L., ZHAO, L., & WEI, L. J. 2014. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, **15**(2), 222–233.

ZHAO, L. 2021. Deep Neural Networks For Predicting Restricted Mean Survival Times. *Bioinformatics*, **36**, 5672–7.

ZHONG, YINGCHAO, & SCHAUBEL, DOUGLAS E. 2022. Restricted mean survival time as a function of restriction time. *Biometrics*, **78**(1), 192–201.

DATA-DRIVEN SMOOTHING PARAMETER SELECTION FOR CIRCULAR DATA ANALYSIS *

Jose Ameijeiras-Alonso [1]

[1] CITMAga, Department of Statistics, Math. An. and Optim., Universidade de Santiago de Compostela, (e-mail: jose.ameijeiras@usc.es)

ABSTRACT: In this talk, we introduce a novel data-based smoothing parameter tailored for circular kernel density estimation and its derivatives. Building upon the plug-in ideas, we replace unknown quantities with appropriate estimates to derive an optimal smoothing parameter. Specifically, we present a circular adaptation of the renowned Sheather and Jones bandwidths through direct and solve-the-equation plug-in rules. The theoretical underpinning of our approach is established, encompassing the asymptotic mean squared error of the density estimator, its derivatives, and its functionals for circular data. We further conduct a simulation study to compare the performance of our proposed selectors with existing data-based smoothing parameters. To illustrate the applicability of our plug-in rules, we apply them to a real data example.

KEYWORDS: circular data, directional statistics, kernel density estimation, plug-in rule

*Supported by Grant PID2020-116587GB-I00 funded by MCIN/AEI/10.13039/501100011033 and the Competitive Reference Groups 2021-2024 (ED431C 2021/24) from the Xunta de Galicia.

VIEW IT DIFFERENTLY: FINDING GROUPS IN MICROBIOME DATA

Laura Anderlucci [1], Silvia Dallari[1] and Angela Montanari[1]

[1] Department of Statistical Sciences, University of Bologna, (e-mail: laura.anderlucci@unibo.it, silvia.dallari2@unibo.it, angela.montanari@unibo.it)

ABSTRACT: Microbiota plays a crucial role in human health. Recently, NGS technologies have enabled the exploration of the microbiome without isolation and culturing. However, analyzing and translating microbiome data into meaningful biological insights is challenging due to the data's compositional nature, high dimensionality, sparseness, and over-dispersion. The gut microbiome can vary from individual to individual, and microbiome communities can be grouped to identify community types linked to environmental or health conditions. Different data features, such as individual profiles, community-based descriptors, or genera interactions within a community, provide different perspectives on microbiome complexity. Combining these perspectives could lead to a more comprehensive understanding of microbiome data.

KEYWORDS: model-based clustering, community diversity measures, network-based clustering, consensus clustering

1 Introduction

Microbiota is largely recognized as being a central player in the human health and in that of all organisms and ecosystems, and subsequently has been the subject of intense study. Recently, Next Generation Sequencing (NGS) technologies have enabled the exploration of microbiome without the need for isolation and culturing. The data we are going to study have been obtained through deep sequencing of 16SrRNA genes and grouping bacteria at a certain level of 16SrRNA gene similarity. The analysis and the translation of microbiome data into meaningful biological insights remain still very challenging, also due to particular data characteristics. Microbiome data, in fact, are taxa counts that are compositional in nature (Gloor *et al.*, 2017), high-dimensional, sparse and over-dispersed. In humans, gut microbiome can vary from individual to individual and individual microbiome communities can be grouped to identify community types whose variability can be differently linked to environmental or health conditions.

According to the literature on microbiome data (Xia *et al.*, 2018), different data features can provide different perspectives on microbiome complexity.

The focus has typically been placed either on individual profiles or on community-based descriptors or on genera interactions within a community. We argue that combining these different perspectives could provide a more comprehensive understanding.

2 Microbiome data views

2.1 Individual profiles

The basic sampling units, over which conclusions are generalized, are biological samples. It is of interest to highlight similarities and differences across these units. The fundamental features with which to describe samples are the counts of bacterial species. For interpretation, it is common to imagine prototypical units which can be used as a point of reference for observed samples. In microbiome analysis, these are called *communities*: different communities have different bacterial signatures.

It is worth noticing that this kind of data structure closely resembles the term-document matrix, typically used in the analysis of textual data, and that microbiome data share many of its pros and cons (Sankaran & Holmes, 2019).

2.2 Diversity measures

Characteristic of biological communities is the *biodiversity*, and it can be described either focusing on within-individual richness of taxa or on inter- individual variability. α-diversity is the diversity within a single sample and can be measured via Shannon-Wiener diversity index H' or via Simpson diversity index D:

$$H' = -\sum_{i=1}^{p} p_i \log p_i, \qquad D = 1 - \sum_{i=1}^{p} p_i^2$$

where p_i is the proportion of individuals (or relative abundance) of species i in the community and p is the total number of species present.

β-diversity evaluates differences between two or more units or local assemblages, thus allowing to describe how many taxa are shared between communities or individuals. Examples are the Bray-Curtis dissimilarity:

$$BC = \frac{\sum_{i=1}^{p} |X_{ij} - X_{ik}|}{\sum_{i=1}^{p} (X_{ij} + X_{ik})}$$

where X_{ij}, X_{ik} are the number of individuals in species i in each sample (j,k) and p is the total number of species in samples, and the UniFrac distance. The unweighted (d^U) and weighted (d^W) UniFrac distances exploit the phylogenetic tree information and can be found for two communities A and B as

$$d^U = \sum_{t=1}^{T} \frac{b_t |I(p_t^A > 0) - I(p_t^B > 0)|}{\sum_{t=1}^{T} b_t}, \qquad d^W = \frac{\sum_{t=1}^{T} b_t |p_t^A - p_t^B|}{\sum_{t=1}^{T} b_t (p_t^A + p_t^B)}$$

where p_t^A and p_t^B are the taxa proportions descending from the branch t for community A and B, respectively, T is the rooted phylogenetic tree's branches and b_t is the length of the branch t.

2.3 Network structures

The interactions among the constituent members of a microbial community play a major role in determining the overall behavior of the community and the abundance levels of its members (Xia *et al.*, 2018). These interactions can be modeled using a network whose nodes represent microbial taxa and edges represent pairwise interactions. It is often unreasonable to expect that a single network is able to account for all the interactions in a community and network clustering can help in detecting microbiome features connected, for instance, with different health and environment condition.

3 Microbiome multi-view clustering

Clustering individual profiles (view 1) can be performed via partitioning and hierarchical methods (such as, e.g., spherical k-means, Partitioning Around Medoids, Ward's method) or via model-based methods such as mixtures of Von Mises-Fisher distributions, Dirichlet Multinomial Mixtures, Latent Dirichlet Allocation (see, for a review, Sankaran & Holmes, 2019).

In view of the analogy between microbiome and textual data, we propose to use here the method proposed in Anderlucci *et al.*, 2019, which models the clustering structure through a cosine distance-based mixture. Specifically, given the cosine dissimilarity $d(\mathbf{x},\xi)$ of a generic sample/document $\mathbf{x}$ from a centroid, say ξ, a distance-based density can be constructed as:

$$f(\mathbf{x};\xi,\lambda) = \psi(\lambda) e^{-\lambda d(\mathbf{x},\xi)}$$

where λ is a positive precision parameter and $\psi(\lambda)$ is a normalization constant. In order to perform clustering, we consider a mixture of K cosine distance-

based density functions:

$$f(\mathbf{x};\xi,\lambda) = \sum_{k=1}^{K} \pi_k \psi(\lambda) e^{-\lambda d(\mathbf{x},\xi_k)}$$

with positive mixture weights π_k, summing to unity and component varying centroid vectors ξ_k.

When the focus is on community diversity (view 2), the different diversity measures can be combined in a Gower's-coefficient-like fashion in order to guide the clustering of the individuals.

Finally, when the aim is to capture the interaction structure between taxa (view 3) network-based clustering via mixtures of Multivariate Poisson Log-Normal distributions can be applied (Tavakoli & Yooseph, 2019).

The clustering results of the three data views will be combined via consensus clustering (Hornik, 2005) or via the Bayesian two-way latent structure model proposed in Swanson *et al.*, 2019. The proposed multi-view clustering method will be applied to real data on gut microbiome described in McDonald *et al.*, 2018.

References

ANDERLUCCI, L., MONTANARI, A., & VIROLI, C. 2019. The importance of being clustered: Uncluttering the trends of statistics from 1970 to 2015. *Statistical Science*, **34**, 280–300.

GLOOR, G.B., MACKLAIM, J.M., PAWLOWSKY-GLAHN, V., & EGOZCUE, J.J. 2017. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.*, **8**, 2224.

HORNIK, K. 2005. A CLUE for CLUster ensembles. *J. Stat. Softw.*, **14**, 1–25.

MCDONALD, D., HYDE, E., *et al.* 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*, **3**(3), e00031–18.

SANKARAN, K., & HOLMES, S.P. 2019. Latent variable modeling for the microbiome. *Biostatistics*, **20**(4), 599–614.

SWANSON, D.M., LIEN, T., BERGHOLTZ, H., SØRLIE, T., & FRIGESSI, A. 2019. A Bayesian two-way latent structure model for genomic data integration reveals few pan-genomic cluster subtypes in a breast cancer cohort. *Bioinformatics*, **35**(23), 4886–4897.

TAVAKOLI, S., & YOOSEPH, S. 2019. Learning a mixture of microbial networks using minorization–maximization. *Bioinformatics*, **35**(14).

XIA, Y., SUN, J., & CHEN, D.G. 2018. *Statistical analysis of microbiome data with R*. Singapore: Springer.

RANDOM-BASED INITIALIZATION FOR CLUSTERING MIXED-TYPE DATA WITH THE K-PROTOTYPES ALGORITHM

Rabea Aschenbruck [1], Gero Szepannek[1] and Adalbert F.X. Wilhelm[2]

[1] Stralsund - University of Applied Sciences, (e-mail: rabea.aschenbruck@hochschule-stralsund.de, gero.szepannek@hochschule-stralsund.de)

[2] Constructor University Bremen gGmbH, (e-mail: awilhelm@constructor.university)

ABSTRACT: One of the most popular partitioning cluster algorithms for mixed-type data is the k-prototypes algorithm. Due to its iterative structure, the algorithm may only converge to a local optimum rather than a global one. Therefore, the resulting cluster partition may suffer from the initialization. In general, there are two ways of achieving an improvement of the initialization: One possibility is to determine concrete initial cluster prototypes, and the other strategy is to repeat the algorithm with different randomly chosen initial objects. Different numbers of algorithm repetitions are analyzed and evaluated comparatively. It is shown that an improvement of the cluster algorithm's target criterion can be achieved by an appropriate choice of repetitions, even with manageable time expenditure.

KEYWORDS: k-prototypes, mixed-type data, cluster analysis, initialization.

1 Introduction to the Problem

In the origin initialization, points to be clustered are chosen randomly as initial cluster prototypes. Subsequent iterations lead to a local optimum of the summed squared error minimization problem, but not necessarily to the global minimum for k-prototypes (Huang, 1997). Therefore, the choice of proper starting points is important. In general, there are three different strategies to receive the initial prototypes: The starting points can be determined based on the knowledge of the clustering use case. Otherwise, one can do a mathematical determination or a random-based choice of objects to be clustered. The latter one is probably the most common way in practice, where k objects are randomly selected. These may or may not be good starting points for the iterative algorithm routine. To increase the probability of reaching a global optimum, one can apply the algorithm multiple times on different, randomly

chosen objects. In the following, different numbers of algorithm repetition are compared and evaluated on different data situations with regard to the adjusted Rand index (*short:* ARI; Hubert & Arabie, 1985) and the computation time.

2 Simulation Study on the Random-based Initialization of the *k*-Prototypes Algorithm

Execution of the Simulation Study The aim of this study is to determine an appropriate number of repetitions to obtain a satisfactory cluster partition but at the same time, the number of algorithm repetitions should be as low as possible because of the increasing computation time. In practice, the number of repetitions can be passed to the R function `kproto` (Szepannek, 2018) via the parameter `nstart`. After the algorithm's application on `nstart` randomly chosen prototypes sets, the partition which minimizes the target criterion is used. The simulation study was executed on a Dell PowerEdge R440 server with two Intel Xeon Silver 4216 processors (2 x 16 cores; 2.1 GHz) and 768 GB RAM.

In the simulation study were included 120 different data situations, differing by the variation of the number of observations (500, 1000, 2000), variables (2, 4, 8) and clusters (2, 4, 8), whether the cluster group sizes were equal or unequal, and the ratio of categorical to numerical variables (0.25, 0.5, 0.75). To mitigate the random influence of the generation process 50 data sets were determined for every of the 120 data situations (see Aschenbruck *et al.*, 2022).

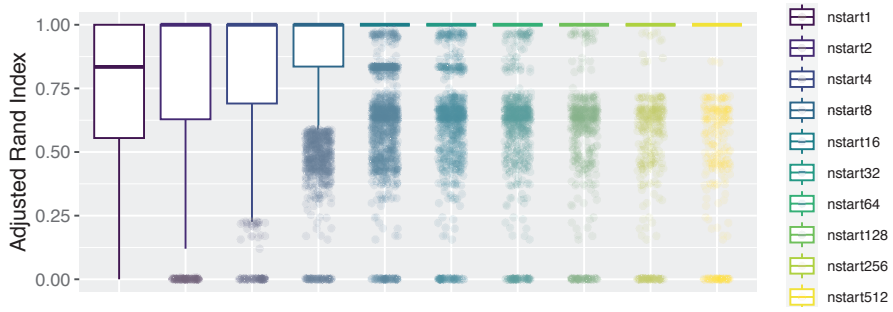


Figure 1. *Boxplots on the adjusted Rand index values of the resulting partitions.*

Comparision of the Different Numbers of Repetitions The higher the number of algorithm repetitions (`nstart`) the higher the rating of the resulting

cluster partitions by the ARI (see Fig. 1). 16 repetitions seem to be a good choice, since for most of the data sets examined, the resulting cluster partition is rated with an adjusted Rand index value of 1, which is the best possible rating and there is virtually not much improvement for more repetitions.

Having a closer look at the rated partitions for the different data situations in Fig. 2, it can be stated that the number of clusters to be determined and whether the clusters are of equal size or not is influencing the need for more repetitions to gain satisfying results. Since an increasing number of algorithm repetitions leads to an increase in computation time, hereafter a determination of the number of repetitions depending on the data situation is proposed.

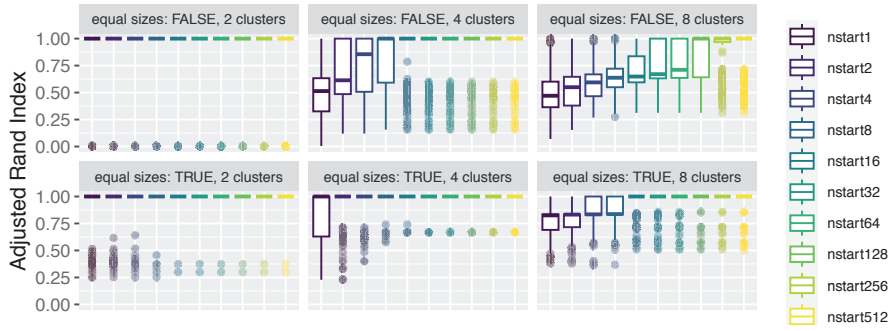


Figure 2. *Boxplots on the ARI of the resulting partitions, shown separately by number of clusters and whether or not the clusters are of the same size.*

Determination of the Number of Repetitions Depending on the Data Situation The data situation based number of repetitions m assures that with a probability of 0.9 at least one of the m sets of initial prototypes contains objects of every cluster group. Considering a geometrically distributed random variable $Z \sim Geo(\pi)$, it follows that the number of repetitions depending on the data situation at hand is

$$m = F_z^{-1}(0.9) \text{ with probability of sucess } \pi^\star = \prod_{i=0}^{k-1} \frac{N - i \cdot \lceil \frac{N}{k} \rceil}{N - i}, \qquad (1)$$

where N is the number of objects to be clustered and k the number of clusters to be determined. Thereby, all clusters are assumed to be of equal size since in practice, the sizes of the clusters to be determined are unknown. Nevertheless, if one suspects a small cluster group it is possible to input k in Eq. (1) as the reciprocal of this size.

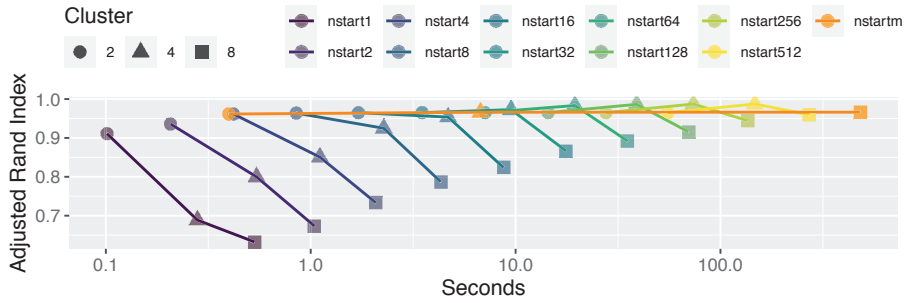


Figure 3. *Relationship between ARI and computation time in seconds (log-scaled), splitted by the initialization approach and the number of clusters in the data.*

Computation Time In Fig. 3 the average computation time for all data situations with the specified number of clusters and the average ARI is given. The influence of the number of repetitions and clusters on the increase in computation time is obvious. The data-based number of algorithm repetitions $m_{k,N}$ ($m_{2,\cdot} = 4$, $m_{4,\cdot} = 24$, $m_{8,500} = 905$, $m_{8,1000} = 931$, $m_{8,2000} = 944$) results in overall good rated partitions while avoiding unnecessary algorithm repetitions.

3 Summary

In this work, a theoretical determination of repetitions was motivated. For a small number of clusters, a few repetitions are sufficient, whereas as that number increases, a strong increase in repetitions is necessary, even at 8 clusters.

References

ASCHENBRUCK, R., SZEPANNEK, G., & WILHELM, A.F.X. 2022. Imputation Strategies for clustering mixed-type data with missing values. *J. Classif.*

HUANG, Z. 1997. Clustering Large Data Sets With Mixed Numeric and Categorical Values. *Pages 21–34 of: Proceedings of the First PAKDD.*

HUBERT, L., & ARABIE, P. 1985. Comparing Partitions. *J. Classif.*, 193–218.

SZEPANNEK, G. 2018. clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *R J.*, **10**(2), 200–208.

POSTERIOR CLUSTERING FOR DIRICHLET PROCESS MIXTURES OF GAUSSIANS WITH CONSTANT DATA

Filippo Ascolani[1] and Valentina Ghidini[2]

[1] Bocconi University, Milan (filippo.ascolani@phd.unibocconi.it)

[2] Bocconi University, Milan (valentina.ghidini@phd.unibocconi.it)

ABSTRACT: Dirichlet process mixtures, obtained by convolving the law of a Dirichlet process with a suitable kernel, are popular methods for density estimation. Due to the almost sure discreteness of the mixing measure, they automatically provide a latent clustering which is often of great interest for applied researchers. However, despite its relevance, little is known about the posterior properties of clustering, even with a large sample. We contribute by considering a simple data generating mechanism and showing the asymptotic properties of the maximum a posteriori clustering with Gaussian kernel.

KEYWORDS: Bayesian nonparametrics; clustering; maximum a posteriori; asymptotic analysis.

1 Introduction

Bayesian nonparametric methodologies have witnessed a growing popularity in the last decades, mainly due to the their flexibility: see Ghosal & Van Der Vaart (2017) for a recent review. A popular model for density estimation is given by Dirichlet process mixtures (Lo (1984)), which can be summarized as follows

$$Y_i \mid \theta_i \sim k(y \mid \theta_i), \quad \theta_i \mid P \overset{\text{i.i.d.}}{\sim} P, \quad P \sim DP(P_0, \alpha), \tag{1}$$

where $k(y \mid \theta)$ is a density function with parameter θ and $DP(P_0, \theta)$ is the law of a Dirichlet process (DP, Ferguson (1973)) with baseline distribution P_0 and concentration parameter $\alpha > 0$. It can be shown that the realizations of P are almost surely discrete probability measures, so that the θ_i's will present ties with positive probability, leading to a latent clustering of the observed datapoints $Y_{1:n} = (Y_1, \dots, Y_n)$.

Models as in (1) are provided with good asymptotic properties in terms of density estimation (Ghosal & Van der Vaart, 2007), when the data are generated i.i.d. from a "true" distribution P^*, but the clustering behavior a posteriori is less understood. As a positive note, it has been shown that, under suitable assumptions, the posterior on the mixing measure converges to the

"true" one in Wasserstein distance (Nguyen, 2013), but the metric is too weak to prove *per se* results on the clustering. More recently, Miller & Harrison (2013, 2014) showed that the posterior distribution on the number of clusters is often inconsistent, in the sense that it places positive mass to a larger number of clusters, even asymptotically. However, such results are not as bad as they sound: indeed, Ascolani *et al.* (2023) suggested that the issue is alleviated by placing a suitable hyperprior on the concentration parameter α, while Wade (2023) empirically showed that different estimators for the partition (obtained by minimizing different loss functions) lead to considerably different estimates for the number of clusters. Beyond this framework, Rajkowski (2019) proved interesting geometric properties of the maximum a posteriori partition.

In this work we consider a Gaussian kernel for model (1) and a purposely simple data generating mechanism, so that computation of posterior quantities becomes easier. We show that in this context the maximum a posteriori clustering converges to the "natural" partition of the observations.

2 Dirichlet process mixtures with Gaussian kernel

As discussed in Section 1, by the discreteness of the DP the set $(\theta_1, \dots, \theta_n)$, corresponding to observations $Y_{1:n}$, yields ties with positive probability. Therefore model (1) induces a distribution over the space of partitions of $[n] = \{1, \dots, n\}$. If $A = \{A_1, \dots, A_s\} \in \tau_s(n)$, where $\tau_s(n)$ is the space of partitions of $[n]$ in s non-empty and disjoint subsets, it is possible to show (Miller & Harrison, 2013; Ascolani *et al.* , 2023) that

$$\mathbb{P}(A \mid Y_{1:n}) \propto \alpha^s \prod_{j=1}^{s} \Gamma(a_j) \prod_{j=1}^{s} m\left(Y_{A_j}\right), \tag{2}$$

where $a_j = |A_j|$, $Y_{A_j} = \{Y_i \mid i \in A_j\}$ and $m\left(Y_{A_j}\right) = \int \prod_{i \in A_j} k(Y_i \mid \theta) P_0(\mathrm{d}\theta)$ denotes the marginal distribution of cluster j. We call the *maximum a posteriori clustering*, the partition $A^*(Y_{1:n})$ which maximizes the above posterior distribution, i.e. $A^*(Y_{1:n}) = \operatorname{argmax}_A \mathbb{P}(A \mid Y_{1:n})$. In this work we assume to observe scalar data points and

$$k(y \mid \theta) = N(\theta, \sigma^2) \quad \text{and} \quad P_0(\mathrm{d}\theta) = N(\mu_0, \sigma_0^2)\mathrm{d}\theta, \tag{3}$$

where $N(\mu, \tau^2)$ denotes the density of a normal distribution with mean μ and variance τ^2, while $(\mu_0, \sigma_0^2, \sigma^2)$ are fixed hyperparameters. With standard com-

putations it is easy to obtain

$$m\left(Y_{A_j}\right) = \sqrt{\frac{\sigma^2}{\sigma_0^2 a_j + \sigma^2}} (2\pi\sigma^2)^{-a_j} e^{\frac{\sigma_0^2 a_j + \sigma^2}{2\sigma^2\sigma_0^2}\left(\frac{\sigma_0^2 a_j}{\sigma_0^2 a_j + \sigma^2}\frac{1}{a_j}\sum_{i\in A_j} Y_i + \frac{\sigma^2}{\sigma_0^2 a_j + \sigma^2}\mu_0\right)^2}. \quad (4)$$

3 Data generating mechanism and main result

As it is commonly done in asymptotic analysis, we assume that the observed datapoints are not generated according to model (1), but rather are independent and identically distributed from a "true" distribution P^*. In the following we assume $P^*(\mathrm{d}y) = \delta_{c^*}(\mathrm{d}y)$, that is all the observations are equal to a fixed real value c^*. This is a stylized setting, where we expect the partition generated by model (1) to converge to $[n]$, i.e. all observations clustered together. However, Theorem 4.1 in Miller & Harrison (2013) implies that the posterior on the number of clusters does not converge to 1 as $n \to \infty$. Notice that Theorem 3 in Ascolani *et al.* (2023) shows instead that consistency holds with a prior on the concentration parameter α. In the following theorem we prove that, even with α fixed, the maximum a posteriori clustering converges to $[n]$, as expected.

Theorem 1. *Consider model (1) with kernel as in (3). Let* $Y_i \overset{i.i.d.}{\sim} \delta_{c^*}$, *with* $i = 1, \dots, n$. *Then, for every* $(\mu_0, \sigma_0^2, \sigma^2)$ *there exists* N *such that for every* $n \geq N$ *it holds* $A^*(Y_{1:n}) = [n]$.

Proof. Fix a triplet $(\mu_0, \sigma_0^2, \sigma^2)$. The statement is proved by showing that there exists N such that for every $n \geq N$ it holds

$$\sup_{2\leq s\leq n} \sup_{A\in\tau_s(n)} \frac{\mathbb{P}(A \mid Y_{1:n})}{\mathbb{P}([n] \mid Y_{1:n})} < 1.$$

By (4) it is easy to show that there exists a constant $K > 0$, which does not depend on s and n, such that $\alpha^{1-s}\prod_{j=1}^{s} m\left(Y_{A_j}\right)/m(Y_{1:n}) \leq e^{Ks}$ for every $A \in \tau_s(n)$. Therefore, by (2) we can give the following bound

$$\sup_{2\leq s\leq n} \sup_{A\in\tau_s(n)} \frac{\mathbb{P}(A \mid Y_{1:n})}{\mathbb{P}([n] \mid Y_{1:n})} \leq \sup_{2\leq s\leq n} \sup_{a\in\sigma_s(n)} e^{Ks} \frac{\prod_{j=1}^{s}\Gamma(a_j)}{\Gamma(n)},$$

where $\sigma_s(n) = \left\{a \in \{1, \dots, n\}^s \mid \sum_{j=1}^{s} a_j = n\right\}$. Moreover, it is not difficult to show that $\sup_{a\in\sigma_s(n)} \prod_{j=1}^{s}\Gamma(a_j) = \Gamma(n-s+1)$, which implies

$$\sup_{2\leq s\leq n} \sup_{A\in\tau_s(n)} \frac{\mathbb{P}(A \mid Y_{1:n})}{\mathbb{P}([n] \mid Y_{1:n})} \leq \sup_{2\leq s\leq n} e^{Ks} \frac{\Gamma(n-s+1)}{\Gamma(n)} =: \sup_{2\leq s\leq n} f(s).$$

Notice that $f(s+1) > f(s)$ if and only if $s > n - e^K$, so that $f(s)$ attains its maximum either at 2 or n. Therefore we conclude

$$\sup_{2\leq s\leq n} \sup_{A\in\tau_s(n)} \frac{\mathbb{P}(A \mid Y_{1:n})}{\mathbb{P}([n] \mid Y_{1:n})} \leq f(2) + f(n) = \frac{e^{2K}}{n-1} + \frac{e^{Kn}}{(n-1)!} \to 0$$

as $n \to \infty$, as desired. □

4 Discussion

We showed that, with constant data, despite inconsistency for the number of clusters, the maximum a posteriori clustering converges to the "true" partition. It would be of great interest to extend this result beyond such simple data generating mechanism, even if the identification of a "true" clustering becomes less clear: see Section 3 of Rajkowski (2019) for some examples. This will be object of future work.

References

Ascolani, F., Lijoi, A., Rebaudo, G., & Zanella, G. 2023. Clustering consistency with Dirichlet process mixtures. *Biometrika*, **forthcoming**.

Ferguson, T. S. 1973. A Bayesian analysis of some nonparametric problems. *Ann. Stat.*, **1**, 209–230.

Ghosal, S., & Van der Vaart, A. W. 2007. Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Stat.*, **35**, 697–723.

Ghosal, S., & Van Der Vaart, A. W. 2017. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Lo, A. Y. 1984. On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.*, **12**, 351–357.

Miller, J. W., & Harrison, M. T. 2013. A simple example of Dirichlet process mixture inconsistency for the number of components. *Adv. Neural Inf. Process. Syst.*, **26**, 199–206.

Miller, J. W., & Harrison, M. T. 2014. Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.*, **15**, 3333–3370.

Nguyen, X. 2013. Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Stat.*, **41**, 370–400.

Rajkowski, L. 2019. Analysis of the Maximal a Posteriori Partition in the Gaussian Dirichlet Process Mixture Model. *Bayesian Anal.*, **14**.

Wade, S. 2023. Bayesian cluster analysis. *Phil. Trans. R. Soc. A*, **381**.

MULTIPLE IMPUTATION FOR CLUSTERING ON INCOMPLETE DATA

Vincent Audigier [1], Ndèye Niang[1]

[1] CEDRIC Lab, MSDMA Teamn CNAM, (e-mail: vincent.audigier@cnam.fr, n-deye.niang_keita@cnam.fr)

ABSTRACT: We present how MI can be considered for addressing missing values in the context of clustering. For achieving this goal, we present a novel imputation method entitled FCS-homo, as well as a pooling method for the set of partitions obtained from each imputed data set. The proposed methodology is evaluated using a simulation study in comparison with state of the arts methods. We start by treating the case where the observations are generated from a gaussian mixture model with missing at random values. The study is completed by experiments based on various real data sets where the distribution of the data is unknown. These first results tend to show that multiple imputation is a efficient method for handling missing data in clustering, especially when the data distribution is unknown.

KEYWORDS: clustering, missing data, multiple imputation

1 Introduction

Among methods for addressing missing values, direct methods (DM) and multiple imputation (MI) are probably the most commonly considered. DM can be described as methods consisting in adapting the analysis methodology to be applied on incomplete data. This can be achieved by optimising a criterion based on incomplete data rather than complete data. DM are theoretically appealing, but they require a dedicated methodology for each analysis method. On the contrary, MI consists in separating the missing data issue to the analysis by proceeding in three steps. The first step is the imputation step, which consists in replacing each missing values by several plausible values. At the end, several imputed datasets are available. The second step consists in analysing each imputed dataset according to the analysis method wished. Finally, the third step consists in pooling the several analysis results to obtain a unique one. By separating the imputation step and the analysis step, MI allows applying any statistical analysis when missing values are imputed and consequently is less analysis method dependent that DM. However, it can also introduce bias if the imputation method is not well chosen in regard to the analysis model.

Several DM have been proposed to perform clustering with missing values. For instance, Marbac *et al.* (2019) proposed an EM algorithm to estimate parameters from a gaussian mixture model, Chi *et al.* (2016) proposed to extend kmeans criterion for accounting for missing values, while Hathaway & Bezdek (2001) extended fuzzy c-means algorithm by an *optimal completion strategy*. However, addressing missing values by MI remains challenging in clustering for at least two reasons. Firstly, because the imputation step requires specific models. Indeed, available imputation methods are generally based on the assumption that observations are drawn from a unique distribution, which is obviously inconsistent with the underlying assumptions made in cluster analysis. The second reason is that the way to pool partitions obtained at the second MI step is unclear. Indeed, the pooling rules in MI are theoretically applied on the parameters from a generalized linear model and not on a categorical variable as a partition of observations. Thus, addressing missing values in clustering by MI is not straightforward.

In this work, we propose a novel methodology for addressing missing values in clustering by MI. It consists in a novel imputation method entitled FCS-homo as well as a novel pooling rule.

2 Method

2.1 FCS-homo

Fully conditional specification (FCS) (van Buuren *et al.*, 2006) consists in imputing missing data by assuming a distribution for each variable conditionally to the others and then impute each variable sequentially according to each ones. FCS methods are often used in practice since they allow a better fit of the imputation model. More precisely, let $P(X_j|X_{-j};\zeta_j)$ be the distribution of X_j $(1 \leq j \leq p)$ conditionally to other variables, denoted X_{-j}, and parameterized by ζ_j. For instance, $P(X_j|X_{-j};\zeta_j) = \mathcal{N}(X_{-j}\beta,\sigma^2)$ with $\zeta_j = (\beta,\sigma)$. Then, FCS methods impute the mth data set as follows:

- initialize missing values of $\mathbf{X}$ by random draws from observed values
- for j in 1 ... p
  - a generate ζ_j based on observed individuals on X_j
  - b impute $X_j$according to $P(X_j|X_{-j};\zeta_j)$
- repeat until convergence

In a context of cluster analysis, we propose a FCS method which accounts for the cluster data structure. To achieve this goal, each regression model is

conditional to a supplementary variable W indicating the cluster of each observation. Let $Z = (W, X)$ be the incomplete data set gathering the cluster variable W, which is unknown and considered as fully missing, and X the incomplete data set. Then, the algorithm involves two main steps: imputation of Z given W and vice versa. Generating Z given W is performed using regression models including an intercept specific to each cluster $P(Z_j|Z_{-j}, W; \zeta_j) = \mathcal{N}(Z_{-j}\beta + \mu_w, \sigma^2) \quad \zeta_j = (\beta, \sigma, \mu_w)$ while generating W given Z is performed using linear discriminant analysis (see Audigier *et al.* (2021) for more details).

2.2 Pooling

Given M imputed data set, we denote Ψ_m the partition obtained from the data set m. This partition can be obtained from any clustering algorithm (e.g. k-means). The set $(\Psi_m)_{1 \leq m \leq M}$ is pooling using Non Negative matrix Factorization which consists in looking at the partition $\bar{\Psi}$ such as

$$\bar{\Psi} = argmin_{\Psi} \sum_{m=1}^{M} \delta(\Psi, \Psi_m) \tag{1}$$

with $\delta(\Psi, \Psi_m)$ the number of disagreements * between Ψ and Ψ_m. An associated instability can also be computed as proposed in Audigier & Niang (2022).

3 Results

The proposed methodology is evaluated by comparison with DM approaches under MAR mechanisms. For this purpose, we focus on three clustering techniques: the Gaussian mixture model, the k-means and the fuzzy c-means. The study is first carried out on data simulated according to a Gaussian mixture model in which we vary the separability of the clusters, their number, their size and their correlation structure. Missing data are generated according to different mechanisms varying by their nature (MCAR or MAR) and by the rate of missing values. In a second step, both approaches are compared on different real data sets where the distribution is not known but where a cluster structure is well identified. In both cases, the three clustering techniques are applied using the theoretical number of clusters and the missing data are handled either directly or by multiple imputation. The resulting partitions are

*$\delta(\Psi, \Psi') = \sum_{(i,i')} \delta_{ii'}$ with $\delta_{ii'} = 1$ if individuals i and i' are in the same cluster for a given partition and not for the second, while $\delta_{ii'} = 0$ otherwise

then compared to the expected partition according to the adjusted Rand index (ARI).

4 Discussion

The study illustrates that the use of multiple imputation for handling missing values in clustering generally improves the partition quality for geometric clustering methods, namely k-means and fuzzy c-means, compared to direct k-pod and optimal completion strategy approaches (respectively). As for the results on the parametric Gaussian model approach, similar performances are observed when the data are derived from a Gaussian mixture. Nevertheless, significant differences are observed on real data where the direct methods often lead to lower ARI.

Thus, these first results tend to show that MI is a efficient method for handling missing data in clustering, especially when the data distribution is unknown. Moreover, this technique allows to apply any clustering method on incomplete data, whereas direct methods remain specific to the clustering technique considered.

References

AUDIGIER, V., & NIANG, N. 2022. Clustering with missing data: which equivalent for Rubin's rules? *Advances in Data Analysis and Classification*, Sept.

AUDIGIER, V., NIANG, N., & RESCHE-RIGON, M. 2021. *Clustering with missing data: which imputation model for which cluster analysis method?*

CHI, J. T., CHI, E. C., & BARANIUK, R. G. 2016. k-POD: A Method for k-Means Clustering of Missing Data. *The American Statistician*, **70**(1), 91–99.

HATHAWAY, R. J., & BEZDEK, J. C. 2001. Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, **31**(5), 735–744.

MARBAC, M., SEDKI, M., & PATIN, T. 2019. Variable selection for mixed data clustering: application in human population genomics. *Journal of Classification*, 1–19.

VAN BUUREN, S., BRAND, J., GROOTHUIS-OUDSHOORN, C., & RUBIN, D. 2006. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, **76**, 1049–1064.

INTEGRATIVE FACTOR MODELS FOR BIOMEDICAL APPLICATIONS

Alejandra Avalos-Pacheco [1] and Roberta De Vito [2]

[1] Applied Statistics Research Unit, TU Wien, Vienna, Austria, and
Harvard-MIT Center for Regulatory Science, Harvard University, Boston, MA,
(e-mail: alejandra.avalos@tuwien.ac.at)

[2] Dept. of Biostatistics, School of Public Health, Brown University, Providence, RI,
(e-mail: roberta_devito@brown.edu)

ABSTRACT: Data-integration of multiple studies is key to understanding and gaining knowledge in statistical research. However, such data present artifactual sources of variation, also known as covariate effects. Covariate effects can be complex and can lead to systematic biases. If not corrected, these biases may lead to unreliable inferences. Here, we will present novel sparse latent factor regression and multi-study factor regression models to integrate heterogeneous data.

KEYWORDS: factor regression, multi-study factor analysis, sparsity, non-local priors, scalable algorithms

1 Introduction

Data integration is crucial when separate data sources are collected on the same phenomenon. For instance, different economical studies may test the efficacy of several policy-making interventions; clinical trials may analyze various treatments using data gathered at different times. Integrative models provide gains in statistical power and help to take accurate decisions sooner. However, a lack of appropriate integration tools could lead to unreliable inference.

Data integration in biomedicine is particularly challenging as some measurements reappear across different studies. However, high throughput experiments display both biological and artifactual sources of variation. Here, we will present novel sparse factor regression and multi-study factor regression models to integrate such heterogeneous data.

The factor regression (FR) model (Avalos-Pacheco, 2018, Avalos-Pacheco *et al.*, 2022) provides a tool for data exploration via dimensionality reduction and sparse low-rank covariance estimation while correcting for a range of covariate, or artifactual, effects, such as batch effects. A limitation of FR models is the inability to isolate the study-specific latent structure.

The multi-study factor analysis (De Vito *et al.*, 2019, De Vito *et al.*, 2021) is able to handle multiple high-throughput experiments, simultaneously achieving two goals: a) to capture common component(s) across studies and b) to isolate the sources of variation that are unique of each study. We generalize the multi-study factor analysis by adopting a factor regression approach. Our proposed multi-study factor regression (MSFR) will enable us to jointly obtain the group-specific covariances and the common component.

In the conference presentation, we will discuss the use of several sparse priors, local and non-local (Johnson & Rossell, 2010), for learning the dimension of the latent factors. Our approaches provide a flexible methodology for sparse factor regression, which is not limited to data with covariate effects. Our models are fitted via scalable expectation–maximization (EM) algorithms.

We will also show the usefulness of our methods by presenting several examples, with a focus on bioinformatics applications. For all the examples, we give a visual representation of the latent factors of the data. Thereafter, in the case of cancer genomics data sets, we provide survival predictions leveraging the obtained factors; in the case of a Hispanic community health nutritional-data study, we obtain dietary patterns, associating each factor with a measure of overall diet quality related to cardiometabolic disease risk.

2 Model specification

We follow the model proposed in Avalos-Pacheco *et al.*, 2022. We consider vectors $\mathtt{x}_{is} = (x_{i1s}, x_{i2s}, \ldots, x_{ips})^\top \in \mathbb{R}^p$, observed for $i = 1, \ldots, n$ individuals in study $s, s = 1, \ldots, S$. The factor regression model defines $\mathtt{x}_{is}$ as a regression on p_v observed covariates $\mathtt{v}_{is} \in \mathbb{R}^{p_v}$, and q low-dimensional latent variables $\mathtt{f}_{is} \in \mathbb{R}^q$, also known as latent coordinates or factors $\mathtt{x}_{is} = \theta \mathtt{v}_{is} + \Phi \mathtt{f}_{is} + \mathtt{e}_{is}$, where $\theta \in \mathbb{R}^{p \times p_v}$ is the matrix of regression coefficients, $\Phi \in \mathbb{R}^{p \times q}, q \ll p$, is the loading matrix, $\mathtt{e}_{is} \in \mathbb{R}^p$ is the error, distributed as $\mathtt{e}_{is} \sim N(0, \mathcal{T}_s^{-1})$ independently across $i = 1, \ldots, n$, with $\mathcal{T}_s^{-1} = \text{diag}\{1/\tau_{ls}, l = 1, \ldots, p\}$ as the idiosyncratic precision matrix for study s. Factors are assumed to be standard normal, $\mathtt{f}_{is} \sim N(0, \mathbf{I})$, independent across $i = 1, \ldots, n$ and independent of $\mathtt{e}_{is}$.

We first set priors for the precisions $\tau_{ls} \mid \eta, \xi \sim$ Gamma$(\eta/2, \eta\xi/2)$,, and regression parameters $\theta \sim N(0, \psi\mathbf{I})$. The loadings $\Phi = \{\phi_{jk}, j = 1, \ldots, p, k = 1, \ldots, q\}$ play a key part in factor models as they allow us to improve shrinkage and simplify interpretation. Here, we set a non-local spike-and-slab prior on ϕ_{jk}, as in Avalos-Pacheco *et al.*, 2022. This prior distinguishes the loading elements that should be included, modelled by the slab component, from those that should be excluded, modelled by the spike component. We consider a mix-

ture distribution with a product moment non-local prior (Johnson & Rossell, 2010) for the slab components and a normal prior for the spike components:

$$\mathrm{p}(\phi_{jk} \mid \gamma_{jk}) = (1-\gamma_{jk})\mathrm{N}(\phi_{jk};0,\lambda_0) + \gamma_{jk}\frac{\phi_{jk}^2}{\lambda_1}\mathrm{N}(\phi_{jk};0,\lambda_0). \tag{1}$$

We set a hierarchical prior over the latent indicator $\gamma_{jk} \mid \zeta_k \sim \text{Bernoulli}(\zeta_k)$, $\gamma_{jk} \mid \zeta_k \sim \text{Beta}\left(\frac{a_\zeta}{k}, b_\zeta\right), j = 1,\ldots,p, k = 1,\ldots,q$.

Inference is done by an efficient EM algorithm with closed-form expressions. We refer to Avalos-Pacheco *et al.*, 2022, for details, prior elicitation, parameter initialization, post-processing and description of the EM algorithm.

3 Pancreatic cancer

To quantify the effectiveness of our approach, we study an unpublished gene expression data set for individuals with pancreatic cancer. We analyze two studies collected under different experimental conditions and sizes ($n_1 = 27$ and $n_2 = 183$). We select the 5% genes with the highest total variance across all samples ($p = 1,177$ genes). We normalize the data to have zero mean and unit variance and included the type of tissue (normal or tumour) and a study indicator as covariates for our model. In order to evaluate the effect of the non-local prior, we compare our model (FR-NLSS) with methods that use a normal spike-and-slab prior (George & McCulloch, 1993) (FR-LSS), instead of our proposed non-local spike-and-slab prior, and that do not leverage any sparse inducing priors (FR-NS). Since the data generating ground truth is unknown, we assess the performance of our estimators by evaluating the cross-validated log likelihood. Table 1 presents the results from 10 independent runs of 10-fold cross-validation. It displays the selected number of factors $\widehat{q}$, the number of estimated non-zero loadings $||\widehat{\Phi}||_0$ and the cross-validated loglikelihood.

Table 1. *Cross-validated log-likelihood analysis for pancreatic cancer dataset.*

	$\widehat{q}$	$\|\|\widehat{\Phi}\|\|_0$	Log-likelihood
FR-NS	100.0	117,700	-1,644.8
FR-LSS	63.0	74,151	-1,622.0
FR-NLSS	19.0	22,363	-1,157.6

The results in Table 1 show that our proposed FR-NLSS obtained a better out-of-sample log-likelihood with fewer factors and sparser Φ than our competitors. Thus, we conclude that FR-NLSS reconstructed the data better than the other methods.

4 Extensions

We extend the FR Model to the Multi-study factor setting (De Vito *et al.*, 2019, De Vito *et al.*, 2021). We refer to this generalization as the Multi-study factor regression (MSFR) (De Vito & Avalos-Pacheco, 2023+).

Marginally, the underlying covariance of x_{is} of the FR Model is $\Sigma_s = \Phi\Phi^\top + \mathcal{T}_s^{-1}$. In the MSFR setting, the Σ_s becomes

$$\Sigma_s = \Phi\Phi^\top + \Lambda_s\Lambda_s^\top + \mathcal{T}_s^{-1}, \tag{2}$$

where $\Lambda_s \in \mathbb{R}^{p\times q_s}, q_s \ll p$, is the study-specific loading matrix. The new Σ_s allows to explain the total variance into the variance of the common factors, the variance of the study-specific factors and the idiosyncratic error. In the conference presentation, we will discuss the FR and MSFR in detail, and we will apply our models to different gene expression and nutritional epidemiology data sets. Both our FR and MSFR will be demonstrated to be valuable to visually depict the underlying factors of the data; and to make survival predictions or to identify dietary patterns and study the embedded risk of cardiometabolic disease. We refer to De Vito & Avalos-Pacheco, 2023+ for further details.

References

AVALOS-PACHECO, A. 2018. *Factor regression for dimensionality reduction and data integration techniques with applications to cancer data*. University of Warwick, PhD thesis.

AVALOS-PACHECO, A., ROSSELL, D., & SAVAGE, R. S. 2022. Heterogeneous large datasets integration using Bayesian factor regression. *Bayesian analysis*, **17**(1), 33–66.

DE VITO, R., & AVALOS-PACHECO, A. 2023+. Multi-study factor regression model: An application in nutritional epidemiology. *arXiv:2304.13077*.

DE VITO, R., BELLIO, R., TRIPPA, L., & PARMIGIANI, G. 2019. Multi-study factor analysis. *Biometrics*, **75**(1), 337–346.

DE VITO, R., BELLIO, R., TRIPPA, L., & PARMIGIANI, G. 2021. Bayesian multistudy factor analysis for high-throughput biological data. *The annals of applied statistics*, **15**(4), 1723–1741.

GEORGE, E., & MCCULLOCH, R. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**(423), 881–889.

JOHNSON, V. E., & ROSSELL, D. 2010. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **72**(2), 143–170.

TEST EQUATING WITH EVOLVING LATENT ABILITY

Silvia Bacci[1], Bruno Bertaccini[1], Carla Galluccio[1], Leonardo Grilli[1] and Carla Rampichini[1]

[1] Department of Statistics, Computer Science, Applications "G. Parenti", (e-mail: silvia.bacci@unifi.it, bruno.bertaccini@unifi.it, carla.galluccio@unifi.it, leonardo.grilli@unifi.it, carla.rampichini@unifi.it)

ABSTRACT: In large-scale assessments, students' ability is usually evaluated using multiple test forms, which require the use of several items. In this context, calibrating items before the official tests can be difficult for different reasons. A solution is to calibrate items during the first test administration and then use these estimates in the subsequent ones. However, this approach does not consider that the populations could be significantly different in terms of average ability, which is particularly problematic when the final output of this process is a merit ranking. Our findings show that, on one side, calibrating item parameters on populations with differences in ability does not affect the final merit ranking and, on the other side, the differences in item parameter estimates are significant.

KEYWORDS: large-scale assessment, bifactor model, test equalisation

1 Introduction

In large-scale assessments, it is common practice to construct multiple test forms so as to increase test security and allow for tests to be implemented on different exam dates and times (van der Linden & Adema, 1998). This organisation requires using many items (in order to construct parallel versions of the same test for each group) and the application of equalisation methods to make the scores obtained on different test forms comparable, a relevant concern when the final output of this process is a merit ranking. Additionally, field trials are usually required to calibrate test items when using IRT models (Hambleton *et al.*, 1991) to assess subjects' ability in large-scale assessments.

It is worth understanding that calibrating items before official tests could be problematic for different reasons. Among these, the set of items used is usually not large enough to give the possibility of using in advance items that should then constitute the official tests. A possible approach is to calibrate

items on the first group of subjects and then use the item parameter estimates to assess the ability of students who are subsequently administered the tests.

A problem that might arise in some contexts, such as university entrance tests, is that the population on which items are calibrated at the baseline may differ significantly from those at subsequent administrations. For example, it is reasonable to assume that students who took the test at the baseline have lower abilities than those who took it later, at least because they had more time to study and became familiar with the type of test.

The present work aims at answering the following research questions:

RQ1 Is tests equalisation, and consequently the merit ranking, affected by differences in the population average ability?

RQ2 How does calibrating items on a certain population affect estimates of ability in a population that differ for the average ability levels?

2 Statistical Model

In certain contexts, the structure of the test is characterised by the presence of subsets of questions concerning the same topic (referred to as testlets), which implies a violation of the hypothesis of local independence of the items (van der Linden & Hambleton, 2013). Thus, models capable of managing the multidimensionality of the latent trait are required.

In multidimensional item response theory (Reckase, 2009), the bifactor (BF) model (Holzinger & Swineford, 1937) is often used due to its good performance on different kinds of data. In the BF model, a common (i.e., generic, primary) latent variable is assumed to underlie all test items. In addition, specific latent variables (one for each testlet) account for the residual dependence remaining after considering the primary latent construct and due to the presence of the testlets. Primary and specific latent variables are orthogonal.

Let us consider a set of individuals $i = 1,\ldots,n$ taking a test with $j = 1,\ldots,J$ items divided into $s = 1,\ldots,S$ sections. In the two-parameter BF model for dichotomous items Y_{ijs}, the probability that test taker i correctly answer item j of section s is defined as

$$P(Y_{ijs} = 1|\theta_{0i},\theta_{si}) = \frac{1}{1+\exp(-[d_j + a_{0j}\theta_{0i} + a_{sj}\theta_{si}])},$$

where θ_0 is the primary latent variable, θ_s is the s-th specific latent variable, d_j denotes the difficulty parameter of item j, a_{0j} and a_{sj} represent the discrimination parameters of item j on the primary and specific constructs, respectively. If item j loads on specific factor s, $a_{sj} \neq 0$, otherwise $a_{sj} = 0$.

3 Simulation Study

To answer the two research questions RQ1 and RQ2, we performed a simulation study. A test with 50 dichotomously-scored items was generated for the study, including four testlets composed of 7, 15, 15, and 13 items, respectively. Parameters a_{0j} were sampled from a log-normal distribution $logN(0, 0.5)$ constrained to $[0.5, 2]$. Moreover, for each testlet parameters a_{sj} were sampled from a uniform distribution $[0.5, 0.7]$, corresponding to a moderate degree of local dependence between items. Difficulty parameters d_j were sampled from a normal distribution $N(0, 1)$. We assume that the same set of items is administered in two different time occasions.

The generic and the specific latent abilities θ_0 and θ_s were generated from a mixture of two independent Gaussian distributions:

$$f(\theta) = \pi_A f_A(\theta) + \pi_B f_B(\theta)$$

where $f(.)$ is the normal density and π_A and π_B are the mixture component weights, with $\pi_A + \pi_B = 1$. The mean of the mixture is $\mu_M = \pi_A\mu_A + \pi_B\mu_B$, and its variance is $\sigma^2_M = \pi_A\sigma^2_A + \pi_B\sigma^2_B + \left[\pi_A\mu^2_A + \pi_B\mu^2_B - (\pi_A\mu_A + \pi_B\mu_B)^2\right]$.

We assume the mixture components f_A and f_B have mean $\mu_A = -2$ and $\mu_B = 2$ respectively, and common variance $\sigma^2 = 1$. We simulate two groups of subjects with different ability distributions: the baseline group (group 1) with 80% of subjects from the first component and 20% from the second one, and a second time occasion group (group 2) with 20% of subjects from the first component and 80% from the second one. Note that with this configuration, the mixture distributions of groups 1 and 2 have different means but equal variance. For each group, $N = 10,000$ response patterns were simulated. In addition, a set of 500 subjects was assumed to repeat the test, and thus, they are present in both groups, with an ability improvement of 0.5 in group 2 compared to group 1. Parameters estimation was carried out through the EM algorithm implemented in the R package `mirt`.

4 Results

To investigate RQ1, we considered the merit ranking obtained by estimating a BF model under three different strategies: (i) considering the two groups separately; (ii) considering the two groups together; (iii) using for the second group the item parameters estimated on the first one. The merit ranking resulting from each strategy was compared to the true ranking by using the Pearson correlation coefficient.

The correlation coefficients are equal to 0.86, 0.96 and 0.96, respectively, showing no differences in estimating subjects' abilities θ_0 for the two groups together or using the parameters estimated on the first groups in the second one in terms of merit ranking. Conversely, the coefficient obtained when the two models are estimated separately (strategy *i*) is remarkably lower. This result is in line with the literature on equalisation methods with non-equivalent groups.

To answer RQ2, we compared some constrained BF models. We first assessed a base (unconstrained) model (Model 0), in which the 30% of items in each testlet were in common and the other ones were considered as different, so that different parameters were estimated for the same item administered in the two time occasions. Then, four models (Model 1-4) nested in the base one were estimated, where the items within each testlet were constrained to have equal parameters across the two time occasions. We compared the constrained models with the base one using BIC, AIC, and the log-likelihood. Results provide evidence in favor of the base model, recognising an effect on item parameter estimation when populations present remarkable differences in ability.

5 Conclusions

Preliminary results above presented advice against separately calibrating tests administered in different occasions and outline the presence of an effect of populations with different ability distributions on the item parameters. Future work will focus on extending the simulation study to more general scenarios, such as different mixtures of populations and tests with only a sub-set of common items.

References

HAMBLETON, R.K., SWAMINATHAN, H., & ROGERS, H.J. 1991. *Fundamentals of Item Response Theory*. Vol. 2. Sage, California.

HOLZINGER, K.J., & SWINEFORD, F. 1937. The bi-factor method. *Psychometrika*, **2**, 41–54.

RECKASE, M.D. 2009. *Multidimensional Item Response Theory Models*. Springer, New York.

VAN DER LINDEN, W.J., & ADEMA, J.J. 1998. Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, **35**(3), 185–198.

VAN DER LINDEN, W.J., & HAMBLETON, R.K. 2013. *Handbook of Modern Item Response Theory*. Springer Science & Business Media, New York.

CAUSAL INFERENCE ON THE IMPACT OF EXTREME AMBIENT TEMPERATURES ON POPULATION HEALTH

Michela Baccini[1], Alessandra Mattei[1], Elena Degli Innocenti[1], Giulio Biscardi[1], Aitana Lertxundi[2,3]

[1] Department of Statistics, Computer Science, Applications, University of Florence, Italy (e-mail: michela.baccini@unifi.it)

[2] Department of Preventive Medicine and Public Health, University of Basque Country, Leioa, Spain

[3] Biodonostia Health Research Institute, San Sebastian, Spain

ABSTRACT: A potential outcome approach to causal inference is used to infer the average exposure-response curve describing the relationship between daily temperature and daily mortality in the city of San Sebastian (Spain) for the period 2010-2015. The analysis relies on the estimate of the generalized propensity score and specification of a model for potential outcomes. The impact of extreme temperatures on population health is also provided, in terms of attributable deaths.

KEYWORDS: temperature, mortality, dose-response curve, generalized propensity score, health impact assessment.

1 Introduction

Climate change is now regarded as the greatest challenge of the 21st century. Extreme temperature levels are one of its consequences. Many studies, based on the analysis of daily time series through regression approaches, have identified a U-, V- or J-shaped relationship between environmental temperature and mortality, indicating that heat and cold are associated with death counts. For the first time, this study estimates this relationship by using a potential outcome approach to causal inference. The method proposed is based on the generalized propensity score and uses a semi-parametric specification for the outcome model. We ground on the method used in Forastiere et al. (2020) for the analysis of the short term effect of air pollution on mortality in the city of Milan (Italy).

2 Data

The health and exposure data used in this paper have been collected for the city of San Sebastian (Basque Country region of Spain) for the period 2010-2015. They include the daily number of deaths from natural causes, cardiovascular and respiratory causes, grouped by age (0-64, 65-84, 85+); meteorological variables

(temperature and humidity); and several known confounders of the temperature-mortality relationship (average pollutant levels and an indicator of influenza epidemics).

3 Methods

The analyses are performed separately for cold and warm season.
According to the potential outcome framework, under the Stable Unit Treatment Value Assumption (Imbens and Rubin 2015), we denote by Y_i (z) the potential number of deaths in day i (i = 1, 2,…, n) if z were the level of temperature in that day. For each day we only observe one potential outcome, that is, the one corresponding to the actual exposure of that day, Z_i, all the other potential outcomes with $z \neq Z_i$ being missing. We denote the observed outcome with Y_i, while $\mathbf{X}_i = (x_{1i}, x_{2i}, \ldots, x_{Ki})$ is the vector of the K covariates measured on day i.
We are interested in the average Dose Response Function (aDRF), defined as:

$$\mu(z)=n^{-1}\sum_i Y_i(z). \qquad [1]$$

Under the unconfoundedness assumption, we fill in missing potential outcomes in [1] following the procedure described in Hirano and Imbens (2004), which requires the specification of a model for the exposure, used for GPS estimation, and a model for the outcome.
The model for the exposure is a log-Normal model on the daily average temperature Z_i, given the confounders ($\mathbf{X}_i$) and seasonality terms. The confounders are included in the model through flexible functions and interactions are allowed. The GPS for day i at the level of exposure z is then defined as the value of the density function for log(Z), derived from the estimated model:

$$r(z, \mathbf{X}_i)=(2\pi s)^{-1}\exp[-(\log(z)-m_i)^2/(2s^2)],$$

where m_i is the value of log(Z) predicted by the model for day i, and s^2 is the estimated error variance.
The model for the outcome is a Poisson regression model on the daily mortality Y_i, given both daily average temperature Z_i and the value of GPS estimated for $z = Z_i$, $R_i = r(Z_i, \mathbf{X}_i)$. Different specifications of the outcome model can be adopted: we define a bivariate spline on temperature and GPS.
Once the two models have been estimated, there is the phase of prediction and potential outcome imputation. After defining a grid of temperatures, we calculate, for each day, the GPS on each value z^* of the grid. Then, we plug in z^* and the corresponding GPS, $r(z^*, \mathbf{X}_i)$, in the estimated outcome model, in order to predict the mortality level $Y_i(z^*)$ that would be observed if the temperature in day i were equal to z^*. Finally, for each z*, the predicted potential outcomes are averaged over the days, so that an exposure-response curve is obtained.

We estimate the aDRF for mortality from all causes and by cause of death, for all age and separately by age group. Also, we estimate, in terms of attributable deaths, the impact of temperatures higher or lower than specific thresholds on population mortality. Confidence intervals are obtained through a block-bootstrap procedure.
A crucial point in the analysis is the specification of the exposure model. The validity of the specification adopted for the exposure model is assessed by checking the covariate balance as described in Hirano and Imbens (2004).

4 Results

Extreme temperatures, both cold and warm, have a detrimental effect on health. The so-called `turning point', defined as the temperature where the aDRF is minimum, is found to be around 19.5° C. The analysis by age group confirms these effects for people over 65 years of age, while negligible effects are observed for younger people (0-64).
Taking the value of 19.5°C as an optimal threshold for health, we estimate that, in the warm season, exceeding it has caused 115 deaths (90% CI: 22.39, 229.31) during the study period. In the cold season, staying below the same threshold is estimated to have caused 483 deaths (90% CI: 97.21, 836.64).

5 Discussion

This study states the existence of a causal relationship between temperature and mortality and provides an approach to estimate the average dose-response function, as well as the impact of extreme exposures. Extensions of the method could allow the estimation of an entire curve on the whole year and the investigation of the delayed effect of temperature on mortality.

References

FORASTIERE, L., CARUGNO, M., & BACCINI, M. 2020. Assessing short-term impact of PM10 on mortality using a semiparametric generalized propensity score approach. *Env Health.* **19**, 46.

IMBENS, G., & RUBIN, D. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

IRANO, K. & IMBENS, G.W. 2004. The propensity score with continuous treatments. In: Gelman, A., Meng, X.L., editors. Applied Bayesian Modelling and CausalInference from Missing Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family. p. 73–84.

MEASUREMENT INVARIANCE TESTING OF LATENT CLASS MODELS USING RESIDUAL STATISTICS AND LIKELIHOOD RATIO TEST

Zsuzsa Bakk [1]

ABSTRACT: In latent class (LC) analysis a standard assumption is conditional independence, that is the indicators of the LC are independent of the covariates given the LCs. We compare the likelihood ratio based MIMIC test to residual statistics (BVR and $EPC_{interest}$) for identifying nonuniform direct effects (DEs) of covariates on the indicators of the LC model. The simulation study results show that the LR test and $EPC_{interest}$ correctly identifies direct effects more often than the BVR.

KEYWORDS: latent class analysis, measurement invariance, bivariate residuals, EPC, likelihood ratio test

1 Introduction

An often violated basic assumption of latent class modeling is the conditional independence assumption, also known as measurement equivalence. That is the association between the indicators of the LC model and the covariates are conditionally independent given the latent classes. Measurement equivalence can be tested by likelihood ratio based tests that compare the measurement equivalent model to models where direct effects (uniform or nonuniform) of covariates are allowed on the indicators of the LC model. An alternative approach for detecting missfit of the conditional independence model is to use residual statistics that can show violations of the conditional independence assumption. In this presentation we compare the power of the likelihood ratio based MIMIC model (Masyn, 2017) and that of two residual statistics ($EPC_{interest}$ and BVR) to detect the most complex type of measurement invariance, nonuniform direct effect. We first introduce the simple LC model, followed by a short presentation of the 3 approaches to detect missfit, compare them via a simulation experiment and conclude.

2 Latent class model

Consider the vector of responses $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})$, where Y_{ik} denotes the response of individual i on one of the K categorical indicator variables, with $1 \leq k \leq K$ and $1 \leq i \leq N$. Latent class (LC) analysis assumes that respondents belong to one of the T categories ("latent classes") of an underlying categorical latent variable X which affects the responses (Goodman, 1974). The measurement model for $\mathbf{Y}_i$ can then be written as:

$$p(\mathbf{Y}_i|Z) = \sum_{t=1}^{T} p(X=t|Z) \prod_{k=1}^{K} p(Y_{ik}|X=t). \quad (2)$$

Usually the conditional class membership probabilities $P(X|Z)$ are parameterized using a multinomial logistic regression parametrization:

$$P(X=t|Z=z_i) = \frac{\exp(\alpha_t + \beta Z_i)}{1+\sum_{t=2}^{T}\exp(\alpha_t + \beta Z_i)}. \quad (3)$$

The model defined in Equation 2 assumes that given the LC variable X there is no direct relationship between Z and $\mathbf{Y}$ - a fairly common assumption in LV modeling known as measurement invariance. This assumption can be relaxed:

$$p(\mathbf{Y}_i|Z_i) = \sum_{t=1}^{T} p(X=t|Z_i) \prod_{k=1}^{K} p(Y_{ik}|X=t, Z_i). \quad (4)$$

The most complex nonuniform DE can be parameterized as:

$$P(Y=y|Z=z_i) = \frac{\exp(\alpha_t + \beta \mathbf{X_t} + \beta Z_i|X)}{1+\sum_{t=2}^{T}\exp(\alpha_t + \beta \mathbf{X_t} + \beta Z_i|X)}. \quad (5)$$

The simpler uniform DE would mean dropping the class specific formulation of the effect of Z.

3 Identifying direct effects in LC models

3.1 Residual statistics

The BVR evaluates the residual association between each possible pair of observed variables (j, j') using a χ^2 test with 1 degree of freedom. The statistics can be formally defined as:

$$BVR_{jj'} = 1/P \sum_{j}\sum_{j'} \frac{(n_{jj'} - En_{jj'})^2}{n_{jj'}} \quad (6)$$

where the expected association $En_{jj'}$ for the covariate- indicator association is defined based on equation 2 in such a way that given the LC variable X there is no association between Z and Y. A downside of BVR is that the assumption of χ^2 distribution with 1 df does not hold (Oberski *et al.*, 2017).

.

Based on equations 5 we can see that the test of measurement invariance often takes the form of restricting a set of parameters to 0. In our case this refers to $\beta Z|X$. Let us consider a restriction on a vector of such logit coefficients as $\psi = \mathbf{0}$. In a general form the $EPC_{interest}$ can be formulated as:

$$EPC_{interest} = \mathbf{P}(\frac{\partial\theta}{\partial\psi'})(\psi - \psi') \tag{7}$$

where $\mathbf{P}$ is a matrix selecting the parameters of interest and θ is the vector of free model parameters. $EPC_{interest}$ can be seen as a linear approximation of the relationship between the free and fixed parameters of interest (Oberski *et al.*, 2017).

3.2 Likelihood ratio based stepwise multiple indicator multiple cause (MIMIC) modeling

The likelihood ratio based MIMIC approach (Masyn, 2017) is a multistage approach where nested models are compared with the goal to find the least restrictive well fitting model. The approach starts by comparing the latent class model with covariate (see Eq 2) to the model including all possible nonuniform DEs (see Eq 5). In case the LR test of the 2 nested models shows better fit of the all-DE model, the assumption of no DE is rejected, and a stepwise approach follows to identify the source of misfit. In the 2nd step an item by item testing of non uniform DE is performed, followed by an item by item testing of uniform DE for items for which a non uniform DE was confirmed in step 2. The approach has in total 7 possible steps, but we focus only on first 2 steps that focus on identifying nonuniform DE.

4 Simulation study

Table 1. *Percentage of correctly(T) and wrongly (F) identified DE with BVR, EPC and LR test separately for the low and high DE condition per latent class separation condition averaged over all sample sizes*

Class	High DE						Low DE					
sep	BVR_T	BVR_F	EPC_T	EPC_F	LR_T	LR_F	BVR_T	BVR_F	EPC_T	EPC_F	LR_T	I
high	,18	,00	,98	,16	,97	,18	,00	,00	,41	,10	,63	,
med	,22	,00	,86	,20	,83	,29	,00	,00	,44	,12	,60	,
low	,03	,00	,41	,15	,52	,34	,00	,00	,37	,15	,42	,

To test the ability of the 3 approaches to identify the presence of non uniform DE we run a simulation study with a LC model with 3 equal sized classes (class 1 low on all indicators, class3 high on all, class 2 low on first 3, high on last 3 indicators) measured by 6 indicators and regressed on a covariate. A full factorial design crossing sample size (250,500,1000,200), class separation ($Y|X$: .70,.80,.90), strength of DE (low, $\beta Z|X = .25$; high, .75) was used. DE was allowed on items 1 and 6.

When comparing the LR test for all nonuniform DE vs no DE model in all simulated conditions the more complex model was chosen, as such results are not detailed. The results in Table 1 show that the BVR is not a good statistic to identify a nonuniform DE, while the performance of $EPC_{interest}$ and LR test is better, their ability to identify a DE strongly depends not only on the strength of the DE, but also on the quality of the measurement model. With weaker measurement models all statistics fail to have a nominal rate close to the 95%.

5 Discussion

In a simulation experiment we compared $EPC_{interest}$, BVR and LR tests to identify a nonuniform DE. The results show that the $EPC_{interest}$ and LR test are more reliable, yet only in a few conditions meat the nominal 95% true-positive rate while maintaining a high false positive rate (between .16% to .29%). The BVR test was the most unreliable. We can conclude that nonuniform DE in most conditions is under identified by all estimators.

References

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.

MASYN, K. E. 2017. Measurement Invariance and Differential Item Functioning in Latent Class Analysis With Stepwise Multiple Indicator Multiple Cause Modeling. *Structural Equation Modeling*, **24**(2), 180–197.

OBERSKI, DANIEL L., VERMUNT, JEROEN K., & MOORS, GUY B. D. 2017. Evaluating Measurement Invariance in Categorical Data Latent Variable Models with the EPC-Interest. *Political Analysis*, **23**(4), 550–563.

NETWORK INTERFERENCE AND EFFECT MODIFICATION

Falco J. Bargagli-Stoffi*[1], Costanza Tortú *[2] and Laura Forastiere[3]

[1] Harvard University, (e-mail: fbargaglistoffi@hsph.harvard.edu)

[2] Sant'Anna School for Advanced Studies, (e-mail: costanza.tortu@santannapisa.it)

[3] Yale University (e-mail: laura.forastiere@yale.edu)

ABSTRACT: While most of causal inference studies typically disregard interference between units, it's important to recognize that agents often interact through social, physical, or virtual connections, and the effect of the intervention can propagate from one unit to other connected individuals in the network. In this work, we propose an innovative machine learning algorithm called Network Causal Tree (NCT), which combines a tree-based methodology with a Horvitz-Thompson estimator to assess the heterogeneity of treatment and spillover effects with respect to individual and network characteristics, in the presence of clustered network interference. Using NCT, we examine the heterogeneous effects of information sessions on the adoption of a new insurance policy in rural China.

KEYWORDS: causal inference, interference, networks.

1 Introduction

According Cox (1958), *inteference* occurs when the treatment assignment of one unit affects the outcome of other units. In the context of policy interventions, interference can arise from many types of interactions, such as social, physical, or virtual connections. The standard Rubin Causal Model for causal inference studies (Rubin, 1986) assumes no interference. However, when interference is likely to occur but is ignored, it introduces bias into the estimates (Forastiere *et al.* , 2021). Furthermore, understanding spillover effects is crucial for measuring the overall impact of an intervention and enhancing the efficiency of treatment assignment mechanisms. As a result, recent research has developed innovative methodologies to address interference.(see, e.g., Sobel, 2006; Rosenbaum, 2007). In parallel to this field of research on interference, researchers have developed machine learning algorithms to assess the heterogeneity of treatment effects with respect to individual characteristics (Athey &

Imbens, 2016). The intuition behind these algorithms is that sub-populations are partitioned by iteratively separating those groups whose estimated average treatment effect deviates the most.

In this study, we present a cutting-edge integration of the aforementioned two topics in the field of causal inference through the introduction of a novel machine learning algorithm, named *Network Causal Tree* (NCT), that investigates the heterogeneity of both treatment and spillover effects with respect to individual, neighborhood and network characteristics, in randomized settings. NCT works in the presence of *clustered network interference*, where agents belong to separate clusters and spillover mechanisms occur only within clusters, according to the links of a cluster-specific network. Conditional effects are estimated by using an extended version of the Horvitz-Thompson estimator (Aronow & Samii, 2017) to allow for clustered network interference. We showcase the NCT methodology to assess the effect of intensive training sessions to promote the uptake of a new weather insurance policy for rice farmers living in villages of rural China (Cai *et al.* , 2015). In this setting, interference is likely to arise, since treated households may share what they have learned with the interfering untreated households.

2 Methods

2.1 Clustered Network Interference

We examine a sample $\mathcal{V}$ consisting of N units distributed across K distinct clusters. Each cluster is represented by the indicator $k \in \mathcal{K} = [1,\dots,K]$, and within each cluster k, units are identified by the index $i = 1,\dots,n_k$.. These units interact within a clustered network structure $\boldsymbol{G}$, where units belonging to the same cluster may share connections, while connections between different clusters are absent. Essentially, $\boldsymbol{G}$ can be seen as a collection of K separate subgraphs, denoted as G_k. The assignment of units to the intervention is random, and we use the binary variable $W_{ik} \in 0,1$ to represent the treatment assigned to unit i in cluster k. The observed outcome for each unit is indicated by Y_{ik}. Additionally, for each unit ik, we have access to a set of individual or network characteristics denoted as $\mathbf{X}_{ik}$.

To define the potential outcomes (Rubin, 1986), we have to rely on some assumptions on the interference mechanism. Here, we assume that *Clustered Network Interference* (CNI) takes place. Under CNI i) the spillover mechanism is confined to units within the same cluster, and ii) an individual's outcome is influenced by the treatment status of units directly connected to her/him

based on the cluster-specific network. Potential outcomes are indexed with respect to the individual intervention W_{ik} and to the neighborhood treatment G_{ik}, which represents a numerical synthesis of the treatment assignment vector of the neighbors: $Y_{ik}(W_{ik}, G_{ik})$. Here, the variable G_{ik} represents a binary network exposure based on a threshold function applied to the number of treated neighbors: G_{ik} equals 1 if the unit ik has at least one treated neighbor, 0 otherwise. We outline four estimands of interest $\tau_{(w,g;w',g')}$, two treatment effects and two spillover effects, where treatment (spillvoer) effects are defined by comparing average potential outcomes under different levels of the individual (neighborhood) treatment status, while keeping as fixed the level of the neighborhood (individual) treatment.

2.2 The Network Causal Tree algorithm

The NCT algorithm is designed to detect and estimate heterogeneous treatment and spillover effects in randomized settings, under CNI. NCT is also able to discover the heterogeneity with respect to more than one estimand simultaneously. NCT takes as inputs the observed data $\{W_{ik}, Y_{ik}, \mathbf{X}_{ik}\}_{ik \in \mathcal{V}}$, the global network $\boldsymbol{G}$, the experimental design and a vector of weights $\omega(w, g; w', g')$ ruling the extent of which estimands contribute to the criterion function, while it returns as output a partition Π of the covariate space, together with point estimates and standard errors of the conditional average causal effects:

The algorithm provides three main steps. In the first step, NCT randomly splits clusters in two sets - the discovery set and the estimation set. In the second step, using the discovery set, NCT sprouts a tree according to the in-sample splitting function and stops when a stopping criterion is met (reached maximum depth, insufficient sample size in the leafs). In the third step, NCT computes the estimated effects and the corresponding standard errors, in all the partitions identified at the first step.

3 Empirical results

Data include 4,586 households living in specific villages of rural China and provide information on the friendship networks connecting households in the same village. Some households are randomly assigned to receive intensive information sessions on a new weather insurance policy, while the remaining households receive *simple* sessions. Households who have at least one treated household in their neighborhood are assumed to receive an indirect exposure to the intervention. The outcome indicates whether the household decides to take

up the insurance policy. The heterogeneity of treatment and spillover effects is evaluated with respect to variables that refer either to characteristics of the household (production area, size) or to characteristics the of the household's owner (sex, age, level of education, risk aversion, perceived probability of disaster, trust in the government).
Results suggest that the most important heterogeneity drivers of the treatment effect are the production area, the risk aversion and the trust in the government. Spillover effects are not statistically significant.

References

ARONOW, PETER M, & SAMII, CYRUS. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, **11**(4), 1912–1947.

ATHEY, SUSAN, & IMBENS, GUIDO. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, **113**(27), 7353–7360.

CAI, JING, DE JANVRY, ALAIN, & SADOULET, ELISABETH. 2015. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, **7**(2), 81–108.

COX, DAVID ROXBEE. 1958. *Planning of experiments.* Wiley: New York.

FORASTIERE, LAURA, AIROLDI, EDOARDO M, & MEALLI, FABRIZIA. 2021. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, **116**(534), 901–918.

HORVITZ, DANIEL G, & THOMPSON, DONOVAN J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**(260), 663–685.

IMBENS, GUIDO W, & RUBIN, DONALD B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

ROSENBAUM, PAUL R. 2007. Interference between units in randomized experiments. *Journal of the American Statistical Association*, **102**(477), 191–200.

RUBIN, DONALD B. 1986. Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, **81**(396), 961–962.

SOBEL, MICHAEL E. 2006. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association*, **101**(476), 1398–1407.

FLEXIBLE MODELLING OF HETEROGENEOUS POPULATIONS OF NETWORKS: A BAYESIAN NONPARAMETRIC APPROACH

Francesco Barile [1], Simonón Lunagómez [2] and Bernardo Nipoti[1]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy
(e-mail: francesco.barile@unimib.it, bernardo.nipoti@unimib.it)

[2] Department of Statistics, Instituto Tecnológico Autónomo de México, México
(e-mail: simon.lunagomez@itam.mx)

ABSTRACT: A popular approach to the problem of clustering multiple network data makes use of distance metrics that measure the similarity among networks based on some of their global or local characteristics. In this context, we propose a novel Bayesian nonparametric approach to model undirected labelled graphs sharing the same set of vertices, which allows us to identify clusters of networks characterized by similar patterns in the connectivity of nodes. Our construction relies on the definition of a location-scale Dirichlet process mixture of centered Erdős–Rényi kernels. An efficient Markov chain Monte Carlo scheme is proposed to carry out posterior inference and provide a convenient clustering of the multiple network data, while the number of clusters in the population is not set a priori but inferred from the data. The performance of our approach is investigated by means of an extensive simulation study and illustrated with the analysis of a dataset on brain networks.

KEYWORDS: Bayesian nonparametrics, centered Erdős–Rényi model, Dirichlet process, mixture model, multiple network data.

NORMALIZED LATENT MEASURE FACTOR MODELS

Mario Beraha [1] and Jim E. Griffin[2]

[1] Department of Economics and Statistics, University of Torino,
(e-mail: mario.beraha@unito.it)

[2] Department of Statistical Sciences, University College London
(e-mail: j.griffin@ucl.ac.uk)

ABSTRACT: Building on dependent normalized random measures, we consider a prior distribution for a collection of discrete random measures where each measure is a linear combination of a set of *latent* measures, interpretable as characteristic traits shared by different distributions, with positive random weights. The model is non-identified and a method for post-processing posterior samples to achieve identified inference is developed. This uses Riemannian optimization to solve a non-trivial optimization problem over a Lie group of matrices. Our approach leads to interesting insights for populations and easily interpretable posterior inference.

KEYWORDS: comparing probability distributions, dependent random measures, latent factor models, normalized random measures, Riemannian optimization

1 Introduction

In this short paper, we review the methodology for modeling and comparing probability distributions discussed in Beraha and Griffin (2022). Modeling a collection of random probability measures is an old problem that has received considerable attention in the Bayesian nonparametric literature, see, e.g. Quintana et al. (2022) for a recent review. We consider here specifically the case where data are naturally divided into groups or subpopulations, and data are partially exchangeable. Let $(y_1, \dots, y_g)$ denote a sample of observations divided into g groups where $y_j = (y_{j1}, \dots, y_{jn_j})$. By de Finetti's theorem, partial exchangeability is tantamount to assuming that there is a vector of random probability measures $(p_1, \dots, p_g) \sim Q$ such that

$$\begin{aligned} y_{j1}, \dots, y_{jn_j} \mid p_j &\overset{\text{iid}}{\sim} p_j, \qquad j = 1, \dots, g \\ p_1, \dots, p_g &\sim Q \end{aligned} \tag{1}$$

and independence holds across groups. In particular, we focus here on mixture models of the kind

$$p_j(y) = \int_\Theta f(y \mid \theta) \widetilde{p}_j(\mathrm{d}\theta)$$

where the $\widetilde{p}_j$'s are almost surely discrete random probability measures.

The construction of a flexible prior Q that can suitably model heterogeneity while borrowing information across different groups has been thoroughly studied in Bayesian nonparametrics. See Quintana et al. (2022) for a recent review of such constructions. Previously proposed approaches consider constructing $\widetilde{p}_1, \dots, \widetilde{p}_g$ in a hierarchical model fashion (Teh et al., 2006; Camerlenghi et al., 2019; Bassetti et al., 2020; Argiento et al., 2019), considering convex combinations of shared and group-specific random measures (Müller et al., 2004), and starting from additive processes (Griffin et al., 2013; Lijoi et al., 2014).

Within this setting, our goal is to propose a flexible model that, in addition to combining heterogeneous sources of data, gives an efficient way of representing the difference in distribution across populations. In particular, we are interested in the situation when the number of groups g is large relative to the sample size in each group n_j. Then, it is likely that the dataset cannot inform the huge number of parameters that are associated with extremely flexible models and we

advocate for a more parsimonious model where substantial sharing of information is encouraged across different groups of data. The setting "large g, small n_j" is somewhat reminiscent of high-dimensional data analysis, where the dimension of each observation is large relative to the sample size. In this case, latent factor models (see, e.g., Arminger and Muthén, 1998) provide a powerful tool. In a latent factor model, it is assumed that each observation $x_i \in \mathbb{R}^p$ is a linear combination of a set of H d-dimensional latent factors weighted by observation-specific scores, plus an isotropic error term. We follow this analogy and propose *normalized latent measure factor models*, a class of prior distributions for a vector of random probability measures $\widetilde{p}_1, \ldots, \widetilde{p}_g$. Informally, our model amounts to considering $\widetilde{p}_j$ as a convex combination of a set of latent random probability measures.

2 The Model

As already mentioned in the Introduction, we assume

$$y_{j1}, \ldots, y_{jn_j} \mid \widetilde{p}_j \overset{\text{iid}}{\sim} p_j := \int_\Theta f(\cdot \mid \theta)\widetilde{p}_j(\mathrm{d}\theta)$$

and that each $\widetilde{p}_j$ is a normalized random measure, that is

$$\widetilde{p}_j(\cdot) = \frac{\widetilde{\mu}_j(\cdot)}{\widetilde{\mu}(\Theta)}, \qquad j = 1, \ldots, g.$$

Then, the model is specified by a choice of the mixture kernel $f(\cdot \mid \cdot)$ and a prior distribution for $(\widetilde{\mu}_1, \ldots, \widetilde{\mu}_g)$. Let $(\mu_1^*, \ldots, \mu_H^*)$ be a completely random vector (i.e., a vector of completely random measures). Let λ_{jh}, $j = 1, \ldots, g$, $h = 1, \ldots, H$ be a double sequence of almost surely positive random variables (specific choices of the distribution of the λ_{jh}'s are discussed later). We assume

$$\widetilde{\mu}_j(\cdot) = \sum_{h=1}^{H} \lambda_{jh}\mu_h^*(\cdot). \tag{2}$$

Note that (2) generalizes the construction in Griffin et al. (2013) and Lijoi et al. (2014).

A suitable model for our applications arises when $\mu_1^*, \ldots, \mu_H^*$ share their support points. In particular, we will assume that $\mu_1^*, \ldots, \mu_H^*$ is a compound random measure (CoRM, Griffin and Leisen, 2017). That is,

$$\mu_h^*(\cdot) = \sum_{k\geq 1} m_{hk}J_k\delta_{\theta_k^*}(\cdot),$$

where m_{hk} are positive random variables such that $m_k = (m_{1k}, \ldots, m_{Hk})$, $k \geq 1$, are independent and identically distributed from a probability measure on $\mathbb{R}_+^H$, and $\eta = \sum_{k\geq 1} J_k\delta_{\theta_k^*}$ is a completely random measure with Lévy intensity $\nu^*(\mathrm{d}z)\alpha(\mathrm{d}x)$. In this case we can write

$$\widetilde{\mu}_j(\cdot) = \sum_{k\geq 1} (\Lambda M)_{jk}J_k\delta_{\theta_k^*}(\cdot), \tag{3}$$

where Λ is the $J \times H$ matrix with entries λ_{jh}, M is a $H \times \infty$ matrix, so that $\Gamma = \Lambda M$ is a $g \times \infty$ matrix with entries γ_{jk}, $j = 1, \ldots, g$, $k \geq 1$. Note that, in analogy to CoRMs, our model includes shared weights J_k for all the measures $\widetilde{\mu}_j$. We find that the additional borrowing of strength obtained through the J_k's is useful in practice since, in our applications, the $\widetilde{\mu}_j$'s are usually similar. Suitable prior distributions for all the parameters will be specified in later sections.

Equations (2) and (3) share analogies with latent factor models, where the observed variable is $X \in \mathbb{R}^p$ and its ℓ-th entry is modeled as $X_\ell \approx \sum_{h=1}^H \omega_{\ell h}Z_h$, for $Z = (Z_1, \ldots, Z_H)$ an H-dimensional random variable. In particular, we could consider $\mu_1^*, \ldots, \mu_H^*$ to be measure-valued factor loadings and the λ_{jh}'s to be factor scores. This yields an interpretation similar to functional factor models (Montagna et al., 2012). On the other hand, we could consider the measure-valued vector $(\widetilde{\mu}_1, \ldots, \widetilde{\mu}_g)$ as a single high-dimensional observation and model it as a linear combination of measure-valued factors with loadings λ_{jh}'s. Both interpretations make sense and lead to interesting analogies. We use the latter and call Λ the loadings matrix and the μ_h^*'s the latent measures.

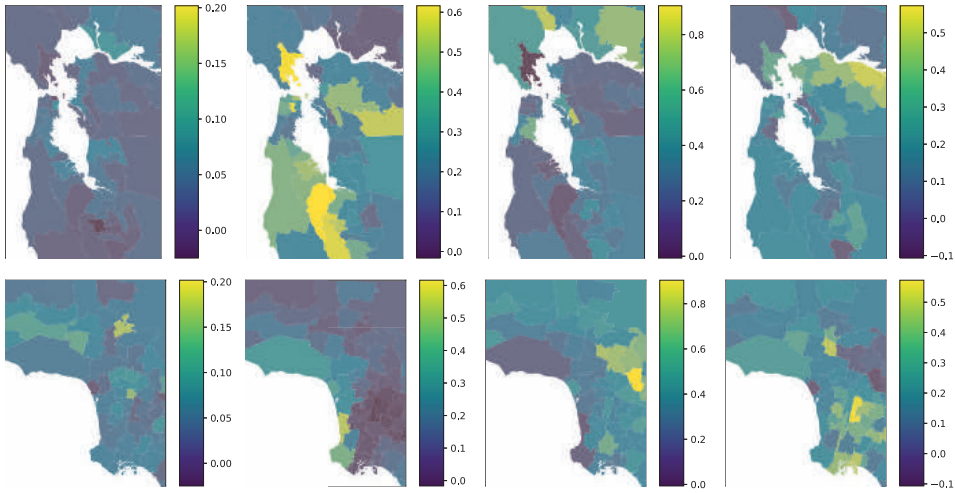


Figure 1: Spatial distribution of the scores in the Californian income dataset. Top row: San Francisco area. Bottom row: Los Angeles ares. Columns represent overall average, high, median, and low income prevalence, respectively.

3 Some details about posterior inference and post-processing

We perform posterior inference by proposing an ad-hoc Markov chain Monte Carlo algorithm, which combines Gibbs updates (when the full conditionals belong to known parametric families) with Hamiltonian Monte Carlo steps (when they do not). For computational convenience, we truncate the support of the CoRM to K atoms, but a slice sampler could be alternatively employed. Software is implemented using the `JAX` Python package.

Our model is not identifiable due to the multiplicative relationship between Λ and $(\mu_1^*, \dots, \mu_h^*)$. This is not surprising, as the same holds for common latent factor models (Geweke and Singleton, 1980), where the likelihood is invariant to the action of orthogonal matrices. The non-identifiability in our model is more severe than the one of common latent factor models. In fact, for any Q s.t. Q^{-1} is well defined, the likelihood is invariant when considering $\Lambda' = \Lambda Q^{-1}$ and $M' = QM$. Nevertheless, the constraints that $\Lambda' \geq 0$ (element-wise) and $M' \geq 0$ greatly reduce the number of matrices Q that can cause non-identifiability. In particular, we do not need to worry about sign ambiguity.

As in Poworoznek et al. (2021), we propose to find an optimal Q via an ad-hoc post processing that aims to maximally separate the latent measure μ_h^*'s, according to some notion of distance between measures. We formalize the post-processing into a constrained optimization problem over the special linear group, that is the set of matrices with determinant equal to one. The special linear group is not a linear space, but can take advantage of its differential structure, and tackle the problem via a Riemannian augmented Lagrangian method that leverages recent advances in Riemannian optimization (França et al., 2021).

4 Analysis of Californian Income Data

We consider the 2021 American Community Survey census data publicly available at `https://www.census.gov/programs-surveys`. Specifically, we consider the `PINCP` variable that represents the personal income of the survey responders and restrict to the citizens of the state of California. For privacy reasons, data are grouped into geographical units, denoted PUMA,

roughly corresponding to 100,000 inhabitants. There are 265 PUMAs in California. We consider $y_{j,i}$ to be the logarithm of the income of the i-th person in the j-th PUMA. The total number of responders is 43,380, with the median number of observations per PUMA being 164.

We assume independent log-Gaussian Markov random field priors for each column of Λ, beta priors for the J's and gamma priors for the m_{hk}'s and fix $H = 4$. Although not shown here, the four latent measures can be interpreted as representing *average incomes* (i.e. the distribution is equal to the whole population distribution), *high incomes*, *median incomes* (i.e., the distribution is concentrated on median values) and *low incomes*. Figure 1 shows the values of Λ for the four latent measures associated with the PUMAs in the San Francisco and Los Angeles areas. In particular, we note that the second factor is highly represented in Palo Alto, home to several tech tycoons, and San Rafael, home to entertainers. Finally, note that the fourth factor (associated with the lowest incomes) has a high weight in some areas in Los Angeles. In particular, the PUMA around the port and the one corresponding to the "south LA" neighborhoods going from University Park to Green Meadows. This is in agreement with the 2008 *Concentrated Poverty in Los Angeles* report, which estimates that the percentage of households in poverty is typically greater than 40% in those areas.

References

Argiento, R., A. Cremaschi, and M. Vannucci (2019). Hierarchical normalized completely random measures to cluster grouped data. *J. Am. Stat. Assoc. 115*(529), 318–333.

Arminger, G. and B. O. Muthén (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika 63*(3), 271–300.

Bassetti, F., R. Casarin, and L. Rossini (2020). Hierarchical species sampling models. *Bayesian Anal. 15*(3), 809–838.

Beraha, M. and J. E. Griffin (2022). Normalized latent measure factor models. *arXiv preprint arXiv:2205.15654*.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *Ann. Stat. 47*(1), 67–92.

França, G., A. Barp, M. Girolami, and M. I. Jordan (2021). Optimization on manifolds: A symplectic approach. *arXiv preprint arXiv:2107.11231*.

Geweke, J. F. and K. J. Singleton (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *J. Am. Stat. Assoc. 75*(369), 133–137.

Griffin, J. E., M. Kolossiatis, and M. F. J. Steel (2013). Comparing distributions by using dependent normalized random-measure mixtures. *J. R. Statist. Soc. B 75*(3), 499–529.

Griffin, J. E. and F. Leisen (2017). Compound random measures and their use in Bayesian nonparametrics. *J. R. Statist. Soc. B 79*(2), 525–545.

Lijoi, A., B. Nipoti, and I. Prünster (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli 20*(3), 1260–1291.

Montagna, S., S. T. Tokdar, B. Neelon, and D. B. Dunson (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics 68*(4), 1064–1073.

Müller, P., F. Quintana, and G. Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. B 66*(3), 735–749.

Poworoznek, E., F. Ferrari, and D. Dunson (2021). Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching. *arXiv preprint arXiv:2107.13783*.

Quintana, F. A., P. Müller, A. Jara, and S. N. MacEachern (2022). The dependent Dirichlet process and related models. *Stat. Sci. 37*(1), 24–41.

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet Processes. *J. Am. Stat. Assoc. 101*(476), 1566–1581.

ESTIMATION ISSUES IN MULTIVARIATE PANEL DATA

Silvia Bianconcini[1] and Silvia Cagnone[1]

[1] Department of Statistical Sciences, University of Bologna
(e-mail: silvia.bianconcini@unibo.it, silvia.cagnone@unibo.it)

ABSTRACT: Latent variable models are a powerful tool in various research fields when the constructs of interest are not directly observable. However, the likelihood-based model estimation can be problematic when dealing with many latent variables and/or random effects since the integrals involved in the likelihood function do not have analytical solutions. In the literature, several approaches have been proposed to overcome this issue. Among them, the pairwise likelihood method and the dimension-wise quadrature have emerged as effective solutions that produce estimators with desirable properties. In this study, we compare a weighted version of the pairwise likelihood method with the dimension-wise quadrature for a latent variable model for binary longitudinal data by means of a simulation study.

KEYWORDS: latent variables, binary data, weighted pairwise likelihood, dimension-wise quadrature

1 Latent variable models for longitudinal binary data

Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$ be vectors of p binary observed variables each of them observed at T different occasions, $z_1, z_2, \ldots, z_T$ latent variables that account for the associations among the p items at each time point. Let $u_1, u_2, \ldots, u_p$ be p random effects that account for the associations of the same item at different time points. The joint density of the observed variables can be defined as

$$f(\mathbf{y}) = \int_{R^q} g(\mathbf{y} \mid \mathbf{z}, \mathbf{u}) h(\mathbf{z}, \mathbf{u}) d\mathbf{z} d\mathbf{u}$$

where $g(\mathbf{y} \mid \mathbf{z}, \mathbf{u})$ is referred to as measurement part of the model and $h(\mathbf{z}, \mathbf{u})$ as structural part of the model. The dimension of the integral is $q = p + T$.
The measurement part of the model is defined as a generalized linear model with the random component given by

$$g(\mathbf{y}|\mathbf{z}, \mathbf{u}) = \prod_{t=1}^{T} \prod_{j=1}^{p} g(y_{tj}|z_t, u_j) = \prod_{t=1}^{T} \prod_{j=1}^{p} \pi_{tj}(z_t, u_j)^{y_{tj}} (1 - \pi_{tj}(z_t, u_j))^{(1-y_{tj})},$$

where the first equality comes from the conditional independence assumption between items and over time. Each $g(y_{tj}|z_t,u_j)$ follows a Bernoulli distribution of parameter $\pi_{tj}(z_t,u_j)$, that is the probability of success of item j at time t. The systematic component defines the linear predictor $\eta_{tj}=\alpha_{0tj}+\alpha_{tj}z_t+u_j$ where α_{0tj}'s are item and time-dependent intercepts and α_{tj}'s are item and time-dependent factor loadings. We consider the logit as the link function between the systematic component and the conditional means of the random component.
As for the structural part of the model, we assume that the latent variables follow an autoregressive process of the first order (Cagnone *et al.* , 2009) as follows

$$z_t=\phi z_{t-1}+\delta_t \tag{1}$$

where ϕ is the autoregressive coefficient, $\delta_t\sim N(0,1)$ and $z_1\sim N(0,\sigma_1^2)$.
Moreover, the joint density $h(\mathbf{z},\mathbf{u})$ is a multivariate normal with zero mean vector and block diagonal covariance matrix Ψ that contains the matrices $\Omega=diag_{j=1,\ldots,p}\{\sigma_{uj}^2\}$ and the autocovariance matrix Γ of the latent variables.

2 Model estimation

Model estimation is usually performed by using a full maximum likelihood method. Given a sample of size n, the log-likelihood is given by

$$L(\theta)=\sum_{i=1}^{n}\log f(\mathbf{y}_i,\theta)=\sum_{i=1}^{n}\log\int_{R^q}g(\mathbf{y}_i\mid\mathbf{z}_i,\mathbf{u}_i)h(\mathbf{z}_i,\mathbf{u}_l)d\mathbf{z}_i d\mathbf{u}_i \tag{2}$$

where θ is the vector of parameters to be estimated. A problem related to the maximization of the log-likelihood is that, in general, the multidimensional integral in (2) is not solvable analytically. Recent solutions proposed in the literature to solve this problem include the pairwise likelihood (PL) approach (Lindsay, 1988) and the dimension-wise (DW) quadrature method (Bianconcini *et al.* , 2017). In this work, we compare DW with a weighted version of PL (Varin & Czado, 2010).
The PL estimator is obtained by maximizing bivariate likelihood products that contain the greatest quantity of model parameter information. In the latent variable model for longitudinal binary data described in Section 2, the bivariate density for a pair of responses is

$$f(y_{ijt},y_{ij't'};\theta)=\int g(y_{ijt}|z_{it},u_{ij})g(y_{ij't'}|z_{it'},u_{ij'})h(z_t,z_{t'},u_j,u_{j'})dz_t dz_{t'}du_j du_{j'}.$$

The dimension of the integrals involved in the expression of $f(y_{ijt}, y_{ij't'}; \theta)$ is four and if $j = j'$ or $t = t'$ it reduces to three. Thus, they can be easily approximated using the Gauss Hermite (GH) quadrature method. As close pairs are more informative, we use a PL likelihood constructed from marginal probabilities of observed pairs less distant than $d \geq 0$ time points. This produces a weighted log PL likelihood of order d defined as

$$pl^{(d)}(\theta; \mathbf{y}) = \sum_i \sum_{j,j',t,t'} \log f(y_{ijt}, y_{ij't'}; \theta) I_{[0,d]}(t' - t). \quad (3)$$

$I_{[0,d]}$ is the indicator function, equal to 1 if $(t' - t) \in [0, d]$ and 0 otherwise.
The DW method is based on the following representation of the marginal density function

$$\begin{aligned} f(\mathbf{y}; \theta) &= |\mathbf{C}_{mo}| \int_{R^q} \frac{\prod_{j=1}^p g(y_j | \mathbf{C}_{mo}\mathbf{b}^* + \mathbf{b}_{mo}) h(\mathbf{C}_{mo}\mathbf{b}^* + \mathbf{b}_{mo})}{\phi(\mathbf{b}^*; \mathbf{0}, \mathbf{I})} \phi(\mathbf{b}^*; \mathbf{0}, \mathbf{I}) d\mathbf{b}^* = \\ &= |\mathbf{C}_{mo}| \int_{R^q} m(\mathbf{b}^*) \phi(\mathbf{b}^*; \mathbf{0}, \mathbf{I}) d\mathbf{b}^* = |\mathbf{C}_{mo}| E_\phi[m(\mathbf{b}^*)] \quad (4) \end{aligned}$$

where $\mathbf{b} = (\mathbf{z}, \mathbf{u})$, $\Sigma_{mo} = \mathbf{C}_{mo}\mathbf{C}'_{mo}$ and $\phi(\cdot)$ is the normal density function.
DW consists in approximating the function $m(\mathbf{b}^*)$ as follows (Bianconcini *et al.* , 2017)

$$\hat{m}(\mathbf{b}^*) = \sum_{l=0}^{s} (-1)^l \begin{pmatrix} q-s+l-1 \\ l \end{pmatrix} m_{s-l}(\mathbf{b}^*) = \sum_{l=0}^{s} A_l m_{s-l}(\mathbf{b}^*) \quad (5)$$

where $m_{s-l}(\mathbf{b}^*) = m(0, \cdots, b^*_{k_1}, 0 \cdots, b^*_{k_{s-l}}, \cdots, 0)$ and $A_l = (-1)^l \begin{pmatrix} q-s+l-1 \\ l \end{pmatrix}$.
Replacing (5) in (4) we obtain the approximate density function

$$f_a(\mathbf{y}; \theta) = f_L + |\mathbf{C}_{mo}| \left[\sum_{l=0}^{s-1} A_l \int_{R^{s-l}} \sum_{k_1 < \ldots < k_{s-l}} m_{s-l}(\mathbf{b}^*) \phi(b^*_{k_1}) \cdots \phi(b^*_{k_{s-l}}) db^*_{k_1} .. db^*_{k_{s-l}} \right] . \quad (6)$$

where f_L denotes the classical Laplace approximation of the integral when $s = 0$. The dimension of the integrals in expression (6) depends on the choice of s. With low values of s, the integrals can be easily approximated using the GH quadrature. In the extreme cases of $s = 0$ and $s = q$, we obtain the classical Laplace and the adaptive GH quadrature method respectively.

3 Simulation study: preliminary results

We perform a simulation study with $p = 3$, $T = 6$, $n = 200$. We consider the UnWeighted (UW) PL function where all the pairs are involved and the PL of order $d = 1, 2, 3$. As for DW, we set $s = 0, 1, 2$. For both methods, the number of quadrature points of GH is fixed at 8. 500 replications are generated for each condition of the study. From the results (Table 1) it is evident that DW with $s = 2$ shows the best performance for almost all the parameter estimates. As for PL, in this design, we don't observe relevant differences for different d and UW. We will further explore the effect of T on the PL method by increasing it.

Table 1. *Estimated bias and rmse (in brackets),* $p = 3$ *and* $T = 6$*,* $n = 200$*.*

True	PL				DW		
	UW	$d=1$	$d=2$	$d=3$	$s=0$	$s=1$	$s=2$
$\alpha_1 = 1.00$							
$\alpha_2 = 0.96$	−0.11(0.55)	0.08(0.39)	0.16(0.58)	0.13(0.49)	−0.21(0.24)	−0.12(0.22)	−0.02(0.18)
$\alpha_3 = 1.07$	−0.02(0.33)	0.05(0.40)	0.14(0.55)	0.09(0.44)	−0.27(0.30)	−0.24(0.29)	−0.09(0.21)
$\phi = 0.50$	0.01(0.11)	−0.02(0.11)	−0.02(0.11)	−0.02(0.10)	0.07(0.10)	0.01(0.09)	−0.02(0.09)
$\sigma^2_1 = 2$	−0.19(1.22)	0.26(1.24)	0.21(1.13)	0.17(1.10)	0.23(0.93)	0.46(1.11)	0.29(0.93)
$\sigma^2_{u1} = 1$	−0.02(0.29)	0.02(0.30)	0.03(0.32)	0.02(0.31)	−0.30(0.42)	−0.26(0.41)	−0.08(0.31)
$\sigma^2_{u2} = 1$	−0.07(0.38)	0.07(0.39)	0.08(0.40)	0.07(0.42)	−0.20(0.34)	−0.14(0.34)	−0.01(0.33)
$\sigma^2_{u3} = 2$	0.01(0.63)	0.09(0.74)	0.12(0.78)	0.09(0.71)	−0.40(0.57)	−0.30(0.51)	−0.09(0.47)

References

BARTHOLOMEW, D, KNOTT, M, & MOUSTAKI, I. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley series in Probability and Statistics.

BIANCONCINI, S, CAGNONE, S, & RIZOPOULOS, D. 2017. Approximate likelihood inference in generalized linear latent variable models based on the dimension-wise quadrature. *Electronic Journal of Statistics.*, **11**, 4404–4423.

CAGNONE, S, MOUSTAKI, I, & VASDEKIS, V. 2009. Latent variable models for multivariate longitudinal ordinal responses. *British Journal of Mathematical and Statistical Psychology.*, **62**, 401–415.

LINDSAY, B. 1988. *Statistical inference from stochastic processes*. Providence: Am. Math. Soc. Chap. Composite likelihood methods, pages 221–239.

VARIN, C, & CZADO, C. 2010. A mixed autoregressive probit model for ordinal longitudinal data. *Biostatistics.*, **11**, 127–138.

THE NEXUS BETWEEN ESG AND INITIAL COIN OFFERINGS: EVIDENCE FROM TEXT ANALYSIS

Alessandro Bitetto [1], Paola Cerchiello[1]

[1] Department of Economics and Management, University of Pavia, (e-mail: alessandro.bitetto@unipv.it, paola.cerchiello@unipv.it)

ABSTRACT: Initial Coin Offerings (aka ICOs) have gained a prominent interest in the FinTech world as an alternative way to fundraising for innovative and cutting-edge business ideas. So far, academics have studied drivers of success without posing specific attention to the products or activities proposed by the ICOs. In this paper, we investigate the possible nexus between ICOs and Environmental, Social and Governance (ESG) indicators, by studying a set of 621 ICOs. Specifically, we extract keywords related to ESG from whitepapers associated with each ICO and build a variable which acts as a signal of attention to sustainability topics. Our research hypothesis concerns the evaluation of whether ICOs oriented towards ESG are more likely to raise expected funds successfully. Preliminary results confirm such a hypothesis.

KEYWORDS: Initial coin offering (ICO), ESG, Sustainability, Blockchain-based crowdfunding, Machine learning

1 Introduction

Nowadays themes like Environment, Social Change, and Governance are becoming more and more important. We could state that, for a company, Environmental, Social, Governance investments and reporting represent one of the ways to keep up with the market. As a matter of fact, companies with stronger ESG propositions tend to have higher growth, higher worker efficiency, lower volatility, cost decrease, and fewer institutional interventions. Furthermore, in recent years, start-ups and the most innovative businesses turn to alternative sources of capital instead of classic channels, such as Initial Coin Offerings (ICOs). An ICO is a new way to fund businesses and initiatives, it is one of the blockchain-based processes that allow the emission of a utility token rather than a security or equity token. The growing popularity of the ICOs is clearly due to several related benefits, such as the high level of offered return on investment, high liquidity, fast financing, cost minimization and high availability, which are increasingly encouraging innovative investors and busi-

nesses to abandon traditional financing methods. However, it is also a young and ever-changing market full of significant risks.

Our paper puts a special emphasis on whether ESG dimensions influence ICOs performances. Thus, we propose to investigate the role played by an ESG flag covariate, appropriately built as described in the following section, in predicting the probability of success when collecting the expected amount of funds during the funding round. To this end, we use textual analysis techniques for creating a proper sustainability flag variable; afterwards, we fit logistic models with several specifications along the ESG dimensions and controls.

2 Data

For the database, we scraped data from the website `ICOmarks.com` and downloaded 7574 Initial Coin Offerings (ICO). The available information includes *ICO details*, such as Website, Whitepaper, Whitelist and MVP, Bounty and Bonus, start/end date, country, *ICO classification*, such as Category (Tech, Finance, Energy, Infrastructure), *ICOmarks rating*, *Token details*, such as Ticker, Platform, Amount available for sale, Technology involved, *Financials*, such as ICO's Token price, (crypto)-currency accepted, Total funds raised, Hard/Soft cap for the funding round, *Team and Advisors size* and *Social Media details*, such as media on which the ICO is advertised or where the investors can discuss. We decided to focus only on ICO and we downloaded all the available whitepapers. Then, we cleaned the downloaded data because of typos and different decimal/thousands separator and we converted all ICO prices reported in fiat or crypto money or in terms of ICO's tokens to U.S. Dollars, using the average FX rate of the ICO's start date.

Our target variable *ICOSUCCESS*, similar to previous literature (for example Meoli & Vismara, 2022), is the binary flag of ICOs success/failure, evaluated as the ratio of raised funds and the hard cap, i.e. the maximum amount of funding expected to be raised. If the ratio is above 0.5, we assign success, failure otherwise.

The whitepapers have been analyzed through advanced textual analysis techniques based on Bidirectional Encoder Representations from Transformers (BERT) architecture (Devlin *et al.*, 2019), to extract information about the characteristics of the proposed business idea. In particular, we use pre-trained models specifically tailored to ESG indicators and financial-related vocabularies (Huang *et al.*, n.d.). The outcome of the model is a probability score for each classification class, e.g. Environmental, Social, Governance, estimating how much pertinent the whitepaper's text is to the topic. Such a step is cru-

cial for building the *ESGFLAG* covariate used in the analysis: we assign the value of 1 if at least one of the three probabilities (E, S or G) is greater than the probability of non-relevance with the topics. Additionally, the length of the whitepaper *LOGWORDS* indicates the logarithm of the number of words in each paper.

3 Methodology and Results

We fit a logit model with OLS estimation, taking into account year-quarter, country and sector fixed effects, as well as clustering the error by country. Given the imbalance in the target variable, we opt for a weighted logit model, to mitigate the impact of the "failure" class. Table 1 reports the results. Results are stable over the two scenarios. In particular, we observe that the success of an ICO is promoted when the project shows an interest in the ESG topic.

Thus, preliminary results appear to confirm the nexus between ICOs' success and ESG. The attention towards sustainability-related topics in general seems to favour fundraising activities. This is in line with a public audience's tendency in evaluating better every activity connected to ethics and responsible behaviour. Such analysis will be further improved and robustified by enlarging the dataset and evaluating more control variables and scenarios.

References

DEVLIN, JACOB, CHANG, MING-WEI, LEE, KENTON, & TOUTANOVA, KRISTINA. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Pages 4171–4186 of: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

HUANG, ALLEN H., WANG, HUI, & YANG, YI. FinBERT: A Large Language Model for Extracting Information from Financial Text*. *Contemporary Accounting Research*, **n/a**(n/a).

MEOLI, MICHELE, & VISMARA, SILVIO. 2022. Machine-learning forecasting of successful ICOs. *Journal of Economics and Business*, **121**, 106071. Scaling up the fintech business: competition, regulation & data management.

Table 1: Predicting ICOs success with logistic model.

Variable	1	2
ESGFLAG	0.129**	0.159**
	(0.0727)	(0.0723)
DURATION	-0.461**	-0.442*
	(0.192)	(0.230)
RATING	0.126***	0.105**
	(0.0403)	(0.0473)
TEAMSIZE	0.0461***	0.0362***
	(0.00756)	(0.00787)
ADVISORSIZE	0.0441***	0.0356*
	(0.0144)	(0.0186)
WHITELIST	0.149***	0.162
	(0.0553)	(0.138)
BOUNTY	-0.238**	-0.195
	(0.109)	(0.121)
BONUS	-0.0916	-0.0281
	(0.148)	(0.116)
PRESALE	-0.273***	-0.0950
	(0.0894)	(0.0801)
LOGWORDS	0.177	0.194
	(0.162)	(0.156)
Observations	871	869
Pseudo R^2	0.043	0.067
Quarter-Year effects	No	Yes
Country effects	No	Yes
Category effects	No	Yes
Clustered Std. Err.	Country	Country

Notes: The table reports coefficients and their standard error (in parentheses). The outcome variable is the binary flag of ICO's success/failure and all variables are defined in Section 2. Data span over the period 2014-2019. Estimation method is OLS with standard errors clustered by ICO's country. The bottom part of the table reports which fixed effects are used in each model specification. The *, ** and *** symbols denote the p-values at 10^{th}, 5^{th} and 1^{st} significance level, respectively.

A CLUSTERING MODEL FOR THREE-WAY ASYMMETRIC PROXIMITY DATA

Laura Bocci[1] and Donatella Vicari[2]

[1] Department of Social and Economic Sciences, Sapienza University of Rome
(e-mail: laura.bocci@uniroma1.it)

[2] Department of Statistical Sciences, Sapienza University of Rome
(e-mail: donatella.vicari@uniroma1.it)

ABSTRACT: This paper presents a model for clustering three-way asymmetric proximity data which represent flows or exchanges between objects observed at different occasions. In order to account for systematic differences between occasions, the asymmetric data are assumed to subsume two clustering structures common to all occasions: the first defines a standard partitioning of all objects which fits the average amount of the exchanges; the second one, which fits the imbalances, defines an "incomplete" partitioning of the objects, where some of them are allowed to remain unassigned. The model is fitted in a least-squares framework and an efficient Alternating Least Squares algorithm is given.

KEYWORDS: Asymmetric dissimilarities, three-way data, partition.

1 Introduction

In many real-world applications, information is measured or observed in the form of several pairwise asymmetric proximity (similarity or dissimilarity) matrices related to the same N objects observed at H occasions (i.e., times, subjects, scenarios). Such kind of data represent three-way two-mode asymmetric proximity data which, without loss of generality, can derive from mobility flows, brand-switching, import/export exchanges or other type of transactions or trade. For example, international student mobility between countries over several years gives rise to a three-way asymmetric proximity array where in each year-matrix the rows correspond to the origins and the columns to the destinations of the mobile students.

In the analysis of asymmetric data, the asymmetry has often been ignored by symmetrizing the proximities (i.e., averaging the two different values for any pair of objects). Nonetheless, if one hypothesizes that the asymmetries are meaningful and systematic across occasions, special models are needed (see Saito & Yadohisa, 2005, Bove et al., 2021, for extended reviews).

Clustering three-way asymmetric proximity data is a complex task since each proximity data matrix generally subsumes a (more or less) different clustering of objects due to the heterogeneity of the occasions and the asymmetry may incorporate some important information about clustering. Chaturvedi and Carroll (1994) generalized the INDCLUS model to asymmetric proximities by identifying two

different sets of (overlapping) clusters of the N objects (for rows and columns, respectively) common to all occasions, while the three-way heterogeneity is accounted for by occasion-specific weights for the clusters.

In order to extract as much information as possible from the three-way asymmetries taking into account the heterogeneity of the occasions, we present here a generalization to asymmetric three-way data of the model proposed by Vicari (2020) for clustering an asymmetric dissimilarity matrix. To account for the asymmetric structure of the data, the model relies on the decomposition of the asymmetric matrices into the sum of their symmetric and skew-symmetric components which are jointly modelled. The asymmetric dissimilarities are assumed to subsume two clustering structures common to all occasions: the first defines a standard partitioning of all objects which fits the symmetric component of the exchanges; the second one, which fits the imbalances, defines an *incomplete* partitioning of the objects, where some of them are allowed to remain unassigned. Objects within the same clusters in both clustering structures share the same behaviours in terms of exchanges directed to the other clusters and identify "origin" and "destination" clusters. Note that the partition to fit the imbalances is allowed to be incomplete to better identify the directions of the exchanges, so those objects not assigned to any cluster (incomplete partitioning) qualify objects with "small" asymmetries. Moreover, to account for the heterogeneity of the occasions, occasion-specific sets of weights are estimated which account for both the average amounts and the directions of the exchanges.

In Section 2, the model is formalized in a least-squares framework and an appropriate Alternating Least Squares algorithm is given.

2 The model

Let $\mathbf{X}$ be a three-way two-mode asymmetric array of size $(N \times N \times H)$, where the H frontal slices consist of square asymmetric matrices $\mathbf{X}_h$ $(h = 1, \dots, H)$ of pairwise dissimilarities between N objects observed in H occasions and where the generic element x_{ijh} is generally different from x_{jih}.

The model proposed here aims at clustering the N objects by decomposing the observed asymmetries into symmetric and skew-symmetric effects, modelled as functions of two nested partitions of the objects which subsume clustering structures common to all occasions. Specifically, all occasions are supposed to share the same partition of the N objects into J disjoint clusters $\{C_1, \dots, C_J\}$ uniquely identified by an $(N \times J)$ binary membership matrix $\mathbf{U} = [u_{ij}]$ $(u_{ij} = \{0,1\}$ for $i = 1, \dots, N$ and $j = 1, \dots, J$ and $\sum_{j=1}^{J} u_{ij} = 1$ for $i = 1, \dots, N)$, where $u_{ij} = 1$ if object i belongs to cluster C_j and $u_{ij} = 0$ otherwise. Since any object is required to be assigned to some cluster C_j, such a partition is referred to as *complete partition*. Furthermore, a second partition of the N objects into J clusters $\{G_1, \dots, G_J\}$ common to all occasions is identified by an $(N \times J)$ binary membership matrix $\mathbf{V} = [v_{ij}]$ $(v = \{0,1\}$ for $i = 1, \dots, N$ and $j = 1, \dots, J)$, where any object i is allowed either not to be assigned to any cluster or to

belong to cluster G_j if it belongs to cluster C_j in the complete partition, i.e. $v_{ij} \leq u_{ij}$ ($i = 1, \dots, N$ and $j = 1, \dots, J$). The partition identified by $\mathbf{V}$ is referred to as an *incomplete partition* because a number N_0 ($N_0 \leq N$) out of N objects are allowed to remain unassigned to any cluster. Moreover, the *complete* and the *incomplete* partitions are *common* to all occasions and linked each other, the latter being constrained to be nested into the former one ($G_j \subseteq C_j$ for $j = 1, \dots, J$).

Hereafter, $\mathbf{I}_N$ denotes the identity matrix of size N, $\mathbf{1}_{AB}$ and $\mathbf{1}_A$ denote the matrix of size ($A \times B$) of all ones and the column vector with A ones, respectively.

Let us recall that any square matrix $\mathbf{X}_h$ ($h = 1, \dots, H$) can be uniquely decomposed into a sum of a symmetric matrix $\mathbf{S}_h$ and a skew-symmetric matrix $\mathbf{K}_h$, which are orthogonal to each other (i.e., $trace(\mathbf{S}_h\mathbf{K}_h) = 0$), as

$$\mathbf{X}_h = \mathbf{S}_h + \mathbf{K}_h = \tfrac{1}{2}(\mathbf{X}_h + \mathbf{X}_h') + \tfrac{1}{2}(\mathbf{X}_h - \mathbf{X}_h') , \qquad (h = 1, \dots, H). \tag{1}$$

Both components in $\mathbf{X}_h$ can be modeled by defining two clustering structures depending on matrices $\mathbf{U}$ and $\mathbf{V}$, respectively, as introduced in Vicari (2020) for a two-way asymmetric dissimilarity matrix.

Specifically, the symmetric component $\mathbf{S}_h$ and the skew-symmetric component $\mathbf{K}_h$ for occasion h are modeled by the two clustering structures introduced in Vicari (2014, 2018) and depend on the *common complete* membership matrix $\mathbf{U}$ and the *common incomplete* membership matrix $\mathbf{V}$, respectively, as

$$\mathbf{S}_h = \mathbf{U}\mathbf{C}_h\big(\mathbf{1}_{NJ} - \mathbf{U}\big)' + \big(\mathbf{1}_{NJ} - \mathbf{U}\big)\mathbf{C}_h\mathbf{U}' + \mathbf{E}_{hS} , \qquad (h = 1, \dots, H), \tag{2}$$

$$\mathbf{K}_h = \mathbf{V}\mathbf{D}_h\big(\mathbf{1}_{NJ} - \mathbf{V}\big)' - \big(\mathbf{1}_{NJ} - \mathbf{V}\big)\mathbf{D}_h\mathbf{V}' + \mathbf{E}_{hK} , \qquad (h = 1, \dots, H), \tag{3}$$

where $\mathbf{C}_h$ and $\mathbf{D}_h$ are ($J \times J$) occasion-specific diagonal weight matrices associated with the clusters of the complete and incomplete partition, respectively, and the error terms $\mathbf{E}_{hS}$ and $\mathbf{E}_{hK}$ represent the parts of $\mathbf{S}_h$ and $\mathbf{K}_h$ not accounted for by the model, respectively. For identifiability reasons, any matrix $\mathbf{V}\mathbf{D}_h$ is constrained to sum to zero: $\mathbf{1}_N'(\mathbf{V}\mathbf{D}_h)\mathbf{1}_J = 0$ ($h = 1, \dots, H$).

Models (2) and (3) can be combined in (1) to specify the model accounting for the asymmetric dissimilarities between clusters

$$\begin{aligned}\mathbf{X}_h = &\Big[\mathbf{U}\mathbf{C}_h\big(\mathbf{1}_{NJ} - \mathbf{U}\big)' + \big(\mathbf{1}_{NJ} - \mathbf{U}\big)\mathbf{C}_h\mathbf{U}'\Big] + \Big[\mathbf{V}\mathbf{D}_h\big(\mathbf{1}_{NJ} - \mathbf{V}\big)' - \big(\mathbf{1}_{NJ} - \mathbf{V}\big)\mathbf{D}_h\mathbf{V}'\Big] \\ &+ b_h(\mathbf{1}_{NN} - \mathbf{I}_N) + \mathbf{E}_h , \quad (h = 1, \dots, H),\end{aligned} \tag{4}$$

where b_h is the additive constant term and the general error term $\mathbf{E}_h$ represents the part of $\mathbf{X}_h$ not accounted for by the model.

It is worth noting that all occasions are assumed here to share the same clustering structure but with different patterns of weights which account for the heterogeneity of the occasions. In fact, the occasion-specific diagonal entries of $\mathbf{C}_h$ and $\mathbf{D}_h$ provide quantifications of the exchanges between clusters in terms of amounts and directions and allow to measure at what extent the exchanges vary across occasions.

In model (4), the *complete* and the *incomplete* membership matrices **U** and **V**, the weight matrices $\mathbf{C}_h$ and $\mathbf{D}_h$ $(h = 1, \dots, H)$ and the constants b_h $(h = 1, \dots, H)$ can be estimated by solving the following least-squares fitting problem:

$$\min F(\mathbf{U}, \mathbf{V}, \mathbf{C}_h, \mathbf{D}_h, b_h) = \sum_{h=1}^{H} \left\| \mathbf{X}_h - \left[\mathbf{U}\mathbf{C}_h \left(\mathbf{1}_{NJ} - \mathbf{U} \right)' + \left(\mathbf{1}_{NJ} - \mathbf{U} \right) \mathbf{C}_h \mathbf{U}' \right] - \left[\mathbf{V}\mathbf{D}_h \left(\mathbf{1}_{NJ} - \mathbf{V} \right)' - \left(\mathbf{1}_{NJ} - \mathbf{V} \right) \mathbf{D}_h \mathbf{V}' \right] - b_h (\mathbf{1}_{NN} - \mathbf{I}_N) \right\|^2 \quad (5)$$

subject to

$$u_{ij} = \{0,1\} \quad (i = 1, \dots, N; j = 1, \dots, J) \text{ and } \sum_{j=1}^{J} p_{ij} = 1 \quad (i = 1, \dots, N), \quad (5a)$$

$$v_{ij} = \{0,1\} \quad (i = 1, \dots, N; j = 1, \dots, J) \text{ and } v_{ij} \leq u_{ij} \quad (i = 1, \dots, N), \quad (5b)$$

$$\mathbf{1}_N' (\mathbf{V}\mathbf{D}_h) \mathbf{1}_J = 0 \quad (h = 1, \dots, H). \quad (5c)$$

Problem (5) can be solved by using an Alternating Least-Squares algorithm which alternates the estimation of a set of parameters when all the others are kept fixed. The algorithm proposed here estimates in turn: a) the *complete* and *incomplete* membership matrices **U** and **V** by sequentially solving joint assignment problems for the different rows of **U** and **V**: given any row *i*, setting $u_{ij} = 1$ implies that either $v_{ij} = 0$ or $v_{ij} = u_{ij}$ for $j = 1, \dots, J$; b) the occasion-specific weight matrices $\mathbf{C}_h$ and $\mathbf{D}_h$ $(h = 1, \dots, H)$ by solving regression problems; c) the additive constant b_h $(h = 1, \dots, H)$ by successive residualizations of the three-way data matrix. The main steps are alternated and iterated until convergence and the best solution over different random starts is retained to prevent from local minima.

Results from applications to real data will be presented to show the performance of the algorithm and the capability of the model to identify common clusters of objects which best account for their pairwise dissimilarities.

References

BOVE, G., OKADA, A., & VICARI, D. 2021. *Methods for the Analysis of Asymmetric Proximity Data*. Springer Nature Singapore.

CHATURVEDI, A., & CARROLL, J.D. 1994. An alternating combinatorial optimization approach to fitting the INDCLUS and Generalized INDCLUS models. *Journal of Classification*, **11**, 155–170.

SAITO, T., & YADOHISA, H. 2005. *Data analysis of asymmetric structures. Advanced approaches in computational statistics*. New York: Marcel Dekker.

VICARI, D. 2014. Classification of asymmetric proximity data. *Journal of Classification*, **31**(3), 386–420.

VICARI, D. 2018. CLUXEXT: CLUstering model for Skew-symmetric data including EXTernal information. *Adv Data Anal Classif*, **12**, 43–64.

VICARI, D. 2020. Modeling Asymmetric Exchanges Between Clusters In: T. Imaizumi, A. Nakayama and S. Yokoyama (Eds), Advanced Studies in Behaviormetrics and Data Science. Springer Nature Singapore, 297-313.

CLUSTER ANALYSIS FOR NETWORKS USING A FUZZY APPROACH

Ilaria Bombelli [1 2], Ichcha Manipur [3] and Maria Brigida Ferraro [2]

[1] Italian National Institute of Statistics

[2] Department of Statistical Sciences, Sapienza University of Rome, (e-mail: ilaria.bombelli@uniroma1.it, mariabrigida.ferraro@uniroma1.it)

[3] Institute for High-Performance Computing and Networking, National Research Council, (e-mail: ichcha.manipur@icar.cnr.it)

ABSTRACT: As the network representation is widely used to describe problems in an increasing number of disciplines, novel methodologies are needed to handle such complexity. In particular, cluster analysis is an interesting and challenging task in the network framework. In this work, we focus on how to represent networks for fuzzy clustering and how to apply standard fuzzy algorithms for clustering multiple networks on synthetic data.

KEYWORDS: Ensembles of Networks, Fuzzy Clustering, Networks Clustering, Whole-graph Embedding.

1 Introduction

Networks represent a powerful model for problems in different scientific and technological fields, such as neuroscience, molecular biology, biomedicine, sociology, social network analysis and political science. The increasing number of network applications leads research on clustering analysis develop rapidly.

In a network framework, a well-known approach to the clustering problem is the detection of clusters of nodes (or *communities*). A new approach to the clustering problem is to consider a single network as the unit of interest and to detect clusters of networks.

What is proposed here is to apply fuzzy cluster analysis techniques to identify clusters of networks by choosing an adequate representation. The novelty here lies in the usage of a fuzzy approach: indeed, related works use only a hard approach to clustering, meaning that each network can belong to one cluster only. However, networks may have characteristics in common to more than one cluster, and therefore in such situations, a more flexible approach is

more adequate. In this sense, the fuzzy approach guarantees major flexibility than the hard approach, by allowing each network to belong to all clusters according to different membership degrees.

2 Network representation

To cluster networks, we need to find an adequate representation. In the early proposals on this topic, networks have been represented using some topological characteristics, but very different networks might be represented by the same values of the chosen features, making the data analysis difficult. Moreover, the well-known *adjacency matrix* representation does not account for differences in specific parts of the network and therefore ignores its topological characteristics.

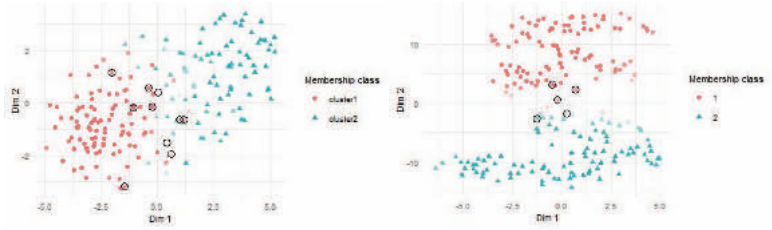
To overcome these limits, we study two types of network representations: a probabilistic representation of graphs (either Node Distance Distribution or Transition Matrices, see Granata *et al.*, 2020 for details) and a whole-graph embedding representation (Joint Embeddings by Wang *et al.*, 2021). By using the probabilistic representation, the Jensen-Shannon (JS) Divergence is then used to compute pairwise distances between networks and finally to obtain a distance matrix; instead, the embedding techniques provide a vector space representation of the networks to identify a space that is optimal with respect to some characteristics; the output is therefore a units by variables matrix, where units are networks and variables are networks' features.

3 Algorithms for fuzzy clustering

Once we have chosen how to adequately represent the networks, it is possible to apply fuzzy clustering algorithms. We use Non-Euclidean Fuzzy Relational Clustering, introduced by Davé & Sen, 2002, when the networks are represented by a matrix of distances; instead, we applied the Fuzzy k-Means (Bezdek, 1981), when they are in form of a feature matrix.

4 Simulation

We empirically analyze our proposal on synthetic dataset. In detail, the simulated networks are generated using the Multiple Random Eigen Graphs (MREG) model, defined in Wang *et al.*, 2021. Particularly, an MREG dataset with 200 graphs having 100 nodes each was generated using the MREG model. The



(a) NEFRC applied on $\mathcal{M}^N$ (b) [FkM applied on JE$_{d=2}$

Figure 1: t-SNE representation of clustering results of NEFRC, FkM (MREG networks). Misclassified units are circled in black. The intensity of the colors is given by the membership degree of each network to the corresponding assigned cluster.

graphs belong to 2 classes, with 100 graphs in each class. The clustering task consists of grouping networks with a similar distribution of edges.

Here, for the sake of brevity, we show two applications of fuzzy clustering algorithms (NEFRC and FkM) to two networks' representations: $\mathcal{M}^N$, i.e. the distance matrix obtained by applying JS divergence on Node Distance Distribution representation of networks; JE, i.e. the feature matrix resulting from Joint Embedding technique. Table 1 shows the algorithm's performance using

Table 1: Main results of the application of NEFRC the Distance Matrix $\mathcal{M}^N$ and FkM to Feature Matrix JE (MREG networks)

	NEFRC $\mathcal{M}^N$		FkM JE	
	ARI	AMI	ARI	AMI
Median	**0.81**	0.72	**0.9**	0.83
IQR	0	0	0	0
SD	0.01	0.02	0.01	0.01

the clustering validity indices. In detail, high ARI and AMI indices values show that most of the networks are correctly assigned to their original clusters.

The graphical representation allows us to explore the results more in-depth in Figure 1. Figure 1 shows that the two clusters are well separated; misclassified networks are highlighted by the circled points. The fuzzy membership

degrees allows us to deeply study the misclassified units. By applying NEFRC to distance matrices, we notice that, on average, approximately 40% of misclassified networks are in the middle of the two cluster prototypes, having membership degrees close to 0.5 and being represented by blurry colors in Figure 1 (a). Regarding the application of F*k*M to JE, we notice that 20% of misclassified units are represented by very blurry colors in Figure 1 (b) and are softly assigned to both the clusters. Therefore, membership degrees allow us to consider the uncertainty of an assignment of a unit to a cluster and then possibly add information on clustering interpretation: this represents one of the main advantages of a fuzzy approach.

5 Final Remarks

This study explores clustering analysis when the statistical units are networks. To this extent, we focus on different methodologies that can provide a suitable representation of the sample of the networks for subsequent data analysis. We applied fuzzy clustering algorithms on such representations, using standard metrics to evaluate their performance on synthetic datasets. Our analysis provides valuable hints for cluster analysis in a network framework.

References

BEZDEK, JAMES C. 1981. *Pattern recognition with fuzzy objective function algorithm*. Plenum Press, New York.

DAVÉ, RAJESH N, & SEN, SUMIT. 2002. Robust fuzzy clustering of relational data. *IEEE Transactions on Fuzzy Systems*, **10**(6), 713–727.

GRANATA, ILARIA, GUARRACINO, MARIO ROSARIO, MADDALENA, LUCIA, & MANIPUR, ICHCHA. 2020. Network Distances for Weighted Digraphs. *Pages 389–408 of: International Conference on Mathematical Optimization Theory and Operations Research*. Springer.

WANG, SHANGSI, ARROYO, JESÚS, VOGELSTEIN, JOSHUA T, & PRIEBE, CAREY E. 2021. Joint embedding of graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 1324–1336.

VISUALIZING ANOMALIES IN CIRCULAR DATA

Davide Buttarazzi [1] and Giovanni C Porzio[2]

[1] Sample Survey and Data Analysis Lab, University of Cassino and Southern Lazio, (e-mail: davidebuttarazzi@outlook.com)

[2] Department of Economics and Law, University of Cassino and Southern Lazio, (e-mail: porzio@unicas.it)

ABSTRACT: Anomaly detection has a long history in Statistics, with one of the most effective approaches being robustness. First, a model describing the majority of the data is assumed. Second, its parameters are robustly estimated. Then, the distance of all the points from such a model is evaluated. Eventually, extremely far (i.e., unlikely) observations are flagged as outliers. Visually, this procedure is well described by the well-worn Tukey's box-and-whisker plot. Thanks to its robustness properties, it is probably the graphical tool mostly adopted to highlight anomalies in univariate data sets.

This work aims at investigating if the same strategy can be exploited in circular data analysis, i.e., for data lying on the boundary of the unit circle. For this kind of data, a specific boxplot has been designed. However, its first formulation did not focus on anomaly detection. It was rather conceived as an exploratory tool to display the main features of a circular data set. Reliyng on a non-robust estimate of the data dispersion, it will be simply misleading if used to visualize anomalies. A robust circular boxplot is then introduced. It will be able to correctly identify circular outliers under a specific parametric model.

KEYWORDS: Circular boxplot, directional statistics, von Mises distribution.

MODEL-BASED CLUSTERING OF RIGHT-CENSORED LIFETIME DATA WITH FRAILTIES AND RANDOM COVARIATES

Andrea Cappozzo [1], Chiara Masci[1], Francesca Ieva[12] and Anna Maria Paganoni[1]

[1] MOX, Department of Mathematics, Politecnico di Milano, (e-mail: andrea.cappozzo@polimi.it, chiara.masci@polimi.it, francesca.ieva@polimi.it, anna.paganoni@polimi.it)

[2] Health Data Science Center, Human Technopole, Milano

ABSTRACT: We introduce a new parametric approach for clustering multilevel survival data that accounts for the heterogeneity at baseline and random distributions of the explanatory variables. The proposed method aims to identify clusters of patients with different survival patterns and uncover the impact of the known hierarchy on survival within each cluster. The objective function is maximized using a stochastic EM algorithm tailored to right-censored lifetime data. The proposed methodology can be seen as a generalization of multilevel cluster-weighted modeling for time-to-event outcomes. Promising results are showcased on synthetic data.

KEYWORDS: model-based clustering, survival data, frailty models, EM algorithm, cluster-weighted models

1 Introduction and model formulation

The paper proposes an approach for clustering survival data in which the procedure takes advantage of cluster-wise different random covariates. Additionally, the heterogeneity at the baseline due to a known hierarchy present in the sample (e.g., patients within hospitals) is accounted for in the time-to-event outcome by means of a parametric frailty model. In details, in our proposal a statistical unit is identified by the triplet $(y_{ij}, \delta_{ij}, \mathbf{x}_{ij})$ where:

- y_{ij} is the minimum between the survival time t_{ij} and censoring time c_{ij} for subject i in hospital j,
- $\delta_{ij} = I(t_{ij} \leq c_{ij})$ is the event indicator,
- $\mathbf{x}_{ij} = (\mathbf{u}_{ij}, \mathbf{v}_{ij})$ denotes the vector of covariates with $\mathbf{u}_{ij}$ and $\mathbf{v}_{ij}$ respectively indicating the subset of continuous and categorical predictors for the ij-th unit.

The entire sample is therefore composed by $N = \sum_{j=1}^{J} n_j$ observations among the J hospitals. We further assume that the observed data can be partitioned into G latent clusters independently of the known J groups. The resulting log-likelihood for the considered model reads as follows:

$$\begin{aligned}
&\ell\left(\{\tau_g, \boldsymbol{\beta}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \lambda_g, \theta_g\}_{g=1}^{G}\right) = \\
&\sum_{g=1}^{G}\left\{\sum_{j=1}^{J}\sum_{i\in R_{jg}}\log\tau_g + \sum_{j=1}^{J}\left[\sum_{i\in R_{jg}}\delta_{ij}\left(\log h_0(y_{ij}) + \mathbf{x}_{ij}'\boldsymbol{\beta}_g\right) + \right.\right. \\
&+\log\left[(-1)^{d_{jg}}\mathcal{L}^{(d_{jg})}\left(\sum_{i\in R_{jg}} H_0(y_{ij})\exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta}_g\right);\theta_g\right)\right]\Bigg] + \\
&\left.\left.+\sum_{j=1}^{J}\sum_{i\in R_{jg}}\log\phi(\mathbf{u}_{ij};\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) + \sum_{j=1}^{J}\sum_{i\in R_{jg}}\log\psi(\mathbf{v}_{ij};\lambda_g)\right\}.
\end{aligned} \tag{1}$$

The quantities $h_0(\cdot)$ and $H_0(\cdot)$ denote the baseline hazard and cumulative hazard functions, and $\mathcal{L}^{(q)}$ is the q-th derivative of the Laplace transform of the frailty distribution. Depending on the chosen baseline and/or frailty term, the formulation in (1) encompasses a general family of parametric mixture frailty models. With $\phi(\cdot)$ and $\psi(\cdot)$ we respectively identify the densities of a multivariate Gaussian and independent multinomial distributions (one for each categorical variable), needed to incorporate the cluster-wise different contribution of the covariates. Further, d_{jg} is the total number of observed events assigned to cluster g belonging to hospital j, and R_{jg} contains the indexes of the observations in cluster g and hospital j. The remaining terms are model parameters that need to be estimated from the sample. In details, τ_g represents the mixing proportion for cluster g, with $\tau_g \geq 0$ for all g and $\sum_{g=1}^{G}\tau_g = 1$. The vector of regression coefficients is denoted with $\boldsymbol{\beta}_g$, while θ_g is the heterogeneity parameter for $g = 1,\ldots,G$. Lastly, parameters for the conditionally independent multinomial distributions within each cluster are compactly identified with λ_g, and $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$ denote the mean vector and the covariance matrix of the continuous covariates.

Maximization of (1) is carried out by means of a stochastic EM algorithm tailored to right-censored lifetime data (Bordes & Chauveau, 2016). The proposed methodology extends the work in Berta & Vinciotti, 2019 by considering a time-to-event outcome, leveraging on recent advances in the efficient estimation of parametric frailty models (Munda *et al.*, 2012). To this extent, the goal of the proposed procedure is twofold. On the one hand, we aim to identify G clusters of patients with different survival patterns. On the other hand, within

each cluster we wish to uncover the different impact the known hierarchy has on the survival. Promising results are reported for synthetic data, as described in the next section.

Table 1. *BIC and ARI for several choices of baseline, number of clusters and frailty densities in the Multilevel time-to-event cluster-weighted model.*

G	Baseline	Frailty	BIC	ARI
2	Exponential	None	-3795.60	0.84
2	Exponential	Gamma	-3756.93	0.84
2	Weibull	None	-3279.04	0.93
2	Weibull	Gamma	**-2586.46**	**0.95**
3	Exponential	None	-3767.83	0.73
3	Exponential	Gamma	-3597.16	0.67
3	Weibull	None	-3193.45	0.82
3	Weibull	Gamma	-2810.98	0.80

2 Results on simulated data

We assess the performance of the proposed procedure on a two components ($G = 2$) synthetic population simulated with the `genfrail` function of the `frailtySurv R` package (Monaco *et al.*, 2018). The data generating process includes $n_j = 40$ for all $j = 1,\dots,J$ and $J = 10$, resulting in a sample whose size is equal to $N = 400$. The baseline hazard has a parametric Weibull distribution, while a Gamma density is used to simulate the frailty term in the equally sized clusters. The survival time depends on two continuous covariates, whose distribution is multivariate Gaussian with cluster-wise different mean vectors and equal covariance matrix. Model results are reported in Table 1 in which several specifications for the baseline and frailty densities are considered. The comparison includes also an option with fixed effects only, denoted with Frailty equals to None in the table. We observe that the model with Weibull baseline, Gamma frailty and true number of clusters outperforms the competing methods in both goodness of fit and clustering performance, showcasing higher values in both Bayesian Information Criterion and Adjusted Rand Index metrics.

3 Conclusion

The proposed approach provides a flexible method for analyzing right-censored lifetime data with random covariates and frailties, making it a valuable tool for applications in personalized medicine and hospitals evaluation. Some analyses are currently being accomplished on this regard and they will be the object of future work.

References

BERTA, PAOLO, & VINCIOTTI, VERONICA. 2019. Multilevel logistic cluster-weighted model for outcome evaluation in health care. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **12**(5), 434–443.

BORDES, LAURENT, & CHAUVEAU, DIDIER. 2016. Stochastic EM algorithms for parametric and semiparametric mixture models for right-censored lifetime data. *Computational Statistics*, **31**(4), 1513–1538.

MONACO, JOHN V., GORFINE, MALKA, & HSU, LI. 2018. General Semiparametric Shared Frailty Model: Estimation and Simulation with frailtySurv. *Journal of Statistical Software*, **86**(4).

MUNDA, MARCO, ROTOLO, FEDERICO, & LEGRAND, CATHERINE. 2012. parfm : Parametric Frailty Models in R. *Journal of Statistical Software*, **51**(11).

CLUSTERING IMBALANCED FUNCTIONAL DATA

Michelle Carey [1], Catherine Higgins[1]

[1] School of Mathematics and Statistics, University College Dublin, (e-mail: michelle.carey@ucd.ie, catherine.higgins1@ucdconnect.ie)

ABSTRACT: Class imbalance is a common problem in functional clustering where some clusters have significantly more curves than other clusters. In such cases, most clustering algorithms tend to prioritize the majority class, resulting in sub-optimal cluster assignments. We propose a functional iterative hierarchical clustering approach to address the issue of class imbalance in functional data clustering. The performance of the proposed approach is compared with existing approaches. The proposed approach yields more accurate cluster assignments and a more precise approximation of the average trajectory of the curves within each cluster.

KEYWORDS: functional data, unsupervised clustering, class imbalance

1 Introduction

Unsupervised functional clustering techniques classify a sample of curves into homogeneous groups of curves, without prior knowledge of the true underlying clustering structure. The two common approaches for clustering functional data are: to obtain an approximation of the functional data in a finite-dimensional space and then use traditional clustering tools to cluster the resulting vectors (Chen *et al.*, 2012 and Wang & Xu, 2017) or to perform functional model-based clustering (Bouveyron & Jacques, 2011, Bouveyron *et al.*, 2015, and Centofanti *et al.*, 2023). See Jacques & Preda, 2014 for detailed reviews of functional clustering methods.

The problem of class imbalance occurs when the number of curves in one cluster significantly exceeds the number of curves in another cluster, posing a difficult challenge for most functional clustering algorithms. Minor clusters are often classified incorrectly into major clusters, which results in inaccurate cluster assignments and a poor approximation of the average trajectory of the curves within each cluster.

By extending Carey *et al.*, 2016 iterative hierarchical clustering method to a functional data context, we provide an approach for clustering imbalanced functional data.

2 Functional Iterative hierarchical clustering

The observations of the behavior of the curves at discrete points are subject to measurement error, that is $y_{i,j} = x_i(t_{i,j}) + \varepsilon_{i,j}$, where $t_{i,j}$ denotes the finite set of times from which one samples the i^{th} curve and the errors $\varepsilon_{i,j}$ are assumed to be independently distributed with mean 0 and a constant variance σ^2. Given the observed values $y_{i,j}$ for $i = 1,\ldots,N$ and $j = 1,\ldots,M_i$. The functional IHC algorithm performs the following steps:

1. **Reconstruct the functional form from the discrete observations**: Approximate the curves via a basis function expansion, that is, $\hat{x}_i(t) = \sum_{k=1}^{K} c_k \phi_k(t)$, and estimate the coefficients of the basis function expansion $\{c_k : k = 1,\ldots,K\}$ using the standard penalized least squares smoothing approach of Ramsay & Silverman, 2005.
2. **Cluster the first derivative of the curves:** The estimated first derivative of the curves evaluated at the points $\boldsymbol{t} = [t_{1,1},\ldots,t_{N,M}]$ are then given by the $N \times M$ matrix $\mathbf{A} = \sum_{k=1}^{K} \hat{c}_k D\phi_k(\boldsymbol{t})$, where $M = max(M_i)$ for $i = 1,\ldots,N$. Let α_{min} and α_{max} be the minimum and maximum of the Spearman rank correlation between all the possible pairs of the rows of $\mathbf{A}$. Define $[\alpha_{min},\ldots,\alpha_{max}]$, as a grid of Q equally spaced values from α_{min} to α_{max}. Cluster the rows of $\mathbf{A}$ using the iterative hierarchical clustering method proposed in Carey *et al.*, 2016. Select the optimal α_{opt} so that the value of the Davies-Bouldin index is minimized.

3 Simulations

The simulated sample curves X_i are realizations of a Gaussian process with the Matérn covariance function $C(s,t) = 0.2 \times \exp(-0.3\|s-t\|)$, over the domain $I = [0,15]$. To obtain six simulated groups of curves we define six different mean functions: $\sin(2\pi t)$, $\cos(2\pi t)$, $\sin(4\pi t + \pi/2)$, $\sin(4\pi t - \pi/2)$, $\sin(3\pi t + \pi/3)$ and $\sin(6\pi t - \pi)$. The six clusters are large, medium, and small in size, that is $N = 500, 500, 200, 15, 10, 3$. The sample data are given by, $Y_{i,j} = X_i(t_{i,j}) + \varepsilon_{i,j}$ for $i = 1,\ldots,N$ and $j = 1,\ldots,M_i$, where $\varepsilon_{i,j}$ is a normally distributed random variables with mean 0 and standard deviation σ_ε. We assume that $t_{i,j}$ are obtained from an equally spaced discretization of the domain and that this is the same for all curves.

The functional IHC is compared with the following eight state-of-the-art functional clustering methods: funFEM (Bouveyron *et al.*, 2015), funHDDC (Bouveyron & Jacques, 2011); SaS-Funclust (Centofanti *et al.*, 2023), functional EMCluster (Chen *et al.*, 2012), functional kCFC (Chiou & Li, 2007),

FADPclust1 and FADPclust2 (Wang & Xu, 2017). Table (1) presents the accuracy of the clustering methods measured by the average Adjusted Rand Index (μ_{ARI}) and the average Davies-Bouldin index (μ_{DB}). The adjusted rand index

Table 1. *The average Adjusted Rand Index (μ_{ARI}); and Davies-Bouldin index μ_{DB} for all eight functional clustering methods*

Method	μ_{ARI}	μ_{DB}	μ_{ARI}	μ_{DB}	μ_{ARI}	μ_{DB}	μ_{ARI}	μ_{DB}
	M=15				M=200			
	$\sigma_\varepsilon = 0.05$		$\sigma_\varepsilon = 0.15$		$\sigma_\varepsilon = 0.05$		$\sigma_\varepsilon = 0.15$	
funIHC	1.00	0.84	0.99	0.86	0.99	0.82	0.99	0.89
funFEM	0.55	6.18	0.55	0.99	0.53	0.86	0.52	0.95
funHDDC	0.53	1.08	0.47	1.11	0.30	3.29	0.28	3.41
SaS-Funclust	0.53	1.08	0.53	0.97	0.35	3.67	0.15	3.11
Functional EMCluster	0.94	1.08	0.96	1.01	0.68	1.76	0.69	1.90
Functional kCFC	0.97	1.08	0.90	1.52	0.54	1.64	0.64	2.10
FADPclust1	0.68	1.66	0.71	1.14	0.69	1.10	0.71	1.15
FADPclust2	0.68	1.02	0.68	1.16	0.74	0.94	0.71	1.06

ranges from 0 to 1 and measures the similarity between the clustering assignment and the true group structure. Clustering assignments are more accurate when the value is larger. The Davies-Bouldin index is based on the ratio of within-cluster distances to between-cluster distances. Clusters that are farther apart and less dispersed will result in a lower index. The funIHC obtains the highest adjusted rand index and the lowest Davies-Bouldin index. FunIHC is the only approach to correctly identify the number of curves in each cluster and the true average temporal pattern. FunFEM, FunHDDC, Sasfunclust, Functional EMCluster, and Functional kCFC provide a good approximation of the average temporal patterns for the larger clusters but provide a poor approximation for ($N < 200$). FADPclust1 and FADPclust2 miss-classifies the small and medium clusters into the larger clusters resulting in poor approximations of the average temporal pattern for all clusters.

4 Conclusion

A functional iterative hierarchical clustering approach is proposed that can effectively address the issue of class imbalance in functional data clustering. The proposed approach is shown to outperform existing approaches in terms of the

accuracy in the cluster assignments and the approximations of the average temporal pattern of the cluster members.

References

BOUVEYRON, CHARLES, & JACQUES, JULIEN. 2011. Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, **5**(4), 281–300.

BOUVEYRON, CHARLES, CÔME, ETIENNE, & JACQUES, JULIEN. 2015. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, **9**(4), 1726–1760.

CAREY, MICHELLE, WU, SHUANG, GAN, GUOJUN, & WU, HULIN. 2016. Correlation-based iterative clustering methods for time course data: the identification of temporal gene response modules for influenza infection in humans. *Infectious Disease Modelling*, **1**(1), 28–39.

CENTOFANTI, FABIO, LEPORE, ANTONIO, & PALUMBO, BIAGIO. 2023. Sparse and smooth functional data clustering. *Statistical Papers*, 1–31.

CHEN, WC, MAITRA, R, & MELNYKOV, V. 2012. EMCluster: EM algorithm for model-based clustering of finite mixture Gaussian distribution. *R Package, URL http://cran. r-project. org/package= EMCluster*.

CHIOU, JENG-MIN, & LI, PAI-LING. 2007. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(4), 679–699.

DAVIES, DAVID L, & BOULDIN, DONALD W. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 224–227.

JACQUES, JULIEN, & PREDA, CRISTIAN. 2013. Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, **112**, 164–171.

JACQUES, JULIEN, & PREDA, CRISTIAN. 2014. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, **8**(3), 231–255.

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

RODRIGUEZ, ALEX, & LAIO, ALESSANDRO. 2014. Clustering by fast search and find of density peaks. *science*, **344**(6191), 1492–1496.

WANG, XIAO-FENG, & XU, YIFAN. 2017. Fast clustering using adaptive density peak detection. *Statistical methods in medical research*, **26**(6), 2800–2811.

PARTIAL MEMBERSHIP MODELS FOR HIGH-DIMENSIONAL SPECTROSCOPY DATA

Alessandro Casa[1], Thomas Brendan Murphy[2] and Michael Fop[2]

[1] Faculty of Economics and Management, Free University of Bozen-Bolzano,
(e-mail: alessandro.casa@unibz.it)

[2] School of Mathematics and Statistics, University College Dublin,
(e-mail: michael.fop@ucd.ie, brendan.murphy@ucd.ie)

ABSTRACT: The demand for detecting food adulteration has recently grown, due to its economic and health implications. Infrared spectroscopy provides an efficient method of collecting data for use in food authenticity analyses. Statistical methods are routinely employed to analyze spectroscopy data in order to effectively detect adulterants in different food items and ensure food authenticity. This work presents a novel partial membership model for mid-infrared spectral data. Our approach not only detects the level of adulteration but also provides information on the spectral regions most affected by the adulterant. These insights can be used in combination with subject-matter expertise to characterize the chemical impact of the adulteration.

KEYWORDS: partial membership, latent variable models, food authentication, shrinkage prior

1 Introduction

Expensive foods are often subject to fraud and food adulteration, with some of the original components being removed or replaced by cheaper alternatives, to lower their prices or to increase their bulk. On one hand, this can represent an economic problem for food producers. On the other hand, it might also lead to health issues for the consumers. Therefore, food authenticity studies, which aim to determine if a sample has been adulterated or not, are increasingly important. In this work, we examine Fourier transform mid-infrared (MIR) spectroscopy data, which have been previously used effectively to tackle the aforementioned problem. To the task, we propose a novel partial membership model for spectroscopy data. The model introduces a more sophisticated authentication tool, capable of not only identifying the presence of potential adulterants in food, but also of determining the percentage of contamination. The proposed model also enables the identification of which wavelengths are more

impacted by the adulterant, constituting a starting point for further chemical analysis. In Section 2, we introduce this new model and outline the adopted estimation approach. Section 3 reports an application to spectrometry data of Irish honey samples.

2 Model definition and estimation

Individual-level mixture models generalize standard model-based clustering by encompassing situations where units can belong to multiple groups simultaneously, with varying degrees of membership. This idea has been developed in two directions, namely mixed membership and partial membership models (PMM), with the latter being the focus of this work; see Airoldi *et al.* , 2014 for a discussion. Let $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be the observed data. When $\mathbf{y_i} \in \mathbb{R}^p$, a multivariate Gaussian distribution is often assumed for the K component densities. Therefore, according to PMM, $\mathbf{y}_i$ is conditionally distributed as

$$(\mathbf{y}_i|g_i,\Theta) \sim N_p\left(\left(\sum_{k=1}^{K} g_{ik}\Sigma_k^{-1}\right)^{-1}\left(\sum_{k=1}^{K} g_{ik}\Sigma_k^{-1}\mu_k\right), \left(\sum_{k=1}^{K} g_{ik}\Sigma_k^{-1}\right)^{-1}\right) \quad (1)$$

where $\Theta = \{\mu_k, \Sigma_k\}_{k=1}^{K}$ denotes mixture component means and covariance matrices, while $\mathbf{g}_i = (g_{i1}, \ldots, g_{iK})$ is the partial membership vector for the i-th observation with $g_{ik} \in [0,1]$, for $k = 1, \ldots, K$, and $\sum_k g_{ik} = 1$. For food authentication purposes, we consider $K = 2$, with the two components corresponding to the pure food item and the adulterant, respectively. Moreover, we assume that the adulterant has an additive and wavelength-specific effect. As such, we have that

$$\begin{aligned} \mu_1 &= \mu^{pure} = (\mu_1^{pure}, \ldots, \mu_p^{pure}) \\ \mu_2 &= \mu^{ad} = (\mu_1^{pure} + \delta_1, \ldots, \mu_p^{pure} + \delta_p) \end{aligned}$$

where δ_j, for $j = 1, \ldots, p$, represents the mean-shift induced by the adulterant on the j-th wavelength. Pairing this specification with shrinkage or penalization strategies for δ_j's can lead to the detection of the spectral regions most influenced by the adulterant. Assuming $\Sigma_1 = \Sigma_2 = \Sigma$, model (1) reads as

$$(\mathbf{y}_i|g_i,\Theta) \sim N_p\left(\mu^{pure} + g_{i2}\delta, \Sigma\right) \quad (2)$$

where g_{i2} is the percentage of adulterant in the i-th sample and $\delta = (\delta_1, \ldots, \delta_p)$. When dealing with spectroscopy data, the high number of variables can jeopardize the practical usefulness of model (2). For this reason, simplifying assumptions would consider a factor analytic or a diagonal structure for Σ.

Two alternative ways to estimate model (2) are explored. The first, heuristic, estimation procedure can be used to obtain a naive and fast first model evaluation. More specifically, it aims to maximize iteratively the following quantity

$$SS_f = \sum_{i=1}^{n} (\mathbf{y}_i - \mu^{pure} - g_{i2}\delta)^2$$

with respect to μ^{pure}, δ and g_{i2}. As it is, this procedure does not account for the correlation structure among wavelengths and does not induce shrinkage on δ. Interestingly, it can be used to provide initial values for a Bayesian estimation procedure adopting a Dirichlet prior distribution for the membership vectors $\mathbf{g}_i, i = 1,\ldots,n$ while, for the δ_j's, $j = 1,\ldots,p$, an horseshoe prior (Carvalho *et al.* , 2010) is employed, thus imposing sparsity on the mean-shifts. Lastly, standard conjugate priors are assumed for μ^{pure}, for the diagonal entries of Σ, or for Λ and Ψ, if a factor analytic structure is considered. The model is estimated via MCMC algorithm, by means of the `NIMBLE` software. Note that some degree of supervision can be introduced in the estimation. In particular, for some spectra, g_{i2} can be assumed known, since it is often possible to augment the observed data with experimental data with a controlled amount of adulteration. Unreported analyses showed the beneficial impact of small amount of supervision.

3 Application to honey data

Our proposal is tested on MIR spectral data comprising samples from pure honey and samples contaminated with different adulterants (Kelly *et al.* , 2006). The data have $n = 410$ spectra, $n_H = 290$ from pure honey and $n_B = 120$ adulterated with beet sucrose in different percentages (10%, 20% and 30%). Prior to running the analysis, a data aggregation step has been performed to reduce the overall computational cost. Consequently, the original $p = 285$ wavelengths have been reduced to $p^* = 57$ aggregated ones. Some supervision has been imposed, assuming prior knowledge of the adulteration level for 40 spectra. A diagonal structure for Σ has been considered and the hyperparameters of the horseshoe prior have been selected following suggestions from Piironen & Vehtari, 2017. An excerpt of the results is reported in Figure 1. Here, it is shown how the proposed method is able to precisely estimate the spectrum for the most adulterated samples, with the 95% credible interval always containing the true observed values. A closer inspection for the estimated δ_j's shows

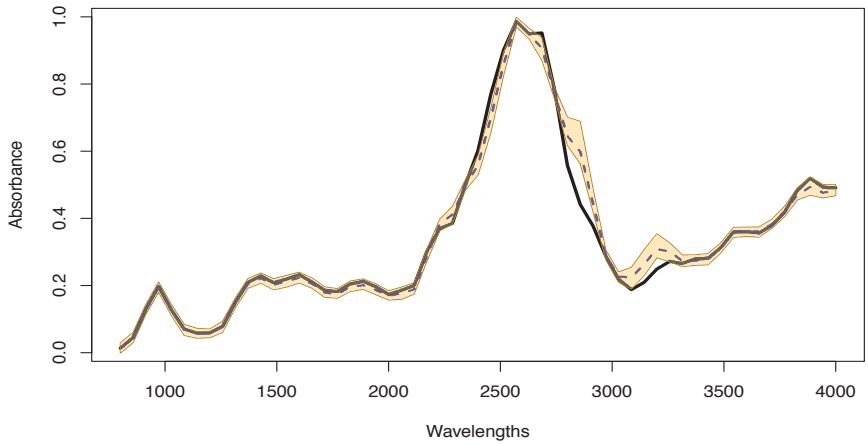


Figure 1. *In black the estimated μ^{pure}. Dashed blue line depicts the observed average spectrum for the most adulterated samples, while the gold shaded area represents the estimated 95% credible interval for the same quantity.*

how beet sucrose seems to have a non negligible impact only on 10 aggregated wavelengths in the region from 2377.46 cm^{-1} to 3166.19 cm^{-1}. These results, if paired with subject-matter knowledge, can shed light on the chemical mechanism underlying the adulteration process.

References

AIROLDI, E.M., BLEI, D., EROSHEVA, E.A., & FIENBERG, S.E. 2014. *Handbook of mixed membership models and their applications*. CRC press.

CARVALHO, C.M., POLSON, N.G., & SCOTT, J.G. 2010. The horseshoe estimator for sparse signals. *Biometrika*, **97**(2), 465–480.

KELLY, J.D., PETISCO, C., & DOWNEY, G. 2006. Application of Fourier transform midinfrared spectroscopy to the discrimination between Irish artisanal honey and such honey adulterated with various sugar syrups. *Journal of Agricultural and Food Chemistry*, **54**(17), 6166–6171.

PIIRONEN, J., & VEHTARI, A. 2017. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *Pages 905–913 of: Artificial Intelligence and Statistics*. PMLR.

SPARSE CLUSTERING FOR FUNCTIONAL DATA

Fabio Centofanti [1], Antonio Lepore[1] Biagio Palumbo[1]

[1] Department of Industrial Engineering, University of Naples Federico II, Piazzale Tecchio 80, Napoli (e-mail: fabio.centofanti@unina.it, antonio.lepore@unina.it, biagio.palumbo@unina.it)

ABSTRACT: The sparse and smooth functional clustering (SaS-Funclust) method is presented for sparse clustering of functional data, i.e., to split a sample of curves into homogeneous groups while jointly detecting the most informative portions of the domain. SaS-Funclust relies on a functional adaptive pairwise fusion penalty and a roughness penalty. The former allows identifying the noninformative portion of the domain, whereas the latter improves the interpretability by imposing some degree of smoothing to the cluster means. The practical advantages of the SaS-Funclust method are illustrated through a real-data example in the analysis of the Berkeley growth study dataset. The SaS-Funclust method is implemented in the R package sasfunclust, available on CRAN.

KEYWORDS: functional data analysis, functional clustering, model-based clustering, penalized likelihood, sparse clustering

1 Introduction

In the last years, due to recent developments in technology and computational power, the majority of the data gathered by practitioners and scientists in many fields contain information about curves or surfaces that are apt to be modelled as functional data, i.e., continuous random functions defined on a compact domain (Ramsay & Silverman, 2005). Cluster analysis is a key tool in functional data analysis, just as it is in the multivariate (non-functional) statistical literature, with applications in several fields. Functional clustering main goal is to classify a sample of functional data into homogenous groups of curves with no explicit information on the actual underlying clustering structure (Capezza *et al.*, 2021). However, as stated in many multivariate data applications, some characteristics could be entirely unhelpful in revealing the desired clustering structure. In this setting, to achieve more accurate group identification, it is important to determine the features in which respect true clusters differ the most, or equivalently noninformative features that may conceal the true clustering structure. More in general, the methods capable of selecting informative

features and eliminating noninformative ones are referred to as *sparse* (Witten & Tibshirani, 2010; Pan & Shen, 2007; Guo *et al.*, 2010). Recently, the notion of sparseness has been translated into a functional data clustering framework. Sparse functional clustering methods have appeared in literature with the aim of clustering functional data while jointly detecting the most informative portion of the domain and improving both the accuracy and the interpretability of the analysis (Floriello & Vitelli, 2017; Vitelli, 2023). In this article, we present the model-based procedure for the sparse clustering of functional data, which has been recently proposed by Centofanti *et al.*, 2023, and referred to as sparse and smooth functional clustering (SaS-Funclust). The SaS-Funclust procedure is implemented in the R package sasfunclust and is openly available on CRAN.

2 The SaS-Funclust method

Suppose that N vectors $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$, of size n_i, $i = 1, \ldots, N$, of observed values of a function f_i over the time points $t_{i1}, \ldots, t_{in_i}$ are spread among $g = 1, \ldots, G$ unknown clusters and the probability of each observation to belong to the gth cluster is π_g. The function f_i is assumed a Gaussian random process with mean μ_g, covariance ω_g, and values in $L^2(\mathcal{T})$, which denotes the separable Hilbert space of square-integrable functions defined on the compact domain $\mathcal{T}$. We assume that, conditionally on the cluster membership, $\boldsymbol{Y}_i$ is modelled as

$$\boldsymbol{Y}_i = \boldsymbol{f}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, N,$$

where $\boldsymbol{f}_i = (f_i(t_{i1}), \ldots, f_i(t_{in_i}))^T$ contains the values of the function f_i at t_{i1}, $\ldots, t_{in_i}$ and $\boldsymbol{\varepsilon}_i$ is a vector of random errors zero mean and constant variance σ_e^2. In this setting, the SaS-Funclust solution (Centofanti *et al.*, 2023) is obtained by maximizing the following penalized log-likelihood

$$L_p(\boldsymbol{\Theta} | \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N) = \sum_{g=1}^{G} \pi_g \psi\left(\boldsymbol{Y}_i; \boldsymbol{\mu}_{gi}, \boldsymbol{\Omega}_{gi} + \boldsymbol{I}\sigma_e^2\right) - \mathcal{P}(\mu_1, \ldots, \mu_G), \quad (1)$$

where $\boldsymbol{\Theta} = \{\pi_g, \mu_g, \omega_g, \sigma_e^2\}_{g=1,\ldots,G}$ is the parameter set of interest, $\boldsymbol{\mu}_{gi} = (\mu_g(t_{i1}), \ldots, \mu_g(t_{in_i}))^T$, $\boldsymbol{\Omega}_{gi} = \{\omega_g(t_{ki}, t_{li})\}_{k,l=1\ldots,n_i}$, $\psi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate Gaussian density distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and $\mathcal{P}(\cdot)$ is a penalty

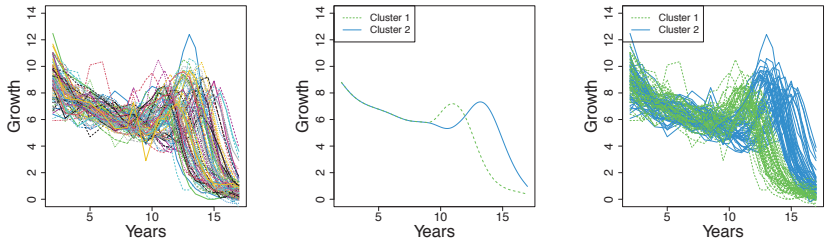
function defined as

$$\mathcal{P}(\mu_1,\ldots,\mu_G) = \lambda_L \sum_{1\le g\le g'\le G} \int_{\mathcal{T}} \tau_{g,g'}(t)\,|\mu_g(t)-\mu_{g'}(t)\,|dt + \lambda_s \sum_{g=1}^{G} \int_{\mathcal{T}} \left(\mu_g^{(s)}(t)\right)^2 dt, \tag{2}$$

where $\lambda_L, \lambda_s \geq 0$ are tuning parameters, $\tau_{g,g'}$ are prespecified weight functions, and $\mu_g^{(s)}(\cdot)$ denotes the sth-order derivative of μ_g. The first element of the right-hand side of Equation (2) is the functional adaptive pairwise fusion penalty (FAPFP). It allows the pair of cluster means to be equal over a specific portion of the domain that is considered noninformative for separating the cluster means. Thus, the SaS-Funclust method is able to detect, for each cluster pair, the portion of the domain that is noninformative for the cluster analysis, i.e., the portion of the domain where the corresponding cluster means are not fused. The last term in Equation (2) is a roughness penalty, applied on the cluster means to further improve the interpretability of the analysis by constraining, with a magnitude quantified by λ_s, the cluster means to own a certain degree of smoothness, measured by the derivative order s. A specific expectation-conditional maximization (ECM) algorithm is used to maximize the objective function in Equation (1), after some structure is imposed on f_i. Then a cross-validation procedure is proposed to select the appropriate model parameters. Further details are in Centofanti *et al.*, 2023.

3 A Real-data Example: Berkeley Growth Study Data

In this section, the SaS-Funclust method is applied to the growth dataset from the Berkeley growth study. In this study, 31 height measurements of 54 girls and 39 boys are available from ages 1 through 18. The aim of the analysis is to cluster growth velocities from age 2 to 17. Figure 1 shows (a) the interpolating growth velocity curves for all the individuals, (b) the estimated cluster means, and (c) the clustered growth curves for the SaS-Funclust method. The estimated cluster means are fused over the first portion of the domain, whereas they are separated over the remaining portion. This implies that on average, the two identified clusters do not differ over the first portion of the domain, which can be, thus, regarded as noninformative. The separation between the two groups arises over the remaining informative portion of the domain, where two sharp peaks of growth velocity arise, instead. The latter peaks are referred to as pubertal spurts in the medical literature and in this regard, the obtained results highlight two primary timing/duration groupings. The male pubertal spurt occurs later and lasts longer than the female one. The estimated cluster



(a) (b) (c)

Figure 1: (a) Growth velocities, (b) estimated cluster curve means, and (c) curve clusters for the SaS-Funclust in the Berkeley growth study dataset.

means from some competing methods do not allow for a similar straightforward interpretation.

References

CAPEZZA, C., CENTOFANTI, F., LEPORE, A., & PALUMBO, B. 2021. Functional clustering methods for resistance spot welding process data in the automotive industry. *Applied Stochastic Models in Business and Industry*, **37**(5), 908–925.

CENTOFANTI, F., LEPORE, A., & PALUMBO, B. 2023. Sparse and smooth functional data clustering. *Statistical Papers*, 1–31.

FLORIELLO, D., & VITELLI, V. 2017. Sparse clustering of functional data. *Journal of Multivariate Analysis*, **154**, 1–18.

GUO, J., LEVINA, E., MICHAILIDIS, G., & ZHU, J. 2010. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, **66**(3), 793–804.

PAN, W:, & SHEN, X. 2007. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, **8**(May), 1145–1164.

RAMSAY, J. O., & SILVERMAN, B. W. 2005. *Functional data analysis*. Wiley Online Library.

VITELLI, V. 2023. A novel framework for joint sparse clustering and alignment of functional data. *Journal of Nonparametric Statistics*, 1–30.

WITTEN, D. M., & TIBSHIRANI, R. 2010. A framework for feature selection in clustering. *Journal of the American Statistical Association*, **105**(490), 713–726.

INTERPRETABLE AND ACCURATE SCALING IN LARGE-SCALE ASSESSMENT: A VARIABLE SELECTION APPROACH TO LATENT REGRESSION

Yunxiao Chen[1], Motonori Oka[1] and Matthias von Davier[2]

[1] Department of Statistics, London School of Economics and Political Science, (e-mail: y.chen186@lse.ac.uk, m.oka1@lse.ac.uk)

[2] Lynch School of Education and Human Development, Boston College, (e-mail: matthias.vondavier@bc.edu)

ABSTRACT: This paper concerns the construction of scaling models for large-scale assessments in education. A scaling model, which makes use of information from both responses to cognitive assessment and background survey items, produces plausible values for individual students. There are two major challenges when building a scaling model – (1) a large number of background variables and (2) many missing values in the background survey data. To tackle these challenges, we propose a variable selection approach to latent regression modelling. The proposed approach handles missing data by iterative imputation and controls variable selection error by a data-splitting procedure.

KEYWORDS: Latent regression, large-scale assessment, variable selection, missing data, imputation

1 Problem Setup

Consider data collected from N students, where data from different students are independent. For each student i, the data can be divided into two parts – (1) responses to cognitive items and (2) non-cognitive predictors. We use a random vector $\mathbf{Y}_i$ to denote student i's cognitive responses. Due to the matrix sampling design for cognitive items in international large-scale assessments (ILSAs), the length of $\mathbf{Y}_i$ can vary across students. More precisely, we use $\mathcal{B}_i$ to denote the set of cognitive items that student i is assigned. Then $\mathbf{Y}_i = \{Y_{ij} : j \in \mathcal{B}_i\}$. For simplicity, we assume all the items are binary, i.e., $Y_{ij} \in \{0,1\}$. In addition, consider p predictors collected via non-cognitive survey questions. Let $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ip})^\top$ denote the complete predictor vector for student i. Often, there are missing values in $\mathbf{Z}_i$. Let $\mathcal{A}_i$ denote the set of observed predictors for student i, and let $\mathbf{Z}_i^{\text{obs}} = \{Z_{ij} : j \in \mathcal{A}_i\}$ and $\mathbf{Z}_i^{\text{mis}} = \{Z_{ij} : j \notin \mathcal{A}_i\}$. The predictors

are of mixed types. Here, binary, categorical (ordinal/nominal), and continuous predictors are considered. Note that an ordinal variable will be treated as a nominal one here for simplicity. In what follows, we introduce a latent regression model, which can be decomposed into (1) a measurement model, (2) a structural model and (3) a predictor model.

Measurement model. We introduce a latent variable θ_i as the latent construct, which is measured by the cognitive items. The measurement model is an IRT model that specifies the conditional distribution of $\mathbf{Y}_i$ given θ_i. More specifically, this model assumes local independence, an assumption that is commonly adopted in IRT models (Embretson & Reise, 2000). That is, Y_{ij}, $j \in \mathcal{B}_i$, are conditionally independent given θ_i. For a dichotomous item j, the conditional distribution of Y_{ij} given θ_i is assumed to follow a two-parameter logistic model (2PL, Birnbaum, 1968). That is,

$$\mathbb{P}(Y_{ij} = 1|\theta_i) = \frac{\exp(a_j\theta_i + b_j)}{1+\exp(a_j\theta_i + b_j)}, \tag{1}$$

where a_j and b_j are two item-specific parameters. We assume that all the item parameters are known – they are fixed to be the pre-calibrated values.

Structural model. The structural model regresses the latent construct θ_i onto the complete-data predictors Z_{i1}, ..., Z_{ip}. A linear regression model is assumed for θ_i given Z_{i1}, ..., Z_{ip}. More specifically, for each variable j, we introduce a transformation $g_j(Z_j)$. When Z_j is an ordinal variable with categories $\{0,...,K_j\}$, the transformation function g_j creates K_j dummy variables, i.e., $g_j(Z_j) = (\mathbb{I}(\{Z_j = 1\}),...,\mathbb{I}(\{Z_j = K_j\}))^\top$. For continuous and binary variables, g_j is an identity link, i.e., $g_j(Z_j) = Z_j$. We assume $\theta_i|\mathbf{Z}_i \sim N(\beta_0 + \beta_1^\top g_1(Z_{i1}) + \cdots + \beta_p^\top g_p(Z_{ip}), \sigma^2)$, where β_0 is the intercept, β_1, ..., β_p are the slope parameters, and σ^2 is the residual variance. Note that β_j is a scalar when predictor j is continuous or binary and is a vector when the predictor is ordinal. Here, β_0, β_1, ..., β_p, and σ are unknown and will be estimated from the model. The main goal of our analysis is to find predictors for which $\|\beta_j\| \neq 0$.

Predictor model. To handle missing values in Z_{ij}s, we impose a joint model for the predictors. Although different models may be imposed here, we assume a Second-Order Exponential (SOE) model, under which missing data imputation and parameter estimation can be carried out in a computationally efficient

way. More precisely, we let $(\theta_i, \mathbf{Z}_i)$ be i.i.d., following an SOE model. Under this model, the conditional distribution of θ_i given $\mathbf{Z}_i$ is the linear regression model in the above structural model. The conditional distribution of Z_{ij} given $(\theta_i, Z_{i,-j})$ takes the following forms:

- A linear regression model (with normal residual), if variable j is continuous;
- A logistic regression model, if variable j is binary;
- A multinomial logistic regression model if variable j is categorical.

These conditional distributions will be used later for missing data imputation and parameter estimation. We remark that except for the parameters of the structural model, the rest of the parameters in the SOE can be viewed as nuisance parameters, as they are not of interest to us. The predictor model and these nuisance parameters are introduced to handle the missing values in the predictors.

2 Estimation and Variable Selection

The model introduced in the previous section implies a joint distribution of complete data, which further implies the distribution of observed data under the Missing At Random (MAR) assumption. We estimate the model and conduct variable selection based on this implied distribution for observed data. More specifically, we estimate the model parameters using an iterative imputation algorithm. According to Liu *et al.* , 2014, the estimate produced by this algorithm is asymptotically equivalent to a full Bayesian posterior-mean estimator based on the observed data likelihood. Thanks to the connection between the frequentist and Bayesian estimation provided by the Bernstein-von Mises Theorem (Van der Vaart, 2000, Chapter 10), this estimate also enjoys the desired frequentist properties, such as consistency and asymptotic normality.

Furthermore, we adopt a data-splitting method for controlled variable selection. More specifically, we combine the data-splitting method (Dai *et al.* , 2022) and the iterative imputation method to select the non-null predictors in the structural model of latent regression. Thanks to the properties of the iterative imputation method, this method has the theoretical guarantee to control the asymptotically false discovery rate for variable selection. The theoretical properties of the proposed method are confirmed by simulation results.

3 Discussions

Traditionally, a PCA-based latent regression model is used for the scaling of large-scale assessment data, in which the missing values are handled by a missing indicator approach, and the high dimensionality of the background variables and their missing indicators is reduced by Principal Component Analysis (PCA). However, this approach has three drawbacks: (1) the missing indicator approach does not perform well under certain data missingness patterns, (2) PCA may introduce spurious dependence between the achievement traits and background variables, and (3) the resulting model lacks interpretability due to the involvement of hard-to-interpret principal component scores. The proposed method does not suffer from these issues. It handles missing values more properly using iterative imputation. Furthermore, the FDR-controlled variable selection result is more interpretable and better characterises the relationship between the achievement traits and the background variables. Thus, this approach may be more suitable than the PCA-based approach in practice for scaling large-scale assessment data.

References

BIRNBAUM, ALLAN. 1968. Some latent trait models. *Pages 397–424 of:* LORD, F. M., & NOVICK, M. R. (eds), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

DAI, CHENGUANG, LIN, BUYU, XING, XIN, & LIU, JUN S. 2022. False discovery rate control via data splitting. *Journal of the American Statistical Association*, 1–38.

EMBRETSON, SUSAN E, & REISE, STEVEN P. 2000. *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

LIU, JINGCHEN, GELMAN, ANDREW, HILL, JENNIFER, SU, YU-SUNG, & KROPKO, JONATHAN. 2014. On the stationary distribution of iterative imputations. *Biometrika*, **101**, 155–173.

VAN DER VAART, AAD W. 2000. *Asymptotic statistics*. Cambridge, England: Cambridge University Press.

CLUSTERING THREE-WAY DATA WITH OUTLIERS

Katharine M. Clark [1] and Paul D. McNicholas [1]

[1] Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada (e-mail: clarkkm2@mcmaster.ca, paul@math.mcmaster.ca)

ABSTRACT: An approach for clustering three-way data is discussed. The approach, which is based on mixtures of matrix-variate distributions, uses an iterative subset log-likelihood approach to detect and trim outliers.

KEYWORDS: clustering, matrix-variate, mixture models, outliers, three-way data.

1 Introduction

Grubbs (1969) describes an outlier as an observation "that appears to deviate markedly from other members of the sample in which it occurs." Outliers, and their treatment, is a long-studied topic in the field of applied statistics. The problem of handling outliers in multivariate clustering has been studied in several contexts including work by García-Escudero *et al.* (2008), Punzo & McNicholas (2016), Punzo *et al.* (2020), and Clark & McNicholas (2023). The approach of Clark & McNicholas (2023) is extended to the matrix-variate paradigm, i.e., to account for three-way data such as multivariate longitudinal data. The OCLUST algorithm introduced in Clark & McNicholas (2023), and supported by the R package `oclust` (Clark & McNicholas, 2022), is based on the mixture model-based clustering framework (see, e.g., McNicholas, 2016) and uses an iterative subset log-likelihood approach to detect and trim outliers. An analogue of the OCLUST algorithm is developed for three-way data.

2 Background

The density of a finite mixture model is $f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$, where $\boldsymbol{\vartheta} = \{\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_G\}$, $\pi_g > 0$ is the gth mixing proportion with $\sum_{g=1}^{G} \pi_g = 1$, and $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the gth component density with parameters $\boldsymbol{\theta}_g$. Most (mixture) model-based clustering methods assume, either explicitly or implicitly, that the data are free of outliers. Outlier algorithms in (multivariate) model-based clustering usually fall into either one of two paradigms: outlier-inclusion and outlier trimming. Focusing on the latter, Cuesta-Albertos *et al.* (1997)

developed an impartial trimming approach for k-means clustering; however, this method maintains the drawbacks of k-means clustering, where the clusters are spherical with equal — or, in practice, similar — radii. García-Escudero *et al.* (2008) improved upon trimmed k-means with the TCLUST algorithm. TCLUST places a restriction on the eigenvalue ratio of the covariance matrix, as well as implementing a weight on the clusters, allowing for clusters of various elliptical shapes and sizes. An obvious challenge with these methods is that the eigenvalue ratio must also be known *a priori*. There exists an estimation scheme for the proportion of outliers but it is heavily influenced by the choices for number of clusters and eigenvalue ratio.

The OCLUST algorithm (Clark & McNicholas, 2023) uses the fact that the Mahalanobis distance is χ^2_p for p-dimensional multivariate normal data (Mardia *et al.*, 1979) to derive the distribution of subset log-likelihoods for clustering multivariate normal data. A subset log-likelihood is considered to be the log-likelihood of a model fitted with $n-1$ of the data points. There are n such subsets. The OCLUST algorithm uses the subset log-likelihoods and their distribution to identify and trim outliers.

Two-way data can be regarded as the observation of n vectors, whereas three-way data can be considered the observation of n matrices. Mixtures of matrix-variate distributions have been used to cluster three-way data (e.g., Viroli, 2011; Anderlucci & Viroli, 2015; Gallaugher & McNicholas, 2018). An $r \times c$ random matrix $\mathcal{X}$ comes from a matrix-variate normal distribution if its density is of the form

$$\phi_{r\times c}(\boldsymbol{X} \mid \boldsymbol{M}, \boldsymbol{V}, \boldsymbol{U}) = \frac{1}{(2\pi)^{\frac{rc}{2}} |\boldsymbol{V}|^{\frac{r}{2}} |\boldsymbol{U}|^{\frac{c}{2}}} \exp\left\{ -\frac{1}{2} \text{tr}\big(\boldsymbol{V}^{-1}(\boldsymbol{X}-\boldsymbol{M})^\top \boldsymbol{U}^{-1}(\boldsymbol{X}-\boldsymbol{M})\big) \right\}, \tag{1}$$

where $\boldsymbol{M}$ is the $r \times c$ mean matrix, $\boldsymbol{U}$ is the $r \times r$ row covariance matrix, and $\boldsymbol{V}$ is the $c \times c$ column covariance matrix. Note that there is an identifiability issue with regard to the parameters $\boldsymbol{U}$ and $\boldsymbol{V}$, i.e., if k is a strictly positive constant, then replacing $\boldsymbol{U}$ and $\boldsymbol{V}$ by $(1/k)\boldsymbol{U}$ and $k\boldsymbol{V}$, respectively, leaves (1) unchanged. Various different solutions have been proposed to resolve this issue, including setting $\text{tr}(\boldsymbol{U}) = r$ or $\boldsymbol{U}_{11} = 1$.

For multivariate normal data, the Mahalanobis distance can be expressed as $\mathcal{D}(\mathbf{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$. Pocuca *et al.* (2023) derive a similar expression for matrix-variate normal data:

$$\mathcal{D}_M(\boldsymbol{X}_i, \boldsymbol{M}, \boldsymbol{V}, \boldsymbol{U}) = \text{tr}\left\{ \boldsymbol{U}^{-1}(\boldsymbol{X}_i - \boldsymbol{M})\boldsymbol{V}^{-1}(\boldsymbol{X}_i - \boldsymbol{M})^\top \right\}, \tag{2}$$

and prove that if a Kronecker product structure exists for $\mathbf{\Sigma}$, then

$$\mathcal{D}_M(\boldsymbol{X}_i, \hat{\boldsymbol{M}}, \hat{\boldsymbol{U}}, \hat{\boldsymbol{V}}) \xrightarrow{P} \mathcal{D}_M(\boldsymbol{X}_i, \boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V}), \tag{3}$$

where $\xrightarrow{P}$ denotes convergence in probability.

3 Methodology

As in the multivariate case, consider a subset log-likelihood in the matrix-variate case to be the log-likelihood of a model fitted with $n-1$ of the data points. Formally, if we denote our complete dataset as $\mathcal{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$, then the jth subset is defined as the complete dataset with the jth point removed, i.e., $\mathcal{X} \setminus \boldsymbol{X}_j = \{\boldsymbol{X}, \ldots, \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, \ldots, \boldsymbol{X}_n\}$. Analogous to the multivariate case, treat point $\boldsymbol{X}_k$, whose absence produced the largest subset log-likelihood, as our candidate outlier, ie.

Definition 1 (Candidate Outlier)**.** *We define our candidate outlier as $\boldsymbol{X}_k$, where*

$$k = \arg\max_{j \in [1,n]} \ell_{\mathcal{X} \setminus \boldsymbol{X}_j},$$

and $\ell_{\mathcal{X} \setminus \boldsymbol{X}_j}$ is the log-likelihood of the subset model with the jth point removed.

Remove candidate outliers one-by-one until we obtain our best model, which is determined by the distribution of our subset log-likelihoods, stated in Proposition 1.

Proposition 1. *For a point $\boldsymbol{X}_j$ belonging to the hth cluster, if $Q_{\mathcal{X}}$ is a simplified log-likelihood and $Y_j = Q_{\mathcal{X} \setminus \boldsymbol{X}_j} - Q_{\mathcal{X}}$, then Y_j has an approximate shifted gamma density*

$$Y_j \sim f_{gamma}\left(y_j - k \;\middle|\; \alpha = \frac{p}{2}, 1\right), \tag{4}$$

for $y_j - k \geq 0, \alpha > 0$, where

$$k = -\log \pi_h + \frac{rc}{2}\log(2\pi) + \frac{c}{2}\log|\boldsymbol{U}_h| + \frac{r}{2}\log|\boldsymbol{V}_h|,$$

n_h is the number of points in cluster h, and $\pi_h = n_h/n$.

The mathematical results for this proposition will be given in the full paper, along with other technical details as well as illustrations via real and simulated data.

References

ANDERLUCCI, L., & VIROLI, C.. 2015. Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics*, **9**(2), 777–800.

CLARK, K. M., & MCNICHOLAS, P. D. 2022. *oclust: Gaussian model-based clustering with outliers*. R package version 0.2.0.

CLARK, K. M., & MCNICHOLAS, P. D. 2023. *Using subset log-likelihoods to trim outliers in Gaussian mixture models*. arXiv preprint arXiv:1907.01136v4.

CUESTA-ALBERTOS, J. A., GORDALIZA, A., & MATRÁN, C. 1997. Trimmed *k*-means: an attempt to robustify quantizers. *The Annals of Statistics*, **25**(2), 553–576.

GALLAUGHER, M. P. B., & MCNICHOLAS, P. D. 2018. Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, **80**, 83 – 93.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2008. A General Trimming Approach to Robust Cluster Analysis. *The Annals of Statistics*, **36**(3), 1324–1345.

GRUBBS, F. E. 1969. Procedures for detecting outlying observations in samples. *Technometrics*, **11**(1), 1–21.

MARDIA, K. V., KENT, J. T., & BIBBY, J. M. 1979. *Multivariate Analysis*. London: Academic Press.

MCNICHOLAS, P. D. 2016. Model-Based Clustering. *Journal of Classification*, **33**(3), 331–373.

POCUCA, N., GALLAUGHER, M. P. B., CLARK, K. M., & MCNICHOLAS, P. D. 2023. Assessing and visualizing matrix-variate normality. *Australian and New Zealand Journal of Statistics*. In press

PUNZO, A., BLOSTEIN, M., & MCNICHOLAS, P. D. 2020. High-dimensional unsupervised classification via parsimonious contaminated mixtures. *Pattern Recognition*, **98**, 107031.

PUNZO, A., & MCNICHOLAS, P. D. 2016. Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**(6), 1506–1537.

VIROLI, C. 2011. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and computing*, **21**(4), 511–522.

A TWO-COMPONENT MARKOV SWITCHING REGRESSION MODEL*

Roberto Colombi [1] and Sabrina Giordano[2]

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

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

ABSTRACT: The proposed approach addresses the issue of response styles in longitudinal ordered categorical data, where respondents tend to endorse certain options on a Likert scale regardless of the item's content. These response styles, including middle, extremes, acquiescence, and disacquiescence, can introduce bias in the results. To tackle this, the approach uses a Markov switching logit model with two latent components. One component captures serial dependence and respondent's unobserved heterogeneity, while the other accommodates the responding attitude (RS or no-RS). The responses' dependence on covariates is modeled using a flexible stereotype logit model with parameters varying based on the two latent components.

KEYWORDS: Latent variables; Response styles; Stereotype logit models.

Motivation and Models

Longitudinal ordered categorical data is susceptible to response styles, where respondents, when asked to assess items using Likert scales at various time points, tend to consistently select only a few specific options on the rating scale, regardless of the item's actual content.

Numerous studies in psychometrics and statistics have explored different types of response styles (RS) and their consequences (Van Vaerenbergh & Thomas, 2013). The commonly recognized response styles include acquiescence, disacquiescence, extreme and middle RS as described in Baumgartner & Steenkamp, 2001, among many others.

*The author Sabrina Giordano received partial financial support by MUR, grant number 2022XRHT8R - The SMILE project and by COST Action CA19130 Fintech and Artificial Intelligence in Finance - Towards a transparent financial industry (FinAI), funded by COST (European Cooperation in Science and Technology).

Ignoring the RS mechanism can introduce response heterogeneity, biases the estimated model parameters, and consequently leads to inaccurate results (e.g., Colombi *et al.*, 2021).

In this study, our goal is to account for the temporal evolution of the RS behavior, which contrasts with previous approaches that either ignore it or model it as a time-invariant latent trait or random effect (Billiet & Davidov, 2008; Schauberger & Tutz, 2022).

Recognizing the significance of RS and its temporal dynamics, we propose a Markov switching model (Fruhwirth Schnatter, 2001) driven by a bivariate latent Markov chain. One component of this chain has k states or regimes, known as the k-regime switching indicator, which captures serial dependence and respondent heterogeneity due to unobserved covariates. The other binary component, called the response style regime switching indicator, dictates whether respondents answer according to an RS or use the rating scale appropriately to accurately represent their feelings.

Given the k-regime switching indicator, the observed categorical responses' dependence on time-varying and subject-specific covariates under the no-RS regime is modeled using a stereotype logit model (Anderson, 1984), while under the RS regime, it is modeled using a parallel local logit model with restricted intercepts, accommodating the tendency of respondents to select categories due to RS.

This approach adds a contribution to the literature on multivariate Markov chains in the context of Markov switching models (e.g., Pohle *et al.*, 2021, among others). The novelty of our approach, which extends existing models for longitudinal categorical data (e.g., Bartolucci *et al.*, 2012), lies in providing a Markov switching regression model for ordered responses that simultaneously considers attitude towards response styles, unobserved heterogeneity, serial dependence, and the impact of time-varying covariates.

The fundamental assumption in our current approach is that transition probabilities remain identical across subjects, as unit-specific covariate effects are accounted for at the observation level. However, an alternative scenario has been explored by Colombi *et al.*, 2023, where a non-homogeneous latent process is considered. In this case, the initial probabilities can be influenced by time-invariant regressors, while the transition probabilities by time-specific covariates. This approach considers the restriction of subject and time-invariant observation probability functions. In the mentioned paper, the observed variables are treated as indicators of a latent construct of interest, allowing covariates to naturally affect only the latent component of the model. The primary focus in the proposed work is centered on logit regression models fea-

turing time-varying parameters for the observable variables, with the k-regime switching indicator serving as a tool to model both unit heterogeneity and time dependence arising from unobserved covariates. For a deeper comprehension of the proposed model and real-world applications, refer to Colombi & Giordano, 2023's work, which provides extensive details on both aspects.

Our approach has potential applications in various longitudinal surveys that collect opinions on health status, risk of illness, economic difficulties, the impact of climatic events, discriminatory and racist beliefs, and political attitudes. These surveys may reveal biased perceptions due to response styles that vary over time, reflecting the ever-changing nature of human behavior.

References

ANDERSON, J. A. 1984. Regression and ordered categorical variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **46**(1), 1–22.

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

BAUMGARTNER, H., & STEENKAMP, J. B. E. M. 2001. Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research*, **38**, 143–156.

BILLIET, J. B., & DAVIDOV, E. 2008. Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research*, **36**(4), 542–562.

COLOMBI, R., & GIORDANO, S. 2023. Markov Switching Stereotype Logit Models for Longitudinal Ordinal Data Affected by Unobserved Heterogeneity in Responding Behavior. *Submitted (second revision)*.

COLOMBI, R., GIORDANO, S., & TUTZ, G. 2021. A Rating Scale Mixture Model to Account for the Tendency to Middle and Extreme Categories. *Journal of Educational and Behavioral Statistics*, **46**(6), 682–716.

COLOMBI, R., GIORDANO, S., & KATERI, M. 2023. Hidden Markov Models for Longitudinal Rating Data with Dynamic Response Styles. *Statistical Methods and Applications (in press)*.

FRUHWIRTH SCHNATTER, S. 2001. *Finite Mixture and Markov Switching Models*. Springer.

POHLE, J., LANGROCK, R., SCHAAR, M. VAN DER, KING, R., & JENSEN, F. H. 2021. A primer on coupled state-switching models for multiple interacting time series. *Statistical Modelling*, **21**(3), 264–285.

SCHAUBERGER, G., & TUTZ, G. 2022. Multivariate ordinal random effects models including subject and group specific response style effects. *Statistical Modelling*, **22**(5), 409–429.

VAN VAERENBERGH, Y., & THOMAS, T. D. 2013. Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, **25**(2), 195–217.

THE BROAD PHENOTYPE-SPECIFIC APPLICATIONS OF THE NETWORK-BASED SWIM TOOL

Federica Conte [1] and Paola Paci [2]

[1] Institute for Systems Analysis and Computer Science "A. Ruberti" (IASI) National Research Council (CNR) of Rome, Rome, Italy, (e-mail: federica.conte@iasi.cnr.it)

[2] Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy, and Karolinska Institutet, 17177 Stockholm, Sweden (e-mail: paci@diag.uniroma1.it)

ABSTRACT: SWIM is a recently developed network-based tool that fulfils the criteria of the new quickly emerging field of Network Medicine in finding disease-associated genes, called switch genes. Here, a brief summary of the promising results obtained by applying SWIM in different biological contexts is presented.

KEYWORDS: network medicine, network theory, disease genes

1 Introduction

Recently, we developed a new promising methodology, called SWIM (SWItch Miner), which integrates different network-based methods to analyse the correlation network arising from large-scale gene expression data Paci *et al.*, 2017. Considering the topological properties of the nodes and assessing their functional roles according to their ability to convey information within and between modules in the network, SWIM identifies a small pool of genes (called switch genes) that are associated with intriguing patterns of molecular co-abundance and play a crucial role in the observed phenotype.

The phenotype-specific applications of SWIM are broad and include the identification of switch genes in grapevine berry maturation Palumbo *et al.*, 2014, in human cancers Paci *et al.*, 2017, including glioblastoma multiforme (GBM) Fiscon *et al.*, 2018 and in chronic obstructive pulmonary disease (COPD) Paci *et al.*, 2020. More recently, SWIM has been applied within the framework of Network Medicine to study the interplay between switch genes and human diseases in the human interactome (i.e., the cellular network of all physical molecular interactions) Paci *et al.*, 2021.

In the following, a detailed description of the more recent applications of SWIM to complex diseases is provided.

2 Methods

SWIM is a freely downloadable network-based tool, developed both in MATLAB Paci *et al.*, 2017 and in R language Paci & Fiscon, 2022, which predicts important (switch) genes that are strongly associated with drastic changes in cell phenotype. SWIM encompasses several steps detailed in Paci *et al.*, 2017.

3 Results and Discussions

3.1 Glioblastoma

Glioblastoma is the most aggressive and frequent brain tumour, with a median survival time of 12–15 months from diagnosis Young *et al.*, 2015. This tumour is resistant to the standard therapies and its aggressiveness seems to be due to the presence of cancer stem-like cells Gimple *et al.*, 2019. Thus, targeting cancer stem-like cells could pave the way for new therapeutic strategies.

A recent study identified 19 neurodevelopmental transcription factors (TFs) that are selectively expressed in glioblastoma stem-like cells to maintain their stem-like phenotype and prevent differentiation Suva *et al.*, 2014. A subset of only four of them (named 4-core TFs), SOX2, OLIG2, POU3F2, and SALL2, has been shown to be sufficient to fully reprogram differentiated cells into glioblastoma stem-like cells Suva *et al.*, 2014.

In order to identify switch genes related to the stem-like phenotype, SWIM was applied to GBM dataset of Suva *et al.*, 2014 and then the further dataset of Schulte *et al.*, 2011 was used to validate the results Fiscon *et al.*, 2018. Among the common switch genes obtained by running SWIM on the these two GBM datasets, there is FOSL1. It is up-regulated in differentiated glioblastoma cells and this up-regulation highly correlates with the over-expression of genes involved in cell-cell communications. It is down-regulated in stem-like cells and this down-regulation highly correlates with the up-regulation of the 4-core of TFs. To investigate a possible co-regulation of the 4-core of TFs, their promoter regions were inspected to search for enriched motifs and they were found to harbour a consensus binding site for FOSL1.

Altogether these findings suggest FOSL1 as possible therapeutic biomarker of glioblastoma, which could promote the differentiation of cancer stem-like cells by repressing the 4-core TFs. This hypothesis has been partially experimentally validated in Pecce *et al.*, 2021.

3.2 COPD

COPD is a heterogeneous and complex syndrome influenced by both genetic and environmental determinants, and is one of the main causes of morbidity and mortality worldwide.

By applying SWIM on COPD Paci *et al.*, 2020, the correlation network turned out to be formed by three well-characterised modules: i) one populated by switch genes, all up-regulated in COPD cases and involved in COPD-related pathways, like B cell receptor signalling pathway; ii) one populated by negative interactors of switch genes, down-regulated in COPD cases, including well-known GWAS genes like AGER and CAVIN1; iii) one populated by well-recognised immune signature genes, all up-regulated in COPD cases. Switch genes appear to form localised connected subnetworks displaying an intriguingly common pattern of up-regulation in COPD cases compared with controls. A more sophisticated analysis revealed that they were not only topologically related, but also functionally relevant to the observed phenotype as witnessed by their enrichment in the regulation of inflammatory and immune responses. Finally, SWIM was applied on another severe lung disease with an inflammatory component, i.e., the acute respiratory distress syndrome (ARDS), demonstrating that, even though different diseases can share similar endophenotypes, the molecular network determinants responsible for them are disease-specific.

3.3 Network Medicine

Network Medicine is a new emerging paradigm in medicine, where disease proteins are assumed not to be randomly scattered, but agglomerate in specific regions of the molecular interactome, suggesting the existence of specific disease network modules for each disease Barabási *et al.*, 2011. To quantify the interplay between switch genes and human diseases in the human interactome, the results obtained by the pan-cancer Paci *et al.*, 2017 and COPD Paci *et al.*, 2020 SWIM-based analysis were complemented with the application of SWIM tool on two cardiac disorders (i.e., ischemic and non-ischemic cardiomyopathy) and on Alzheimer's disease (AD) Paci *et al.*, 2021. Switch genes associated with specific disorders were found to be not randomly scattered but they form localised connected subnetworks. These subnetworks overlap between similar diseases (like cancers or cardiac disorders) and are situated in different neighbourhoods for pathologically distinct phenotypes (like AD and COPD), showing a direct relation between the pathobiological similarity of diseases and their relative distance in the human interactome. Finally, the first SWIM-

informed Human Disease Network was built, where nodes correspond to distinct disorders and a link occurs between two diseases if they share a substantial number of switch genes. Clustering of nodes belonging to the same disease class means that similar pathophenotypes have a higher probability of sharing switch genes than do pathophenotypes that belong to different disease classes. These findings support the hypothesis that SWIM-based correlation network, when integrated with an interactome-based network analysis, not only identifies novel candidate disease genes, but also may offer useful tool by which to elucidate the molecular underpinnings of human disease and reveal commonalities between seemingly unrelated diseases.

References

BARABÁSI, ALBERT-LÁSZLÓ, GULBAHCE, NATALI, & LOSCALZO, JOSEPH. 2011. Network medicine: a network-based approach to human disease. *Nature Reviews. Genetics*, **12**(1), 56–68.

FISCON, GIULIA, CONTE, FEDERICA, LICURSI, VALERIO, NASI, SERGIO, & PACI, PAOLA. 2018. Computational identification of specific genes for glioblastoma stem-like cells identity. *Scientific Reports*, **8**(1), 7769.

GIMPLE, RYAN C., BHARGAVA, SHRUTI, DIXIT, DEOBRAT, & RICH, JEREMY N. 2019. Glioblastoma stem cells: lessons from the tumor hierarchy in a lethal cancer. *Genes & Development*, **33**(11-12), 591–609.

PACI, PAOLA, & FISCON, GIULIA. 2022. SWIMmeR: an R-based software to unveiling crucial nodes in complex biological networks. *Bioinformatics*, **38**(2), 586–588.

PACI, PAOLA, COLOMBO, TERESA, FISCON, GIULIA, GURTNER, AYMONE, PAVESI, GIULIO, & FARINA, LORENZO. 2017. SWIM: a computational tool to unveiling crucial nodes in complex biological networks. *Scientific Reports*, **7**(Mar.), srep44797.

PACI, PAOLA, FISCON, GIULIA, CONTE, FEDERICA, LICURSI, VALERIO, & OTHERS. 2020. Integrated transcriptomic correlation network analysis identifies COPD molecular determinants. *Scientific Reports*, **10**(1), 1–18. Number: 1 Publisher: Nature Publishing Group.

PACI, PAOLA, FISCON, GIULIA, CONTE, FEDERICA, WANG, RUI-SHENG, FARINA, LORENZO, & LOSCALZO, JOSEPH. 2021. Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. *npj Systems Biology and Applications*, **7**(1), 1–11. Number: 1 Publisher: Nature Publishing Group.

PALUMBO, MARIA CONCETTA, ZENONI, SARA, FASOLI, MARIANNA, MASSONNET, MÉLANIE, FARINA, LORENZO, CASTIGLIONE, FILIPPO, PEZZOTTI, MARIO, & PACI, PAOLA. 2014. Integrated Network Analysis Identifies Fight-Club Nodes as a Class of Hubs Encompassing Key Putative Switch Genes That Induce Major Transcriptome Reprogramming during Grapevine Development. *The Plant Cell Online*, Dec., tpc.114.133710.

PECCE, VALERIA, VERRIENTI, ANTONELLA, FISCON, GIULIA, SPONZIELLO, MARIALUISA, CONTE, FEDERICA, DURANTE, COSIMO, FARINA, LORENZO, FILETTI, SEBASTIANO, & PACI, PAOLA. 2021. FOSL1 in the stemness mechanism: An in vitro study. *Scientific Reports*.

SCHULTE, ALEXANDER, GÜNTHER, HAUKE S., PHILLIPS, HEIDI S., & OTHERS. 2011. A distinct subset of glioma cell lines with stem cell-like properties reflects the transcriptional phenotype of glioblastomas and overexpresses CXCR4 as therapeutic target. *Glia*, **59**(4), 590–602.

SUVA, MARIO L, RHEINBAY, ESTHER, GILLESPIE, SHAWN M, & OTHERS. 2014. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell*, **157**(3), 580–594.

YOUNG, RICHARD M, JAMSHIDI, ARIA, DAVIS, GREGORY, & SHERMAN, JONATHAN H. 2015. Current trends in the surgical management and treatment of adult glioblastoma. *Annals of translational medicine*, **3**(9).

TESTING GRAPH CLUSTERABILITY: A DENSITY BASED STATISTICAL TEST FOR DIRECTED GRAPHS

Houyem Demni[1], Pierre Miasnikof[2], Alexander Y. Shestopaloff [3] , Cristián Bravo [4] and Yuri Lawryshyn [2]

[1] Department of Economics and Law, University of Cassino and Southern Lazio, Cassino, Italy, (e-mail: houyem.demni@unicas.it, mariorosario.guarracino@unicas.it)

[2] University of Toronto, Toronto, ON, Canada, (e-mail: p.miasnikof@utoronto.ca, yuri.lawryshyn@utoronto.ca)

[3] Queen Mary University of London, London, UK and Memorial University of Newfoundland, St. John's, NL, Canada (e-mail: a.shestopaloff@qmul.ac.uk)

[4] University of Western Ontario, London, ON, Canada (e-mail: cbravoro@uwo.ca)

ABSTRACT: In this work, we extend a recent statistical test for graph clusterability to directed graphs. Graph clustering, or network community detection, is a pivotal topic in network science. It consists of labeling nodes so they form subsets that display a greater similarity to each other than to the remaining vertices on the graph. Here, node similarity is measured in connection probability or edge density. Similar nodes have a greater connection probability to each other than to other vertices. However, not all graph have a clustered structure. While the goal of graph clustering is to offer a meaningful summary of a graph through vertex clusters, not all graphs can be summarized in this way. In cases where a graph is not clusterable, clustering is not only a waste of time, it inevitably leads to misleading conclusions. We tailor a statistical test developed for undirected networks to directed ones. The test is based on measuring the heterogeneity of local densities. It does not assume any particular graph generative model or edge probability distribution. The test only rests on the hypothesis that a clusterable graph must display a mean local (induced subgraph) density that is significantly greater than the graph's overall density. We posit that this inequality is a necessary (but not sufficient) condition for a graph to have a clustered structure. After highlighting the probabilistic nature of local and global densities, we offer a statistical test to assess the significance of this inequality in densities. This test is also based on sampling node neighborhoods and is thus well suited to very large data sets. We have validated our test on several synthetic graph structures and real world networks. We have also compared our test to other recent statistical tests. Our findings show that our test is more responsive to networks structure than its alternatives.

KEYWORDS: Clustering, global densities, local densities, networks

DEEP NEURAL NETWORK IN THE MODELING OF THE DEPENDENCE STRUCTURE IN RISK AGGREGATION

Anna Denkowska[1], Krystian Szczęsny[1], Joao Vieito[2] and Stanisław Wanat[1]

[1] Department of Mathematics, Cracow University of Economics,
(e-mail: anna.denkowska@uek.krakow.pl, krystian.szczesny@o2.pl, wanats@uek.krakow.pl)

[2] School of Business Studies, Polytechnic Institute of Viana do Castelo,
(e-mail: joaovieito@esce.ipvc.pt)

ABSTRACT: We model the dependency structure in the premium and reserve risk sub-module determining the Solvency Capital Requirement (SCR) and the diversification effect (DE). We use the Deep Neural Network (DNN) to estimate marginal distributions modeling the premium and reserve risk of non-life insurance segments, and a copula defining the multidimensional dependency between segments. We use the energy distance to evaluate the error of fitting the copula to the real data. The determined DE when modeling dependencies using the copula method estimated by the use of DNN is compared with DE when modeling dependencies using the method proposed in the Solvency II Directive and using C-vine copulas. The obtained test results indicate that the use of DNN allows for more accurate modeling of the dependency structure, and the determined DE is at the appropriate level.

KEYWORDS: deep neural network, C-vine copulas, Solvency Capital Requirement, diversification effect, risk aggregation.

1 Introduction

Pursuant to the Solvency II Directive (CDR, 2016), each insurance company is obliged to meet the SCR, which is determined by aggregating the risk factors to which the insurance company is exposed. The Directive provides a Standard Formula (SF) where the variance-covariance method is used for risk aggregation and the risk factor correlation matrix is predetermined in the Directive. An alternative to SF are the internal models that better reflect the insurer's business profile. EIOPA (2020) launched a pan-European benchmarking study on diversification in internal models. The aim is to better understand the relationship between dependency modeling and risk aggregation and their diversification benefits. Our dependency modeling method is a proposal for use in internal models. Studies on the impact of the dependency between aggregated risks on the estimated SCR have been conducted by Bermúdez et al. (2013), Cifuentes and Charlin (2016), Mittnik (2020), and Szczęsny (2022a). Eling and Jung (2020) and Szczęsny (2022b) model the structure of dependencies using C-vine copulas introduced in (Bedford and Cooke,

2002). The disadvantages of this approach are discussed by Acar et al. (2012) and Haff (2013). In the literature, the use of DNN for dependency modeling can be found, among others, in Sun et al., (2019), Hassan and Abraham, (2016), and Yunos et al., (2016). We conduct research on real data obtained from Solvency and Financial Condition Reports (SFCRs) of Polish property insurers. We model dependencies based on C-vine copulas (as in (Szczęsny, 2022b)) and a copula fitted to real data using DNN (We generalize the method proposed by (Zeng and Wang, 2022) to determine the four-dimensional distribution). We assess the accuracy of fitted models to real data based on energy distance (Gneiting and Raftery, 2007).

2 Methodology

The solvency requirement for premium and reserve risk in non-life insurance is not greater than the sum of capitals needed to hedge against the risk of each segment separately. The resulting difference is called the Diversification Effect (CDR, 2016). The size of this effect is assessed using the diversification ratio depending on the capital requirements for each segment and the solvency requirement for premium and reserve risk. The key issue in assessing the diversification effect at the appropriate level is both the selection of methods to determine the capital requirement for the L_i risk of individual segments and for the aggregated risk variable L, which depends mainly on the selection of the aggregation function ψ. In the standard formula, it is assumed that L is the sum of dependent random variables L_i with normal distributions with parameters determined by standard deviation for individual segments, where the dependency between L_i is described by the correlation matrix (variance-covariance aggregation is used). It is assumed that the risk variable L has a normal distribution, which usually does not reflect reality.

We estimate the L_i marginal distributions for selected segments and copulas in two ways: using a parametric approach by means of a C-vine copulas as in the paper (Szczęsny 2022b) and innovatively using DNN. Having one-dimensional marginal distributions and the copula function, we invoke Sklar's theorem and combine the determined marginal distributions into a four-dimensional distribution.

3 Empirical result

Due to the limited availability of complete data, we select segments C0020 (insurance against loss of income), C0040 (motor liability insurance), C0050 (other motor insurance), and C0070 (fire and other property damage insurance) for the analysis. We model the risk of these segments using complex factors L_i (i=1,..,4) (Schubert, Grießmann 2007).

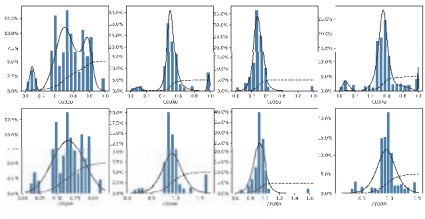
Table 1 presents the actual linear correlation matrix determined differs from the matrix given in the directive (CDR, AnnexIV).

Table 1. Estimated correlations between segments

	C0020	*C0040*	*C0050*	*C0070*
C0020	1	0.2286	0.0034	0.3580
C0040	0.2286	1	0.3028	0.8171
C0050	0.0034	0.3028	1	0.0378
C0070	0.3580	0.8171	0.0378	1

We start the study by estimating the marginal distributions L_i in the two previously given ways. The results for individual segments are presented in Figure 1. Thanks to the method in which we use DNN to estimate probability distributions, we obtain a better fit to real data than in the parametric method. (The upper part of Figure 1.).

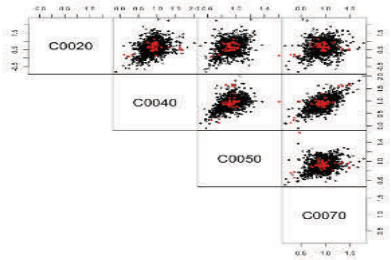
Figure 1. Estimated marginal distributions



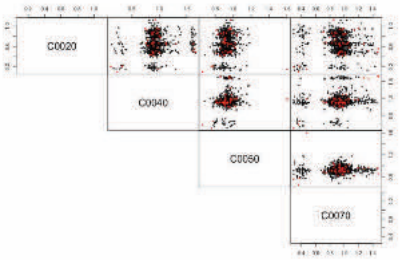
Next, we model the dependency structure between the segments. Figure 2 presents charts for pairs of segments. Red points represent pairs of real observations and black points are pairs of realizations drawn from the four-dimensional joint distribution.

Figure 2. Copulas for pairs of segments

C-vine method



DNN method



Having determined the density distributions for individual segments and the distribution of the copula density, we determine the density of the four-dimensional distribution. Then, after estimating four-dimensional distributions in two ways, using energy distance, we evaluate which one better describes the real data. This distance between each vector of actual observations and the realization vectors drawn from the estimated distributions for the distribution determined using DNN is 0.256471 and for the distribution determined using the parametric approach it is 0.2589487. Which indicates a more accurate match using the DNN method.

Finally, we determine the diversification effects obtained in three ways: the first in accordance with the SF contained in the directive, the second based on the results obtained by applying the C-vine copulas and the third by using the DNN. For the three methods, we get the following values: 0.31, 0.35, 0.53. In the analyzed case, a more accurate determination of the dependency structure through the use of a copula

causes the value of DE to be higher than the value of DE obtained by the variance-covariance method in the Solvency II Directive.

The obtained results show that the proposed deep neural network architecture is a good candidate for the estimation of marginal distributions as well as the identification of the copula used to describe the structure of dependencies between risk types in internal models.

References

ACA, E. F., GENEST, C., & NEŠLEHOVÁ, J. 2012. Beyond simplified pair-copula constructions'. *Journal of Multivariate Analysis.*, **110**, 74-90.

BEDFORD, T., & COOKE, R. M. 2002. Vines - A new graphical model for dependent random variables. *Annals of Statistics.*, **30(4)**, 1031-1068.

BERMÚDEZ, L., FERRI, A., & GUILLÉN, M. 2013. A correlation sensitivity analysis of non-life underwriting risk in solvency capital requirement estimation. *ASTIN Bulletin*., 21-37.

CIFUENTES, A., & CHARLIN, V. 2016. Operational risk and the Solvency II capital aggregation formula: Implications of the hidden correlation assumptions. *Journal of Operational Risk.*, **11(4)**, 23-33.

CDR. 2016. Comission Delegated Regulation (EU) 2015/35 of 10 October 2014 supplementing Directive 2009/138/EC of the European Parliament and of the Council on the taking-up and pursuit of the business of insurance and reinsurance (Solvency II).

ELING, M., & JUNG, K. 2020. Risk aggregation in non-life insurance: Standard models vs. internal models. *Insurance: Mathematics and Economics*., **95**, 183-198.

GNEITING, T., & RAFTERY, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association.*, **102(477)**, 359-378.

SUN, Y., CUESTA-INFANTE, A., & VEERAMACHANENI, K. 2019. Learning vine copula models for synthetic data generation. *AAAI 2019*., 5049-5057.

SZCZĘSNY, K. 2022a. Wpływ błędnej specyfikacji struktury zależności w procesie agregacji ryzyka na efekt dywersyfikacji w Solvency II. W: E. Sojka, J. Acedański (red.), *Problemy gospodarcze i społeczne Polski i Europy*. Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, 98-112.

SZCZĘSNY, K. 2022b. Wykorzystanie kaskad kopuli w agregacji ryzyka w procesie wyznaczania kapitałowych wymogów wypłacalności w Solvency II. W: M. Lemkowska, M. Wojtkowiak (red.), *Sektor ubezpieczeń w obliczu wyzwań współczesności*. Poznań: Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, 98-117.

YUNOS, Z. M. *et al.* 2016. Predictive modeling for motor insurance claims using artificial neural networks. *International Journal of Advances in Soft Computing and its Applications.*, **8(3)**, 160-172.

ZENG, Z., & WANG, T. 2022. Neural Copula : A unified framework for estimating generic high-dimensional Copula functions. https://arxiv.org/abs/2205.15031

CIRCULAR REGRESSION WITH MEASUREMENT ERRORS

Marco Di Marzio [1], Chiara Passamonti[1] and Charles Taylor[2]

[1] DSFPEQ, University of Chieti-Pescara (e-mail: marco.dimarzio@unich.it, chiara.passamonti@unich.it)

[2] Department of Statistics, University of Leeds (e-mail: charles@maths.leeds.ac.uk)

ABSTRACT: We propose techniques for estimating a regression function when the predictor is circular. A case study on Carbon monoxide pollution is presented.

KEYWORDS: Characteristic function, deconvolution kernels, measurement errors.

1 Introduction

We propose a nonparametric regression estimator that is consistent in the presence of measurement error when predictor data are circular. Following the approach of Carroll & Hall, 1988 and Carroll *et al.*, 1995, we introduce a deconvolution-type estimator.

Some facts on the characteristic functions are worth to be recalled. The characteristic function of a circular random variable Θ, denoted as $\varphi_\Theta(\ell) = \alpha_\ell + i\beta_\ell$ satisfies $\varphi_\Theta(\ell) = \varphi_{\Theta+2\pi}(\ell), \ell \in \mathbb{Z}$, being zero elsewhere. Moreover, $\alpha_\ell = E[\cos(\ell\Theta)]$ and $\beta_\ell = E[\sin(\ell\Theta)]$, both are the coefficients in the Fourier series representation of f_Θ, and correspond to the ℓth *trigonometric moment* of Θ. Finally, $\beta_\ell = 0$ when f_Θ is symmetric. If f_Θ is square integrable on $[0, 2\pi)$, one can represent $f_\Theta(\theta)$, $\theta \in [0, 2\pi)$, as

$$\frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\varphi_\Theta(\ell)\exp(-i\ell\theta) = \frac{1}{2\pi}\left\{1 + 2\sum_{\ell=1}^{\infty}(\alpha_\ell\cos(\ell\theta) + \beta_\ell\sin(\ell\theta))\right\}. \quad (1)$$

Our estimator is described in Section 2. In Section 3, we model the carbon monoxide propagation due to wind direction in a region near Huston (Texas).

2 Model and estimator

We consider the case of a circular predictor and linear response. Given the random sample $(\Psi_1, Y_1), \ldots, (\Psi_n, Y_n)$, assume the regression model $Y_i = m(\Psi_i) +$

$\sigma(\Psi_i)e_i$, but it is available the sample $(\Phi_1, Y_1), \ldots, (\Phi_n, Y_n)$, modelled according $\Phi = (\Psi + \varepsilon) \mathsf{mod}(2\pi)$. Here we have that

- the e_is are i.i.d. real-random variables with zero mean and unit variance, and $\sigma^2(\cdot)$ is the conditional variance of Y;
- the Ψ_is are independent copies of the circular latent variable Ψ with density function f_Ψ;
- the ε_is are i.i.d. circular random variables independent of the (Ψ_i, e_i)'s, with a known density function f_ε which is symmetric around zero.

We assume that f_ε, f_Ψ and f_Φ are square integrable, and f_ε is a circular density allowing an absolutely convergent Fourier series representation.

A local estimator for m at $\psi \in [0, 2\pi)$, denoted by $\tilde{m}(\psi; \kappa)$, can be obtained by employing a *circular* deconvolution kernel. Using the inversion formula (1), and considering that for a symmetric function $\beta_\ell = 0$ for any ℓ, we have

$$\tilde{K}_\kappa(\phi) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{\ell=1}^{\infty} \frac{\gamma_\ell(\kappa)}{\lambda_\ell(\kappa_\varepsilon)} \cos(\ell\phi) \right\}, \tag{2}$$

with smoothing parameter $\kappa > 0$, where $\gamma_\ell(\kappa)$ and $\lambda_\ell(\kappa_\varepsilon)$, for $\ell \in \mathbb{Z}$, respectively are the ℓth Fourier coefficient of the periodic weight function K_κ and the error density f_ε whose concentration is κ_ε. The estimator is well defined when the error density has nonvanishing Fourier coefficients, $\gamma_\ell(\kappa)$ is not identically zero and $\sum_{\ell=1}^{\infty} \mid \gamma_\ell(\kappa)/\lambda_\ell(\kappa_\varepsilon) \mid < \infty$ for all $(\kappa, \kappa_\varepsilon) \in \mathbb{R}_+^2$, which, in turn, imply that both K_κ and $\tilde{K}_\kappa$ are square integrable functions.

The local constant estimator for m is defined by

$$\tilde{m}(\psi; \kappa) = \frac{\sum_{i=1}^{n} \tilde{K}_\kappa(\Phi_i - \psi) Y_i}{\sum_{i=1}^{n} \tilde{K}_\kappa(\Phi_i - \psi)}, \tag{3}$$

where $\tilde{K}_\kappa$ is a circular deconvolution kernel.

Theorem 1. *Given the* $[0, 2\pi) \times \mathbb{R}$*-valued random sample* $(\Psi_1, Y_1), \ldots, (\Psi_n, Y_n)$*, consider the local constant estimator. If*

i) K_κ *is a second sin-order kernel admitting a convergent Fourier series representation* $1/(2\pi)\{1 + 2\sum_{\ell=1}^{\infty} \gamma_\ell(\kappa) \cos(\ell\theta)\}$*, with* κ *increasing with n in such a way that, for* $\ell \in \mathbb{Z}^+$*,*
$\lim_{n\to\infty} \frac{1-\gamma_\ell(\kappa)}{1-\gamma_2(\kappa)} = \frac{\ell^2}{4}$,
$\lim_{n\to\infty} \gamma_\ell(\kappa) = 1$ *and* $\lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{\infty} \gamma_\ell^2(\kappa) = 0$,

ii) the second derivative of the regression function m is continuous,
iii) the conditional variance σ^2 is continuous, and the density f_Ψ is continuously differentiable,

then

$$\mathsf{E}[\hat{m}(\psi;\kappa)] - m(\psi) = \frac{(1-\gamma_2(\kappa))}{4}\left\{ m''(\psi) + \frac{2m'(\psi)f'_\Psi(\psi)}{f_\Psi(\psi)} \right\} + o(1-\gamma_2(\kappa)),$$

$$\mathsf{Var}[\hat{m}(\psi;\kappa)] = \frac{\left(1+2\sum_{\ell=1}^{\infty}\gamma_\ell^2(\kappa)\right)}{2\pi n f_\Psi(\psi)}\sigma^2(\psi) + o\left(\frac{\sum_{\ell=1}^{\infty}\gamma_\ell^2(\kappa)}{n}\right).$$

We notice that, as in the Euclidean setting, the measurement error has no effect on the asymptotic bias of the estimator, which, when the predictor observed with error is circular (linear respectively), depends only on the second moment of the classical kernel K_κ (K_h resp.). The asymptotic variance, similarly to the Euclidean setting, depends on the Fourier coefficients (characteristic function resp.) of the error density appearing in roughness of the deconvolution kernel $\tilde{K}_\kappa$ ($\tilde{K}_h$ resp.).

3 Pollution and surface wind data

Usually, air pollution in a region strongly depends on wind direction. We consider data from the Texas Commission on Environmental Quality, where the response variable is the amount of carbon monoxide (CO) while the explanatory variable is the wind direction. We have selected a site near Houston ("North Loop") in Harris County at Latitude: 29.81^o North and Longitude: -95.39^o West using data from 2018*. The data are collected hourly, but we have calculated the average daily wind direction (using the directional average), and the average daily CO (in parts per million). These daily averages were "thinned" to reduce serial correlation resulting in 183 observations from alternate days. We initially fit a parametric model in which CO (y) is related to wind direction (ϕ) using a sine-cosine model $Y_i = \beta_0 + \beta_1 \sin\Phi_i + \beta_2 \cos\Phi_i + e_i$. This gives fitted values $\hat{\beta}_0 = 0.568, \hat{\beta}_1 = -0.173, \hat{\beta}_2 = 0.074$. The CO pollution is highest when the wind is coming from the south (2.73 radians). Then, we fit a standard circular-linear nonparametric regression, in which the measurements are treated as error free. The smoothing parameter (chosen by leave-one-out

*https://www.tceq.texas.gov/

cross-validation) was selected as $\kappa = 7.77$ for a von Mises kernel. For this model, the maximum CO occurs at 2.11 radians.

Finally, in this circular-linear case, we use a error-in-variables model for the observed wind direction which can be approximated by a wrapped Normal error with zero mean and concentration equal to 0.9. The estimated CO is then given using equation (3), in which κ was found by leave-one-out cross-validation to be 3.35. The three curves, depicted in Figure 1, show that, in the last case, the curve appears to be somewhat less smooth than the error-free model estimate. The nonparametric errors-in-variables model has residual sum of squares equal to 1.91, whereas the parametric model is slightly larger (2.40) and the error-free model very similar (1.99). The maximum estimated CO occurs at $\phi = 2.17$ for the errors-in-variables model.

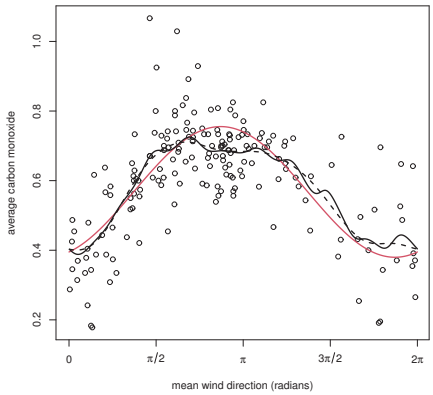


Figure 1. *Carbon monoxide vs wind direction at Houston North Loop monitoring station — alternate daily averages for 2018. Parametric sin/cos model (red), fitted nonparametric errors in variables model (black) and standard circular-linear (no error model) kernel regression (dashed).*

References

CARROLL, R.J., & HALL, P. 1988. Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, **83**, 1184–1186.

CARROLL, R.J., RUPPERT, D., & STEFANSKI, L.A. 1995. *Measurement Error in Nonlinear Models*. New York: Chapman and Hall.

CLASSIFICATION TREE TO IMPROVE DATA QUALITY IN OFFICIAL STATISTICS

Di Zio Marco[1], Filippini Romina[1], Rocchetti Gaia[1] and Toti Simona[1]

[1] Italian National Institute of Statistics - Istat
(e-mail: dizio@istat.it, filippini@istat.it, grocchetti@istat.it, toti@istat.it)

ABSTRACT: Errors inevitably affect data. Statistical Institutes deal with them with rules designing data relationships and imputation for missing information. In this paper, we report the use of classification trees to improve this process when using longitudinal administrative data in the estimation process.

KEYWORDS: Data editing, official statistics, imputation, non-sampling errors.

1 Introduction

National Statistical Institutes (NSI) generally deal with data affected by errors. In the emerging multisource data production system, survey data are integrated with administrative data, thus non-sampling errors assume a relevant role for the quality of statistical figures produced by NSI. A '*statistical data editing and imputation*' step for correcting errors and missing data is performed (Unece 2019). In this process, logical and statistical rules connecting variables are used for treating errors, and imputation of missing data is performed. Rules are proposed by subject matter experts, but of course sometimes may not be precise enough. Machine learning may improve the set of rules and decisions adopted for checking and imputing data by exploiting the longitudinal characteristics of the data. In fact, rules can be learned from the data itself. In this paper, we study the case of the estimation of the Attained Level of Education (ALE) of the Permanent Italian Census. It is strongly based on the use of administrative data, but they are characterized by some gaps and delay with respect to the reference time. To produce ALE estimates for all the Italian resident population, Istat adopted a mass imputation approach integrating administrative, survey and the 2011 census data (Di Zio et al., 2019). After the data time lag is passed, the availability of updated information is an opportunity for an evaluation and a tuning of the procedure. In this paper, we report a study concerning the application of classification trees to the updated information to improve the data editing and imputation process.

2 Structure of available information

Administrative data from the Ministry of Education, Universities and Research (MIUR) refer to people enrolled in a school course from 2011 onwards. For this subset of population, information on the course that people is attending and other important characteristics are available. Unfortunately, this information is not updated, in 2019 the last available information is referred to 2017 (t-2 lag). Moreover, MIUR data trace only students enrolled in an educational course hold in Italy and they do not include qualification courses like Fine Arts, Drama, Dance and Music academic diplomas and more training and vocational careers managed by Italian Regions that are not required to provide data to MIUR. The main consequence is an underestimation of ALE in the administrative source. To estimate ALE at time *t,* an imputation/prediction procedure is adopted. It divides the people in MIUR data into two main subgroups:

- "Inactive" people: not enrolled in any course covered by MIUR in the last available academic year (i.e. *t-2*).
- "Active" people: attending a course in the last available academic year with longitudinal information on school enrolment. Information on ALE and year attendance of educational courses in previous academic years are available for all individuals.

Among "Active" people, it is possible to identify two subgroups of individuals:

- "No-change": people with a zero probability of changing their educational level from *t-2* to *t*.
- "Change": people with a non-zero probability of obtaining a higher qualification than that held in the previous year.

"No-Change" is identified by one of the following conditions:

- Attending 1st years of the course (cannot conclude the course after 1/2 years);
- Already owning the educational level of the course they are attending;
- With a PhD (highest educational level).

"Change" is composed of units that do not meet any of the above conditions. For the "No-Change", the predicted ALE at time *t* is equal to ALE in *t-1* or *t-2*.

For each unit in "Change", the probability of achieving a new qualification is estimated through log-linear models. Models are estimated on data of the previous years, given the course in which the person is enrolled and other individual characteristics, and applied to the last available information to predict the ALE at the time of interest. The underlying hypothesis is that the probability of obtaining a higher qualification does not change between subsequent years.

3 Model assessment and results

When the data time lag is passed, Istat gathers updated administrative data, so we have the opportunity to analyse the differences between the produced ALE estimates

and the updated administrative data. This allows a more precise evaluation and possibly a fine tuning of the procedure, improving the quality of data and final estimates.

The deterministic rules adopted to split the population into Change/No-change is evaluated through Classification And Regression Trees (CART).

The official ALE estimates for *t*=2019 based on *t-2* administrative information, are compared with the ones obtained with updated 2019 administrative data. This is considered as target variable in the experimentation, and variables involved in the rules to identify the No-Change population are used as covariates, specifically ALE in *t-2* (2017) and school attendance in academic year *t-2/t-1* (2017/2018).

The resulting tree shows very picked probability distributions associated to some leafs, coherently with the rules adopted. However, in some cases the resulting tree suggests the introduction of additional deterministic rules. Conditions that characterise these leafs can be evaluated as additional rules for the identification of the No-Change subgroup, specifically:

- people attending the third year of an upper secondary education
- people not enrolled in any course with a Master's degree
- people not enrolled in any course with a Lower secondary education.

The last two conditions suggest rules also for a subset of individuals that belong to the "Inactive" group for which a non-zero probability of obtaining a new educational level was assumed. Individuals with a Master's degree, who wish to improve their educational level, have to enrol in a PhD course. PhD courses are covered by MIUR. Therefore, it is intuitive that individuals with a Master's degree who are not enrolled in any course covered by MIUR can be added to the No-Change group. On the other hand, individuals with a Lower secondary education who wish to improve their educational level have different options: they can enrol in school courses that can be either covered by MIUR or not covered by MIUR.

An experiment was carried out by introducing the new deterministic rules in the 2019 data. Official estimates (Off.Est) are compared with new estimates (New.Est). Table 1. shows also estimates obtained with sample survey data (Survey) and administrative data without any treatment (Raw.admin). Administrative and sample survey data are particularly different for the Lower and Upper secondary education, confirming the problems related to administrative sources. Off.est is closer to the sample survey data estimate, while New.est is closer to the administrative distribution. The average difference between estimated and administrative data distribution goes from 6.7% of the official estimates to 2.7% of the new estimates. The main difference regards the Lower secondary education that is overestimated if we consider only administrative sources.

We observe that, to set an estimation procedure, it is always important to have in depth information on the context on which we are working keeping in mind the importance of metadata to obtain results coherent with the objective of the estimates.

Table 1. Comparison between ALE 2019 distributions obtained from Administrative data (Raw.Admin), weighted sample survey data (Survey), official estimates (Off.est) and new experimental estimates (New.est)

ALE	Raw.Admin	Survey	Off.est	New.est
Illiterate	-	0.1%	-	-
Literate but no att.	8.2%	7.2%	8.2%	8.2%
Primary education	13.9%	13.6%	13.5%	13.5%
Lower secondary	29.1%	25.9%	26.3%	29.3%
Upper secondary	30.5%	34.0%	32.7%	29.7%
Bachelor's degree	8.6%	8.5%	8.4%	8.4%
Master's degree	9.2%	10.1%	10.2%	10.3%
PhD	0.5%	0.6%	0.6%	0.5%

To conclude, we have discussed a real and important case when machine learning can be usefully applied to improve data quality by learning editing rules from data. The situation, based on longitudinal information, is quite common when using administrative data, which by now have entered the production systems of official statistics. The setting describes the case of time lagged administrative data for which, after some years, the updated information is available. The study focuses onto the learning phase of rules describing data relationships that are used to manage data. Further studies will be devoted to the use of machine learning for the imputation phase. In fact, De Fausti et al. (2022) have compared machine learning methods with the official ones, but they have not described how to deal with cases when, with a certain delay, true data are available, and the possibility of automatically learning and improving the modelling that is indeed an important feature of the machine learning approach.

References

DE FAUSTI, F., DI ZIO, M., FILIPPINI, R., TOTI, S., & ZARDETTO, D. 2022. Multilayer perceptron models for the estimation of the attained level of education in the Italian Permanent Census. *Statistical Journal of the IAOS*, (Preprint), 1-10.

DE WAAL, T., PANNEKOEK, J., & SCHOLTUS, S. 2003. *Handbook of statistical data editing and imputation* (Vol. 563). John Wiley & Sons.

DI ZIO, M., FILIPPINI R., ROCCHETTI G. 2019. An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data. *Rivista di Statistica Ufficiale*, N. 2-3/2019.

Unece 2019. Generic Statistical Business Process Model, Version 5.1, *United Nations Economic Commission for Europe*

UNCERTAINTY AND RESPONSE STYLE IN LATENT TRAIT MODELS TO ASSESS EMOTIONAL INTELLIGENCE OF ELITE SWIMMERS

Rosa Fabbricatore[1] and Maria Iannario[2]

[1] Department of Social Sciences, University of Naples Federico II, (e-mail: rosa.fabbricatore@unina.it

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

ABSTRACT: Emotional intelligence is a key factor for success in sporting competitions, arousing great interest in the psychological assessment of athletes. When the evaluation relies on Likert-type psychometric scales, individuals could tend to respond to items regardless of their content, compromising the measurement process. In this vein, the present contribution aims to address measurement issues regarding uncertainty and response style during the assessment of emotional intelligence of elite swimmers by exploiting latent trait models. Results provide evidence in favor of models accounting for response behavior.

KEYWORDS: Elite swimmers, emotional intelligence, latent trait models, response style, uncertainty

1 Introduction

Data concerning athletes' performance and their behavior are the essential core for competitive sports. Recently, an increasing interest has been devoted to understanding the psychological behaviour of some athletes and how personality traits influence their performance. Among them, emotional intelligence (EI) stands out, affecting athletes' ability to properly perceive and manage their emotions during competitions and thus allowing them to perform at their best.

From a modeling point of view, EI can be conceived as a personal latent trait that can be measured through a set of manifest indicators, such as multi-item psychometric scales. Therefore, latent variable models represent a relevant statistical framework to detect the underlying latent trait. When categorical observed variables are considered (as for Likert-type measurement scales), item response theory (IRT) models are the main reference (Bartolucci *et al.*, 2015). In particular, the Partial Credit model (PCM) is considered for the current application among the IRT models developed for polytomous items.

Nevertheless, it should be noted that individual responses to items could be affected by factors unrelated to the measured latent trait, especially when dealing with sensitive issues. For example, some works (e.g., Tutz *et al.*, 2018) highlighted different response styles during response behavior, including a tendency to select middle or extreme categories, irrespective of item content, and random answers. Other studies (e.g., Tutz & Schauberger, 2020) focused instead on response behavior driven by different degrees of uncertainty in choosing the preferred category. It is demonstrated that ignoring such subject heterogeneity may yield biased estimates of model parameters.

Herein, latent trait models that extend the PCM to account for athletes' uncertainty and response styles when responding to Likert-type scales measuring EI are analysed. In particular, the PCM Response style (PCMRS) and the Uncertainty PCM (UPCMRS) are considered. Moreover, uncertainty and the underlying trait are linked to explanatory variables concerning age, gender, and Big Five personality traits.

2 PCM with response style and uncertainty

The PCM represents the generalization of the Rasch model in the context of ordinal data. Let $Y_{ij} \in \{0,1,2,\ldots,m\}$ be the response on a Likert scale of individual i to an item j ($j \in \{1,2,\ldots,J\}$). The probability of observing a response category r can be parametrized, according to the PCM, as:

$$P(Y_{ij} = r) = \frac{\exp\left(\sum_{l=1}^{r}(\theta_i - \delta_{jl})\right)}{\sum_{s=0}^{m}\exp\left(\sum_{l=1}^{s}(\theta_i - \delta_{jl})\right)}, \quad r = 1,\ldots,m, \tag{1}$$

where θ_i is the person parameter and δ_{jl} the item-step *difficulty* parameter.

The extended PCM with response style (PCMRS; Tutz *et al.*, 2018) modifies the item-step difficulty parameter δ_{jl} to model the tendency to extreme or middle categories. In particular, the new difficulty parameter $\tilde{\delta}_{jl}$ has the form:

$$\tilde{\delta}_{jl} = \delta_{jl} - (k - l + c)\gamma_i, \tag{2}$$

where γ_i is an additional person parameter accounting for the shifting of thresholds, $k = m/2$ denotes the middle category of the response scale, and c determines the centering of the response style. For $c = 0.5$ there is symmetry around the middle category, ensuring a local Rasch model for adjacent categories. Regarding the person parameter γ_i, positive and negative values indicate a tendency to middle or extreme categories, respectively.

In the extended version of PCM accounting for uncertainty (UPCM; Tutz & Schauberger, 2020), the new predictor $\eta_{ijr} = e^{\alpha_i}(\theta_i - \delta_{jl})$ is introduced, which contains the additional subject-specific parameter α_i. The added parameter discriminates between uncertain or non-uncertain respondents, considering e^{α_i} the *uncertainty effect*. In particular, for ordered thresholds $\delta_{jr} \leq \delta_{j,r+1}$, it follows that: (i) if $\alpha_i = 0$, the classic PCM is obtained; (ii) for decreasing values of α_i, one comes closer to a uniform distribution across categories, whatever the parameter θ_i is (random responding); (iii) for increasing α_i, the selection for categories becomes very distinct depending on the value of θ_i.

3 Elite swimmers' response behavior during EI assessment

In this contribution, a practical definition of EI validated for sports was considered, whose assessment relays on 30 items with a 7-point Likert response scale (from "strongly disagree" to "strongly agree") measuring the four dimensions of well-being, sociability, emotionality, and self-control (Petrides, 2009). In what follows, only the results for the *emotionality* subscale are presented.

The study involved $n = 205$ elite swimmers enrolled in the Italian Swimming Federation, predominantly males (61%) with a mean age of 16.8 (sd = 3.6). In addition to EI, the Big Five personality traits (Extraversion, Emotional stability, Openness, Agreeableness, and Conscientiousness) were assessed.

A simple PCM and the extended version of PCMRS and UPCM were fitted to the data. The variance of the random effect for the trait parameters in the PCM was estimated to be $\sigma^2 = 0.07$. When fitting the PCMRS and UPCM without covariates the following covariance matrices resulted:

$$\hat{\Sigma}_{PCMRS} = \begin{pmatrix} 0.06 & 0.04 \\ 0.04 & 0.16 \end{pmatrix}, \qquad \hat{\Sigma}_{UPCM} = \begin{pmatrix} 0.04 & 0.03 \\ 0.03 & 0.87 \end{pmatrix}.$$

The latter report the estimate for the variance of the trait and response style (or uncertainty) effects on the main diagonal and their covariance out of the diagonal. Figure 1 displays the estimates of the item parameters for simple PCM, extended PCMRS, and UPCM. It can be seen that for all items the estimates of item thresholds differ between the considered models, especially the extreme responses. BIC indexes (6110, 5459, and 5607 for PCM, PCMRS, and UPCM, respectively) support the selection of the model with the response style component, followed by the model accounting for individual uncertainty. Regarding the UPCM model with covariates, results reported a significant effect (p-value< 0.05) of Agreeableness on uncertainty ($\beta = 0.23$) and EI ($\beta = 0.09$), and a significant effect of Conscientiousness on uncertainty ($\beta = -0.28$).

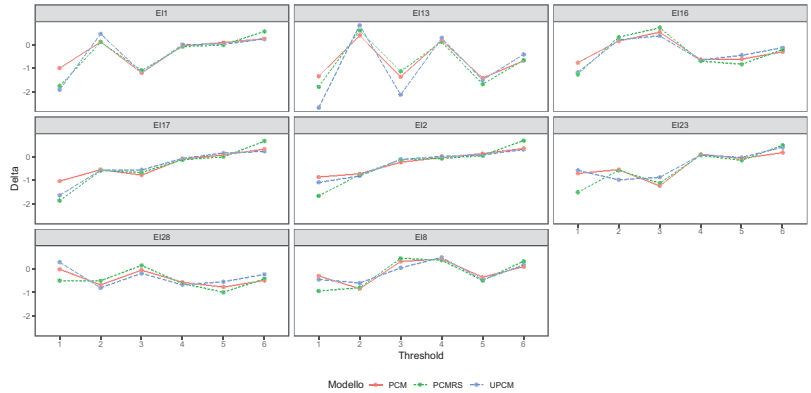


Figure 1. *Item parameter estimates for PCM, PCMRS, and UPCM.*

4 Conclusion

The study provides some evidence regarding the effect of response style and uncertainty in the assessment of the EI of swimmers. Improving the accuracy of parameter estimation by exploiting sophisticated statistical models, as those herein employed, allows for disentangling the latent trait component and the response behavior. Moreover, accounting for the effect of covariates makes it possible to identify subgroups that differ in uncertainty and the underlying trait to better promote successful factors, such as EI, in sports.

References

BARTOLUCCI, FRANCESCO, BACCI, SILVIA, & GNALDI, MICHELA. 2015. *Statistical analysis of questionnaires: A unified approach based on R and Stata*. Vol. 34. CRC press.

PETRIDES, KONSTANTINOS V. 2009. Psychometric properties of the trait emotional intelligence questionnaire (TEIQue). *Pages 85–101 of: Assessing emotional intelligence: Theory, research, and applications*. Springer.

TUTZ, GERHARD, & SCHAUBERGER, GUNTHER. 2020. Uncertainty in latent trait models. *Applied Psychological Measurement*, **44**(6), 447–464.

TUTZ, GERHARD, SCHAUBERGER, GUNTHER, & BERGER, MORITZ. 2018. Response styles in the partial credit model. *Applied Psychological Measurement*, **42**(6), 407–427.

THREE-STEP RECTANGULAR LATENT MARKOV MODELING BASED ON ML CORRECTION

Rosa Fabbricatore [1], Roberto Di Mari[2], Zsuzsa Bakk[3], Mark de Rooij[3] and Francesco Palumbo[4]

[1] Department of Social Sciences, University of Naples Federico II, (e-mail: rosa.fabbricatore@unina.it)

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

[3] Department of Methodology and Statistics, Leiden University, (e-mail: z.bakk@fsw.leidenuniv.nl, rooijm@fsw.leidenuniv.nl)

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

ABSTRACT: Rectangular latent Markov (LM) models have been recently introduced to account for different numbers of latent states over time. This contribution proposes a three-step estimation procedure for such models, which proved useful in the LM modeling framework for flexibility. Specifically, a bias-adjusted maximum likelihood (ML) estimator is introduced for the third step. A simulation study provided preliminary encouraging results regarding the efficacy and effectiveness of the method.

KEYWORDS: Rectangular LM models, three-step estimation, ML-based correction

1 Introduction

Latent Markov (LM) models represent a primary reference to study change over time in the framework of non-parametric latent variable models (Bartolucci *et al.*, 2014). Given a set of response variables repeatedly measured at different time points, LM models allow analyzing individuals' transitions across latent states over time, assuming a first-order Markov chain for the latent process. Three types of parameters characterize LM models: *initial state probabilities*, namely state proportion at the first time point; *transition probabilities*, describing the transition from one state to another at each subsequent time point; *class-conditional parameters* accounting for the relation between latent states and observed indicators. Moreover, the effect of individual covariates on initial and transition probabilities can be considered.

One-step and multi-step approaches have been proposed for model parameter estimation. Due to their flexibility and high feasibility, step-wise ap-

proaches are usually preferred in practice. Among them, a bias-adjusted three-step approach exploiting a maximum likelihood-based (ML) correction was proposed (Di Mari *et al.*, 2016).

Rectangular LM models have been recently introduced to address the issue of possible different numbers of latent states for the considered time points (Anderson *et al.*, 2019). Indeed, over time the nature and number of latent classes tend to vary; therefore, a unique overall definition of the latent classes, as typical in classical LM models, might result in either too restrictive or redundant. Currently, only a one-step estimation procedure for this model has been proposed (Anderson *et al.*, 2019). In this vein, the present contribution aims to further generalize the bias-adjusted three-step approach based on ML correction to the case of LM models with rectangular transition matrices.

The following section outlines the proposed three-step approach. Section 3 presents the simulation study carried out to obtain a preliminary evaluation of the developed estimator. Section 4 reports some conclusions.

2 Three-step rectangular LM modeling

Let $\mathbf{Y}_s^{(t)} = (Y_{s1}^{(t)}, \ldots, Y_{sK_t}^{(t)})'$ be the vector of responses for individual $s = 1, \ldots, N$ on the K_t indicators measured at time point $t = 1, \ldots, T$, with a realization $\mathbf{y}_s^{(t)}$. It is worth noting that the set and the number of indicators K_t varies over time. Denote with $X_s^{(t)}$ the categorical latent variable at time t taking value $i = 1, \ldots, I_t$, producing rectangular transition matrices wherever $I_{t-1} \neq I_t$.

In Step 1, the measurement part of the model is estimated for each time point exploiting a latent class model. This step connects the latent states $i = 1, \ldots, I_t$ to the response variables $\mathbf{Y}_s^{(t)}$, providing for each individual s and time t, the posterior class probability $P(X_s^{(t)} = i | \mathbf{Y}_s^{(t)} = \mathbf{y}_s^{(t)})$. In Step 2, state membership $W_s^{(t)}$ is obtained according to the modal assignment rule, namely allocating individuals in the class for which they present the largest posterior probability. Accordingly, the classification error probabilities included in the time-specific $\mathbf{D}^{(t)}$ matrix are defined as the conditional probability of the estimated class value conditional on the true one $P(W_s^{(t)} = g | X_s^{(t)} = i)$, with $g, i = 1, \ldots, I_t$. In Step 3, a rectangular LM model is estimated with the vector of class assignments $\mathbf{W}_s = (W_s^{(1)}, \ldots, W_s^{(T)})$ as single indicators and known error probabilities included in the $\mathbf{D}^{(t)}$ matrices. Keeping out of consideration the effect of covariates, the third-step log-likelihood is $\ell(\eta) = \sum_{s=1}^{N} \log\{P(\mathbf{W}_s)\}$, where η is the vector of free model parameters. The probability $P(\mathbf{W}_s)$ can be

expressed for rectangular transition matrices as

$$P(\mathbf{W}_s) = \sum_{i^{(1)}=1}^{I_1} \sum_{i^{(2)}=1}^{I_2} \cdots \sum_{i^{(T)}=1}^{I_T} P\big(X_s^{(1)} = i^{(1)}\big) \prod_{t=2}^{T} P\big(X_s^{(t)} = i^{(t)} | X_s^{(t-1)} = i^{(t-1)}\big) \prod_{t=1}^{T} P\big(W_s^{(t)} = g^{(t)} | X_s^{(t)} = i^{(t)}\big),$$

where the state-dependent distributions (given by classification errors) are considered fixed parameters, and thus they are not estimated.

A generalization of the Baum–Welch algorithm (Rabiner, 1989) for rectangular LM, which exploits forward and backward probabilities during estimation, was implemented in the R statistical software. The proposed estimator allows for both time-varying and time-invariant measurement models.

3 Simulation study for the developed bias-adjusted estimator

A simulation study was carried out to evaluate the performance of the bias-adjusted maximum likelihood estimator. Different scenarios were considered, mainly concerning class separation and sample size. In particular, three simple latent class models (one per time point) with 3-3-2 latent classes were considered for the measurement part of the model. Class separation was modeled through response probabilities to ten dichotomously-scored items, setting a probability of 0.8 and 0.9 for the most likely responses in the case of moderate and large class separation, respectively. Four sample sizes were considered: 200, 500, 2000, and 10000 observations. Finally, equal size was imposed for initial probabilities and persistent Markov chains for transition probabilities. For each condition, 500 replications were carried out. The bias in the model parameters estimates (initial and transition probabilities) was used to compare the estimator's performance under different conditions.

The results support the overall good performance of the proposed third-step bias-adjusted estimator. The data log-likelihood increases monotonically according to the number of iterations and the algorithm reaches convergence within 20 iterations. The variability of the estimated bias distribution for both initial and transition probabilities becomes smaller as class separation and sample size increase. Figure 1 shows an example of the estimated bias for the transition matrix from Time 2 to Time 3, with $\gamma_{tij} = log\frac{P(X_s^{(t)}=i|X_s^{(t-1)}=j)}{P(X_s^{(t)}=j|X_s^{(t-1)}=j)}$. Note that more accuracy for initial probabilities estimates emerged, which reported an average bias close to 0 in all the considered conditions. Conversely, as the

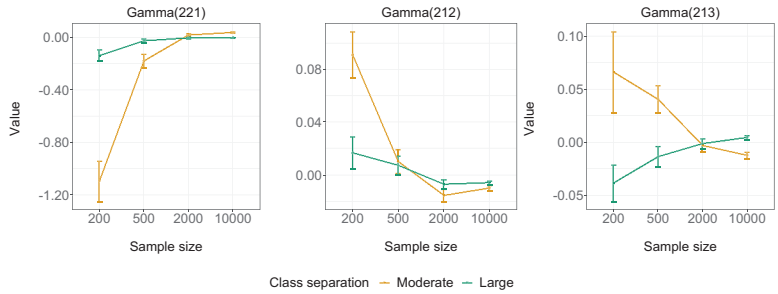


Figure 1. *Mean and standard error of transition probabilities bias (T2 to T3).*

figure shows, a small sample size ($n = 200$) strictly affects transition probabilities estimation due to the presence of very small probabilities in the transition matrix cells that can easily end up in an estimate close to the boundary. Of course, this rarely happens with large samples.

4 Conclusion

A bias-adjusted three-step rectangular LM modeling approach was proposed. In particular, a new estimator for an ML-based correction was developed for the third step. The proposed estimator proved to perform well asymptotically, with a larger estimation bias for small samples and lower class separation. Current developments aim at also considering the covariates' effect on initial and transition probabilities. Empirical applications could provide further insights into the practical advantages of the proposed method.

References

ANDERSON, G., FARCOMENI, A., PITTAU, M. G., & ZELLI, R. 2019. Rectangular latent Markov models for time-specific clustering, with an analysis of the well-being of nations. *J Roy Stat Soc C-App*, **68**, 603–621.

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2014. Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test*, **23**, 433–465.

DI MARI, R., OBERSKI, D. L., & VERMUNT, J. K. 2016. Bias-adjusted three-step latent Markov modeling with covariates. *Struct Equ Modeling*, **23**(5), 649–660.

RABINER, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc of the IEEE*, **77**(2), 257–286.

MID-QUANTILE REGRESSION FOR DISCRETE PANEL DATA

Alessio Farcomeni[1], Alfonso Russo[1] and Marco Geraci[2]

[1] Department of Economics and Finance, University of Rome "Tor Vergata", (e-mail: alessio.farcomeni@uniroma2.it, alfonso.russo@uniroma2.it)

[2] Department of Methods and Models for Economics, Territory and Finance, Sapienza - University of Rome (e-mail: marco.geraci@uniroma1.it)

ABSTRACT: We propose a novel method for quantile regression for discrete longitudinal data. The approach is based on the notion of conditional mid-quantiles, which have good theoretical properties even in the presence of ties, and a Ridge-type penalised framework to accommodate dependent data. We illustrate the methods with a simulation study and an original application to the use of macroprudential policies in more than one hundred countries over a period of fifteen years.

KEYWORDS: mid-quantile regression, macroprudential policies, fixed effects, random effects, longitudinal data

1 Introduction

Quantile Regression (QR) involves modeling effects of predictors at specific quantiles of an endogenous variable. Most QR methodologies are restricted to continuous outcomes, with some notable exceptions (Machado & Santos Silva, 2005; Frumento & Salvati, 2021). Recently, Geraci & Farcomeni, 2022 proposed a method based on conditional mid-quantiles (see also Ma *et al.*, 2011). We extend their approach to the case of discrete panel data. Our approach can also be seen as an extension to discrete outcomes of the penalised framework for QR for continuous panel data (Koenker, 2004). We develop a collect of methods that are based on a two-step algorithm. At the first step, the conditional mid-quantile function is estimated through a semiparametric approach; at the second step we optimise a possibly penalised objective function to obtain parameter estimates. We illustrate the methods by means of a simulation study, and an original application to macroprudential policies in a panel of countries.

2 Penalized mid-quantile regression

Let $y_{it}, t = 1,\dots,T_i$ and $i = 1,\dots,n$, denote a discrete/ordered outcome and $\tilde{x}_{it}$ an associated vector of covariates. Measurements are repeatedly taken $T_i \geq 1$ times for each unit, with $T_i > 1$ at least for one subject. We define the conditional mid-CDF of Y as $G_{Y|X}(y|x) = F_{Y|X}(y|x) - 0.5 \cdot \mathrm{m}_{Y|X}(y|x)$ and $G^C_{Y|X}(y|x)$ the continuous function that interpolates $G_{Y|X}(y|x)$, where $F_{Y|X}(y|x) = \mathbb{P}(Y \leq y|X = x)$ and $\mathrm{m}_{Y|X}(y|x) = \mathbb{P}(Y = y|X = x)$. Let $p \in (0,1)$. The conditional mid-quantile function is the generalised inverse $H_{Y|X}(p) = G^{-1}_{Y|X}(y|x)$.

We assume a p-specific model that is linear on the scale of a link function $h(\cdot)$:

$$h\{\eta_{it}(p)\} = \alpha_i(p) + \tilde{x}^T_{it}\beta(p) = H_{h(Y)|X}(p) \tag{1}$$

Estimation, as in Geraci & Farcomeni, 2022, proceeds in two steps. In the first step, one obtains estimates of the conditional mid-CDF. Similarly to Peracchi, 2002, we define outcome variables $1\{y_{it} \leq c\}$ at appropriate cut-points. We then estimate logistic regression models with either (1) fixed subject-specific, (2) random subject-specific, or (3) homogeneous intercepts. For a fixed penalty $\lambda > 0$, our objective function for the second step is given by

$$\psi_n[\theta(p); p] = \sum_{i=1}^{n}\sum_{t=1}^{T}\left\{p - \hat{G}^c_{Y|X}(\eta_{it}|\tilde{x}_{it})\right\}^2 + \lambda\sum_{i=1}^{n}\alpha_i^2. \tag{2}$$

The optimum is available in closed form as a Ridge-type estimator. For selection of the penalty parameter we use an heuristic reasoning as in Ruppert *et al.*, 2003. As long as $\min_i T_i > 1$ it is possible also to set $\lambda = 0$; and it is also possible to set $\lambda \to \infty$, therefore obtaining homogenenous intercepts $\alpha_i = \alpha$. In summary we are proposing three possible routes for estimation of the conditional CDF and three possible routes for the second step. The case with homogeneous intercepts at the first step and $\lambda \to \infty$ recovers the methodology in Geraci & Farcomeni, 2022.

3 Simulation study

In Figure 1 we show mean squared error (500 replicates) for regression coefficient estimates for ten alternative model specifications, reported by quantiles $(0.2, 0.5, 0.8)$ and two sample sizes. The first nine model specifications involve our proposed class, where at the first step intercepts can be homogenous (*HMG*), treated as fixed (*FE*), or random (*RE*). Each specification is

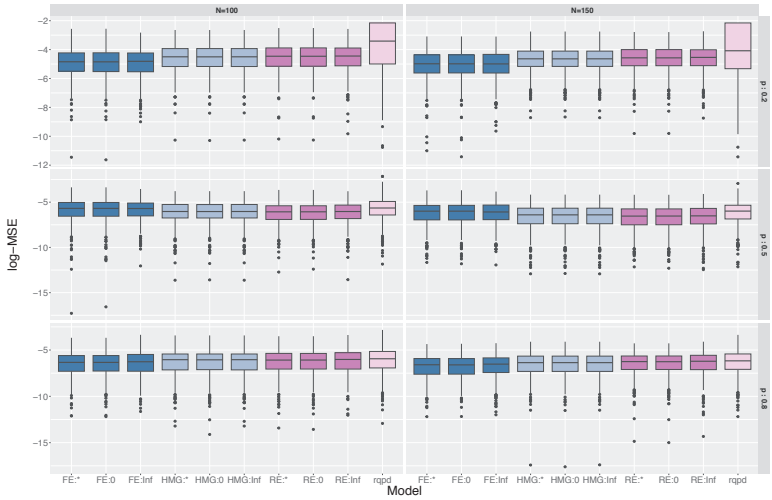


Figure 1. *Log Mean Squared Error of parameter estimates for a Poisson response with two continuous covariates.*

paired, at the second step, with three different choices for the penalty parameter $\lambda \to 0, \lambda \to \infty$, or $\lambda = \lambda^*$. Finally *rqpd* denotes the quantile regression procedure in Koenker, 2004. At the data generation stage we simulate Poisson responses with two Gaussian covariates. Several other settings are available in the accompanying paper. The general conclusions that can be drawn are that (i) our method outperforms *rqpd*, which does not take into account the discrete nature of the outcome, and (ii) the MSE decreases at the expected rate.

4 Real data example

Macroprudential policies (MP) (Galati & Moessner, 2013) are used by central banks to protect macroeconomic performance from the drawbacks of externalities, market failures, excessive procyclicality and other factors. They involve currency instruments, limits to bank exposure, and similar requirements. In this work our focus is on the determinants of the use of MP. Our endogenous variable is the number (up to twelve) of different MP used by a country in a given year. We collect data on a panel of $n = 115$ countries over $T = 18$ years starting from 2001. Predictors include World Bank label for the economy, debt to gdp ratio, unemployment rate, trade as % of GDP. All covariates are lagged by one year.

Results for optimal model specification selected through 10-fold cross val-

Table 1. *Macroprudential policy determinants in 115 countries from 2001 to 2017. Parameter estimates (95% CI in parenthesis) at different quantiles p.*

	$p = 0.2$	$p = 0.5$	$p = 0.8$	p
Trade-to-GDP	$0.09(0.05, 0.12)$	$0.06(0.03, 0.09)$	$0.06(0.03, 0.09)$	$0.05(0$
Unempl.	$-0.03(-0.06, -0.01)$	$-0.03(-0.05, -0.00)$	$-0.02(-0.04, 0.00)$	$-0.02(-$
Debt-to-GDP	$0.03(0.01, 0.05)$	$0.02(0.00, 0.04)$	$0.02(0.01, 0.04)$	$0.03(0$
High income	$0.31(0.24, 0.37)$	$0.38(0.32, 0.44)$	$0.31(0.26, 0.36)$	$0.30(0$
Up-Mid Income	$0.54(0.46, 0.61)$	$0.60(0.53, 0.66)$	$0.47(0.42, 0.53)$	$0.45(0$
Low-Mid Income	$0.29(0.23, 0.35)$	$0.34(0.29, 0.40)$	$0.28(0.23, 0.32)$	$0.26(0$
Time	$0.05(0.04, 0.05)$	$0.04(0.04, 0.05)$	$0.04(0.04, 0.04)$	$0.04(0$

idation are reported in Table 1. Consistently with the literature upper-middle income countries tend to use more MP. Effects are quantile-dependent, with high trade-to-GDP and debt-to-GDP prompting larger use at low quantiles.

References

FRUMENTO, P., & SALVATI, N. 2021. Parametric modeling of quantile regression coefficient functions with count data. *Statistical Methods & Applications*, **30**, 1237–1258.

GALATI, G., & MOESSNER, R. 2013. Macroprudential Policy - a literature review. *Journal of Economic Surveys*, **27**, 846–878.

GERACI, M, & FARCOMENI, A. 2022. Mid-quantile regression for discrete responses. *Statistical Methods in Medical Research*, **31**(5), 821–838.

KOENKER, R. 2004. Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91**(1), 74–89.

MA, Y., GENTON, M. G., & PARZEN, E. 2011. Asymptotic properties of sample quantiles of discrete distributions. *Annals of the Institute of Statistical Mathematics*, **63**(2), 227–243.

MACHADO, J. A. F., & SANTOS SILVA, J. M. C. 2005. Quantiles for counts. *Journal of the American Statistical Association*, **100**(472), 1226–1237.

PERACCHI, F. 2002. On estimating conditional quantiles and distribution functions. *Computational Statistics & Data Analysis*, **38**(4), 433–447.

RUPPERT, D., WAND, M. P., & CARROLL, R. J. 2003. *Semiparametric regression*. Cambridge university press.

TRIMMED FACTORIAL K-MEANS

Matteo Farnè [1]

ABSTRACT: This paper provides the definition of trimmed factorial k-means (TFKM) algorithm. TFKM is a robust version of factorial k-means, where a robust covariance matrix input is used, and outliers in the identified reduced space are iteratively removed via a trimming procedure. The selected latent rank, number of clusters and outlier proportion are those which maximize Hartigan's statistic.

KEYWORDS: robust clustering, dimension reduction, factorial k-means, trimming

1 Introduction

Clustering high-dimensional data with many objects is a challenging task for several reasons. First, a high dimension and a large sample size make agglomerative hierarchical methods like Ward's one (Ward, 1963) computationally intractable. Second, hierarchical partitioning methods like k-means algorithm (MacQueen, 1967) may become very unstable in high dimensions, due to numerical instability and multicollinearity. Third, any non-robust methodology applied to a large dataset is likely to be affected by outliers, so that there is the need to develop and apply robust versions of traditional methods to prevent the identification of uninformative partitions, like trimmed k-means (TKM) (Cuesta-Albertos *et al.* , 1997).

In order to approach dimension reduction, Vichi & Kiers, 2001 proposed factorial k-means (FKM), a method to identify the latent space most able to maximize the distinctiveness of projected objects. The strong consistency of FKM was proved in Terada, 2015. In this paper, we present a robust version of factorial k-means, named trimmed factorial k-means (TFKM), where outliers are iteratively removed in the reduced space, thus simultaneously identifying radial outliers and designing better shaped clusters. This is obtained by minimising the trimmed least squares criterion in the reduced space. A preliminary version of TFKM was first described in Farnè & Vouldis, 2021. Here, we employ MCD (Minimum Covariance Determinant, see Rousseeuw & Driessen, 1999) or ROBPCA (Hubert *et al.* , 2005) to robustly estimate the input covariance matrix, and we then iteratively apply the trimming procedure to estimated factor scores.

[1] [0] Department of Statistical Sciences, University of Bologna, (e-mail: matteo.farne@unibo.it)

2 Trimmed factorial k-means algorithm

Let us consider a $n \times p$ data matrix $\mathbf{X}$. The trimmed factorial k-means of Vichi & Kiers, 2001 assumes that

$$\mathbf{XAA}' = \mathbf{U}\overline{\mathbf{Y}}\mathbf{A}' + \mathbf{E}, \tag{1}$$

where $\mathbf{A}$ is a $p \times r$ semi-orthogonal *coefficient matrix*, such that $\mathbf{A}'\mathbf{A} = \mathbf{I}_r$; $\mathbf{U}$ is a $n \times c$ *membership matrix* such that $\mathbf{U}_{ij} = 1$, $i = 1,\ldots,n$, $j = 1,\ldots,c$, if observation i belongs to cluster j; $\overline{\mathbf{Y}}$ is a $c \times r$ *centroid matrix*; r is the *latent rank* and c is the *number of clusters*. Model (1) assumes that the variable space is approximately isomorphic to a latent linear space, spanned by the same variables, on which the projected data vectors are maximally apart. It is recovered by minimizing $FKM(\mathbf{A},\mathbf{U},\overline{\mathbf{Y}}) = \|\mathbf{XAA}' - \mathbf{U}\overline{\mathbf{Y}}\mathbf{A}'\|^2 = \|\mathbf{XA} - \mathbf{U}\overline{\mathbf{Y}}\|^2$, which is the deviance within clusters in the reduced space, where by least squares we can obtain $\overline{\mathbf{Y}} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{XA}$.

Denoting the $n \times r$ factor score matrix by $\mathbf{F} = \mathbf{XA}$, in this paper we assume that $(100\alpha)\%$ of the n true factor score vectors, with $\alpha \in [0, 0.5]$, are arbitrarily distant from the bulk of the rest of factor score vectors. Therefore, in this situation it is appropriate to minimize $FKM(\mathbf{A},\mathbf{U},\overline{\mathbf{Y}})$ under the constraint $\sum_{i=1}^{n}\sum_{j=1}^{c}\mathbf{U}_{ij} = [(1-\alpha)n]$, with $\sum_{j=1}^{c}\mathbf{U}_{ij} = \{0,1\}$, for each $i = 1,\ldots,n$. This problem can be numerically solved by adapting the original Alternated Least Squares (ALS) algorithm of Vichi & Kiers, 2001 to the framework of Rousseeuw & Van Driessen, 2000 (see also Farnè & Vouldis, 2021). In particular, once initialized, $\mathbf{A}$, $\mathbf{U}$, and $\overline{\mathbf{Y}}$ are first recovered by the original ALS algorithm, which is the H-step, and a trimming procedure is subsequently applied by excluding the $[\alpha n]$ observations most apart from the respective cluster centroids in the reduced space, which is the C-step.

The algorithm input is the Minimum Covariance Determinant (MCD) covariance matrix estimate, if $n \geq 2p$, or the ROBPCA-based reduced covariance matrix with fixed rank $p/10$, otherwise. We call the algorithm input $\mathbf{C}$. We then fix the latent rank r, the number of clusters c, and the outlier proportion α, and we apply the following procedure.

- **Step 0**. We derive the best r-ranked approximation of $\mathbf{C}$ as $\mathbf{C}_r = \mathbf{V}_r\mathbf{D}_r\mathbf{V}_r'$ by extracting the top r principal components of $\mathbf{C}$. We generate a permutation square matrix of size p, $\mathbf{P}$, we orthogonalize it by Gram-Schmidt algorithm, getting $\tilde{\mathbf{P}}$, and we obtain the initial coefficient matrix as $\mathbf{A}_0 = \tilde{\mathbf{P}}\mathbf{V}_r$. Then, we calculate $\mathbf{F}_0 = \mathbf{XA}_0$, the mean factor score $\overline{\mathbf{F}}_0$, and the distances $\mathbf{d}_{i,0} = \mathbf{F}_{i,0} - \overline{\mathbf{F}}_0$, for $i = 1,\ldots,n$. We derive for each i a T-score

as follows: $T_{i,0} = n\mathbf{d}'_{i,0}\mathbf{C}^{-1}_{F,0}\mathbf{d}_{i,0}$, where $\mathbf{C}_{F,0}$ is the $r \times r$ covariance matrix of $\mathbf{F}_0$. Then, we calculate the $2c$ quantiles of $T_{i,0}$, and we allocate each object to the closest quantile among the first, the third, ..., the $(c-1)$-th. We thus obtain the initial membership matrix $\mathbf{U}_0$, and the initial centroid matrix $\overline{\mathbf{Y}}_0 = (\mathbf{U}'_0\mathbf{U}_0)^{-1}\mathbf{U}'_0\mathbf{X}\mathbf{A}_0$. We set $k = 1$, and we proceed as follows.

- **Step 1**. We minimize $FKM(\mathbf{A}_{k-1}, \mathbf{U}_k, \overline{\mathbf{Y}}_{k-1})$ with respect to $\mathbf{U}_k$ given the values of $\mathbf{A}_{k-1}$ and $\overline{\mathbf{Y}}_{k-1}$. For each row i of $\mathbf{U}_k$, we first impose for each $\nu = 1, \ldots, c$ that $\mathbf{U}_{i\nu,k} = 1$, and we then set $\mathbf{U}_{ij,k} = 1$ if and only if

$$\arg \min_{\nu=1,\ldots,c} FKM(\mathbf{A}_{k-1}, \mathbf{U}_{i\nu,k}, \overline{\mathbf{Y}}_{k-1}) = j.$$

- **Step 2**. We calculate $\mathbf{F}_{k-1} = \mathbf{X}\mathbf{A}_{k-1}$, and the distances $\mathbf{d}_{i,k} = \mathbf{F}_{i,k-1} - \overline{\mathbf{Y}}_{l_i,k-1}$, where l_i is s.t. $\mathbf{U}_{il_i,k} = 1$. Then, we derive for each object a T-score as follows: $T_{i,k} = n\mathbf{d}'_{i,k}\mathbf{C}^{-1}_{F,k-1}\mathbf{d}_{i,k}$, $i = 1, \ldots, n$, where $\mathbf{C}_{F,k-1}$ is the $r \times r$ covariance matrix of $\mathbf{F}_{k-1}$. At this stage, we derive the $(1-\alpha)$-quantile of $\mathbf{T}_k$, $T_{1-\alpha,k}$, and we set $\mathbf{U}_{il_i,k} = 0$ if $T_{i,k} > T_{1-\alpha,k}$.
- **Step 3**. $FKM(\mathbf{A}_k, \mathbf{U}_k, \overline{\mathbf{Y}}_k)$ is minimized keeping fixed $\mathbf{U}_k$, to jointly update $\mathbf{A}_k$ and $\overline{\mathbf{Y}}_k$. Among all the linear combinations of $\mathbf{X}$, the ones closer to the centroids (in the transformed space) are derived by taking the first r eigenvectors of $\mathbf{X}'(\mathbf{U}_k(\mathbf{U}'_k\mathbf{U}_k)^{-1}\mathbf{U}'_k - \mathbf{I}_n)\mathbf{X}$ (see Ten Berge, 1993). Based on the optimal $\mathbf{A}_k$, we can then update $\overline{\mathbf{Y}}_k$ using the expression $(\mathbf{U}'_k\mathbf{U}_k)^{-1}\mathbf{U}'_k\mathbf{X}\mathbf{A}_k$.
- **Step 4** $FKM(\mathbf{A}_k, \mathbf{U}_k, \overline{\mathbf{Y}}_k)$ is computed for the current values of $\mathbf{U}_k$, $\mathbf{A}_k$, and $\overline{\mathbf{Y}}_k$. If $FKM(\mathbf{A}_k, \mathbf{U}_k, \overline{\mathbf{Y}}_k) < FKM(\mathbf{A}_{k-1}, \mathbf{U}_{k-1}, \overline{\mathbf{Y}}_{k-1})$, we increase k by 1 and we go again with Steps 1, 2 and 3. Otherwise, the process has converged, we set $k^* = k-1$ and we retain as solutions $\mathbf{U}_{k^*}$, $\mathbf{A}_{k^*}$, and $\overline{\mathbf{Y}}_{k^*}$.

The reported algorithm is repeated $N = 1000$ times, and the final solution is chosen as the one with minimum objective $FKM(\mathbf{A}_{k^*}, \mathbf{U}_{k^*}, \overline{\mathbf{Y}}_{k^*})$ across the N trials.

A grid of possible values for the latent rank r, the number of clusters c and the outlier proportion α is specified. Given that $\overline{\mathbf{Y}} = (\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}$, and $\text{rk}((\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X}\mathbf{A}) = \min(c-1, r)$, we cannot explore any combination violating the condition $r \leq c-1$, to avoid singularity in the reduced space. We denote the solutions for each triple (r, c, α) as $\mathbf{U}(r, c, \alpha)$, $\mathbf{A}(r, c, \alpha)$, $\overline{\mathbf{Y}}(r, c, \alpha)$, obtained under the constraint $\sum_{i=1}^{n}\sum_{j=1}^{c}\mathbf{U}_{ij} = [(1-\alpha)n]$, with $\sum_{j=1}^{c}\mathbf{U}_{ij} = \{0, 1\}$, for each $i = 1, \ldots, n$. The optimal values of r, c, and α are then identified by employing Hartigan's statistic (1975), which can be obtained as follows.

First, within clusters deviance is computed for each triple (r, c, α) as $W(r, c, \alpha) = \sum_{i=1}^{n} \|\mathbf{d}_i\|$, where $\mathbf{d}_i = \mathbf{F}_i(r, c, \alpha) - \overline{\mathbf{Y}}_{l_i}(r, c, \alpha)$, l_i is such that

$\mathbf{U}_{il_i}(r,c,\alpha) = 1$, $\mathbf{F}(r,c,\alpha) = \mathbf{XA}(r,c,\alpha)$,
$\overline{\mathbf{Y}}(r,c,\alpha) = (\mathbf{U}(r,c,\alpha)'\mathbf{U}(r,c,\alpha))^{-1}\mathbf{U}(r,c,\alpha)'\mathbf{XA}(r,c,\alpha)$. Second, Hartigan's statistic $H(r,c,\alpha)$ is obtained as

$$H(r,c,\alpha) = (p-c-1)\left(\frac{W(r,c,\alpha)}{W(r,c-1,\alpha)} - 1\right).$$

Finally, we select the triple (r^*,c^*,α^*) returning the maximum $H(r,c,\alpha)$ across selected grid values.

References

CUESTA-ALBERTOS, J., GORDALIZA, A., & MATRÁN, C. 1997. Trimmed k -means: an attempt to robustify quantizers. *The Annals Of Statistics*, **25**, 553–576.

FARNÈ, M., & VOULDIS, A. 2021. Banks' business models in the euro area: a cluster analysis in high dimensions. *Annals Of Operations Research*, **305**, 23–57.

HARTIGAN, J. 1975. Clustering algorithms. *John Wiley & Sons*.

HUBERT, M., ROUSSEEUW, P., & VANDEN BRANDEN, K. 2005. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, **47**, 64–79.

MACQUEEN, J. 1967. Classification and analysis of multivariate observations. *5th Berkeley Symp. Math. Statist. Probability*, 281–297.

ROUSSEEUW, P., & DRIESSEN, K. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

ROUSSEEUW, P., & VAN DRIESSEN, K. 2000. An algorithm for positive-breakdown regression based on concentration steps. *Data Analysis*, 335–346.

TEN BERGE, J. 1993. Least squares optimization in multivariate analysis. *Leiden University Leiden*.

TERADA, Y. 2015. Strong consistency of factorial k-means clustering. *Annals Of The Institute Of Statistical Mathematics*, **67**, 335–357.

VICHI, M., & KIERS, H. 2001. Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, **37**, 49–64.

WARD, J. 1963. Hierarchical grouping to optimize an objective function. *Journal Of The American Statistical Association*, **58**, 236–244.

ESTIMATION OF TEAM'S STRENGTH FOR HANDBALL GAMES PREDICTIONS

Florian Felice[1,2], Christophe Ley[1]

[1] Department of Mathematics, University of Luxembourg, (e-mail: florian.felice@uni.lu, christophe.ley@uni.lu)

[2] Work not related to Amazon

ABSTRACT: In this work, we present Statistically Enhanced Learning (SEL), a general feature extraction approach to improve any type of learning technique, whether it is statistical or machine learning. By adding highly informative covariates, which are obtained as statistical estimates rather than directly observed, SEL improves model learning for any type of data (tabular, computer vision, text). We will discuss the general idea and refer to existing feature extraction methods that actually can be shown to fall under the umbrella of SEL. In particular, we will see how SEL allows improved predictions of handball tournaments and discuss how it can be used to derive a metric for teams' or players' strengths.

KEYWORDS: Statistically Enhanced Learning, Feature extraction, Handball, Team's strength

1 Introduction

Statistically Enhanced Learning (Felice *et al.*, 2023) is a framework that aims to formalize the feature extraction step of the data processing in a machine learning project. Classified in different categories, SEL approaches can include proxy variables (Wooldridge, 2009) as well as statistical features that represent non-measurable quantities (Groll *et al.*, 2019). In particular, in sports predictions, factors such as the strength of the opposing teams are crucial elements but can not be measured objectively. Ley *et al.*, 2019 proposed a bivariate Poisson model to represent the outcome of football games. They estimate the location parameter λ for each team via Maximum Likelihood Estimation approach. They assume that one can derive the ability of the opposing teams using the formula $\lambda = \beta_0 + r_i - r_j$ where β_0 is a constant intercept, r_i and r_j are the abilities for the home and away teams. These parameters are later included in the training data set.

In the context of handball, Groll *et al.*, 2020 analyzed historical games to

determine the best probability distribution to model the number of goals scored in handball matches. Given the level of under-disperson observed, they concluded that a Gaussian distribution with low variance is the most appropriate.

In the following, we will extend the work done by Groll *et al.*, 2020 and consider an additional discrete probability distribution. From this distribution, we will generate a metric representing the strength of a team. This metric can then be considered as a new SEL variable to be added in a training set to predict the outcome of handball games.

2 Modelling handball games with Conway-Maxwell-Poisson

As a fast-paced sport, handball can record a large number of goals during a 60 minute game (on average 27.9 for women and 29.8 for men). To model the number of goals scored, the traditional Poisson distribution assumes equi-distribution (i.e. $\mathbb{E}(X) = \mathbb{V}(X)$), however, historical data rarely satisfy this assumption.
Therefore, we compare here different distributions: the Gaussian and Negative Binomial distributions (as in Groll *et al.*, 2020) and the Conway-Maxwell-Poisson distribution (Sellers, 2023). The latter is a generalization of the common Poisson distribution, which can handle under- and over-dispersion.

Table 1: Comparison of log-likelihood evaluated on scored goals by Metz handball over season 2022/2023.

Distribution	Log-likelihood	AIC
Conway-Maxwell-Poisson	-127.66	259.31
Gaussian	-127.39	258.78
Negative Binomial	-127.36	258.72

As we can observe in Table 1, the three distributions seem to equivalently fit our data. However, we can notice the slight superiority of the Conway-Maxwell-Poisson distribution. Thus, given the continuous nature of the Gaussian distribution, we decide to discard it. Indeed, it can return non-integer values but, more problematically, it is defined on the real line which includes negative values. Furthermore, in our experiments, the Conway-Maxwell-Poisson distribution consistently showed superiority over the Negative Binomial.
As a result, and considering its flexibility to handle any of the under-, over-, and equi-dispersion, we consider the Conway-Maxwell-Poisson distribution for the rest of this work.

3 Estimation of team's strength

Adopting the selected Conway-Maxwell-Poisson (CMP) distribution, we use its parameters to represent the strength of a team both for attack and defense. With the distribution of scored goals following a CMP distribution, $Y_a \sim CMP(\lambda_a, \nu_a)$, the parameter $\lambda_a > 0$ can act as a location parameter and $\nu_a \geq 0$ as the dispersion parameter. We define the attack strength of a team as:

$$s_a = \frac{\log(\lambda_a)}{\nu_a}. \tag{1}$$

We want to penalize for irregular performances, hence we use ν_a as the denominator so the higher the irregularities the lower the attack strength score.
Similar to attack, the distribution of goals conceded by a team follows a CMP distribution, $Y_d \sim CMP(\lambda_d, \nu_d)$. However, the strength of a team's defense is inversely proportional to the goals it concedes. Thus, we define the defense strength s_d as:

$$s_d = \frac{\nu_d}{\log(\lambda_d)}. \tag{2}$$

A team is considered strong when it can perform well in attack and defense. We can consider the overall strength of a team as the product of attack and defense strengths, formally:

$$s = s_a \cdot s_d = \frac{\log(\lambda_a) \cdot \nu_d}{\nu_a \cdot \log(\lambda_d)}. \tag{3}$$

Empirically, we illustrate these results in Table 2 with European female clubs. We can observe that the teams considered the strongest are the leading clubs in their country and strong competitors in the European Champions League.

4 Conclusion

Using the parameters from the fitted Conway-Maxwell-Poisson distribution, we can estimate the strength of a team. The estimated parameters constitute new SEL variables to be added to the training set for match predictions (Felice, 2023). These estimations can also be applied, with a similar logic, to player's performance and derive the player's strength. Our results and conclusions are derived from men and women club's data, but we note that they also apply to national teams for international competitions.

Table 2: Top 10 strongest female teams in Europe for season 2022/2023.

Team	Avg. scored	Avg. conceded	Attack strength	Defense strength	Strength
Győri Audi ETO KC	33.32	24.32	3.49	3.16	11.00
Vipers Kristiansand	37.62	26.38	3.57	3.07	10.96
Podravka Vegeta	30.50	21.75	3.39	3.21	10.89
Metz handball	33.58	24.00	3.47	3.12	10.85
Team Esbjerg	33.33	24.67	3.48	3.11	10.83
SG BBM Bietigheim	34.63	25.21	3.54	3.05	10.80
HC Dunajskà Streda	29.37	22.62	3.38	3.15	10.63
Herning-Ikast Håndbold	28.71	23.29	3.39	3.13	10.61
DVSC Schaeffler	30.83	24.11	3.39	3.12	10.59
CSM București	33.13	25.83	3.48	3.05	10.58

References

FELICE, FLORIAN. 2023. Prediction of Handball Games with Statistically Enhanced learning via Estimated Teams' Strength. *arXiv preprint*.

FELICE, FLORIAN, LEY, CHRISTOPHE, GROLL, ANDREAS, & BORDAS, STEPHANE. 2023. Statistically Enhanced Learning - a Formalization Framework of Feature Extraction Techniques. *arXiv preprint*.

GROLL, ANDREAS, LEY, CHRISTOPHE, SCHAUBERGER, GUNTHER, & VAN EETVELDE, HANS. 2019. A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, **15**(4), 271–287.

GROLL, ANDREAS, HEINER, JONAS, SCHAUBERGER, GUNTHER, & UHRMEISTER, JÖRN. 2020. Prediction of the 2019 IHF World Men's Handball Championship–A sparse Gaussian approximation model. *Journal of Sports Analytics*, **6**(3), 187–197.

LEY, CHRISTOPHE, VAN DE WIELE, TOM, & VAN EETVELDE, HANS. 2019. Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, **19**(1), 55–73.

SELLERS, KIMBERLY F. 2023. *The Conway-Maxwell-Poisson Distribution*. Cambridge University Press.

WOOLDRIDGE, JEFFREY M. 2009. On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, **104**(3), 112–114.

OUTLIER EXPLANATION BASED ON SHAPLEY VALUES FOR VECTOR- AND MATRIX-VALUED OBSERVATIONS

Peter Filzmoser [1] and Marcus Mayrhofer[1]

[1] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Austria, (e-mail: Peter.Filzmoser@tuwien.ac.at, Marcus.Mayrhofer@tuwien.ac.at)

ABSTRACT: Shapley values are a practical tool from Explainable AI used to interpret model outcomes on the observation level. Their usefulness has also been demonstrated in the context of multivariate outlier detection, where the contributions of single variables to the overall outlyingness are evaluated. This allows for an alternative view to cellwise outlyingness, where the interest is in identifying deviating cells of a data matrix. The concept of outlier explanation based on Shapley values can be extended to outlyingness for matrix-valued observations, which is an interesting new topic in robustness by itself.

KEYWORDS: Anomaly explanation, Shapley value, Mahalanobis distance.

1 Shapley Values for Vector-valued Observations

Shapley values have been introduced in cooperative game theory, where they evaluate the collective payoff of a coalition of players (Shapley, 1953). In the context of multivariate data, each observation is analyzed separately. A player would be an individual variable, and one can be interested in a subset of variables' effect on an outcome. For example, for a black-box method in classification, we might want to know why an observation has been assigned to a particular class. Shapley values allow evaluating how the variables contributed to the classifier's decision (Lundberg & Lee, 2017).

Also, in the context of multivariate outlier detection, it is of interest why an observation has been declared outlying. A traditional tool for multivariate outlier detection is the Mahalanobis distance (Mahalanobis, 1936). To reliably identify outliers, it is essential to robustly estimate mean and covariance (Rousseeuw & Zomeren, 1990), and one option is to use the Minimum Covariance Determinant (MCD) estimator (Rousseeuw & Driessen, 1999). Shapley values can be adapted to the setting of squared Mahalanobis distances: One can obtain a decomposition of this distance measure into an outlyingness score for each variable, which can be interpreted as the average marginal contribution

to the outlyingness of an observation (Mayrhofer & Filzmoser, 2023). The sum of all these contributions is identical to the squared Mahalanobis distance of the observation. Another interesting feature is that the computational complexity of determining the Shapley values reduces to a very simple problem in the context of Mahalanobis distances, and thus the computations are very time-efficient, also in higher dimensions.

While the Shapley values inform about the contribution of the variables to the outlyingness of an observation, they do not inform about the values these cells would have if the observation would not be contaminated. This, however, is the goal of cellwise outlyingness methods (Rousseeuw & Bossche, 2018). A modification in the calculations of Shapley values also allows getting this information by which amount a cell needs to be modified to make the observation non-outlying (Mayrhofer & Filzmoser, 2023). As an outcome, one can obtain diagnostics regarding cellwise outlyingness.

2 Shapley Values for Matrix-valued Observations

Another important class of data structures are matrix-valued observations. Thus, the information is represented in the rows and columns of a matrix, and a prominent example are image data. Often, matrix-valued observations are vectorized; for example, the pixel information of an image can be arranged in a long vector, which then forms one row of a "traditional" data matrix. This leads to very high-dimensional data in which the neighborhood relationship of the pixels is lost.

The concept of matrix-valued data is not new at all, and a prominent distribution in this context is the matrix normal distribution (Dawid, 1981). There are different proposals in the literature on how to estimate the parameters of this distribution (Dutilleul, 1999). It is also possible to define a Mahalanobis distance, and the concept of the MCD estimator can be modified to obtain robust estimators. Finally, Shapley values can be used, and their contributions again sum up to an observation's squared Mahalanobis distance. In the context of image data, for example, one can identify outlying images and explain which pixels contribute to this outlyingness. A more detailed background, as well as illustrative examples, will be provided in the presentation.

References

DAWID, A PHILIP. 1981. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, **68**(1), 265–274.

DUTILLEUL, PIERRE. 1999. The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, **64**(2), 105–123.

LUNDBERG, SCOTT M, & LEE, SU-IN. 2017. A Unified Approach to Interpreting Model Predictions. *Pages 4765–4774 of:* GUYON, I., LUXBURG, U. V., BENGIO, S., WALLACH, H., FERGUS, R., VISHWANATHAN, S., & GARNETT, R. (eds), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.

MAHALANOBIS, PRASANTA CHANDRA. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, **2**, 49–55.

MARONNA, RICARDO A, MARTIN, R DOUGLAS, YOHAI, VICTOR J, & SALIBIÁN-BARRERA, MATÍAS. 2019. *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

MAYRHOFER, MARCUS, & FILZMOSER, PETER. 2023. Multivariate outlier explanations using Shapley values and Mahalanobis distances. *Econometrics and Statistics*. To appear.

MOLNAR, CHRISTOPH. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

RAYMAEKERS, JAKOB, & ROUSSEEUW, PETER J. 2022. The Cellwise Minimum Covariance Determinant Estimator. *arXiv preprint arXiv:2207.13493*.

RIBEIRO, MARCO, SINGH, SAMEER, & GUESTRIN, CARLOS. 2016 (02). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

ROUSSEEUW, PETER. 1985. Multivariate Estimation With High Breakdown Point. *Mathematical Statistics and Applications Vol. B*, 01, 283–297.

ROUSSEEUW, PETER, & ZOMEREN, BERT. 1990. Unmasking Multivariate Outliers and Leverage Points. *Journal of The American Statistical Association - J AMER STATIST ASSN*, **85**(06), 633–639.

ROUSSEEUW, PETER J., & BOSSCHE, WANNES VAN DEN. 2018. Detecting Deviating Data Cells. *Technometrics*, **60**(2), 135–145.

ROUSSEEUW, PETER J., & DRIESSEN, KATRIEN VAN. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41**(3), 212–223.

SHAPLEY, LLOYD S. 1953. A value for n-person games. *Contributions to the Theory of Games*, **2**(28), 307–317.

ZHANG, YICHI, SHEN, WEINING, & KONG, DEHAN. 2022. Covariance estimation for matrix-valued data. *Journal of the American Statistical Association*, 1–12.

IDENTIFICATION OF MISOGYNISTIC ACCOUNTS ON TWITTER THROUGH GRAPH CONVOLUTIONAL NETWORKS

Lara Fontanella [1] and Emiliano del Gobbo [2] and Alex Cucco [3]

[1] G. d'Annunzio University, Chieti-Pescara, Italy, (e-mail: lara.fontanella@unich.it)

[2] University of Foggia, Italy, (e-mail: emiliano.delgobbo@unifg.it)

[3] National Heart and Lung Institute, Imperial College, London, UK, (e-mail: a.cucco20@imperial.ac.uk)

ABSTRACT: Misogyny is the hatred, dislike, and mistrust towards women simply because of their gender, accompanied by ingrained prejudice against them. Our study focuses on producers of misogyny on social media platforms, specifically examining content shared in Italian on Twitter. Using a substantial collection of Italian tweets, we analyse textual and relational data from the friend/follower network to classify Twitter accounts based on a binary misogyny scheme. We employ Graph Convolutional Networks to achieve this.

KEYWORDS: misogyny, textual data, relational data, Graph Convolutional Networks

1 Introduction

Cyberspace is often misused to spread offensive and abusive content. Women are among the most targeted groups for online abusive content (Amnesty International Italia, 2022). Hate speech against women is strongly linked to misogyny, which is the cultural attitude of hatred towards females simply because they are female. In our research, we focus on identifying producers of misogynistic content shared in Italian on Twitter. Specifically, we tackle an automatic classification task by utilising textual-based features extracted from the shared content, as well as relational data derived from the network of relationships between Twitter accounts. In hate speech research, studies have focused on automatically detecting abusive online content, while more recently, attention has shifted towards examining the behaviour and relationships of individuals who spread abusive comments on mainstream platforms. Only a few studies have adopted a network modelling approach (Chatzakou *et al.*, 2017; Mishra *et al.*, 2018). These studies have integrated graph-based features from the pro-

ducers' network into a classification model along with textual data to enhance the classification performance. However, as far as we know, networked data has not been utilized to identify misogynistic producers of online content.

2 Materials and methods

2.1 Textual and relational data

To build the textual corpus, we downloaded Italian tweets containing keywords, mentions, and hashtags related to approximately fifty politically-active women, feminists, journalists, influencers, and female television personalities. Tweets were downloaded in real time from August to December 2022, and the downloaded dataset contains 1,002,226 tweets, associated with 204,095 accounts. We filtered out accounts that no longer existed, information providers (e.g., newspapers, radio stations, television channels and programs, news aggregators), and accounts with less than 5 tweets. To ensure a less biased composition of the retrieved network, we down-sampled the accounts with tweets focusing only on Giorgia Meloni. This was necessary because approximately 75% of the total number of tweets mentioned her, which was a result of the electoral campaign and her subsequent role as Prime Minister. The final dataset includes 82,807 tweets from 7,371 accounts, and the friend/follower relations among these accounts were retrieved.
We manually annotated a subset of 942 accounts using a misogyny binary scheme. To select these accounts, we considered node centrality measures to ensure a well-spread sample on the network that included nodes with the highest degree and betweenness indexes. We also considered the distribution of tweets by the women included in the corpus construction to ensure a larger variability in the textual content and higher domain coverage. Finally, we used the revised Hurtlex dictionary (Tontodimamma *et al.*, 2023) to compute an offensiveness score at the producer level. Out of the annotated accounts, 44.6% were flagged as misogynistic.

2.2 Collective classification and Graph Convolutional Networks

Given a network, represented through a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ the set of edges, different information can be associated with each node $v \in \mathcal{V}$. In particular, we might have a set of local features $\mathbf{x}_v$, generally assumed known for the entire network, and a label y_v, which can be observed only on a node subset. In this setting, a collective classification pro-

cedure allows to jointly predict the unobserved labels considering the attributes of the nodes to be predicted in addition to the observed attributes and labels and unobserved labels of neighbouring elements. GCNs (Kipf & Welling, 2017) are a type of neural network that can perform collective node classification by learning a function f that maps a feature description $\mathbf{x}_v$ and the graph structure, represented by an adjacency matrix $\mathbf{A}$, to a node-level output $\mathbf{y}^{(U)}$, where $U \subset \mathcal{V}$ is the unlabelled nodes subset. By jointly considering the feature descriptions and the graph structure, GCNs can improve classification accuracy compared to traditional machine learning models that only use node features.
In our analysis, we utilised a binary scheme for the misogyny classification task. We employed users' textual data to extract node local features, while relational data were derived from the friend/follower users' network.

3 Preliminary results

For these preliminary results, the feature matrix was built through a bag of words approach, where functional words (i.e., pronouns, prepositions, conjunctions) and non specific domain terms, along with a misogynistic tailored lexical dictionary, were included in the document-term matrix. For the implementation of GCNs, we adopted the FastGCN algorithm (Chen *et al.*, 2018). In the classified network the misogynistic accounts amount to the 27.0% of the nodes. Figure 1 highlights some network characteristics of the misogynistic accounts: they are likely to be clustered, tend to follow more people, to be followed by less people, and to have less importance in the network structure.

4 Conclusion and future works

From our preliminary results, collective node classification performed through GCNs shows promising results regarding the prediction of misogynistic accounts. Our findings are in line with previous research on hater networks (Ribeiro *et al.*, 2018; Mathew *et al.*, 2019) that showed how hateful social media users are very densely connected and differ from normal ones in terms of their word usage and network structure. It also results from literature that haters are more likely to have a lower number of followers while following a larger number of accounts.
As future work, we will compare different GCN models using different types of embeddings to derive the feature matrix, and we will also explore masking techniques to ensure cross-domain comparison.

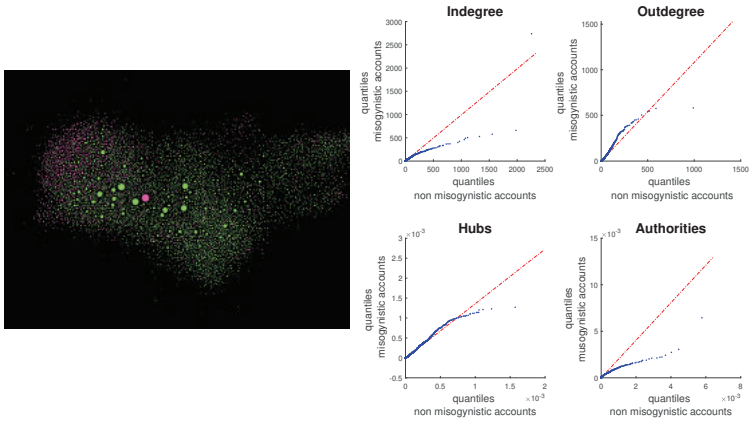


Figure 1. *Classified network - misogynistic accounts are depicted in pink - and centrality measures' qqplots*

Acknowledgements: This work was supported by EU Next Generation, MUR-Fondo Promozione e Sviluppo-DM 737/2021 [*ICOMIC: Identifying and Countering Online Misogyny*]

References

AMNESTY INTERNATIONAL ITALIA. 2022. *Odio in rete: italiani senza cittadinanza tra razzismo e xenofobia*. Tech. rept.

CHATZAKOU, D., KOURTELLIS, N., BLACKBURN, J., DE CRISTOFARO, E., STRINGHINI, G., & VAKALI, A. 2017. Mean Birds: Detecting Aggression and Bullying on Twitter. *In: WebSci '17: Proceedings of the 2017 ACM on Web Science Conference*.

CHEN, J., MA, T., & XIAO, C. 2018. FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling. *In: 6th International Conference on Learning Representations, ICLR 2018*.

KIPF, T. N, & WELLING, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *In: 5th International Conference on Learning Representations, ICLR 2017*.

MATHEW, B., DUTT, R., GOYAL, P., & MUKHERJEE, A. 2019. Spread of hate speech in online social media. *In: WebSci '19: Proceedings of the 10th ACM Conference on Web Science*.

MISHRA, P., DEL TREDICI, M., YANNAKOUDAKIS, H., & SHUTOVA, E. 2018. Author Profiling for Abuse Detection. *In: Proceedings of the 27th International Conference on Computational Linguistics*.

RIBEIRO, M., CALAIS, P., SANTOS, Y., ALMEIDA, V., & MEIRA JR, W. 2018. Characterizing and Detecting Hateful Users on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, **12**(1).

TONTODIMAMMA, A., FONTANELLA, L., ANZANI, S., & BASILE, V. 2023. An Italian lexical resource for incivility detection in online discourses. *Quality and Quantity*, **57**, 3019–3037.

DEPTH FUNCTIONS FOR TREE-INDEXED PROCESSES

Giacomo Francisci and Anand Vidyashankar [1]

[1] Department of Statistics, College of Engineering and Computing, George Mason University, Fairfax, VA (e-mail: gfranci@gmu.edu, avidyash@gmu.edu)

ABSTRACT: Depth functions have long been used to describe the quantiles of multivariate distributions. Also, local depth functions have been used for classification and clustering. These objects have been extended to functional and metric space valued data. In this presentation we describe depth functions for the intensity measure of a point process. When the point process consists of i.i.d. components we obtain the depth of those components. Specifically, we study point processes indexing the edges of Galton-Watson trees and investigate their statistical properties. We use these results in classification of tree-indexed data and develop an analog of the DD-classifier.

ANALYSING THE EFFECT OF DIFFERENT DESIGN CHOICES IN NETWORK-BASED TOPIC DETECTION

Carla Galluccio[1], Matteo Magnani[2], Davide Vega[2],
Giancarlo Ragozini[3] and Alessandra Petrucci[1]

[1] Department of Statistics, Computer Science, Applications "G. Parenti", (e-mail: carla.galluccio@unifi.it, alessandra.petrucci@unifi.it)

[2] Department of Information Technology, Division of Computing Science, (e-mail: matteo.magnani@it.uu.se, davide.vega@it.uu.se)

[3] Department of Political Sciences, (e-mail: giancarlo.ragozini@unina.it)

ABSTRACT: In the literature on topic modelling, network-based procedures for topic detection have become popular as an alternative to classical topic models, showing promising results. However, the lack of a systematic analysis of how the design choices made in text processing and network definition affect the results in terms of topics detected makes using these procedures demanding. Therefore, this work aims to fill this gap by showing how and to what extent the choices made during the analysis influence the features of the topics discovered.

KEYWORDS: text network analysis, community detection, topic detection

1 Introduction

Network-based procedures for topic detection are based on the idea that any text can be represented as a word co-occurrence network, where topics are defined as groups of strongly connected words (Hamm & Odrowski, 2021).

More specifically, a network-based topic discovery process is made up of different steps that could be summarised as follows: i) text preprocessing; ii) definition of the word co-occurrence matrix; iii) network definition and selection of the community detection algorithm.

Even if many works have applied network-based procedures for analysing textual data and discovering topics, none of them focused on how the choices made in the design phase affect the final result in a systematic way.

Thus, this work aims to start filling this gap by studying how and to what extent some of the choices made during the analysis influence the features of topics discovered. In particular, in this work, we focused primarily on the definition of the word co-occurrence matrix and the selection of the community detection algorithm, as these steps are unique to network-based approaches.

2 Method and Materials

We conducted the analysis employing the BBC news article collection, a widely used corpus in the context of textual analysis and topic detection. The collection comprises 2,225 complete news articles collected from 2004 to 2005 regarding five topics: business, entertainment, politics, sport and technology (Greene & Cunningham, 2006).

As text preprocessing, we removed non-alphanumeric characters, numbers and words composed of 1 or 2 characters, divided the text into tokens (unigrams), removed the stopwords using a stoplist provided with the dataset, and finally stemmed the text. Then, we removed words with a value of tf-idf less than 0.01 (Allahyari *et al.*, 2017). At the end of the preprocessing step, the number of unique word tokens was equal to 18,422.

The word co-occurrence matrices were generated by counting the number of times two words co-occur in the same document within a specific window size, that is a set of neighbouring words within a specified distance, respectively equal to 2, 5, 10, 15 and 20 words on the right of the baseline word.

Afterwards, we defined different filters and weighting schemes on the word co-occurrence matrices. The first aspect was examined by removing from the word co-occurrence matrix the 100, 500, and 1000 words with the lowest co-occurrence values and the 50, 100, and 500 words with the highest co-occurrence values.

On the other hand, the second aspect was tested by considering an additional weighting scheme based on word proximity. In this case, we assigned more weight to the words nearest the target one inside the window. For example, for a window size equal to 5, we set a weight equal to 1 to the word adjacent to the target word, a weight equal to 4/5 to the next word and so on, until the last word, which takes a weight equal to 1/5.

Finally, we employed the Louvain community detection algorithm, Newman's leading eigenvector algorithm and the SLPA algorithm to discover topics in text networks obtained from the word co-occurrence matrices (interpreted as weighted adjacency matrices). The first two algorithms are non-overlapping community detection algorithms based on modularity maximisation, while the third is an overlapping community detection algorithm (Blondel *et al.*, 2008, Newman, 2006, Xie *et al.*, 2013).

The choice of using an overlapping community detection algorithm lies in the hypothesis that while non-overlapping community detection algorithms could correctly assign topics' characteristic words, multi-topic words could be arbitrarily assigned to one of the communities they should have been included.

3 Results

Our findings showed that with the increase in the window sizes, the number of communities found by the three algorithms decreases, remaining stable for window sizes greater than 5. In particular, for window sizes greater than 2, the number of communities found by the Louvain algorithm and Newman's algorithm is always greater than the number of communities identified by SLPA, which finds only one community with these settings.

Further to this point, Newman's algorithm generally finds three communities in the different experimental settings, while the Louvain algorithm finds almost always five communities for window sizes greater than 5. Notice that the communities found by the Louvain algorithm are coherent in number and content with the BBC news articles collection's topics. Interestingly, when the Louvain algorithm finds a number of communities greater than five for window sizes greater than 5, they are pretty unbalanced, with five bigger communities coherent with the original topics.

Computing the ARI (Hubert & Arabie, 1985) on the communities found by the Louvain algorithm under different settings, we observed that the ARI is generally high for different window sizes, particularly between the partitions obtained for window sizes greater than 5 (ranging from 0.604 to 0.878). This result shows that even if the algorithm finds the same number of communities, they are not identical.

Filters on words with the lowest degree from the word co-occurrence matrix do not affect the results. Conversely, removing words with the highest word co-occurrence remarkably increases the number of communities found for a window size equal to 2 (ranging from 27 to 112).

Similarly, using a different weighting scheme does seem to affect the number of communities found, which is noticeably higher when we use the proximity weighting scheme. However, also in this case, increasing the window sizes decreases the number of communities found.

4 Conclusions

The results obtained show that different design choices during text preprocessing and network definition affect the features of topics detected, mainly in terms of the number of topics discovered. For future work, we aim to focus on extending the assessment of the effects of these design choices on different kinds of texts, such as textual social media (like Twitter or Facebook).

References

ALLAHYARI, M., POURIYEH, S., ASSEFI, M., SAFAEI, S., TRIPPE, E.D., GUTIERREZ, J.B., & KOCHUT, K. 2017. A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv:1707.02919*, 1–13.

BLONDEL, V.D., GUILLAUME, J., LAMBIOTTE, R., & LEFEBVRE, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 1–12.

GREENE, D., & CUNNINGHAM, P. 2006. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. *Pages 377–384 of:* COHEN, W., & MOORE, A. (eds), *Proc. 23rd International Conference on Machine learning (ICML'06)*. ACM Press, New York.

HAMM, A., & ODROWSKI, S. 2021. Term-Community-Based Topic Detection with Variable Resolution. *Information*, **12**, 221–252.

HUBERT, L., & ARABIE, P. 1985. Comparing partitions. *Journal of classification*, **2**, 193–218.

NEWMAN, M.E.J. 2006. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, **74**, 1–12.

XIE, J., KELLEY, S., & SZYMANSKI, B.K. 2013. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, **45**, 1–35.

A PROPOSAL FOR THE JOINT AUTOMATED DETECTION OF CLUSTERS AND ANOMALIES

Luis A. García-Escudero [1] , Christian Hennig[2] , Agustín Mayo-Iscar[1] , Gianluca Morelli[3] and Marco Riani[3]

[1] Department of Statistics and Operation Research and IMUVA, University of Valladolid, (e-mail: lagarcia@uva.es, agustin.mayo.iscar@uva.es)

[2] Department of Statistical Sciences "Paolo Fortunati", University of Bologna, (e-mail: christian.hennig@unibo.it)

[3] Dipartimento di Scienze Economiche e Aziendali, Universita' degli Studi di Parma, (e-mail: gianluca.morelli@unipr.it, marco.riani@unipr.it))

ABSTRACT: It is known that outliers can be problematic when statistical techniques are applied. This is also the case in Cluster Analysis and, with this in mind, the TCLUST method was introduced as a robust clustering alternative. Given a fixed trimming level α, TCLUST attempts to detect the fraction α of observations that should best be discarded after assuming k normally distributed components. However, the main problem is how to determine reasonable values for k and α for a given data set. An approach was introduced to choose k and α through visual inspection of "classification trimmed likelihood" curves. Theoretical background will be provided for a better understanding of that approach, along with a parametric bootstrap method to reduce subjectivity and produce a small list of sensible robust clustering partitions.

KEYWORDS: clustering, robustness, trimming, outliers

1 Robust clustering and TCLUST

It is well known that outliers can be problematic when applying statistical methods for data analysis, and this also happens in the case of Cluster Analysis. Outliers can affect clustering methods in such a way that main clusters can be joined artificially or clusters formed of few outlying observations are detected (see, e.g., García-Escudero & Gordaliza, 1999). Moreover, it is interesting to apply clustering techniques to deal with outliers since clustered sets of outliers are known to be particularly harmful for many (even robust) statistical procedures. Consequently, different robust clustering methods have been introduced that can be used successfully to jointly deal with clusters and outliers (Ritter, 2014, García-Escudero *et al.*, 2016).

One such approach to robust clustering is based on applying impartial trimming. Given a fixed trimming level α, the term "impartial" means that is the data set itself that indicates what fraction α of observations should be trimmed. The TCLUST method introduced in García-Escudero *et al.*, 2008 is a robust clustering procedure based on that impartial trimming principle and where elliptically contoured clusters are allowed.

Given a sample $\mathcal{X} = \{x_1, ..., x_p\}$ in $\mathbb{R}^p$, the TCLUST method is defined by maximizing

$$\sum_{j=1}^{k} \sum_{i \in R_j} \log(\pi_j \phi(x_i; \mu_j, \Sigma_j)), \tag{1}$$

where $\phi(\cdot; \mu, \Sigma)$ is the density function of the p-variate normal distribution, $\{R_0, R_1, ..., R_k\}$ is a partition of the indexes $\{1, 2, ..., n\}$ such that $\#R_0 = [n\alpha]$. Also, in that maximization, we enforce

$$M_n / m_n \leq c$$

for $M_n = \max_{j=1,...,k} \max_{l=1,...,p} \lambda_l(\Sigma_j)$ and $m_n = \min_{j=1,...,k} \min_{l=1,...,p} \lambda_l(\Sigma_j)$ being, respectively, the largest and the smallest of the eigenvalues of the Σ_j scatter matrices. The constant $c \geq 1$ plays an important role by avoiding uninteresting "spurious clusters" and providing well-defined mathematical problems. The $\pi_j \geq 0$ weights also satisfy $\sum_{j=1}^{k} \pi_j = 1$.

The TCLUST procedure can be implemented using the `tclust` package in R (Fritz *et al.*, 2012) and the `FSDA` Matlab toolbox (Riani *et al.*, 2012). However, TCLUST requires the simultaneous specification of the number of clusters k and the trimming fraction α. Choosing correctly those two parameters for a given data set is not always an easy task because, for instance, a set of close outliers could be considered as "noise" to be trimmed (requiring a higher α) or, alternatively, as an additional cluster (requiring a higher k). Therefore, the determination of k and α is a clearly interrelated problem that requires an unified treatment. Even choosing the number of groups k in Cluster Analysis, without trimming, is already well known to be a very complex problem.

2 Classification trimmed likelihood curves

A graphical procedure for selecting sensible values for k and α for TCLUST (when c is fixed) was introduced in García-Escudero *et al.*, 2011. The procedure was based on the visual inspection of the so-called "classification trimmed likelihood" curves. These curves are defined through

$$(k, \alpha) \mapsto \mathcal{L}^{\Pi}(\alpha, k; \mathcal{X}), \tag{2}$$

where $\mathcal{L}^{\Pi}(\alpha,k;\mathcal{X})$ denotes the maximum value reached in the constrained maximization of (1). García-Escudero *et al.*, 2011 explained that

$$t^n_{k,\alpha} = \mathcal{L}^{\Pi}(\alpha,k+1;\mathcal{X}) - \mathcal{L}^{\Pi}(\alpha,k;\mathcal{X})$$

should not be too small when there is a clear benefit in increasing k to $k+1$ for a trimming level α. This heuristic led to a graphical exploratory tool for choosing reasonable values for k and α.

Given a probability measure P, we can define a population version of the TCLUST problem (García-Escudero *et al.*, 2008). We can also define population versions of the classification trimmed likelihoods appearing in (2), which are denoted as $\mathcal{L}^{\Pi}_{\alpha,k}(P)$. We have that $\mathcal{L}^{\Pi}_{\alpha,k}(P_n) = \mathcal{L}^{\Pi}(\alpha,k;\mathcal{X})$, where P_n denotes the empirical measure corresponding to $\mathcal{X}$ ($\mathcal{X}$ seem as the realization of an i.i.d. sample from P). Given the consistency

$$\mathcal{L}^{\Pi}_{\alpha,k}(P_n) \to \mathcal{L}^{\Pi}_{\alpha,k}(P),$$

and the fact that $t^n_{k,\alpha} = \mathcal{L}^{\Pi}_{\alpha,k+1}(P_n) - \mathcal{L}^{\Pi}_{\alpha,k}(P_n)$, it makes sense to analyse the behaviour of $\mathcal{L}^{\Pi}_{\alpha,k}(P)$ to see under what circumstances $t^n_{k,\alpha}$ should be small. Theoretical have been obtained on the expected changes in $\mathcal{L}^{\Pi}_{\alpha,k}(P)$, when increasing k to $k+1$, depending on the underlying distribution P. These results provide some theoretical background to better understand the key ingredients involved in the classification trimmed likelihood curve and how these curves should be interpreted.

3 Parametric bootstrap automated procedure

In practical applications, it is not always easy to determine sensible values for k and α just from that visual inspection of the classification trimmed likelihood curves. The user must make rather subjective decisions about whether or not $t^n_{k,\alpha}$ can be considered small due to sample variability. A parametric bootstrap procedure will be presented trying to overcome that trouble.

By applying TCLUST to compute $t^n_{k,\alpha}$, we also obtain parameter estimates for the k fitted normal components. These parameters are used to draw B parametric bootstrap samples $\{\mathcal{X}^{*b}\}_{b=1}^{B}$, but also trying to emulate the mechanism generating the fraction α of contaminating observations in $\mathcal{X}$. If k and α are reasonable parameters, then $\{\mathcal{L}(\alpha,k+1;\mathcal{X}^{*b}) - \mathcal{L}(\alpha,k;\mathcal{X}^{*b})\}_{b=1}^{B}$ would allow us to "mimic" the sampling distribution of $t^n_{k,\alpha}$ and compute bootstrap p-values as

$$p_{k,\alpha} = \frac{\#\{b : \mathcal{L}(\alpha,k+1;\mathcal{X}^{*b}) - \mathcal{L}(\alpha,k;\mathcal{X}^{*b}) > t^n_{k,\alpha}\}}{B}.$$

We can use these bootstrap p-values to finally get a reduced list of reasonable k and α values for applying TCLUST in an fully automated way. Users can use this reduced list to choose the robust cluster partition that best meets their ultimate cluster and outlier detection goals, by applying standard cluster validation/visualization tools.

Illustrative and real data examples, together with a simulation study, also seem to justify the interest of the automated selection proposal. Therefore, we consider that the proposal is clearly valuable since it can certainly help the user in the detection of anomalies.

References

FRITZ, H., GARCÍA-ESCUDERO, L.A., & MAYO-ISCAR, A. 2012. tclust: An R Package for a Trimming Approach to Cluster Analysis. *Journal of Statistical Software*, **47**(12).

GARCÍA-ESCUDERO, L.A., & GORDALIZA, A. 1999. Robustness properties of k-means and trimmed k-means. *Journal of the American Statistical Association*, **94**, 956–969.

GARCÍA-ESCUDERO, L.A, GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2008. A general trimming approach to robust Cluster Analysis. *Annals of Statistics*, **36**, 1324–1345.

GARCÍA-ESCUDERO, L.A, GORDALIZA, A., MATRÁN, C., & MAYO-ISCAR, A. 2011. Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, **21**, 585–599.

GARCÍA-ESCUDERO, L.A., GORDALIZA, A., MATRÁN, C., MAYO-ISCAR, A., & HENNIG, C.M. 2016. Robustness and Outliers. *Pages 653 – 678 of:* C. HENNIG, M. MEILA, F. MURTAGH, & R. ROCCI (eds), *Handbook of Cluster Analysis*. Serie Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

RIANI, M., PERROTTA, D., & TORTI, F. 2012. FSDA: A MATLAB toolbox for robust analysis and interactive data exploration,. *Chemometrics and Intelligent Laboratory Systems*, **116**, 17–32.

RITTER, G. 2014. *Cluster Analysis and Variable Selection*. Boca Raton: CRC Press.

MOBILITY ACROSS CRIMES: STATISTICALLY VALIDATED NETWORKS AND TEMPORAL PATTERN RECOGNITION

V. G. Genova[1], C. Edling [23], H. Mondani[245], A. M. Rostami[26], M. Tumminello[1]

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

[2] Department of Sociology, Lund University, Lund, Sweden

[3] Institute for Futures Studies, Stockholm, Sweden

[4] Department of Sociology, Umeå University, Umeå, Sweden

[5] Department of Sociology, Stockholm University, Stockholm, Sweden

[6] Department of Social Work and Criminology, University of Gävle, Gävle, Sweden

ABSTRACT: Criminal careers can be categorised as either general or specialised. A key challenge in studying crime specialisation is determining which crimes should be considered similar and which should be considered distinct from the criminal's perspective. We conducted an empirical study involving a large group of Swedish suspects to address this issue. The primary objective was to investigate generalist and specialist behaviour in crime. By employing directed network analysis, our study aimed to uncover temporal patterns of criminal specialisation. Specifically, we examined the temporal connections between different types of crimes to reveal distinct patterns in criminal behaviour. The findings indicate that individuals who were suspected of at least two crime types within each of the five communities throughout their criminal careers demonstrated varying patterns of specialisation evolution. In contrast, some individuals consistently maintained high levels of generalism. These results highlight the diverse paths individuals take in their criminal behaviour and contribute to our understanding of the dynamics of criminal specialisation over time.

KEYWORDS: communities and criminal specialisation, complex networks, criminal temporal patterns.

1 Introduction

Specialisation in criminal behaviour has significant implications for comprehending the root causes of crime (Piquero, 2000). Theories, such as those focusing on the relationship between brain functioning and delinquency, which

includes brain damage (Neylan, 1999) and low or unstable serotonin levels (Alm *et al.*, 1996), assume that violent crime is the specialisation of aggressive individuals. Similar contemplations are applicable to theories that explore the interactions between genetic and social factors that can lead to violence (Wolfgang *et al.*, 1967). Sutherland's differential association theory (Sutherland *et al.*, 1992) postulates that crime is learned behaviour, thereby suggesting high crime specialisation. Thus, criminal specialisation is a multifaceted issue that is central to criminology, crime prevention, and enforcement (Loeber & Farrington, 1998). In this study, we analyse the Swedish national register of individuals suspected of criminal offences, which comprises about 750,000 individuals that have been suspected of at least 2 crimes in Sweden from 1995 to 2016. The database includes information such as age and sex of the suspects, the types of crimes (521 categories) they have been suspected of, and the date (or period) when the crime was committed. The aim of this work is to discover temporal patterns of criminal specialisation. We utilise directed network analysis to study the temporal association between crime types to reveal these temporal patterns in criminal behaviour.

2 Statistically validated temporal networks

We used a statistically validated network approach for temporal network analysis (SVTN) to assess specialisation in crime. Additionally, to accommodate the database's heterogeneity caused by suspects involved in different numbers of crimes, we partitioned the database into multiple subsets, S_f, according to the number of crimes per criminal. Further details on this can be found in Tumminello *et al.*, 2013. We tested the null hypothesis of random co-occurrence between two crimes, a and b, using the hypergeometric distribution.

$$pvalue(n_{ab}^f) = \sum_{x=n_{ab}^f}^{min(n_a^f, n_b^f)} \frac{\binom{n_a^f}{x}\binom{N^f - n_a^f}{n_b^f - x}}{\binom{N^f}{n_b^f}}, \tag{1}$$

where N^f is the number of criminals in subset S^f, n_a^f (n_b^f), is the number of criminals that committed crime a (b) in the subset S^f, and n_{ab}^f is the number of criminals who were suspected (with a temporal direction) of both crimes, a and b.

The null hypothesis of random co-occurrences exactly takes into account the heterogeneity of both the types of crimes, a and b, by conditioning to n_a^f and n_b^f (Tumminello *et al.*, 2013). We calculated the *p-values* in every subset S^f

to construct a weighted network of crime types based on the excess of co-occurrence. Then, we used the FDR correction for multiple hypothesis testing (Benjamini & Hochberg, 1995) to adjust the *p-values*. In contrast to the study conducted by (Tumminello *et al.*, 2013), this work's novelty lies in incorporating the temporal element in hypothesis testing, enabling us to examine the temporal progression of criminal specialisation. Our SVTN is a weighted directed network with labelled links. Indeed, it comprises four types of links, which are defined as follows:

- **JO:** a and b jointly occur (undirected link) $a - b$;
- **PR:** crime a precedes crime b if a occurred before b, $a \xrightarrow{PR} b$;
- **CO:** b contains a if crime a occurred entirely within the time interval in which crime b was perpetrated, $a \multimap b$;
- **PO:** crime a and b partially overlap if b begins after a and ends after a, $a \xrightarrow{PO} b$.

As seen in the list above, there are two types of links, namely PR and PO, that are temporally directed. These two link types provide information about the temporal progression of criminal activity and are considered separately from the other link types. By concentrating on these links, we developed a weighted directed False Discovery Rate (FDR) network, in which each link corresponds to a statistically significant *p-value* (5% threshold) after the FDR correction, while the weight is determined by the total number of subsets S^f in which the link is significant.

3 Preliminary results

We utilised the MapEquation (Edler *et al.*, 2022) to identify the hierarchical structure of communities within the network. We discovered five primary communities, each consisting of a minimum of 20 crime types that subsequently divided into smaller communities. These large communities represent distinct types of criminal specialisation: 1) fraud, forgery, and taxation (120 crime types); 2) assault, rape, and persecution (198 crime types); 3) drugs, narcotics, attempted homicide, and homicide with a firearm (85 crime types); 4) theft and arson (66 crime types), and 5) violence against unacquainted victims (28 crime types)*. Individuals who were suspected of at least two crime types within each

*In addition, there are three smaller communities that do not further split into more specific ones: environmental crimes (20 crime types), human trafficking for forced labour (2 crime types), and robbery from a shop or a taxi (3 crime types).

community during their criminal career exhibited different patterns of specialisation evolution. Some groups achieved high levels of specialisation early on, while others maintained relatively high levels of generalism even later in their careers.

References

ALM, PO, AF KLINTEBERG, B, HUMBLE, K, LEPPERT, J, SÖRENSEN, S, TEGELMAN, R, THORELL, L-H, & LIDBERG, L. 1996. Criminality and psychopathy as related to thyroid activity in former juvenile delinquents. *Acta Psychiatrica Scandinavica*, **94**(2), 112–117.

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 (Methodological)*, **57**(1), 289–300.

EDLER, DANIEL, HOLMGREN, ANTON, & ROSVALL, MARTIN. 2022. *The MapEquation software package*.

LOEBER, ROLF, & FARRINGTON, DAVID P. 1998. *Serious and violent juvenile offenders: Risk factors and successful interventions*. Sage Publications.

NEYLAN, THOMAS C. 1999. Frontal lobe function: Mr. Phineas Gage's famous injury. *The Journal of neuropsychiatry and clinical neurosciences*, **11**(2), 280–281.

PIQUERO, ALEX. 2000. Frequency, specialization, and violence in offending careers. *Journal of research in crime and delinquency*, **37**(4), 392–418.

SUTHERLAND, EH, CRESSEY, DR, & LUCKENBILL, DF. 1992. *Principles of Criminology. Lanham, MD: General Hall*.

TUMMINELLO, MICHELE, EDLING, CHRISTOFER, LILJEROS, FREDRIK, MANTEGNA, ROSARIO N, & SARNECKI, JERZY. 2013. The phenomenology of specialization of criminal suspects. *PloS One*, **8**(5), e64703.

WOLFGANG, MARVIN E, FERRACUTI, FRANCO, & MANNHEIM, HERMANN. 1967. The subculture of violence: Towards an integrated theory in criminology (Vol. 16). *London: Tavistock Publications*.

A COHORT STUDY ON THE GENDER GAP IN MORTALITY THROUGH THE TUCKER3 MODEL

Paolo Giordani [1], Susanna Levantesi [1], Andrea Nigri [2] and Virginia Zarulli [3]

[1] Department of Statistical Sciences, Sapienza University of Rome, (e-mail: paolo.giordani@uniroma1.it, susanna.levantesi@uniroma1.it)

[2] Department of Economics, Management and Territory, University of Foggia, (e-mail: andrea.nigri@unifg.it)

[3] Interdisciplinary Centre on Population Dynamics, University of Southern Denmark, (e-mail: vzarulli@sdu.dk)

ABSTRACT: In this manuscript, leveraging the Tucker3 model, we investigate the gender gap in mortality considering the ratio of male to female mortality rates, specific for age, cause of death, and cohort. The model is applied to a tensor containing gender gap data by causes, age classes, and non-extinct cohorts.

KEYWORDS: Mortality data, gender gap, cohort study, multi-way data, Tucker3

1 Introduction

Understanding mortality is of great importance for both private and public sectors to design appropriate pension or insurance plans. To this purpose, several interesting applications of multi-way models to mortality data are available in the literature (see, e.g., Cardillo *et al.*, 2023). Generally speaking, in these studies, data usually refer to mortality rates across demographic features such as causes of death, ages, countries, and years. This work represents a further step in mortality analysis by focusing on the gender gap (Zarulli *et al.*, 2021) in causes of death and its evolution by cohort. Limiting our attention to the three-way case, the Tucker3 model is applied to a tensor containing gender gap data in mortality distinguished by causes of death, age classes, and cohorts.

2 Three-way data and models

A three-way array or tensor $\underline{\mathbf{X}}$ of order $(I \times J \times K)$ can be seen as a box containing scores on a set of I observation units with respect to J variables in K different occasions. Observation units, variables and occasions are usually referred to as "modes". The generic element of $\underline{\mathbf{X}}$ is x_{ijk} giving the score

of observation unit i ($i = 1,\ldots,I$) on variable j ($j = 1,\ldots,J$) at occasion k ($k = 1,\ldots,K$). Thus, there are three ways or indices, one for each mode. The array $\underline{\mathbf{X}}$ can be seen as a collection of standard matrices of order $(I \times J)$, one for every occasion.

It is often convenient to summarize $\underline{\mathbf{X}}$ to unravel the relevant information hidden in the data. To this purpose, suitable extensions of Principal Component Analysis for arrays should be considered. One of the most famous models is the Tucker3 one (Tucker, 1966). The Tucker3 model synthesizes $\underline{\mathbf{X}}$ by extracting P ($< I$), Q ($< J$) and R ($< K$) components for the observation units, variables and occasions, respectively, thus allowing different levels of complexity for the three modes. Let $\mathbf{X}_a$ be the matrix of order $(I \times JK)$ obtained by juxtaposing next to each other the standard matrices pertaining to every occasion. The Tucker3 model can be formalized as

$$\mathbf{X}_a = \mathbf{A}\mathbf{G}_a\left(\mathbf{C} \otimes \mathbf{B}\right)^{\mathrm{T}} + \mathbf{E}_a, \tag{1}$$

where $\mathbf{A}$ of order $(I \times P)$, $\mathbf{B}$ of order $(J \times Q)$ and $\mathbf{C}$ of order $(K \times R)$ are the component score matrices for the observation units, the variables and the occasions, respectively. Therefore, each mode is summarized by the corresponding set of components. The triple interactions among such components are measured by the three-way array $\underline{\mathbf{G}}$ of order $(P \times Q \times R)$ called *core*. Finally, $\mathbf{E}_a$ is the error matrix of order $(I \times JK)$ and the symbol $\otimes$ denotes the Kronecker product. Estimation of the model parameters is carried out in the least square sense by

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C},\underline{\mathbf{G}}} ||\mathbf{E}_a||^2, \tag{2}$$

being $||\cdot||$ the Frobenius norm of matrices. An alternating least squares algorithm can be used. It can be shown that the obtained solution is not identifiable. In fact, all component matrices as well as the core array can be rotated. The non-identifiability can be exploited in order to rotate the solution to a simple structure. Given P, Q and R, we can assess the fit percentage of the Tucker3 model as

$$\left(1 - \frac{||\mathbf{E}_a||^2}{||\mathbf{X}_a||^2}\right) 100. \tag{3}$$

The closer to 100, the better the fit of the Tucker3 model. The optimal numbers of components P, Q and R can be found by balancing fit and parsimony, bearing in mind that interpretability is of relevant importance. For further details on the Tucker3 model and related multi-way models, the interested reader may refer to (Kroonenberg, 2008).

3 Results

The analyzed data come from the Human Cause-of-Death Database (HCD) and refer to the mortality rates distinguished by causes of death, age classes and cohorts registered in the United States of America. In particular, we consider the mortality rates of $I = 7$ causes of death (Infectious diseases, Neoplasms, Cardiovascular diseases, Respiratory diseases, Digestive diseases, External causes of death, Other causes of death) distinguished in $J = 7$ five-year age classes from 60 to 90 years for cohorts of people born in $K = 10$ years from 1919 to 1928. In order to deal with fully crossed data, i.e. all observation units have scores on all variables on all occasions, such mortality rates are collected for the years 1979–2018. Letting m^F_{ijk} and m^M_{ijk} be the mortality rates of the cause of death i at age class j for cohort k for females and males, respectively, the generic element of the three-way gender gap data array $\underline{\mathbf{X}}$ is

$$x_{ijk} = \frac{m^M_{ijk}}{m^F_{ijk}}, \tag{4}$$

expressing to what extent the mortality rate for a certain cause of death of a given age and belonging to a specific cohort of males differs from the corresponding rate for females.

To assess whether and how gender differences in mortality are related to causes of death, ages and cohorts, the Tucker3 model with $P = Q = 2$ and $R = 1$ components is used. To motivate this choice, we observe that the fit percentage is rather high (91.60%), despite the low total number of components ($P + Q + R = 5$), and the solution is well interpretable. In this respect, simplicity is achieved by transforming $\mathbf{G}_a$ to the identity matrix and applying the varimax (Kaiser, 1958) rotation to $\mathbf{B}$ compensating it in $\mathbf{A}$. In this way, the components for the causes of death and those for the age classes are related one-to-one.

The component matrix $\mathbf{A}$ for the causes of death is displayed in Figure 1. The component matrix $\mathbf{B}$ for the age classes (not reported here) distinguishes the younger ages (from 60 to 70) with large positive first component scores and the older ages (from 75 to 90) with large positive second component scores. Taking into account that the component scores for the cohorts are all positive and decreasing passing from cohort 1919 to cohort 1928, the main findings are that the gender gap for ages 60–70 increases in connection with Cardiovascular diseases and External causes and decreases with Infectious diseases and Other causes. This especially holds for the oldest cohorts. Conversely, for ages 75–90, the gender gap for Neoplasms and Respiratory diseases is high, whilst

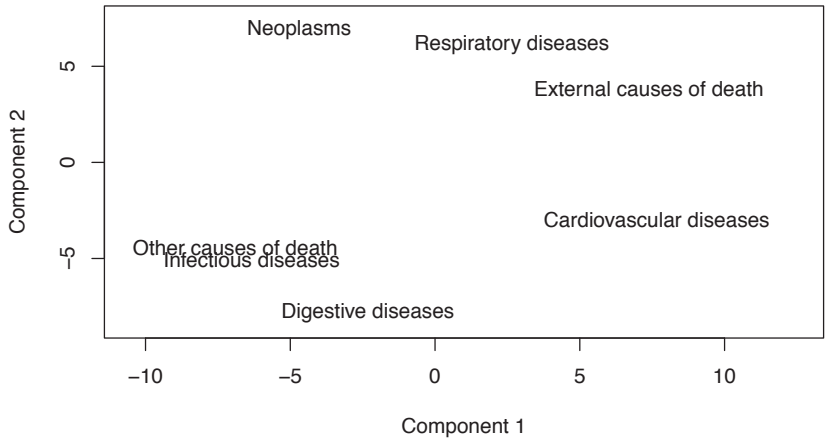


Figure 1. *Component scores for the causes of death.*

the opposite comment holds for Digestive diseases. Further results will be presented during the conference.

References

CARDILLO, G., GIORDANI, P., LEVANTESI, S., NIGRI, A., & SPELTA, A. 2023. Mortality forecasting using the four-way CANDECOMP/PARAFAC decomposition. *Scandinavian Actuarial Journal*, in press.

KAISER, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.

KROONENBERG, P. M. 2008. *Applied Multiway Data Analysis*. Hoboken: Wiley.

TUCKER, L. R. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279–311.

ZARULLI, V., KASHNITSKY, I., & VAUPEL, J.W. 2021. Death rates at specific life stages mold the sex gap in life expectancy. *Proc. Natl. Acad. Sci.*

TRIMMED KERNEL MEAN SHIFT

Luca Greco[1] , Giovanna Menardi[2] and Marco Rudelli[2]

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

[2] Department of Statistical Sciences, University of Padova,
(e-mail: menardi@stat.unipd.it, marco.rudelli@studenti.unipd.it)

ABSTRACT: A robust procedure based on impartial trimming is discussed, aimed to protect nonparametric clustering stemming from kernel mean shift from the deleterious effect of outliers.

KEYWORDS: Clustering, Density estimation, Outliers

1 Introduction

The problem of data contamination, where unexpected points that do not share the pattern of the majority of the data are observed, is known to possibly hinder the validity of inferential procedures. The issue is even more critical in clustering, where the lack of a reference ground-truth to aim at makes even the simplest problem an ill-posed one. Genuine observations forming small clusters can be mistaken with outliers (*swamping*); on the other side, outlying data lying close to each other just by chance can form spurious clusters (*masking*). Moreover, in this setting it is quite difficult to state a working notion of outliers, and robustness is not only data dependent, but rather cluster dependent (Hennig, 2008), which is itself often arbitrary. It then looks clear how contaminated data can compromise or even invalidate unsupervised techniques.

A large amount of work has been done to define robust clustering strategies in the mainstream approaches within the distance- and the model-based approach (see Farcomeni & Greco, 2016, for a review). Conversely, the issue has been largely neglected in the nonparametric framework, where clusters are identified as the domains of attractions of the modes of the underlying density (Stuetzle, 2003). The correspondence between groups and modal regions entails some reasons of attractiveness: clusters are not constrained to predetermined shapes, and resorting to nonparametric methods keeps this flexibility; additionally, the number of clusters is inherent of the data density, hence determined as part of the estimation procedure (see, Menardi, 2016, for a review). However, these very same properties turn out to be pitfalls of nonparametric

methods in the presence of outliers. Actually, outliers can produce spurious modes. In the presence of spurious modes, outliers self-validate themselves, as they can not be declared unlikely with respect to the cluster they have given birth to. Finally: how can one say what is unlikely, with respect to a cluster which can take any shape? In the following, a robust-to-outliers counterpart of the Kernel Mean Shift (KMS, Fukunaga & Hostetler, 1975) for modal detection is discussed, based on an outlyingness criterion specifically designed for the considered framework.

2 Methodology

Let $\mathcal{X} = (x_1, x_2, \ldots, x_n)$ be a sample of size n, with $x_i \in \mathbb{R}^d$, $d \geq 1$. A kernel density estimator is given by $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - x_i)$ where $K_H(x)$ is a d-variate kernel function scaled by a symmetric positive definite $d \times d$ bandwidth matrix H. KMS is an iterative algorithm to identify modal clusters from a kernel density estimate of a set of data. The algorithm recursively shifts each data point to a local weighted mean $m_{K,H}$,

$$m_{K,H}\left(x^{(j)}\right) = x^{(j+1)} - x^{(j)} = \frac{\sum_{i=1}^{n} x_i^{(j)} \nabla K_H(x)\left(x - x_i^{(j)}\right)}{\sum_{i=1}^{n} \nabla K_H(x)\left(x - x_i^{(j)}\right)} \propto \frac{\nabla \hat{f}\left(x_i^{(j)}\right)}{\hat{f}\left(x_i^{(j)}\right)} .$$

until convergence. The weights are normalized gradient vectors of the kernel function. Hence, the mean shift is a gradient ascent algorithm based on a normalised kernel estimator of the gradient.

We propose a robust counterpart of KMS based on impartial trimming (Cuesta-Albertos *et al.*, 1997). The methodology, summarised in Algorithm 1, climbs iteratively via KMS the modes of a trimmed kernel density estimate, obtained by discarding at each iteration a fixed proportion α of data with the lowest densities with respect to the pertaining cluster. Then, the identified clusters allow to update the outlyingness score of each observation and run KMS on a renewed active set. Iterations stop as the trimmed set is not updated. The procedure is impartial since the detection of the trimmed points is a result of the procedure jointly with cluster assignments and it recasts to a trimmed KMS (tKMS). The initial active subset $I^{(0)}$ can be obtained as follows: (a) consider an over-smoothed fitted density; (b) select a proportion of points with the largest fitted densities.

Algorithm 1 Iteration r of tKMS

Optimization Step

Evaluate the kernel density estimate over the active set $I^{(r)}$ of size $n - \lfloor n\alpha \rfloor$

$$\hat{f}^{(r)}(x) = \frac{1}{n - \lfloor n\alpha \rfloor} \sum_{i \in I^{(r)}} K_H(x - x_i)$$

Run KMS to identify the modes of $\hat{f}^{(r)}(x)$, and get a partition of $\mathcal{X}$ in clusters $\left\{ C_m^{(r)} \right\}_m$, each with cardinality $n_m^{(r)}$

Let $m = m_i$ if $x_i \in C_m^{(r)}$

Trimming Step

Compute $\hat{g}_i^{(r)} = \hat{g}_{m_i}^{(r)}(x_i), i = 1, 2, \ldots, n$ with

$$\hat{g}_m^{(r)}(x) = \frac{1}{n_m^{(r)}} \sum_{x_i \in C_m^{(r)}} K_H(x - x_i)$$

Update $I^{(r+1)}$ by ruling out from $\mathcal{X}$ the $\lfloor n\alpha \rfloor$ points with the lowest of $\hat{g}_i^{(r)}$.

3 Examples

We illustrate the effectiveness of the proposed methodology, as well as the drawbacks of classical KMS in the presence of contamination, through some synthetic examples. Figure 1 gives the results from running both KMS and tKMS on a pair of bivariate data structured in three clusters in the presence of background noise. While essentially identifying the true clusters, in both examples KMS also detects spurious modes, wheres tKMS recovers the underlying clustering structure and trimmed points are not assigned to any cluster.

References

CUESTA-ALBERTOS, J. A., GORDALIZA, A., & MATRÁN, C. 1997. Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, **25**(2), 553–576.

FARCOMENI, A., & GRECO, L. 2016. *Robust methods for data reduction*. CRC press.

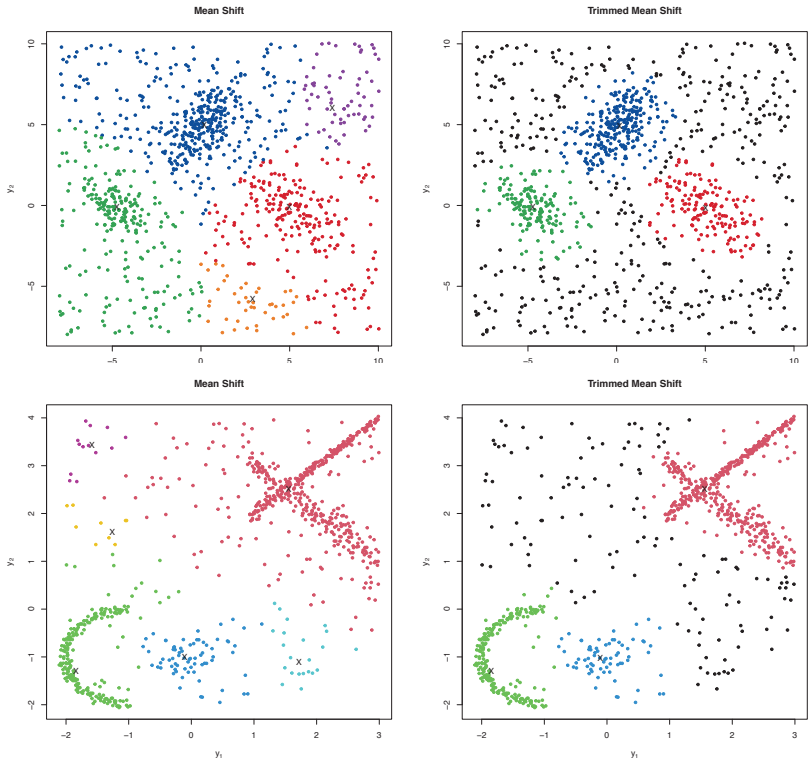


Figure 1. *Classification from KMS (left) and tKMS (right) for two synthetic data exhibiting three variously shaped clusters (one for each row). The identified clusters are denoted by different colors, while the estimated modes are denoted by X. Trimmed points are identified in black.*

FUKUNAGA, K., & HOSTETLER, L. 1975. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, **21**(1), 32–40.

HENNIG, C. 2008. Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of multivariate analysis*, **99**(6), 1154–1176.

MENARDI, G. 2016. A review on modal clustering. *International Statistical Review*, **84**(3), 413–433.

STUETZLE, WERNER. 2003. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of classification*, **20**(1), 25–47.

MODELING ZONE DIAMETER MEASUREMENTS TO INFER ANTIBIOTIC SUCEPTIBILITY OF BACTERIA

Bettina Grün [1], Thomas Petzoldt [2] and Helga Wagner [3]

[1] Institute for Statistics and Mathematics, Vienna University of Economics and Business, (e-mail: Bettina.Gruen@wu.ac.at)

[2] Institute of Hydrobiology, Technische Universität Dresden (e-mail: Thomas.Petzoldt@tu-dresden.de)

[3] Institute of Applied Statistics, Johannes Kepler University Linz (e-mail: Helga.Wagner@jku.at)

ABSTRACT: Disk diffusion tests are employed to determine the susceptibility of bacteria to antibiotics by measuring the zone diameter (ZD) of inhibition. Previous work suggested to use a composite model when modeling minimum inhibitory concentration measurements. This model combines a parametric distribution covering the range of observations from the susceptible component with a non-parametric distribution capturing the range of observations containing also resistant observations. We investigate the use of this model for ZD data and also consider a two-component mixture model combining the parametric distribution with the non-parametric distribution relaxing the split of the support for both distributions. We present maximum likelihood and penalized maximum likelihood estimation of both models using a normal or a skew-normal distribution as parametric distribution while taking the restricted support and the rounding of the data into account. We illustrate the use of these models in a simulation study on artificial data and on data available from the web page of the European Committee on Antimicrobial Susceptibility Testing (EUCAST).

KEYWORDS: antibiotic susceptibility, composite model, disk diffusion test, mixture model, zone diameter measurement.

References

AZZALINI, A. 1985. A Class of Distributions Which Includes the Normal Ones. *Scandinavian Journal of Statistics*, **12**(2), 171–178.

JASPERS, STIJN, AERTS, MARC, VERBEKE, GEERT, & BELOEIL, PIERRE-ALEXANDRE. 2014. Estimation of the Wild-Type Minimum Inhibitory Concentration Value Distribution. *Statistics in Medicine*, **33**(2), 289–303.

CLUSTERING LONGITUDINAL ORDINAL DATA

Julien Jacques [1] and Francesco Amato [1]

[1] Univ Lyon, Univ Lyon 2, ERIC, Lyon, (e-mail: julien.jacques@univ-lyon2.fr, francesco.amato@univ-lyon2.fr)

ABSTRACT: In social sciences, studies are often based on questionnaires asking participants to express ordered responses several times over a study period. We present a model-based clustering algorithm for such longitudinal data. Assuming that an ordinal variable is the discretization of a underlying latent continuous variable, the model relies on a mixture of matrix-variate normal distributions, accounting simultaneously for within- and between-time dependence structures. An EM algorithm is considered for parameter estimation. An evaluation of the model through synthetic data show its estimation abilities and its advantages when compared to competitors. A real-world application concerning preferences for grocery shopping during the Covid-19 pandemic period in France will be presented.

KEYWORDS: Ordinal data, longitudinal data, clustering, matrix variate distribution, EM algorithm

1 The data

Let denote by $y_{i,j,t}$ the observation of the j-th ordinal variable for the i-th unit at time t ($i = 1,...,N$; $j = 1,..,J$ and $t = 1,...,T$). The categories of the j-th ordinal variable are quoted by 1 to C_j. The data are organized in a random-matrix form such that $\mathbf{Y} = \{Y_i\}_{i=1}^N$ is a sample of $J \times T$-variate matrix observations:

$$Y_i = \begin{pmatrix} y_{i,1,1} & \cdots & y_{i,1,t} & \cdots & y_{i,1,T} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ y_{i,j,1} & \cdots & y_{i,j,t} & \cdots & y_{i,j,T} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ y_{i,J,1} & \cdots & y_{i,J,t} & \cdots & y_{i,J,T} \end{pmatrix}$$

2 Latent Gaussian distribution for ordinal variable

We assume that each variable $y_{i,j,t}$ is the manifestation of an underlying latent continuous variable $z_{i,j,t}$ which follows a Gaussian distribution. At this

point, we can assume that each observed ordinal matrix Y_i is indeed the manifestation of a latent continuous random matrix $Z_i \in \mathbb{R}^{J\times T}$, which follows a matrix-normal distribution $\mathcal{MN}_{(J\times T)}(M,\Phi,\Sigma)$, where $M \in \mathbb{R}^{J\times T}$ is the matrix of means, $\Phi \in \mathbb{R}^{T\times T}$ is a covariance matrix containing the variances and covariances between the T occasions or times and $\Sigma \in \mathbb{R}^{J\times J}$ is the covariance matrix containing the variance and covariances of the J variables. The matrix-normal probability density function (pdf) is given by

$$f(Z|M,\Phi,\Sigma) = (2\pi)^{-\frac{TJ}{2}}|\Phi|^{-\frac{J}{2}}|\Sigma|^{-\frac{T}{2}}\exp\left\{-\frac{1}{2}\text{tr}[\Sigma^{-1}(Z-M)\Phi^{-1}(Z-M)^{\intercal}]\right\}.$$

To map from Y_i to Z_i, let γ_j denote a C_{j+1}-dimensional vector of thresholds that partition the real line for the j-th ordinal variable that has C_j levels and let the threshold parameters be constrained such that $-\infty = \gamma_{j,0} \leq \gamma_{j,1} \leq ... \leq \gamma_{j,C_j} = \infty$. If the latent $z_{i,j,t}$ is such that $\gamma_{j,c-1} < z_{i,j,t} < \gamma_{j,c}$ then the observed ordinal response, $y_{i,j,t} = c$.

3 Model-based clustering

When data are heterogeneous, mixture model is an efficient way to perform clustering. In the present case, we consider Mixture of Matrix-Normals (MMN, Viroli, 2011). As usually for mixture models, parameter estimation is done using an EM algorithm. The number of cluster is selected using the BIC criterion.

4 Applications

An evaluation of the model through synthetic data show its estimation abilities and its advantages when compared to competitors. A real-world application concerning preferences for grocery shopping during the Covid-19 pandemic period in France will be presented.

References

VIROLI, CINZIA. 2011. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, **21**(4), 511–522.

"YOU CALL IT A MANIFOLD, I CALL IT A SUBSPACE" - SELECTED EXAMPLES ON THE INTERFACE BETWEEN COMPUTER SCIENCE AND STATISTICS IN THE CONTEXT OF CLUSTERING AND MANIFOLD LEARNING

Daniyal Kazempour [1] and Peer Kröger[1]

[1] Department of Computer Science,
Christian-Albrechts-Universität zu Kiel,
(e-mail: [dka,pkr]@informatik.uni-kiel.de)

ABSTRACT: Do prominent data mining methods in computer science have anything in common with well-established techniques in statistics? Are there any benefits in combining methods from statistics with those from computer science, and if yes, why do we gain such benefits? These and further aspects are approached at the interface between computer science and statistics.

This talk first provides a brief introduction to the clustering and dimensionality reduction tasks from a computer science perspective. Furthermore, a brief introduction to the manifold learning task is given. This foundation is followed by an elaboration on similarities and distinct properties between two seemingly different tasks from different domains (cs and statistics), more specifically the subspace clustering and the manifold learning task.

Pursuing this path on the interface between computer science and statistics, it is elaborated on endeavors of enhancing cluster analysis through manifold learning while investigating why the combination of two methods from different domains (clustering and manifold learning) results in a symbiotic relationship.

In conclusion, this talk aims to sketch selected examples of the potential synergies that can emerge on the interface of computer science and statistics in the context of data mining and machine learning including its challenges and benefits. The examples in this talk do not only target a formal level but also the interdisciplinary experiences gained in collaborations between statisticians and computer scientists encouraging future endeavors between both scientific domains.

SPARSE RULE GENERATING FOLD-CHANGE CLASSIFICATION FOR MOLECULAR HIGH-THROUGHPUT PROFILES

Annika MTU Kestler[1], Nensi Ikonomi[1] , Silke D Werle[1], Julian D Schwab[1], Friedhelm Schwenker[2] and Hans A Kestler[1,3]

[1] Medical Systems Biology, and

[2] Neural Information Processing,

Ulm University, Albert-Einstein-Allee 11, Ulm, 89077, Germany

(e-mail: {annika.kestler, nensi.ikonomi, silke.kuehlwein, julian.schwab, friedhelm.schwenker, hans.kestler}@uni-ulm.de)

[3] Corresponding author

ABSTRACT: Classifying gene expression profiles can be challenging due to their low sample size and high dimensionality. Existing methods employed often lack interpretability or sparsity, and require extensive data preprocessing. Ensemble methods, such as the Set Covering Machine, enable the construction of classifiers depending only on base classifiers. We propose two novel base classifiers that consider relations between features for constructing interpretable decision functions, denoted fold change classifiers. Here, an intrinsic feature selection and a straightforward semantic and syntactic interpretation can be achieved. The proposed classifier no longer depends on equally scaled data since relative measurements within a sample are considered. The applicability of the proposed method is shown in a case study evaluating neuroendocrine tumors.

KEYWORDS: ensemble method, molecular high-dimensional data, set covering machine, fold changes, neuroendocrine tumors

1 Introduction

Classification in light of potentially formulating biological hypotheses often entails classifying high-dimensional data, where the number of samples is greatly outnumbered by the number of dimensions of each sample (Lausser & Kestler, 2013; Marchand & Shah, 2004). Each dimension of a sample is referred to as a feature, and finding distinctions between samples, that only depend on a subset of features, might lead to the formulation of novel biological hypotheses (Lausser & Kestler, 2013; Marchand & Shah, 2004). Moreover,

discarding irrelevant features can be considered essential in order to obtain sparse and interpretable classifiers, with high generalization abilities (Marchand & Shah, 2004).
The *Set Covering Machine* (SCM), enables the construction of a sparse conjunction of base classifiers (Marchand & Shawe-Taylor, 2003), performing an intrinsic feature selection when using a single threshold on one feature as a base classifier. Previous work in this context and these types of classifiers has been done mainly by (Marchand & Shawe-Taylor, 2003, Valiant, 1984, Haussler, 1988), and more recently by others (Drouin *et al.*, 2016; Drouin *et al.*, 2019; Lausser & Kestler, 2013; Schmid *et al.*, 2013; Kestler *et al.*, 2006; Lausser *et al.*, 2020). A resulting interpretable decision function can be of the form "**IF** $f_1 \geq 5$ **AND** $f_2 < 8$ **THEN** the sample belongs to class ...", with f_1 and f_2 being features / genes.
When analyzing high-throughput expression profiles the mere over- or under-expression of single genes might not suffice to identify biologically relevant genes (Shi *et al.*, 2005). Therefore, considering relations between different gene expressions, by pairwise comparing expressions could lead to the identification of global behaviors and point to biological processes involved (Shi *et al.*, 2005). This motivates base classifiers of type $f_1 < f_2$ or $f_1/f_2 \geq t$ where t is a threshold, relating the two features considered. These base classifiers may be less susceptible to noise, as well as exhibit invariance properties (Lausser & Kestler, 2013). This allows the discovery of similar tendencies among different samples, without depending on identical nomalization of the data.

2 Results

Contrary to the originally published SCM, which constructs a classifier depending on a subset of provided samples (Marchand & Shawe-Taylor, 2003), we are able to construct a sparse classifier, depending only on a subset of features, while eliminating concerns about normalization and data-preprocessing. Due to the interpretable decision functions, learnt by our proposed method, a genotype-to-phenotype relation can be established, potentially revealing novel biological mechanisms.
We employed the proposed method in a case study dealing with pancreatic neuroendocrine tumours (PanNETs). PanNETs are rare but quite heterogeneous tumour entities lacking specific biomarkers for disease progression. The resulting decision function, an ensemble of order relations, is sparse and yields perfect reclassification contrary to other classification methods employed on this data. Here, the gene relations involved in the decision functions could be

validated via a literature search, suggesting mechanistic interactions to be further investigated. The restriction to the evaluation of order relations reduces the gained flexibility of the presented base classifiers. This can be further validated by the sparsity of the decision function, implying that the base classifier involved carry much information.

References

DROUIN, ALEXANDRE, GIGUÈRE, SÉBASTIEN, DÉRASPE, MAXIME, MARCHAND, MARIO, TYERS, MICHAEL, LOO, VIVIAN G., BOURGAULT, ANNE-MARIE, LAVIOLETTE, FRANÇOIS, & CORBEIL, JACQUES. 2016. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, **17**(1), 754.

DROUIN, ALEXANDRE, LETARTE, GAËL, RAYMOND, FRÉDÉRIC, MARCHAND, MARIO, CORBEIL, JACQUES, & LAVIOLETTE, FRANÇOIS. 2019. Interpretable genotype-to-phenotype classifiers with performance guarantees. *Scientific Reports*, **9**(1), 4071.

HAUSSLER, DAVID. 1988. Quantifying Inductive Bias: AI Learning Algorithms and Valiant's Learning Framework. *Artificial Intelligence*, **36**(2), 177–221.

KESTLER, HANS A., LINDNER, WOLFGANG, & MÜLLER, ANDRÉ. 2006. Learning and Feature Selection Using the Set Covering Machine with Data-Dependent Rays on Gene Expression Profiles. *Pages 286–297 of:* SCHWENKER, FRIEDHELM, & MARINAI, SIMONE (eds), *Artificial Neural Networks in Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg.

LAUSSER, LUDWIG, & KESTLER, HANS A. 2013. Fold Change Classifiers for the Analysis of Gene Expression Profiles. *Pages 193–202 of:* GAUL, WOLFGANG, VICHI, MAURIZIO, & WEIHS, CLAUS (eds), *Studies in Classification, Data Analysis, and Knowledge Organization*. Heidelberg: Springer.

LAUSSER, LUDWIG, SZEKELY, ROBIN, KLIMMEK, ATTILA, SCHMID, FLORIAN, & KESTLER, HANS A. 2020. Constraining classifiers in molecular analysis: invariance and robustness. *Journal of The Royal Society Interface*, **17**(163), 20190612.

MARCHAND, MARIO, & SHAH, MOHAK. 2004. PAC-Bayes Learning of Conjunctions and Classification of Gene-Expression Data. *Pages 881–*

888 of: SAUL, L., WEISS, Y., & BOTTOU, L. (eds), *Advances in Neural Information Processing Systems*, vol. 17. Boston, MA: MIT Press.

MARCHAND, MARIO, & SHAWE-TAYLOR, JOHN. 2003. The Set Covering Machine. *Journal of Machine Learning Research*, **3**, 723–746.

SCHMID, FLORIAN, LAUSSER, LUDWIG, & KESTLER, HANS A. 2013. Three Transductive Set Covering Machines. *Pages 303–311 of:* GAUL, WOLFGANG, VICHI, MAURIZIO, & WEIHS, CLAUS (eds), *Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin, Heidelberg: Springer International Publishing.

SHI, LEMING, TONG, WEIDA, FANG, HONG, SCHERF, UWE, HAN, JING, PURI, RAJ K., FRUEH, FELIX W., GOODSAID, FEDERICO M., GUO, LEI, SU, ZHENQIANG, HAN, TAO, FUSCOE, JAMES C., XU, Z. AALEX, PATTERSON, TUCKER A., HONG, HUIXIAO, XIE, QIAN, PERKINS, ROGER G., CHEN, JAMES J., & CASCIANO, DANIEL A. 2005. Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics*, **6**(2), S12.

VALIANT, LESLIE G. 1984. A Theory of the Learnable. *Communications of the ACM*, **27**(11), 1134–1142.

COMPARISON OF THE HOUSEHOLDS' WORK INTENSITY IN SLOVAKIA AND CZECHIA THROUGH LEAST SQUARES MEANS ANALYSIS BASED ON GLM

Silvia Komara1[1], Martina Košíková[1], Erik Šoltés[1] and Tatiana Šoltésová[2]

[1] Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava, (e-mail: silvia.komara@euba.sk, martina.kosikova@euba.sk, erik.soltes@euba.sk)

[2] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, University of Economics in Bratislava, (e-mail: tatiana.soltesova@euba.sk)

ABSTRACT: Work intensity (WI) of household is primarily monitored in the context of identifying (quasi-)jobless (QJ) households. QJ households are those whose members use less than 20% of their work potential. Persons in such households, together with income-poor and the severely materially and socially deprived persons are included in the Europe 2030 Strategy as socially excluded persons who need to be targeted by social policies.

The aim of the paper is to assess the impact of relevant factors and their interactions on the WI of households in Slovakia and Czechia. For this purpose, general linear models, contrast analysis and estimates of marginal means are employed. Presented analyses are based on the EU-SILC 2021 survey and carried out separately for Slovakia and Czechia. The paper reveals common and different features of these countries in the WI of households. Particular attention is given to the identification of profiles of persons with a high risk of living in QJ households.

KEYWORDS: work intensity, general linear model, marginal means, contrast analysis.

References

FILANDRI, M., PASQUA, S., & STRUFFOLINO, E. 2020. Being working poor or feeling working poor? The role of work intensity and job stability for subjective poverty. *Social Indicators Research*, **147**(3), 781-803.

HOREMANS, J. 2018. Atypical employment and in-work poverty. In *Handbook on in-work poverty*. Edward Elgar Publishing.

SCHAD, D. J., VASISHTH, S., HOHENSTEIN, S., & KLIEGL, R. 2020. How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, **110**, 104038.

MODEL BASED CLUSTERING PROCEDURES FOR MULTIVARIATE MIXED TYPE LONGITUDINAL DATA

Arnošt Komárek [1]

[1] Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (e-mail: komarek@karlin.mff.cuni.cz)

ABSTRACT: The talk will present model based clustering methods to classify units based on so called multivariate mixed type longitudinal or panel data. The multivariate aspect of data points to the situation when more than one outcome is observed for each unit within a longitudinal study at each measurement occassion. The mixed type data then arise in situations when such multivariate outcomes are not necessarily of the same type, e.g., some of them are numeric, some of them categorical. The talk will provide an overview of clustering approaches for such data developed by author over past about 10 years, partly in cooperation with Lenka Komárková, Jan Vávra, Bettina Grün and Gertraud Malsiner-Walli.

KEYWORDS: classification, finite mixture, generalized linear mixed model, panel data.

1 Introduction

In different types of studies data are nowadays routinely gathered repeatedly over time on the same units leading to *longitudinal* or *panel* data. On top of that, multiple outcomes, both numeric and categorical, i.e., of a *mixed type*, are recorded at each measurement occasion leading to multivariate longitudinal data of a mixed type. An important area of interest is how to suitably model and analyze this kind of data if *unobserved heterogeneity* is suspected. In this case a statistical method is frequently required which forms homogeneous groups of similar units in the study population and develops a classification rule on how to classify not only available but perhaps also future units into those groups using the same type of data.

2 Notation

We are assuming that a dataset suitable for analysis by methods presented in this talk is composed of N units which we want to classify into $K > 1$ groups,

where K, the number of groups, is not necessarily known in advance. We are further assuming that the aim is to classify the units into the groups on basis of $R \geq 1$ of longitudinally gathered outcomes (being possibly of a mixed type). Let for $i = 1, \ldots, N$ and $r = 1, \ldots, R$, $\mathbf{Y}_{i,r} = \left(Y_{i,r,1}, \ldots, Y_{i,r,n_i}\right)$ denote a vector of the values of the rth outcome of the ith unit obtained at n_i occasions at times $\mathbf{t}_i = \left(t_{i,1}, \ldots, t_{i,n_i}\right)$. Further, let $\mathbf{v}_{i,r,1}, \ldots, \mathbf{v}_{i,r,n_i}$ be vectors of additional covariates that may explain random fluctuation of the outcomes $\mathbf{Y}_{i,r}$ we may want to adjust for in the classification procedure. These additional covariates are also allowed to be both numeric and categorical. They may include characteristics that are constant over time for a given unit but may also be time dependent. Furthermore, let $\mathcal{C}_{i,r} = \left\{\mathbf{t}_i, \mathbf{v}_{i,r,1}, \ldots, \mathbf{v}_{i,r,n_i}\right\}$ denote both the measurement times and the covariate values for the rth outcome of the ith panel member. Finally, let

$$\mathbf{Y}_i = \left(\mathbf{Y}_{i,1}, \ldots, \mathbf{Y}_{i,R}\right), \qquad \mathcal{C}_i = \left\{\mathcal{C}_{i,1}, \ldots, \mathcal{C}_{i,R}\right\}$$

denote all information (outcomes and covariate values) available for the ith unit that can be used in the data analysis and exploited for the classification of the units into one of the K groups.

By the fact that the outcomes are of a mixed type, we consider a situation that for different values of r, the elements of the vectors $\mathbf{Y}_{i,r}$ are possibly of a different type. Some of them might be *numeric*, some of them *counts*, *binary*, *ordinal* or general *multinomial*. This reflects a common practical situation of gathering multiple outcomes of different nature in one longitudinal study. Finally, it is obvious that the elements of the vectors $\mathbf{Y}_i$ cannot be assumed to be independent and some modelling of the dependence structure is a must with any realistic modelling approach.

3 Model based clustering

In the talk, several model based clustering approaches developed in Komárek & Komárková, 2013, Komárek & Komárková, 2014, Vávra & Komárek, 2022 and Vávra *et al.*, 2023 will be presented for data having the structure outlined in Section 2. The model behind all clustering procedures is a sort of the mixture of (generalized) linear mixed models. Unknown model parameters are estimated using the Bayesian approach and the Markov chain Monte Carlo (MCMC) methodology. Furthermore, related R software routines will also be discussed. Finally, we show on how to perform not only clustering of units based on available data but also how to classify a new observation into one of the clusters.

References

KOMÁREK, A., & KOMÁRKOVÁ, L. 2013. Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*, **7**(1), 177–200.

KOMÁREK, A., & KOMÁRKOVÁ, L. 2014. Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software*, **59**(12), 1–38.

VÁVRA, J., & KOMÁREK, A. 2022. Classification based on multivariate mixed type longitudinal data with an application to the EU-SILC database. *Advances in Data Analysis and Classification (early access)*.

VÁVRA, J., KOMÁREK, A., GRÜN, B., & MALSINER-WALLI, G. 2023. Clusterwise multivariate regression of mixed-type panel data. *Preprint on Research Square*.

INEQUALITY, POPULISM, AND UNFAIRNESS: A COMPARISON OF UNFAIR INCOME INEQUALITIES IN POLAND AND NORWAY

Tomasz Kwarciński[1], Paweł Ulman[2]

[1] Department of Philosophy, Krakow University of Economics,
(e-mail: kwarcint@uek.krakow.pl)

[2] Department of Statistics, Krakow University of Economics,
(e-mail: ulmanp@uek.krakow.pl)

ABSTRACT: Income inequality has been linked to political polarization and the rise of populist movements. However, the mechanisms through which this happens are not fully understood. The use of the Gini coefficient, which does not reflect the idea of fairness, may contribute to this challenge. To address this issue, the paper introduces the concept of "Unfairness Gini" to measure unfair income inequality, which takes into account the differences between people's real income and their fair income due to factors related to individual effort and choices. The Unfairness Gini draws on luck egalitarianism, which assumes that inequality in society is only justified if it arises from differences of personal effort or choice. The paper uses the methodological approach developed by Cappelen and Tungodden to calculate the Unfairness Gini using EU-SILC data from 2011 and 2019 for Poland and Norway, two countries with different levels of income inequality and populism. The choice of time points offers an opportunity to compare unfairness inequality before and after the populist party began to govern in Poland in 2015.

KEYWORDS: inequality, populism, unfairness, Gini coefficient.

References

ALMÅS, I., CAPPELEN, A. W., LIND, J. T., SØRENSEN, E., & TUNGODDEN, B. (2011). Measuring unfair (in)equality. *Journal of Public Economics*, 95(7–8), 488–499.

COHEN, G. A. (1989). On the Currency of Egalitarian Justice. *Ethics*, 99(4), 906–944.

RAWLS, J. (1999). *A theory of justice: Revised edition*. Harvard University Press.

STOETZER, L. F., KLÜVER, H., & GIESECKE, J. (2023). How does income inequality affect the support for populist parties? *Journal of European Public Policy*, 30.

HURLEY, S. L. (2003). *Justice, luck, and knowledge*. Harvard University Press.

SEGMENTING TOROIDAL TIME SERIES BY NONHOMOGENEOUS HIDDEN SEMI-MARKOV MODELS

Francesco Lagona [1], Marco Mingione[1]

[1] Department of Political Sciences, University of Roma Tre, (e-mail: francesco lagona@uniroma3.it, marco.mingione@uniroma3.it)

ABSTRACT: Motivated by classification issues in marine studies, we propose a hidden semi-Markov model to segment toroidal time series according to a finite number of latent regimes. The time spent in a given regime and the chances of a regime-switching event are separately modeled by a battery of regression models that depend on time-varying covariates.

KEYWORDS: hidden semi-Markov model, toroidal data, model-based classification, wave, wind.

1 Introduction

Bivariate sequences of angles are often referred to as toroidal time series, because the pair of two angles can be represented as a point on a torus. Examples include time series of wind and wave directions and time series of turning angles in studies of animal movement.

We introduce a nonhomogeneous, toroidal hidden semi-Markov model (HSMM) that segments toroidal time series. Precisely, the distribution of toroidal data is approximated by a mixture of toroidal densities, whose parameters evolve according to a latent semi-Markov process with covariate-specific dwell times.

Our proposal extends previous approaches that are based on toroidal hidden Markov models (Lagona & Picone, 2013). Under a toroidal hidden Markov model, the sojourn times of the states of the latent process are distributed according to a geometric distribution. Our proposal relaxes this restrictive assumption by replacing the latent Markov chain with a latent, nonhomogeneous semi-Markov model, where the (non necessarily geometric) time spent in a given regime and the chances of a regime-switching event are separately modeled by a battery of regression models that allow the introduction of covariates.

2 A toroidal hidden semi-Markov model

Let $\boldsymbol{y} = (\boldsymbol{y}_t, t = 1, \dots T)$ be a bivariate time series, where $\boldsymbol{y}_t = (y_{t1}, y_{t2})$ is a vector of two circular observations. Further, let $\boldsymbol{u} = (\boldsymbol{u}_t, t = 1, \dots T)$ be a sequence of latent multinomial random variables $\boldsymbol{u}_t = (u_{t1} \dots u_{tK})$ with one trial and K classes (or states), whose binary components represent class membership at time t. Our proposal is a hierarchical model where the joint distribution of the time series is obtained by

$$f(\boldsymbol{y}) = \sum_{\boldsymbol{u}} f(\boldsymbol{y} \mid \boldsymbol{u}) p(\boldsymbol{u}).$$

The joint distribution $p(\boldsymbol{u})$ of the latent process is described by extending the notion of a Markov chain. If $\mathbf{u}$ is a Markov chain, then $p(\boldsymbol{u})$ is fully known up to a vector of K initial probabilities $\pi_k = P(u_{1k} = 1), k = 1, \dots, K, \sum_k \pi_k = 1$, and a $K \times K$ matrix of transition probabilities

$$\begin{pmatrix} \pi_{11} & \pi_{12} & \dots & \pi_{1K} \\ \pi_{21} & \pi_{22} & \dots & \pi_{2K} \\ \dots & \dots & \dots & \dots \\ \pi_{K1} & \pi_{K2} & \dots & \pi_{KK} \end{pmatrix} = \begin{pmatrix} 1-p_1 & p_1\omega_{12} & \dots & p_1\omega_{1K} \\ p_2\omega_{21} & 1-p_2 & \dots & p_2\omega_{2K} \\ \dots & \dots & \dots & \dots \\ p_K\omega_{K1} & p_K\omega_{K2} & \dots & 1-p_K \end{pmatrix}$$

where $p_k = \sum_{k' \neq k} \pi_{kk'}$ is the probability of a transition from k to a different state and $\omega_{kk'}$ is the conditional probability of a transition to state $k' \neq k$, given a transition from state k. Under this setting, if the process is in state k, the time τ_k up to a transition to a different state is geometric

$$P(\tau_k = \tau) = p_k(1-p_k)^{\tau-1}. \tag{1}$$

More generally, let $S_k(\tau) = P(\tau_k > \tau) = \exp\left(-\int_0^\tau h_k(v)dv\right)$ be the survival function of τ_k, where $h_k(\tau)$ is the associated hazard function. Then

$$p_k(\tau) = P(\tau_k \leq \tau + 1 \mid \tau_k > \tau) = 1 - \exp\left(-\int_\tau^{\tau+1} h_k(v)dv\right),$$

is the conditional probability of a transition at time $t+1$, given that the process has been in state k during a period of length t. Then

$$P(\tau_k = \tau) = p_k(\tau) \prod_{i=1}^{\tau-1} (1 - p_k(\tau)). \tag{2}$$

When the hazard h_k is time-constant, then (2) reduces to (1). Alternatively, (2) can be approximated with the desired accuracy by

$$P(\tau_k = \tau) = p_k(m)(1 - p_k(m))^{\tau - m} \prod_{i=1}^{m-1} (1 - p_k(i)). \tag{3}$$

Parametric hazard functions can be borrowed from the survival analysis literature and some of them are conveniently associated to a link function g that trasforms $p_k(\tau)$ to a linear function of time, say $g(p_k(\tau)) = \beta_{0k} + \beta_{1k}\tau$. Such a specification can be further extended by introducing a vector of q (possibly time-varying) covariates, say $\boldsymbol{x}_t$, which influence the dwell time distribution

$$g(p_k(\tau; \boldsymbol{x}_t)) = \beta_{0k} + \beta_{1k}\tau + \boldsymbol{x}_t^{\mathsf{T}} \boldsymbol{\beta}. \tag{4}$$

Similarly, covariates may be introduced to shape the conditional transition probabilities, say $\omega_{kk'} = \omega_{kk'}(\boldsymbol{x}_t)$, through a multinomial regression equation. The introduction of time-varying covariates makes the latent process nonhomogeneous, extending recent literature proposals.

Our proposal is completed by a conditional independence assumption on the observation process. Precisely,

$$f(\boldsymbol{y} \mid \boldsymbol{u}) = \prod_{t=1}^{T} \prod_{k=1}^{K} \prod_{i=1}^{m} f(\mathbf{y}_t; \boldsymbol{\theta}_k)^{u_{tki}}, \tag{5}$$

where $\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_K$ is a sequence of unknown parameters. Parametric toroidal densities can be borrowed by the proposals available in the directional statistics literature. A convenient specification is for example the bivariate wrapped Cauchy distribution (Kato & Pewsey, 2015). It is unimodal, pointwise symmetric and has a closed-form expression for the conditional distribution. A single dependence parameter controls the relationship between the two component circular variables, ranging from independence to perfect correlation. The remaining four parameters respectively indicate the two marginal means and concentrations.

3 Results

Figure 1 shows the results obtained on a time series of $T = 1326$ semi-hourly wind and wave directions, taken in wintertime by the buoy of Ancona, which is located in the Adriatic Sea at about 30 km from the coast. A 2-state hidden semi-Markov model has been used to segment the data. The model integrates

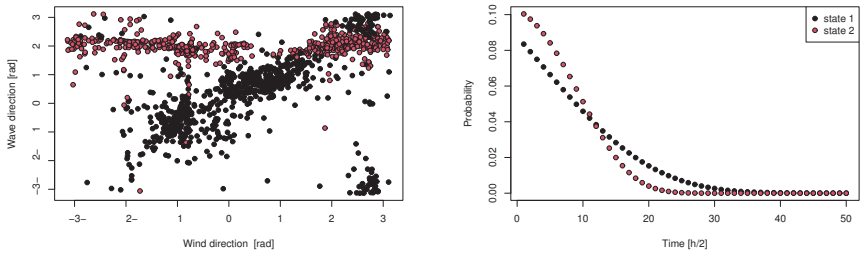


Figure 1. *Left: toroidal data clustered within state 1 (black) and state 2 (red). Right: state-specific dwell time distribution at baseline.*

bivariate wrapped Chaucy densities with dwell time regressions that depend on a baseline Gompertz hazard rate and a time-varying covariate, the *fetch*. The fetch is the closest coastal point following the direction from which the wave comes from and it is computed here by cyclical cubic smoothing splines (Wood, 2017) that appropriately smooth distances across the Adriatic basin.

The model successfully segments the observations according to two clusters, and offers a clear-cut indication of the distribution of the data under each regime. Under state 1, winds appear well syncronized with waves. Under state 2, wind and wave directions are essentially independent. Under state 1, the tail of the baseline dwell time distribution is larger than that one estimated under state 2, indicating that state 1 is more persistent than state 2. The regression coefficient of the fetch is equal to -1.38, indicating that the longer is the distance from the coast, the smaller is the probability of a state transition.

References

KATO, S., & PEWSEY, A. 2015. A Mobius transformation-induced distribution on the torus. *Biometrika*, **102**, 359–370.

LAGONA, F, & PICONE, M. 2013. Maximum likelihood estimation of bivariate circular hidden Markov models from incomplete data. *Journal of Statistical Computation and Simulation*, **83**, 1223–1237.

WOOD, SIMON N. 2017. *Generalized additive models: an introduction with R*. CRC Press.

HOW TO BUILD YOUR LATENT MARKOV MODEL: THE ROLE OF TIME AND SPACE

Roland Langrock[1] and Sina Mews[1]

[1] Department of Business Adminstration and Economics, Bielefeld University, (e-mail: roland.langrock@uni-bielefeld.de, sina.mews@uni-bielefeld.de)

ABSTRACT: In empirical research involving latent Markov models, there is a tendency of research communities building up expertise on one particular class of such models, then shoehorning any given data set into that very model formulation. This talk attempts to overcome this myopia by offering a unifying view on what otherwise are often considered completely separate model classes — from hidden Markov models to Cox processes — thereby providing guidance as to how a latent Markov model formulation can be suitably tailored to the data at hand.

KEYWORDS: Cox process, hidden Markov model, state-space model.

1 Introduction

Over the last two decades, latent Markov models* have taken applied research by storm. This success story can be explained by their intuitive appeal, their mathematical tractability, and the various types of inference they allow for. Yet, while empirical researchers are well-acquainted with the various flavours of regression, the same cannot be said for latent Markov models. Instead, a tendency can be recognised that researchers focus on building expertise on one particular type of such models, and then shoehorn any given data set into the model they happen to know best.

The challenge of identifying a suitable model formulation for a given data set primarily concerns two choices to be made: whether to use a discrete-time or a continuous-time model formulation, and whether to assume a discrete or a continuous state space. Classifying different model classes along these two dimensions, we provide an overview of the most relevant classes of latent

*i.e. stochastic process models for sequential data driven by latent Markovian processes; note these may also be referred to as, *inter alia*, state-space models, hidden Markov models, doubly stochastic processes, or dependent mixture models — we use the label "latent Markov model" as it appears to be a good umbrella term for all special cases considered in this paper

Markov models and emphasise that the inferential methods for these different classes are for the most part effectively identical, such that there is no reason why researchers should focus on any one class of these models.

2 Overview of latent Markov model formulations

Table 1 attempts to classify the main types of latent Markov models according to the type of states, either discrete or continuous, the observed process is assumed to be driven by, and the role of time, i.e. the mathematical operationalisation of the times at which the sequential observations are made.

Table 1. *Classification of six popular classes of latent Markov models according to the respective role of time and space.*

	discrete states	continuous states
discrete time	*(A)* (basic) hidden Markov model	*(B)* (basic) state-space model
continuous time, non-inform. obs. times	*(C)* continuous-time hidden Markov model	*(D)* continuous-time state-space model
continuous time, inform. obs. times	*(E)* Markov-modulated Poisson process	*(F)* Cox process

The simplest case *(A)* arises when the states are discrete and the process is modelled in discrete time, i.e. as a time series $\{X_t\}_{t=1,\ldots,T}$. In its basic dependence structure, the corresponding hidden Markov model (HMM) is defined by an N–state homogeneous Markov chain $\{S_t\}_{t=1,\ldots,T}$ as the state process, specified by the initial distribution $\delta = (\delta_1, \ldots, \delta_N)$, $\delta_i = \Pr(S_1 = i)$, and the $N \times N$ transition probability matrix $\Gamma = (\gamma_{ij})$, as well as the N emission distributions $f_1(x_t), \ldots, f_N(x_t)$, which are selected by the state process. An intuitive example is animal movement, where the observations $x_1, \ldots, x_T$ could be the hourly step lengths of an animal and the states the behavioural modes (cf. Beumer *et al.*, 2020). HMMs are mathematically tractable, as recursive techniques can be used for likelihood evaluation, state decoding, and forecasting.

In many settings, it will however not be reasonable to assume that the state process $\{S_t\}$ is discrete-valued. For example, the volatility underlying share returns evolves gradually over time. In such cases, it is more adequate to model

the discrete-time state process $\{S_t\}_{t=1,\dots,T}$ as an autoregressive process,

$$s_t = \phi(s_{t-1} - \mu) + \mu + \sigma\varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0,1),$$

with long-term mean $\mu \in \mathbb{R}$, persistence parameter $-1 < \phi < 1$ and standard deviation $\sigma > 0$ of the error process, and with the distribution of x_t in some way depending on s_t. Such a model is commonly referred to as *(B)* state-space model (SSM), and there are many different techniques for estimating the associated parameters, ranging from the Kalman filter to MCMC-based methods (Auger-Méthé *et al.*, 2021). While this plethora of estimation techniques can be intimidating for practitioners, it is worth pointing out that SSMs can conveniently be approximated using HMMs with a large state space.

Basic HMMs or SSMs both need to be modified when the intervals between observation times are not of the same length. Such temporally irregular sampling schemes are quite common for example in medical or survey data. If in such cases a discrete state space seems adequate, then $\{S_t\}_{0 \le t \le T}$ can be modelled as a continuous-time Markov chain, specified by the infinitesimal generator matrix $\mathbf{Q} = (q_{ij})$, with state transition intensities

$$q_{ij} = \lim_{\Delta t \to 0} \frac{\Pr(s_{t+\Delta t} = j \mid s_t = i)}{\Delta t},$$

leading to *(C)* a continuous-time HMM. If instead the states ought to be modelled as continuous-valued, then a stochastic differential equation (SDE) can be used, e.g. the Ornstein-Uhlenbeck process

$$ds_t = \theta(\mu - s_t)dt + \sigma dw_t,$$

where w_t is the Brownian motion and $\theta > 0$ controls the strength of reversion to the long-term mean μ. Such a model would most naturally be labelled *(D)* a continuous-time SSM. For inference, recursive techniques similar to the discrete-time case are available (Jackson *et al.*, 2003; Mews *et al.*, 2022b).

Finally, we need to distinguish cases where the observation times themselves are informative, e.g. in medicine, when longitudinal observations are made whenever a patient goes to a doctor, likely indicating sickness. In such cases, *(E)* Markov-modulated Poisson processes (MMPPs) can be used to model a system traversing through a finite state space, with the observation times modelled as a Poisson arrival process with rate λ_{s_t} depending on the state s_t currently active. Such a model can be further extended by including marks, say for modelling biomarkers measured at each consultation (Mews *et al.*, 2022a).

If assuming only finitely many states of such a process is inadequate, then the continuous-time Markov chain model for $\{S_t\}$ can again be replaced by an SDE, leading to the class of *(F)* Cox processes.

3 Conclusion

By classifying latent Markov models according to the assumptions made concerning time and (state) space, we promote a more unified view on what otherwise are often considered fairly separate model classes. This categorisation is far from perfect — for example, as it stands it does not have a place for SDEs driven by latent states — however, we hope that it can provide some guidance for empirical researchers when making their modelling decisions. The main point we are trying to make is that "you should model the process that gives rise to the data, not shoehorn the data into a model you happen to have at hand" (quote by David L. Borchers, pers. communication) — and to be able to do the former, it is important to have a big picture view of the model classes available.

References

AUGER-MÉTHÉ, M., NEWMAN, K., COLE, D., EMPACHER, F., GRYBA, R., KING, A.A., LEOS-BARAJAS, V., MILLS FLEMMING, J., NIELSEN, A., PETRIS, G., & THOMAS, L. 2021. A guide to state–space modeling of ecological time series. *Ecological Monographs*, **91**, e01470.

BEUMER, L.T., POHLE, J., SCHMIDT, N.M., CHIMIENTI, M., DESFORGES, J.-P., HANSEN, L.H., LANGROCK, R., PEDERSEN, S.H., STELVIG, M., & VAN BEEST, F.M. 2020. An application of upscaled optimal foraging theory using hidden Markov modelling: year-round behavioural variation in a large arctic herbivore. *Movement Ecology*, **8**, 1–16.

JACKSON, C.H., SHARPLES, L.D., THOMPSON, S.G., DUFFY, S.W., & COUTO, E. 2003. Multistate Markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **52**, 193–209.

MEWS, S., SURMANN, B., HASEMANN, L., & ELKENKAMP, S. 2022a. Markov-modulated marked Poisson processes for modelling disease dynamics based on medical claims data. *arXiv:2210.13133*.

MEWS, S., LANGROCK, R., ÖTTING, M., YAQINE, H., & REINECKE, J. 2022b. Maximum approximate likelihood estimation of general continuous-time state-space models. *Statistical Modelling*, **in press**, doi:10.1177/1471082X211065785.

THE COMPARATIVE ANALYSIS OF PUBLICATION ACTIVITY IN HUNGARY AND POLAND IN THE FIELD OF ECONOMICS, FINANCE AND BUSINESS

Paweł Lula[1], Zsuzsanna Géring[2], Magdalena Talaga[3], Ildikó Dén-Nagy[2] and Réka Tamássy[2]

[1] Department of Computational Systems, Krakow University of Economics, (e-mail: pawel.lula@uek.krakow.pl)

[2] Future of Higher Education Research Centre, Budapest Business School, (e-mail: gering.zsuzsannamargit@uni-bge.hu, den-nagy.ildiko@uni-bge.hu, tamassy.reka@uni-bge.hu)

[3] Unit for Evaluation and Quality Assurance of Research Activity, Krakow University of Economics, (e-mail: magdalena.talaga@uek.krakow.pl)

ABSTRACT: The analysis of publication activity of Hungarian and Polish researchers in the field of economics, business and finance is the main goal of the presentation. The study covers publication achievements having the form of journal papers, published in the period from 2017 to 2022, and registered in the Scopus database. All papers with at least one author from Hungary or Poland have been taken into account. In the presentation the following issues for both countries will be discussed: publication effectiveness, distribution of citations, structure and internationalization of authors' teams, collaboration networks (similarity to small-world networks and scale-free networks), identification of main science centers related to economics, finance and business, evaluation of main research institutions active in discussed area in both countries, identification of main topics discussed in papers and assessment of their popularity. Next, the cooperation between researchers from Hungary and Poland will be evaluated. All analysis will be performed by authors with the use of data retrieved from the Scopus database and programs prepared in R and Python.

KEYWORDS: bibliometrics, scientometrics, collaboration networks, exploratory text analysis, Latent Dirichlet Allocation method.

References

ARIA, M., & CUCCURULLO, C. 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics.*, **11(4)**, 959-975.

BLEI, D. M., NG, A. Y., & JORDAN, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993-1022.

NEWMAN, M. E. J. 2010. *Networks: an introduction*. New York: Oxford University Press.

AN R PACKAGE FOR MULTILEVEL LATENT CLASS ANALYSIS WITH COVARIATES

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

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

[2] Department of Methodology and Statistics, Leiden University

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

[4] Department of Methodology, London School of Economics

ABSTRACT: In this article we introduce *multilevLCA* - an R package for efficient estimation of single-level and multilevel latent class models with covariates.

KEYWORDS: Multilevel latent class analysis, R package, two-step estimation.

1 Introduction

Latent class (LC) analysis is to create a discrete classification of units based on a set of observed variables, which are taken as observed indicators of an unknown nominal variable with some number of latent classes. Multilevel LCA has been developed to account for hierarchical data structures, i.e., when lower-level units are nested within higher-level ones (e.g., survey respondents nested within countries, pupils within schools). The multilevel LC model can be extended to allow for external covariates as predictors of class membership.

The general recommendation for fitting single-level and multilevel LC models with covariates is to use stepwise estimators. In particular, the two-step (Di Mari *et al.*, 2023) and two-stage approaches (Bakk *et al.*, 2022) for multilevel LCA, and the two-step approach for single-level LCA (Bakk & Kuha, 2018) have some attractive properties with respect to model construction, and estimation efficiency and algorithmic stability.

In the current paper we introduce the R package `multilevLCA` - the first to implement two-step estimation, in a functional and user-friendly way, for single-level and multilevel latent class analysis with covariates.

2 Modelling framework

Let Y_{ijh} denote the observed response of low-level unit (individual) i in high-level unit (group) $j = 1,\ldots,J$ on the categorical indicator variable $h = 1,\ldots,H$. The full response vector for the same unit is denoted $\mathbf{Y}_{ij} = (Y_{ij1},\ldots,Y_{ijH})$. For simplicity of exposition, we focus below on dichotomous indicators, with a conditional Bernoulli distribution, $P(Y_{ih} = y_{ih}|X_i = t) = \phi_{h|t}^{y_{ih}}(1-\phi_{h|t})^{1-y_{ih}}$.

Let W_j be a group-level latent class variable, with possible value $m = 1,\ldots,M$, and probabilities $P(W_j = m) = \omega_m > 0$. Given a realization of W_j, let X_{ij} be a individual-level latent class variable, with possible values $t = 1,\ldots,T$, and conditional probabilities $P(X_i = t|W_j = m) = \pi_{t|m} > 0$.

We assume that individual response probabilities are conditionally independent from each other given low-level class membership (the classical *local independence* assumption). We further assume that individual response probabilities depend on high-level class membership only through X_{ij} (a common assumption in multilevel LCA; Vermunt, 2003; Lukociene *et al.*, 2010). Then, an unconditional multilevel LC model for $\mathbf{Y}_{ij}$ can be specified as follows:

$$P(\mathbf{Y}_{ij}) = \sum_{m=1}^{M} P(W_j = m) \sum_{t=1}^{T} P(X_{ij} = t|W_j = m) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t). \quad (1)$$

High-level and low-level covariates can be included in order to predict class membership. Let $\mathbf{Z}_{ij} = (1, \mathbf{Z}'_{1j}, \mathbf{Z}'_{2ij})'$ be a vector K covariates, with the sub-vector $\mathbf{Z}'_{1j}$ being defined at the high level, and $\mathbf{Z}'_{2ij}$ being defined at the low level. Let $\mathbf{Z}^*_{1j} = (1, \mathbf{Z}'_{1j})'$. For high-level and low-level latent class membership, respectively, we consider the multinomial logistic models

$$P(W_j = m|\mathbf{Z}^*_{1j}) = \frac{\exp(\alpha'_m \mathbf{Z}^*_{1j})}{1 + \sum_{l=2}^{M} \exp(\alpha'_l \mathbf{Z}^*_{1j})}, \quad (2)$$

$$P(X_{it} = t|W_j = m, \mathbf{Z}_{ij}) = \frac{\exp(\gamma'_{tm} \mathbf{Z}_{ij})}{1 + \sum_{s=2}^{T} \exp(\gamma'_{sm} \mathbf{Z}_{ij})}, \quad (3)$$

In Equation (2), α_m are regression coefficients for $m = 2,\ldots,M$, and $m = 1,\ldots,M$. In Equation (3), γ_{tm} is a vector of regression coefficients for each $t = 2,\ldots,T$. When only the intercept is included in Equation (2), or (3), the corresponding vector of regression coefficients is equal to the log-odds of the class proportions (i.e., $\log(\omega_m/\omega_1)$, or $\log(\pi_{t|m}/\pi_{1|m})$).

In addition, we assume that the observed indicators Y_{ijh} are conditionally independent from the covariates given low-level class membership. Thus, the multilevel LC model for $P(\mathbf{Y}_{ij}|\mathbf{Z}_{ij})$ can be written as:

$$P(\mathbf{Y}_{ij}|\mathbf{Z}_{ij}) = \sum_{m=1}^{M} P(W_j = m|\mathbf{Z}^*_{1j}) \sum_{t=1}^{T} P(X_{ij} = t|W_j = m, \mathbf{Z}_{ij}) \prod_{h=1}^{H} P(Y_{ijh}|X_{ij} = t). \quad (4)$$

The class profiles are defined by the measurement parameters $\phi_{h|t}$, $\pi_{t|m}$, and ω_m. The other parameters of interest are the structural parameters α_m, and γ_{tm}. It is straightforward to reduce the multilevel LC structural model in Equation (4) to the multilevel measurement model, the single-level structural model, or the single-level measurement model.

3 Estimating the multilevel LC model in multilevLCA

The default estimator of Equation (4), in the R package `multilevLCA`, is the two-step approach (Di Mari *et al.*, 2023). We add that future versions of the package will relax the assumptions of Equation (4) to allow for local dependencies. Other options are the two-stage (Bakk *et al.*, 2022) and the simultaneous approaches. A basic function call requires the following arguments:

- `data` The input data (matrix or data frame)
- `Y` The names of the item columns
- `iT` The number of low-level latent classes
- `id_high` The name of the high-level id column
- `iM` The number of high-level latent classes
- `Z` The names of the low-level covariates columns
- `Zh` The names of the high-level covariates columns

Estimation is performed via the function `multiLCA`,

```
out = multiLCA(data,Y,iT,id_high,iM,Z,Zh)
```

The list `out` contains a lot of information about class profiles, structural parameters, and estimation details. A summary of this information can be printed by executing `out` in the prompt. To create a plot of the response probabilities, the user types `plot(out)` in the prompt.

In practice, the number of low-level and high-level classes is unknown to the researchers. Selecting these values is a distinct, yet fundamental task. The `multilevLCA` package includes two state-of-the art model selection strategies, namely sequential model selection (Lukociene et al. 2010) and simultaenous model selection. Both approaches implement the BIC selection criterion on low and high level, reporting also the AIC and ICL BIC.

To implement the former, `iT` and (or) `iM` is replaced by a range of values. The latter is implemented in the same way, but with the extra argument `sequential` set to `FALSE`. For example, to perform simultaneous model selection over 1-4 low-level classes, and 3-4 high-level classes, we execute the following call:

```
out = multiLCA(data,Y,iT=1:4,id_high,iM=3:4,
sequential=FALSE)
```

The list `out` contains the model estimation results as if the selected specification had been estimated directly. Note that specifying `Z` and `Zh` is redundant; in `multilevLCA`, model selection is always performed without covariates.

The tools for model selection, and visualization are available for any LC model, i.e., the multilevel structural model, multilevel measurement model, single-level structural model, and single-level measurement model.

References

BAKK, Z., & KUHA, J. 2018. Two-step estimation of models between latent classes and external variables. *Psychometrika.*, **83**, 871–892.

BAKK, Z., DI MARI, R., OSER, J., & KUHA, J. 2022. Two-stage multilevel latent class analysis with covariates in the presence of direct effects. *Structural Equation Modeling: A Multidisciplinary Journal.*, **29**(2), 267–277.

DI MARI, R., BAKK, Z., OSER, J., & KUHA, J. 2023. A two-step estimator for multilevel latent class analysis with covariates. *arXiv preprint arXiv:2303.06091*.

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.

VERMUNT, J. K. 2003. Multilevel latent class models. *Sociological Methodology.*, **33**(1), 213–239.

LONGITUDINAL HIDDEN MARKOV MODELS: PROBLEMS AND METHODS

Mackenzie R. Neal [1] and Paul D. McNicholas[1]

[1] Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada (e-mail: nealm6@mcmaster.ca, paul@math.mcmaster.ca)

ABSTRACT: Methods to handle common data problems for longitudinal hidden Markov models are presented. A missing data mechanism that assumes state-dependent and variable dependent missingness is introduced. High dimensionality is controlled for with the use of an explicit dimension reduction algorithm.

KEYWORDS: mixture model, expectation-maximization, initialization, variable selection, missing data

1 Introduction

Hidden Markov models (HMMs) are dependent mixture models wherein the unobserved process is governed by a Markov process. Traditionally HMMs are used to model time series data and recently have been used to model the movement of subjects across time, i.e., longitudinal data. Due to the abundance of multivariate longitudinal data arising from clinical studies, HMMs have become increasingly useful to the health sciences. This data type, however, is commonly plagued by missing data as individuals miss visits or drop-out of studies. Classically, we account for missing data through one of two means: the inclusion of only individuals with complete data or variable mean imputation. Both of which can introduce bias into analysis results due to reduction in the information provided or by distorting the information provided. Alternative to these pre-processing missing data methods, is the use of model fitting algorithms that can be altered to handle missing data at each iteration. One such algorithm is the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). We adopt this approach and develop a modified EM for longitudinal HMMs with informative missing data. In addition to missing data, approaches for handling high dimensionality and uninformative variables must be developed for longitudinal HMMs. Many implicit and explicit dimension reduction methods exist for independent mixture models. We focus on explicit dimension reduction, and extend the `vscc` algorithm (Andrews & McNicholas, 2014) to longitudinal HMMs.

2 Background

2.1 Longitudinal hidden Markov models

Longitudinal hidden Markov models contain an unobservable first-order Markov chain $S_{it}, i = 1,...,n, t = 1,..,T$ and an observed process $\mathbf{Y}_{it}$ representing the response vector of individual i at time t. The simplest model of this kind can by summarized by

$$Pr(S_{i1}^{t}|\mathbf{S}_{i1}^{t-1}) = Pr(S_t|S_{t-1}), i = 1,...,n, t = 2,3,...,T \tag{1}$$

$$Pr(Y_{it}|\mathbf{Y}_{i1}^{t-1}, \mathbf{S}_{i1}^{t}) = Pr(Y_{it}|S_{it}), i = 1,...,n, t = 1,2,...,T \tag{2}$$

where S_{i1}^{t} represents the history of the unobserved parameter process for individual i, from time 1 to time t, with state space $S = 1,...,m$,and $\mathbf{Y}_{i1}^{t}$ represents the history of the state-dependent process. The HMM parameters include both the parameters from the Markov chain and the state-dependent distribution, often taken to be Gaussian. The Markov chain parameters include the transition matrix Γ where $\gamma_{it\,jk} = P(S_{it} = k|S_{it-1} = j)$ and the initial probabilities δ where $\delta_{ij} = P(S_{i0} = j)$. The simplest model assumes homogeneity, thus $\gamma_{it\,jk} = \gamma_{jk}$ and $\delta_{ij} = \delta_j$. To ease calculation of the likelihood, we introduce forwards and backwards probabilities. The forwards probabilities is defined as such $\alpha_{it}(j) = P(\mathbf{Y}^{(t)}, S_{it} = j) = \delta\mathbf{P}(\mathbf{Y}_{i1})\Gamma\mathbf{P}(\mathbf{Y}_{i2})\ldots\Gamma\mathbf{P}(\mathbf{Y}_{it})$ and the backwards probabilities are defined as $\beta_{it}(j) = P(\mathbf{Y}_{it+1}^{T}, S_{it} = j)$, thus $\beta_{it}^{\top} = \Gamma\mathbf{P}(\mathbf{Y}_{it+1})\ldots\Gamma\mathbf{P}(\mathbf{Y}_{iT})1^{\top}$. The likelihood is as follows

$$L_T = \prod_{i=1}^{n} \delta\mathbf{P}(\mathbf{Y}_{i1})\Gamma\mathbf{P}(\mathbf{Y}_{i2})\ldots\Gamma\mathbf{P}(\mathbf{Y}_{iT})1^{\top} \tag{3}$$

and can be redefined with respect to the forwards or backwards probabilities via $L_T = \prod_{i=1}^{n} \alpha_{it}\beta_{it}^{\top}$ or $L_T = \prod_{i=1}^{n} \alpha_{iT}1^{\top}$.

2.1.1 Model estimation

Various versions of the EM algorithm for HMMs exist, in this paper we use the Baum-Welch algorithm (Baum *et al.*, 1970; Welch, 2003) to obtain maximum likelihood estimates. The Baum-Welch algorithm is based on max-

imization of the complete-data log-likelihood, as seen below

$$l(\vartheta) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m} u_{i0j} \log \delta_j + \sum_{t=1}^{T} \sum_{j=1}^{m} \sum_{k=1}^{m} v_{itjk} \log \gamma_{jk} + \sum_{t=0}^{T} \sum_{j=1}^{m} u_{itj} \log f(y_{it} | S_{it} = j) \right\}. \tag{4}$$

The E-step consists of calculating expectations of the missing data, $u_{itj} = P(S_{it} = j|\mathbf{Y}_{i1}^{T})$ and $v_{itjk} = P(S_{it-1} = j, S_{it} = k|\mathbf{Y}_{i1}^{T},)$. The M-step consists of obtaining the maximum likelihood estimates with respect to the expected complete-data log-likelihood. In particular, the MLE for δ_j and γ_{jk} with respect to u_{itj} and v_{itjk} are

$$\delta_j = \frac{\sum_{i=1}^{n} \hat{u}_{i0j}}{n} \tag{5}$$

and,

$$\gamma_{jk} = \frac{\sum_{i=1}^{n} \sum_{t=1}^{T} \hat{v}_{itjk}}{\sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{k=1}^{m} \hat{v}_{itjk}}. \tag{6}$$

Additionally, the state-dependent distribution parameters are estimated in the M-step, based on the assumed distribution.

2.2 Missing data

Missing data for model-based clustering is a well studied problem, beginning with Eirola *et al.* (2014). The data is first partitioned into the observed and unobserved parts as such $(\mathbf{Y}_i^o, \mathbf{Y}_i^m)$. By assuming the joint distribution of the missing and observed part to be Gaussian, we can obtain the conditional distribution of the missing part given the observed part as

$$(\mathbf{Y}^m|\mathbf{Y}^o) \sim \mathcal{N}(\mu_m + \Sigma_{mo}\Sigma_{oo}^{-1}(\mathbf{Y}^o - \mu_o), \Sigma_{m|o}) \tag{7}$$

(Anderson, 2003). Based on these assumptions the conditional expectation of the missing data and the conditional covariance matrices can be determined and used in the EM algorithm to account for missingness. We extend this method to longitudinal HMMs and add in methods to handle informative missingness.

2.3 Variable selection

The `vscc` algorithm, proposed by Andrews & McNicholas (2014), selects variables based on minimization of within-cluster variance and correlation to the set of selected variables. The `vscc` algorithm tends to be much faster and

perform better than step-wise variable selection methods where model fitting occurs at every inclusion/exclusion step.

3 Methodology

Similar to Sportisse *et al.* (2021), we modify the Baum-Welch algorithm to allow for state-dependent and variable-dependent missingness. We do so by adjusting the definitions of the forwards and backwards probabilities, which are then used to update the E and M steps. Additionally, E and M steps are added to implement conditional mean and covariance imputation and to estimate the missingess parameters.
The modified Baum-Welch algorithm is used within the `vscc` algorithm, to allow for simultaneous handling of missing data and uninformative variables.
The mathematical results and full model estimation algorithm will be given in the full paper, as well as illustrations on real and simulated data

References

ANDERSON, T.W. 2003. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley.

ANDREWS, JEFFREY L, & MCNICHOLAS, PAUL D. 2014. Variable selection for clustering and classification. *Journal of Classification*, **31**(2), 136–153.

BAUM, LEONARD E, PETRIE, TED, SOULES, GEORGE, & WEISS, NORMAN. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, **41**(1), 164–171.

DEMPSTER, ARTHUR P, LAIRD, NAN M, & RUBIN, DONALD B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, **39**(1), 1–22.

EIROLA, EMIL, LENDASSE, AMAURY, VANDEWALLE, VINCENT, & BIERNACKI, CHRISTOPHE. 2014. Mixture of Gaussians for distance estimation with missing data. *Neurocomputing*, **131**, 32–42.

SPORTISSE, AUDE, MARBAC, MATTHIEU, BIERNACKI, CHRISTOPHE, BOYER, CLAIRE, CELEUX, GILLES, JOSSE, JULIE, & LAPORTE, FABIEN. 2021. Model-based clustering with missing not at random data. *arXiv preprint arXiv:2112.10425*.

WELCH, LLOYD R. 2003. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, **53**(4), 10–13.

CLUSTER ANALYSIS FOR THE STUDY OF ONLINE VISUAL COMMUNICATION

Matteo Magnani[1], Matias Piqueras[1], Alexandra Segerberg[2], Davide Vega[1] and Victoria Yantseva[1]

[1] InfoLab, Department of Information Technology, Uppsala University, Sweden (e-mail: matteo.magnani@it.uu.se, matias.piqueras@it.uu.se, davide.vega@it.uu.se, victoria.yantseva@it.uu.se)

[2] Department of Government, Uppsala University, Sweden (e-mail: alexandra.segerberg@statsvet.uu.se)

ABSTRACT: Cluster analysis is a fundamental tool for the study of online communication. In this contribution we focus on the task of clustering online communication networks containing images. We provide an overview of the available approaches to cluster images and how to use image clustering as part of a social data science process. Then we present an approach to cluster online communication networks based on image clustering, that we apply to the study of climate change communication.

KEYWORDS: Online communication, image clustering, networks.

1 Cluster analysis and communication studies

Several online data analysis tasks such as the detection of coordinated inauthentic behaviour used to amplify online disinformation and the study of online polarisation and political persuasion require the summarisation of large numbers of social media posts. Therefore, cluster analysis is a fundamental tool for the study of online communication.

When we consider the type of data generated by people communicating online, visual content plays a fundamental role in all the main social media platforms. Images often carry an important part of the information contained in a social media post; in general, images are also associated to increased spreading, and can be shared across languages (Joo & Steinert-Threlkeld, 2018; Magnani *et al.*, 2013). This is true for image-based (Instagram), video-based (YouTube), and micro-blogging (Twitter) platforms. However, most of the existing computational studies of online communication, including those based on clustering methods, have focused only on either the networks of interaction (e.g. replies and retweets), to discover communities, or the text contained in social media posts, to compute topic models.

In this contribution we provide a road map towards the usage of image clustering to summarise online communication networks, starting from an overview of the available approaches to cluster images. Recent advances in deep learning have added more options to the traditional tools used to extract features from images, providing a number of alternative clustering approaches that still require a thorough comparison.

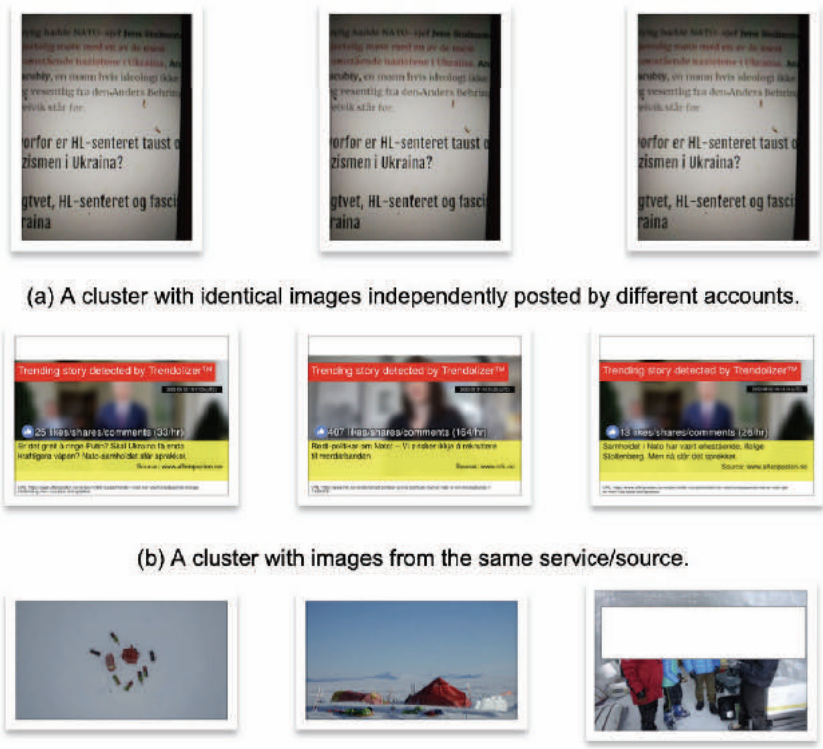


Figure 1. *Three different types of clusters, from Deliverable 5.1, NORdic observatory for digital media and information DISorder (NORDIS)*

The focus of this part of the presentation is on clustering social images (for example, images shared on social media) and how to use image clustering as part of a social data science process (Magnani & Segerberg, 2021; Chen *et al.*, 2021; Giordano *et al.*, 2021; Zhang & Peng, 2022). Figure 1 shows an example of how the same approach, in this case based on clustering the colour histograms of the images, can produce clusters with different functions:

a cluster collecting identical images, a cluster of images coming from the same source, and a cluster of images showing pictures from the same event.

Then we present an approach to cluster online communication networks based on image clustering. This is based our previous work (Vega & Magnani, 2018; Vega & Magnani, 2019) on clustering networks with temporal and textual information (Figure 2), and consists in applying image clustering to organise the posted images into groups, assigning a label to each group or sets of groups to characterise the theme of the interactions, and using these themes to define thematic multiplex networks of social interactions. In these networks, edges represent interactions (e.g. replies, or retweets), and each layer of the multiplex network only includes interactions happened around that theme. Such networks can themselves be clustered using algorithms for multiplex networks (Magnani *et al.*, 2021).

Finally, we conclude by showing an application of image-based communication network clustering to study online visual climate change communication. This is based on a collection of tweets using the #COP*xx* hashtag, with *xx* being the number of the Conference of the Parties (COP) meeting, e.g. COP21 being the meeting held in Paris in 2015. The objective of this case study is both to highlight the importance of different features of image clustering methods within this type of research (e.g. scalability and explainability), and to showcase the possibilities and limitations of this research design.

References

CHEN, YAN, SHERREN, KATE, SMIT, MICHAEL, & LEE, KYUNG YOUNG. 2021. Using social media images as data in social science research. *New Media & Society*, Aug., 146144482110387.

GIORDANO, GIUSEPPE, PRIMERANO, ILARIA, & VITALE, PIERLUIGI. 2021. A Network-based Indicator of Travelers Performativity on Instagram. *SOCIAL INDICATORS RESEARCH*, **156**, 165–179.

JOO, JUNGSEOCK, & STEINERT-THRELKELD, ZACHARY C. 2018. Image as Data: Automated Visual Content Analysis for Political Science. *arXiv:1810.01544 [cs, stat]*. arXiv: 1810.01544.

MAGNANI, MATTEO, & SEGERBERG, ALEXANDRA. 2021. On the conditions for integrating deep learning into the study of visual politics. *In: 13th ACM Web Science Conference*.

MAGNANI, MATTEO, MONTESI, DANILO, & ROSSI, LUCA. 2013. Factors Enabling Information Propagation in a Social Network Site. *Pages 411–426 of:* ÖZYER, TANSEL, ROKNE, JON, WAGNER, GERHARD, &

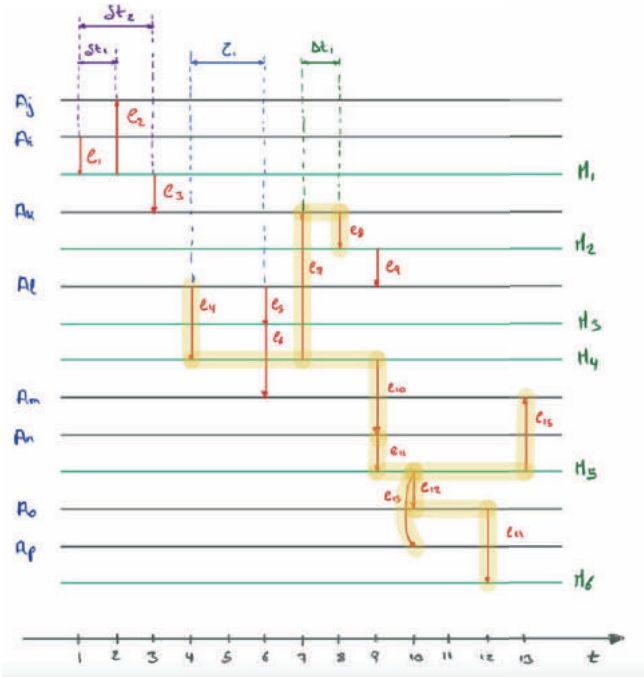


Figure 2. *A model for the analysis of networks and content, from Vega & Magnani, 2019, with A_x representing social media users and M_x representing the content they exchange*

REUSER, ARNO H.P. (eds), *The Influence of Technology on Social Network Analysis and Mining*. Lecture Notes in Social Networks. Vienna: Springer.

MAGNANI, MATTEO, HANTEER, OBAIDA, INTERDONATO, ROBERTO, ROSSI, LUCA, & TAGARELLI, ANDREA. 2021. Community detection in multiplex networks. *ACM Computing Surveys*, **54**(3).

VEGA, DAVIDE, & MAGNANI, MATTEO. 2018. Foundations of Temporal Text Networks. *Applied Network Science*, **3**(1), 25. _eprint: 1803.02592.

VEGA, DAVIDE, & MAGNANI, MATTEO. 2019. Metrics for Temporal Text Networks. Springer, Cham.

ZHANG, HAN, & PENG, YILANG. 2022. Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research. *Sociological Methods & Research*, Apr., 004912412210826.

CLUSTER ANALYSIS OF CANCER METABOLIC NETWORK ENSEMBLES

Ichcha Manipur[1], Ilaria Granata[1], Lucia Maddalena[1]
and Mario R. Guarracino[2]

[1] High-Performance Computing and Networking Institute, Consiglio Nazionale delle Ricerche, (e-mail: ichcha.manipur@icar.cnr.it, ilaria.granata@icar.cnr.it, lucia.maddalena@cnr.it)

[2] Department of Economics and Law, University of Cassino and Southern Lazio, (e-mail: mario.guarracino@unicas.it)

ABSTRACT: Biological networks are representative of the diverse molecular interactions that occur within cells. They model protein-protein interactions, gene regulation, and metabolic pathways. Among these, metabolic networks are of great interest, as they directly influence all physiological processes. Exploration of biochemical pathways using multigraph representation is essential in understanding complex regulatory mechanisms. We present a cluster analysis on tissue-specific metabolic networks for three primary tumor types: breast, lung, and kidney cancer. The metabolic networks integrate genome-scale metabolic models with gene expression data. We empirically proved that network clustering could characterize groups of patients in multiple conditions to explore and characterize the metabolic landscape of tumors.

KEYWORDS: biological network ensembles, network summarization, networks clustering.

Acknowledgements

IM contributed to the present work while she was a PhD student at High-Performance Computing and Networking Institute, Consiglio Nazionale delle Ricerche. The present work is adapted from https://doi.org/10.1186/s12859-020-03564-9, and released under the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

IMPROVING PERFORMANCE IN NEURAL NETWORKS BY DENDRITE-ACTIVATED CONNECTION

Carlo Metta[1], Marco Fantozzi, Andrea Papini [2], Gianluca Amato [3], Matteo Bergamaschi [4], Silvia Giulia Galfrè [5], Alessandro Marchetti[*,3], Michelangelo Vegliò [3], Maurizio Parton [3], and Francesco Morandin [6]

[1] ISTI-CNR Pisa, Italy

[2] Scuola Normale Superiore, Pisa, Italy

[3] University of Chieti-Pescara, Italy

[4] University of Padova, Italy

[5] University of Pisa, Italy

[6] University of Parma, Italy

ABSTRACT: In artificial neural networks, computational units typically compute a linear combination of their inputs and then apply a nonlinear filter, often a ReLU, shifted by some bias. If the inputs come from other units, they have already been filtered with their own biases. In a layer, multiple units share the same inputs, and each input is filtered with a unique bias. This results in output values based on *shared* input biases rather than individually optimized ones. To address this issue, we introduce DAC, a new computational unit that incorporates preactivation and multiple biases. This design allows input signals to undergo independent nonlinear filtering before the linear combination.

In this short note, we sketch the design of this new computational unit. Full theoretical support and empirical evidence, suggesting that DAC could be an improved design for the basic computational unit in neural networks, can be found in Metta *et al.*, 2023. Code at `https://github.com/CuriosAI/dac-dev`

KEYWORDS: preactivation, multi-bias, ResNet, dendritic neural model.

1 Introduction

Historically the structure of the perceptron, the artificial neural network's fundamental computational unit, has rarely been questioned. The biological inspiration is straightforward: input signals from the dendrites are accumulated

[*] Alessandro Marchetti is a PhD student enrolled in the National PhD in Artificial Intelligence, XXXVII cycle, course on Health and life sciences, organized by Campus Bio Medico University of Rome.

at the soma (with a linear combination), and if the result is above the activation threshold (that is, the opposite of some bias) there is a nonlinear reaction, as the neuron fires along the axon (with the activation function).

In time the early sigmoid activation function was replaced by ReLU and variants, and this has brought us to the current situation in which most units output their signal through a nonlinear activation function which effectively destroys some information. In fact, ReLU is not invertible, as it collapses to zero all negative values. Though some of its variants may be formally invertible (ELU Clevert *et al.*, 2016 for example), they overall perform in a way very similar to ReLU. This suggests that their way of compressing negative values leads to the same general properties of the latter.

In this note, we describe a radical rethinking of the standard computational unit, where the output brings its full, uncorrupted information to the next units, and only at this point is the activation function applied, with biases specialized for each unit. From the biological point of view, this is like having the activation at the dendrites instead of at the base of the axon. Thus, we call the new unit 'DAC', for 'Dendrite-Activated Connection'.

An extended version of this note, with implementation details, an efficiency analysis in terms of parameters and FLOPs, empirical evidence that DAC provides several benefits with respect to standard units, a theoretical analysis including a universal approximation theorem, and more, can be found in the full paper Metta *et al.*, 2023.

2 From standard to DAC units

To describe this paper's idea, we look at a neural network as a directed acyclic computational graph. We denote the set of its nodes by I. If $i \in I$ is a node, we denote its parents (in-neighbors) by $I_i \subset I$. In the *standard model* for computational units in a neural network, a bias b_i, a set of weights $w_{i,j}$ for $j \in I_i$ and a nonlinearity φ_i are associated with every node i. In this paper, φ_i is always $\varphi = \text{ReLU}$.

Standard model network flow involves updating node i's value y_i using a nonlinear filter, *activation* $\circ$ *bias*, applied to some information linearly aggregated from node i's parents:

$$\begin{cases} z_i = \sum_{j \in I_i} w_{i,j}\, y_j & \text{linear aggregation} \\ y_i = \varphi(b_i + z_i) & \text{nonlinear filter} \end{cases} \tag{1}$$

Figure 1 exhibits this point of view, emphasizing that parent nodes (white boxes) are themselves filtered with biases and ReLU. For each parent node

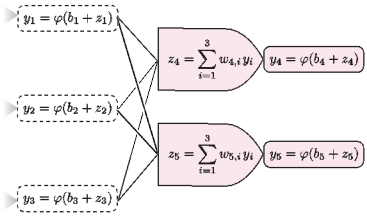


Figure 1: Standard network example. A fully connected layer with 3 input units and 2 computational units. The set of in-neighbors of nodes 4 and 5 is $I_4 = I_5 = \{1,2,3\}$. Bullets and rectangles represent linear aggregation and nonlinear filters (1), respectively. Units 4 and 5 must share the same biases b_1, b_2, b_3 in their inputs, potentially causing outputs y_4 and y_5 to be based on deteriorated information.

$i = 1,2,3$, the bias b_i is uniquely determined. Since children nodes $j = 4,5$ cannot access z_i directly, they must use the filtered versions y_i, sharing the way they are filtered. We refer to this as *shared* biases and argue that it could cause information degradation.

The direct solution to this problem is to apply the nonlinear filter with *non-shared* biases at the input of the unit, before the linear aggregation. We investigate this idea, by studying a new computational unit briefly described as *linearity* $\circ$ (*activation* $\circ$ *non-shared biases*):

$$\begin{cases} y_{i,j} = \varphi(b_{i,j} + z_j) & \text{nonlinear filter} \\ z_i = \sum_{j \in I_i} w_{i,j}\, y_{i,j} & \text{linear aggregation} \end{cases} \tag{2}$$

The artificial neuron described by (2) reflects a recent shift in the understanding of the biological neuron towards a model that incorporates *active* dendrites Larkum, 2022; Magee, 2000. Active dendrites perform a local nonlinear signal modulation before integration at the soma level. Since the biases $b_{i,j}$ in (2) can depend on both input and output nodes, that is, on the edges of the graph, and since these edges correspond to dendrites in biological neurons, we call this new computational unit a *Dendrite-Activated Connection* (DAC) unit. See Metta *et al.*, 2023 for more details on biological inspiration.

One can view DAC as a preactivated unit with multiple *non-shared* input biases, meaning that DAC units sharing the same inputs can filter them with different nonlinearity thresholds, see Figure 2.

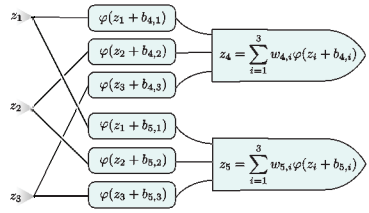


Figure 2: DAC network example: the network from Figure 1 with standard units replaced by DAC units. The input biases b_1, b_2, b_3, contributing to the output values z_4 and z_5, now depend also on the output nodes $4, 5$. We call this feature *non-shared* biases because it allows DAC units sharing the same input to use different (*non-shared*) thresholds instead of a single (*shared*) input bias.

Acknowledgements. Computational resources by CLAI laboratory of Chieti-Pescara. The authors wish to thank Rosa Gini for her important intellectual contribution.

References

CLEVERT, DJORK-ARNÃ©, UNTERTHINER, THOMAS, & HOCHREITER, SEPP. 2016 (Feb.). *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*. arXiv:1511.07289 [cs].

LARKUM, MATTHEW E. 2022. Are Dendrites Conceptually Useful? *Neuroscience*, **489**(May), 4–14.

MAGEE, JEFFREY C. 2000. Dendritic integration of excitatory synaptic input. *Nature Reviews Neuroscience*, **1**(3), 181–190.

METTA, CARLO, FANTOZZI, MARCO, PAPINI, ANDREA, AMATO, GIANLUCA, BERGAMASCHI, MATTEO, GALFRÈ, SILVIA GIULIA, MARCHETTI, ALESSANDRO, VEGLIÒ, MICHELANGELO, PARTON, MAURIZIO, & MORANDIN, FRANCESCO. 2023. *Improving Performance in Neural Networks by Dendrites-Activated Connections*. `https://arxiv.org/abs/2301.00924`.

THE GENERALIZED SHAPLEY MEASURE FOR RANKING PLAYERS IN BASKETBALL: APPLICATIONS AND FUTURE DIRECTIONS

Rodolfo Metulini[1], Francesco Biancalani[2] and Giorgio Gnecco[2]

[1] Department of Economics, University of Bergamo (e-mail: rodolfo.metulini@unibg.it)

[2] Laboratory for the Analysis of CompleX Economic Systems (AXES), IMT School for Advanced Studies (e-mail: francesco.biancalani@imtlucca.it, giorgio.gnecco@imtlucca.it)

ABSTRACT: A wide range of measures have been proposed to quantify a player's marginal contribution to a team. We contributed to this strand of research by proposing, specifically for basketball, a new measure based on a combination of the Shapley value from game theory and the logistic regression, which is based on considering the utility of a player in every single lineup. Some applications where the measure can be useful are presented, such as ranking players, forming lineups, and predicting a remunerative new contract for free agent players. We also discuss possible ideas for future research developments.

KEYWORDS: Sports statistics, Shapley value, logistic regression, players ranking, statistical learning.

1 Introduction & state of the art

Thanks to advancements in technologies and the related increase of available data, measuring the importance of players in team sports to help coaches and staff to win more games is gaining relevance. A wide collection of synthetic indices has been developed in the sport statistics literature to measure each player's contribution to the team win. Among others, we can mention Plus-Minus (PM) and its generalizations (see, e.g., Kubatko *et al.*, 2007; Grassetti *et al.*, 2021), Win-Shares (WS), Wins Above Replacement Player (WARP) and their advances (see, for a review, Sarlis & Tjortjis, 2020, which also highlights pros and cons of such methods). A new measure of players' contribution to the team in basketball has been recently developed (Metulini & Gnecco, 2022). It adopts a combination of a two-step approach based on the logistic regression and the concept of generalized Shapley value (Nowak & Radzik, 1994). This proposal aims to gather most of the advantages (and avoid disadvantages) of industry-standard measures. Recent PM versions moved in the direction of solving some cons, such as just considering only scoring factors,

and multicollinearity. However, those issues still need attention (Terner & Franks, 2021). The measure proposed in Metulini & Gnecco, 2022, similarly to BPM, presents the advantage of being based on both offensive and defensive scoring and non scoring features. Furthermore, the method takes into account probabilities to win the game, which are estimated based on a long time span of box-score synthetic measures (the so-called four Dean's factors, Kubatko *et al.*, 2007) that produce extremely high goodness of fit. Moreover, similarly to what WARP does by introducing the replacement level player, the approach proposed in Metulini & Gnecco, 2022, considering lineups, accounts for marginal utilities of players. This is achieved by explicitly accounting for all the lineups each player has played with. In doing so, considering a proper level for the replacement player is not needed and multicollinearity is avoided.

2 The generalized Shapley measure

The generalized Shapley value for a player in a generalized coalitional game with n players represents his/her average marginal utility to a suitably randomly formed ordered coalition of players. To obtain this measure for basketball players, first, the coefficients of a logistic model applied to game level are computed through the equation $log\frac{P(Y_i=1|\boldsymbol{X})}{P(Y_i=0|\boldsymbol{X})} = \boldsymbol{X}_i\boldsymbol{\beta}$, where the left part of the equation represents the log-odd of Y_i conditional on $\boldsymbol{X}$; $\boldsymbol{Y}$ is the binary response variable representing the outcome of the games, $Y_i \in \{0,1\}$, $i = 1,...,g$, where g is the number of games. $\boldsymbol{X}_i$ is the $i-$th row of the design matrix $\boldsymbol{X}$ with g rows and p columns (p=8, the eight Dean's factors used as explanatory variables computed at the game level). $\boldsymbol{\beta}$ is a vector containing the p regression parameters associated with the explanatory variables. Since the single lineup does not play the full match, to determine the probabilities to win the game for that quintet is not feasible. To deal with this issue, a dataset $\tilde{\boldsymbol{X}}$ where the Dean's factors are computed at the single lineup level (i.e., each row of the dataset corresponds to a lineup) is used and the probabilities to win the game $P(Win)_{L_j}$ is predicted on each lineup L_j by using the vector $\hat{\boldsymbol{\beta}}$ of estimated coefficients from the first step. Let $\tilde{\boldsymbol{X}}_{\boldsymbol{j}}$ be the j-th row of the matrix $\tilde{\boldsymbol{X}}$ with l rows (where l is the number of lineups considered) and p=8 columns (expressing the eight Dean's factors computed at the lineup level). The probabilities to win the game for the lineup L_j is expressed as $P(Win)_{L_j} = \frac{exp(\tilde{\boldsymbol{X}}_{\boldsymbol{j}}\hat{\boldsymbol{\beta}})}{1+exp(\tilde{\boldsymbol{X}}_{\boldsymbol{j}}\hat{\boldsymbol{\beta}})}, j = 1,...,l$. In the third step, one considers two versions (*unweighted* and *weighted*)* of the generalized characteristic function, hence of the (Nowak-Radzik, NR) generalized

*The two differ in terms of taking/not taking in account the time players are on the court.

Shapley value: $\phi_i^{NR}(N,\upsilon) = \frac{1}{n!}\sum_{T\in\mathcal{T}\text{ with }|T|=n}(\upsilon((T(i),i)) - \upsilon(T(i)))$, where $\mathcal{T}$ refers to the set of all ordered coalitions of players, $T(i)$ represents the ordered subcoalition made by the predecessors of i in the permutation T, whereas $(T(i),i)$ is the ordered subcoalition made by $T(i)$ followed by i. $\upsilon : \mathcal{T} \to \mathbb{R}$ (such that $\upsilon(\emptyset) = 0$) is called generalized characteristic function. Metulini & Gnecco, 2022 described two possible choices for the generalized characteristic function $\upsilon(.)$ ($\upsilon_1(.)$ and $\upsilon_2(.)$). When restricted to a lineup, the generalized characteristic function $\upsilon_1(.)$ represents the probability $P(Win)$ to win the game for every specific lineup. At the same time, $\upsilon_2(.)$ is a function of both $P(Win)$ and the probability of occurrence $P(Occ)$ of that lineup on the court. The corresponding generalized Shapley value measures are called unweighted generalized Shapley value (UWGS) and generalized Shapley value (WGS).

3 Applications

The UWGS (WGS) may be used for different purposes. For example, Metulini & Gnecco, 2022, by computing the generalized characteristic function based on all the games (regular season and playoff) from 17 National Basketball Association (NBA) seasons (2004/2005 – 2020/2021)[†], determine (approximations of) the two measures for the Utah Jazz players during season 2020-21, rank players in terms of such measures, and propose a "greedy" algorithm to suggest best lineups conditional to the presence/absence of a specific team player. The algorithm is based on choosing the player with the largest UWGS (WGS), recomputing the UWGS (WGS) of teammates based just on the lineups where the chosen player was in, and repeating the process until five players have been chosen. Since the UWGS and the WGS are composite measures that aim to evaluate a player marginal utility in terms of winning the game, it is reasonable to think that a player may be rewarded with a salary that is proportional to these measures. Biancalani *et al.*, 2023 using income data available at `basketballinsiders.com` and computing such measures for the players of three NBA teams, proposed an instrument to predict the deal of a better contract (compared to the previous year) in the next season based on deviations of estimated salaries (according to a log-linear model) from the true incomes.

4 Possible developments

From a methodological viewpoint, a natural future direction might regard developing a generalized Shapley measure that takes into account players' roles

[†]Features of the logistic model's dependent variable and Dean's factors for both $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ are computed based on the dataset provided by BigDataBall Company (UK) (`www.bigdataball.com`).

as constraints. In fact, with the UWGS (WGS), we might obtain (potentially) that the players with the largest marginal utility are all playing the same role. However, when using the UWGS (WGS) to rank players, forming a lineup with five players of the same role does not make sense. A solution to this issue might be that of classifying players in the same role (by using a cluster analysis), then compute the UWGS (WGS) separately for each role. From an applied point of view, players' popularity retrieved from Google Trends (`trends.google.it/home`) may be exploited to investigate the degrees of relationship between the player's marginal utility and his/her popularity.

Acknowledgements

The authors acknowledge partial support from the INdAM-GNAMPA 2023 project "Sviluppo di metodi di machine learning per la stima del valore Shapley e di sue generalizzazioni", code CUP_E53C22001930001.

References

BIANCALANI, FRANCESCO, GNECCO, GIORGIO, METULINI, RODOLFO, *et al.* 2023. the relationship between players' average marginal contributions and salaries: an application to NBA basketball using the generalized Shapley value. *Statistica Applicata*, 1–29.

GRASSETTI, LUCA, BELLIO, RUGGERO, DI GASPERO, LUCA, FONSECA, GIOVANNI, & VIDONI, PAOLO. 2021. An extended regularized adjusted plus-minus analysis for lineup management in basketball using play-by-play data. *IMA Journal of Management Mathematics*, **32**(4), 385–409.

KUBATKO, JUSTIN, OLIVER, DEAN, PELTON, KEVIN, & ROSENBAUM, DAN T. 2007. A starting point for analyzing basketball statistics. *Journal of quantitative analysis in sports*, **3**(3).

METULINI, RODOLFO, & GNECCO, GIORGIO. 2022. Measuring players' importance in basketball using the generalized Shapley value. *Annals of Operations Research*, 1–25.

NOWAK, ANDRZEJ S, & RADZIK, TADEUSZ. 1994. The Shapley value for *n*-person games in generalized characteristic function form. *Games and Economic Behavior*, **6**(1), 150–161.

SARLIS, VANGELIS, & TJORTJIS, CHRISTOS. 2020. Sports analytics—Evaluation of basketball players and team performance. *Information Systems*, **93**, 101562.

TERNER, ZACHARY, & FRANKS, ALEXANDER. 2021. Modeling player and team performance in basketball. *Annual Review of Statistics and Its Application*, **8**, 1–23.

TREE-BASED REGRESSION WITHIN A HIDDEN MARKOV MODEL FRAMEWORK

Rouven Michels[1], Timo Adam[2] and Marius Ötting[1]

[1] Department of Business Administration and Economics, Bielefeld University, (e-mail: r.michels@uni-bielefeld.de, marius.oetting@uni-bielefeld.de)

[2] Department of Mathematical Sciences, University of Copenhagen, (e-mail: tiad@math.ku.dk)

ABSTRACT: While tree-based regression methods are popular in practice, they miss a time series component. We thus combine regression trees with hidden Markov models (HMMs) and construct a hybrid model that can effectively capture serial correlation and the complex dependencies between the input and output variables, while also providing interpretable results. In a case study, we demonstrate that such an approach offers a powerful and flexible tool for modeling financial data. However, the presented method can be employed in many more fields, e.g. in ecology or sports.

KEYWORDS: hidden Markov model, regression tree, distributional tree, financial markets.

1 Introduction

Tree-based regression models are a popular machine learning tool as they can capture complex interaction effects and yet can be easily interpreted. Combining these models with hidden Markov models (HMMs), which serve for modelling time-series data with serial correlation, is an approach that uses the strengths of both techniques. The scaffold of this model is the assumption that, for each $t = 1, \ldots, T$, the observed time series data $\{Y_t\}_{t=1,\ldots,T}$ is generated by one of N regression trees built by M input variables. Each of these trees corresponds to one of the N states selected by the hidden state process $\{S_t\}_{t=1,\ldots,T}$. We model the latter by an N-state, first-order Markov chain with initial distribution $\delta_i = \Pr(S_1 = i)$ and state transition probabilities $\gamma_{ij} = \Pr(S_t = j \mid S_{t-1} = i),\ i, j = 1, \ldots, N$. Putting these properties together, this results in a model that probabilistically switches between regression trees.

2 Model fitting with the EM algorithm

To fit the model, we use the EM algorithm (Zucchini *et al.*, 2016). We represent the sequence of states $\{S_t\}_{t=1,\ldots,T}$ by the indicator variables $u_i(t) = I(S_t = i)$ and $v_{i,j}(t) = I(S_{t-1} = i, S_t = j)$, $i, j = 1, \ldots N$, $t = 1, \ldots, T$. Then, we can write the joint log-likelihood of the observation process, $\{Y_t\}_{t=1,\ldots,T}$, and the states (i.e. the complete-data log-likelihood) as

$$\begin{aligned} l(\theta) &= \log\left(\delta_{s_1}\prod_{t=2}^{T}\gamma_{s_{t-1},s_t}\prod_{t=1}^{T}\Pr(Y_t = y_t \mid S_t = s_t)\right) \\ &= \sum_{i=1}^{N} u_i(1)\log(\delta_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T} v_{i,j}(t)\log(\gamma_{i,j}) \\ &+ \sum_{i=1}^{N}\sum_{t=1}^{T} u_i(t)\log\big(\Pr(Y_t = y_t \mid S_t = i)\big). \end{aligned}$$

The EM algorithm switches between E- and M-Step, i.e. between estimating the $u_i(t)$'s and $v_{i,j}(t)$'s given the current parameter estimates and maximizing the joint log-likelihood $l(\theta)$. We address the problem of local maxima by running the EM algorithm with different starting values.

Still to discuss is the precise form of $p_i(y_t) = \Pr(Y_t = y_t \mid S_t = i)$. For regular HMMs, this expression is given by the density or probability function of the chosen state-dependent distribution. As we do not make any distributional assumption, we have to find an appropriate expression for regression trees. In the following, we will present two possible procedures: The obvious approach is to employ the CART algorithm (Breiman *et al.*, 1984), to use weights according to the actual state probabilities and to fit regression trees by minimizing the corresponding residual sum of squares (Therneau & Atkinson, 2019). Then, we assume $p_i(y_t)$ to be normally distributed where the mean equals the leaf node's means

$$\mu_t = \frac{1}{n_{\tilde{m}_i}} \sum_{j=1,\ldots,T} I(\mathbf{x}_j \in R_{\tilde{m}_i})\, y_j$$

with $\tilde{m}_i \in 1, \ldots, M_i$ being the node for which $\mathbf{x}_t \in R_{\tilde{m}_i}$ and $n_{\tilde{m}_i}$ denoting the number of observations in region $R_{\tilde{m}_i}$ for the tree of state i. Moreover, the standard deviation σ_t is regarded as a hyperparameter to tune. In the second approach, we do not employ classical regression trees but distributional trees which constitute as a specific form of regression trees. The difference is the way of splitting. While for regression trees the splitting rule only optimizes

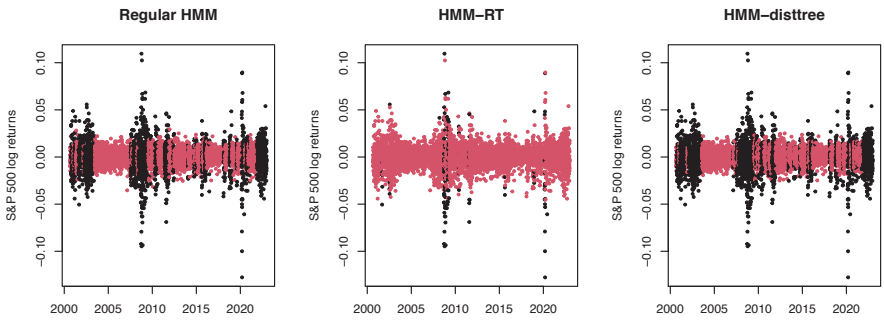


Figure 1. *The time series of log-returns of the S&P 500 from 30th August, 2000 until 30th December, 2022 are displayed. The most likely states under the corresponding model (left panel: Regular HMM; middle panel: HMM-RT; right panel: HMM-disttree) are colorized.*

according to the means between the leaf nodes, for distributional trees the data are split into homogeneous groups with respect to a full parametric distribution (Schlosser *et al.*, 2019). Like in the first approach, we replace $p_i(y_t)$ with the density of a normal distribution, however, the standard deviation is no longer a hyperparameter. We fit such distributional trees using the R package *disttree* (Schlosser *et al.*, 2021).

3 Application to financial data

To illustrate the usefulness of the proposed approach, we consider a case study on financial data. In financial markets, the terms "bullish" and "bearish" describe the overall sentiment of the market participants towards a particular asset or the market as a whole. In an HMM context, we can use these two terms as proxies for latent states. A bullish market is characterized by a calm period of moderately rising prices, while a bearish market is marked by nervousness and oscillating, but mostly falling prices. We apply the presented methods to the daily S&P 500 log-returns from 30.08.2000 – 30.12.2022 as the observed time series and use two input variables, the daily oil and gold log-returns.

After fitting both models to the data, we use the Viterbi algorithm (see Zucchini *et al.*, 2016) for state decoding. We can see in Figure 1 (middle panel) that the classical regression tree approach is not able to capture the bullish and bearish markets as the model switches between states within these market

phases. In contrast, the distributional tree recognizes calm and nervous markets (right panel of Figure 1) which builds the basis for further analysis, e.g. the prediction of future log-returns. When comparing the distributional tree method to a regular HMM with a normal distribution as the state-dependent distribution (left panel of Figure 1), significant similarities can be observed. However, in the presence of more covariates, the distributional tree regression method automatically chooses variables and interactions (see Schlosser *et al.*, 2019) and, thus, circumvents the usual selection problems.

4 Discussion

Using tree-based regression in the framework of HMMs presents a promising approach for modeling complex data sets with a wide range of input variables. Specifically, our findings indicate that employing distributional trees in the EM algorithm outperforms classical regression trees in this context. Differences in other distribution parameters than the mean (such as the standard deviation) can only be captured by distributional trees, which provide much more flexibility without being computationally more costly. In particular, the HMM-RT approach is twice as fast, but also requires cross-validation via the standard deviation, which is why in the end the HMM-disttree method is more efficient.

The approach presented herein should be considered as merely a starting point for establishing connections between HMMs and machine learning algorithms within the regression domain. For instance, the combination of HMMs and random forests could potentially mitigate concerns related to overfitting.

References

BREIMAN, L., FRIEDMAN, J., OLSHEN, R., & STONE, C. 1984. *Classification and Regression Trees*. New York: Wadsworth.

SCHLOSSER, L., HOTHORN, T., STAUFFER, R., & ZEILEIS, A. 2019. Distributional Regression Forests for Probabilistic Precipitation Forecasting in Complex Terrain. *The Annals of Applied Statistics*, **13**(3), 1564 – 1589.

SCHLOSSER, L., LANG, M.N., HOTHORN, T., & ZEILEIS, A. 2021. *disttree: Trees and Forests for Distributional Regression*. R package version 0.2-0.

THERNEAU, T., & ATKINSON, B. 2019. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1–15.

ZUCCHINI, W., MACDONALD, I. L., & LANGROCK, R. 2016. *Hidden Markov Models for Time Series: An Introduction Using R*. New York: CRC press.

SCORING DISTANCES BETWEEN EQUIVALENCE AND PREFERENCE RELATIONS

Boris Mirkin[1,2]

[1]Department of Data Analysis and Artificial Intelligence, National Research University Higher School of Economics, Moscow (e-mail: bmirkin@hse.ru)
[2]Department of Computer Science, Birkbeck University of London (e-mail: mirkin@dcs.bbk.ac.uk)

ABSTRACT: This paper explores association between the notions of similarity and preference by using the framework of the theory of binary relations considered as subsets of the cartesian product of the set of objects by itself. Unordered partitions correspond to the so-called equivalence relations, and ordered partitions, to the so-called weak order relations. We derive a number of properties of the metric space of equivalence and weak order relations. One of them is establishing of the fact that the so-called Kemeny distance between tied rankings is identical to the mismatch distance between corresponding binary relations of weak order.

KEYWORDS: equivalence relation, weak order, distance, contingency table, consensus

1 Introduction

The notions of similarity and preference are usually considered quite different. The former is expressed via the concept of partition, a set of non-overlapping subsets containing "similar" objects, so that different subsets contain 'dissimilar" objects. The latter is expressed via the concept of ordering or, more generally, ordered partition. It is assumed that objects belonging to one part preceding another part are in some sense better than those in this other part. In this sense objects belonging to the same part of an ordered partition are "similar". This association between the notions of similarity and preference can be further elaborated by using the framework of the theory of binary relations considered as subsets of the cartesian product of the set of objects by itself. Unordered partitions correspond to the so-called equivalence relations, and ordered partitions, to the so-called weak order relations. We consider the metric space of binary relations with respect to the so-called matching distance, which is the size of the symmetric difference between relations as subsets of ordered pairs of objects. This allows us to consider both equivalence and weak order relations as part of this metric space and to mathematically explore the separate subspace of equivalence relations and subspace of weak order relations, as well as affinities between these subspaces.

2 Main results

This talk will describe results found within this approach (see also [Mirkin 1979, 2012], Mirkin, Fenner [2019]). Among them are the following.

1. We attend to the Kemeny approach for finding consensus rankings as those minimizing the summary distance to those presented. Here we prove that the Kemeny distance [Kemeny 1959] between rankings is, in fact, the mismatch distance between the corresponding weak-order binary relations. The importance of this result stems from the fact that the former involves Kendall object-to-object matrices with three possible values for the entries: 1 for preceding, -1 for following, and 0 for a tie; whereas the latter involves only two: 1 for the presence and 0 for the absence of a pair in the binary relation [Kendall 1938]. In contrast, the distance between relations involves only 0 (no relation) and 1 (there is relation), with no negative values at all which appear not necessary, in contrast to common sense.

2. We present an explicit statement expressing the Kemeny consensus criterion in terms of the relational consensus matrix, analogous to the so-called consensus matrix in the problem of consensus clustering [Mirkin 2012]. In contrast to the analysis of consensus clustering, however, the (i, j) entry in this consensus matrix is not simply the number of partitions for which elements i and j belong to the same part, but also includes the number of rankings for which i precedes j. The problem, which involves the subtraction of a threshold, is equivalent to maximizing the sum of the consensus matrix entries minus the number of pairs in the corresponding equivalence relation (sometimes referred to as the partition concentration index), weighted with a penalty defined by the threshold. The subtracted part plays the role of a naturally emerging regularizer. The regularizer plays no role, though, when the solution is restricted to a class of ranked partitions like the class of linear rankings with no ties.
3. We test the sensitivity of the Kemeny median concept by applying what we call *Muchnik test* (see [Mirkin 2012] for the case of unordered partitions) to ordered partitions. Specifically, we apply the concept of median to the *Likert scales* popular in Psychology [Likert 1932]. Given an ordered partition $\boldsymbol{R} = (R_1, R_2, \dots, R_p)$, the Likert scale replaces $\boldsymbol{R}$ by the set of binary ordered partitions $\boldsymbol{S}^t$ (t = 1, 2, … , p-1) that separate the union of the first t parts of $\boldsymbol{R}$ from the rest. The question then arises as to whether $\boldsymbol{R}$ is a median for the set of binary rankings $\boldsymbol{S}^t$ (t=1, 2, … , p-1), as one might expect, or not. Perhaps surprisingly, it turns out that it is one of the "coarse" binary rankings $\boldsymbol{S}^t$ that is a median, rather than $\boldsymbol{R}$ itself.
4. We derive explicit formulas for the distance, especially those regarding the relationship between weak orders and their induced equivalence relations, using the ternary relation "between" on the set of binary relations and the notion of "refinement" on the set of tied rankings, as well as the notion of contingency table from statistics. For example, we prove that the mismatch distance between ordered partitions $\boldsymbol{R}$ and $\boldsymbol{R'}$ can be decomposed into ranking and equivalence parts:

$$d(\boldsymbol{R}, \boldsymbol{R'}) = \frac{1}{2} d(\boldsymbol{E}, \boldsymbol{E'}) + d(\boldsymbol{R}*\boldsymbol{R'}, \boldsymbol{R'}*\boldsymbol{R}).$$

where $\boldsymbol{E}, \boldsymbol{E'}$ are equivalence relations corresponding to unordered partitions in $\boldsymbol{R}, \boldsymbol{R'}$and the star $*$ denotes the operation of lexicographic product of two ordered partitions [Mirkin 1979]. The distance between $\boldsymbol{R}*\boldsymbol{R'}$ and $\boldsymbol{R'}*\boldsymbol{R}$ is equal to half of the total of the products of the cardinalities of those parts in the intersection $\boldsymbol{R}\cap\boldsymbol{R'}$ for which the orderings in $\boldsymbol{R}$ and $\boldsymbol{R'}$ are contradictory:

$$d(\boldsymbol{R}* \boldsymbol{R'}, \boldsymbol{R'}*\boldsymbol{R}) = \frac{1}{2}\sum_{s>s'}\sum_{t<t'} N_{st}N_{s't'}$$

Considering the rankings $\boldsymbol{R}$ and $\boldsymbol{R'}$ as unordered partitions, denoted above by $\check{\boldsymbol{R}}$ and $\check{\boldsymbol{R}}'$, respectively, the mismatch distance between the corresponding equivalence relations, $\boldsymbol{E}$ and $\boldsymbol{E'}$, can be expressed as

$$d(\boldsymbol{E}, \boldsymbol{E'}) = \sum_s N_s^2 + \sum_t N'^2_t - 2\sum_{s,t} N_{st}^2$$

where N_s, N'_t , and N_{st} are, as above, the numbers of elements in parts R_s of $\boldsymbol{R}$, R'_t of $\boldsymbol{R'}$ and $R_s\cap R'_t$ of $\boldsymbol{R}\cap\boldsymbol{R'}$, respectively.

3 Conclusion

This shows that, in fact, there is no common ground to simultaneously consider weak orders and equivalence relations, because the lexicographic products are items added to

distances between equivalence relations, which are absent from unordered partitions. Therefore, further advances along the path based on the distance can be made within each ordered partitions (rankings) and unordered partitions, but not in between. Among possible directions for further research, the following two seem quite straightforward. First is the task of numerically solving the problem of consensus ranking by extending the problem of consensus ordering [Charon and Hudry 2007]. For example, the additive structure of the criterion suggests that one might first find an optimal linear ordering and then aggregate some of its parts to form a tied ranking. Second, the failure of the Muchnik test on Likert scales suggests that new ways for formulating more sensitive criteria for consensus are needed.

References

ALESKEROV, F. & MONJARDET, B. 2002. *Utility Maximization, Choice, and Preference*, Stud. Econ. Theory 16, Springer-Verlag, Berlin.

BARTHELEMY, J.P., LECLERC, B. & MONJARDET, B. 1986. On the use of ordered sets in problems of comparison and consensus of classifications. *Journal of Classification*, ***3***(2), 187-224.

CHARON, I. & HUDRY, O. 2007. A survey on the linear ordering problem for weighted or unweighted tournaments. *4OR: A Quarterly Journal of Operations Research*, ***5***(1), 5-60.

DIACONIS, F., & GRAHAM, R. 1977. Spearman's footrule as a measure of disarray, *Journal of Royal Statistics Society*, Series B, **39**, 262-268.

KEMENY, J.G. 1959. Mathematics without numbers, *Daedalus,* **88**, No. 4, *Quantity and Quality*, 577-591.

KENDALL, M.G. 1938. A new measure of rank correlation, *Biometrika*, **30**, 81-93.

LIKERT, R. 1932. A technique for the measurement of attitudes, *Archives of Psychology*, **22(140)**, 55.

MIRKIN, B. 1979. *Group Choice*. Halsted Press: Washington DC. (Translated from Russian *"Problema Gruppovogo Vybora"*, Nauka Physics-Mathematics, Moscow, 1974.)

MIRKIN, B. 2012. *Clustering: A Data Recovery Approach*, CRC Press, Boka-Raton Fl.

MIRKIN, B. & FENNER, T. I. 2019. Distance and consensus for preference relations corresponding to ordered partitions. *Journal of Classification*, **36**, 350-367.

STEELE, K. & STEFANSSON, H. O. 2015. Decision Theory, *The Stanford Encyclopedia of Philosophy.* <http://plato.stanford.edu/archives/win2015/entries/decision-theory/>.

EVALUATION OF THE PERFORMANCE OF A MODULARITY-BASED CONSENSUS COMMUNITY DETECTION ALGORITHM

Fabio Morea [1] and Domenico De Stefano [2]

[1] Area Science Park, Trieste, Italy (e-mail: fabio.morea@areasciencepark.it)

[2] Department of Social and Political Sciences, University of Trieste, Piazzale Europa 1, Trieste, Italy (e-mail: ddestefano@units.it)

ABSTRACT: This paper presents a novel consensus community detection (CCD) performed adopting a modularity-based community detection algorithm that exploits the concept of consensus over N independent trials to generate robust communities and to aggregate marginal nodes into a single community. The algorithm is tested on a class of artificial networks with built-in community structure that can be made to reflect the properties of real-world networks. Preliminary results show that CCD outperforms a single run of the original algorithm in terms of Normalised Mutual Information (NMI), number of communities and community size distribution, and provides an effective tool for community detection in real-world networks and a way to overcome the dependence on random seed of modularity-based algorithms.

KEYWORDS: network analysis, community detection, consensus.

1 Introduction

Community detection algorithms are a powerful tool for understanding the inner structure of natural and social complex systems that can be represented as networks [1]. This is generally an unsupervised learning task, as real-world networks often have no intrinsic labels; thus, the community structure found depends on the definition of "community" that is embedded in the community detection algorithm.

Modularity-based algorithms (a common choice in many fields of research) rely on the definition of community as a set of nodes that are more densely connected to each other than to the rest of the network. Modularity measures the degree to which the nodes within a given community are more densely connected to one another than to nodes in other partitions: the higher the modularity, the better the partition. Such algorithms have the advantage of being fast and providing easily interpretable results but have a relevant issue: using a "greedy" maximization approach, the composition of communities and the number of communities is different at each run.

The intrinsic variability of results may be acceptable if the subsequent analysis is focused on the global structure of the network. However, when the need to be interpreted in terms of individual nodes (e.g. research questions in the form of "do vertices V1 and V2 belong to the same community?") a more robust approach is required.

A common approach in such cases is to repeat the community detection algorithm several times, and to select as “best result” the iteration providing the maximum value of modularity.

We developed a different approach, based on the concept of Modularity-based Consensus Community Detection (MCCD), consists of the following steps:

1. **Independent trials**: The Louvain community detection algorithm [2] is repeated Ni times, with a randomly chosen fraction α of edges assigned a small, but non-zero weight W_0 and a randomly assigned resolution γ. This ensures that the resulting network does not lose connectivity, but edges associated with W_0 are more likely to be assigned to different communities at each iteration.
2. **Consensus**: The consensus algorithm counts how many times a pair of vertices V_i and V_j are assigned to the same community, and assigns a proportion of membership $P_{Vi} \in [0,1]$. Vertices that are strongly connected to one another are always assigned to the same community and have $P_{Vi} = 1$; lower values of P_{Vi} indicate that the vertex is not strongly connected to its neighbours, and it may be assigned to two or more communities with some degree of confidence.
3. **Pruning**: Nodes with $P_{Vi} < 0.5$ and trivially small communities (i.e. communities that have a number of nodes or a weight below a given threshold) are assigned to community "0".

2 Methodology

We evaluated the performance of the MCCD algorithm on artificial benchmark networks with built-in community structure defined by [3] and commonly named LFR after their Authors (Lancichinetti, Fortunato, Radicchi). These networks are characterised by a power-law distribution of the degree of the nodes and the size of the communities, a common feature of real-world networks. We generated a family of 9 benchmark networks with N = 1000 nodes, $\tau_1 = 2$, $\tau_2 = 3$, average degree = 10, and μ values from 0.1 to 0.9. All networks have 37 communities, with community size varying from 20 to 50. Lower values of mixing parameter μ indicate that the communities are clearly separated and are therefore easily identified by community detection algorithms; on the contrary, high values of μ are related to networks with fuzzy communities that are hard to identify.

To measure the performance of the clustering algorithm, we calculated the normalised mutual information (NMI) between the built-in partition of the graph and the one detected by the algorithm as a function of μ. Moreover, we compared the *number of communities* and the *community size distribution* of MCCD results with the original network.

3 Results and discussion

Comparative analysis (*Figure 1)* shows that the MCCD algorithm consistently outperforms the repetition of Louvain algorithm. For low values of μ (clearly defined communities) NMI is close to 1.0, with small differences between the methods; as μ increases (fuzzy communities), the consensus algorithm identifies communities that are more closely related to the original community structure. The parameter α does not significantly influence the values of NMI.

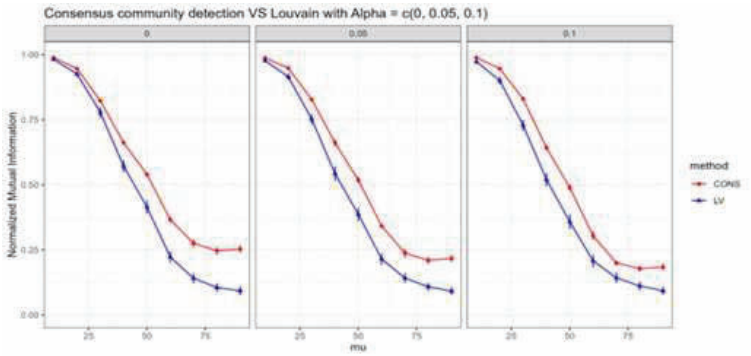


Figure 1: Comparative analysis of the performance of the Modularity-based Consensus Community Detection (MCCD) algorithm on LFR benchmark network with 1000 nodes constructed with the following parameters: degree exponents $\tau_1 = 2$ *and* $\tau_2 = 3$*, average degree* $k_{avg} = 20$*, maximum, degree* $k_{max} = 10$*, minimum community size* $c_{min} = 20$*, maximum community size* $c_{max} = 50$*.*

As of community size, for high values of μ the Louvain algorithm partitions the network in larger communities, while MCCD, with appropriate values of parameter α produces a more robust results as shown in *Figure 2*.

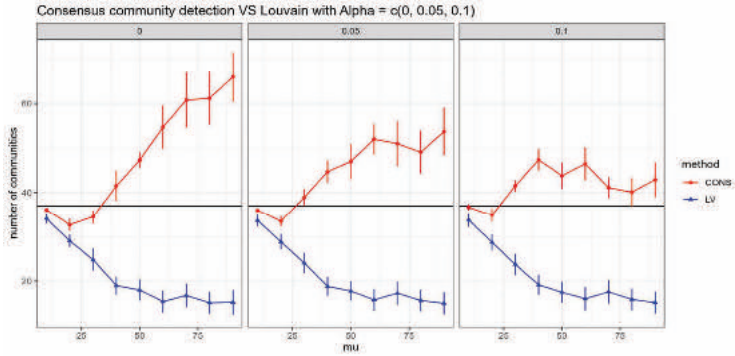


Figure 2: Comparative analysis of the performance of the Modularity-based Consensus Community Detection (MCCD) algorithm on LFR benchmark network (same parameters as in previous figure)

Community size distribution varies significantly at each trial of the Louvain algorithm, and is consistently improved by applying MCCD, as shown in Figure 3. In the plot a single trial is represented by a vertical line, and a marker for each community. The black horizontal lines highlight the original community size distribution between 20 and 50. The Louvain algorithm (*blue lines*) identifies fewer, larger communities than the original ones, while the MCCD algorithm produces results that are much closer to the original community size distribution (*red lines*) and have higher NMI scores.

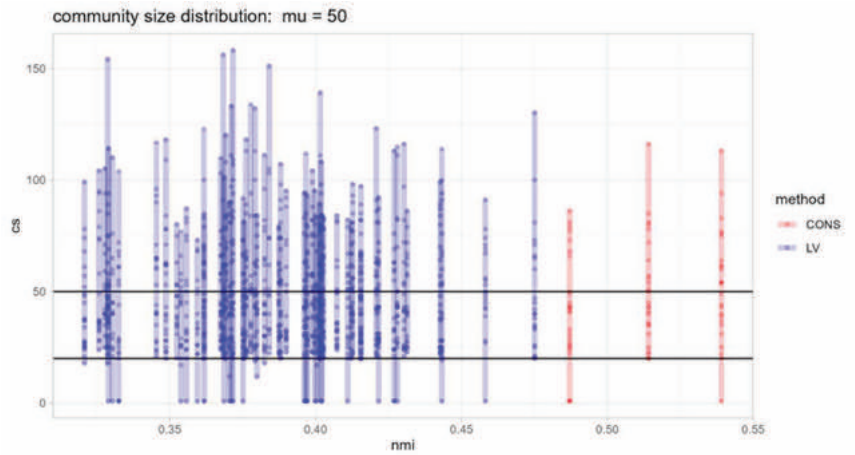


Figure 3: Comparative analysis of the performance of the Modularity-based Consensus Community Detection (MCCD) algorithm on LFR benchmark network (same parameters as in previous figures).

Further research should focus on of the influence of parameters (independent trials, consensus and pruning), on LFR benchmark of different size and structure, as well as the application of MCCD to real world networks of different types.

References

[1] JIN D., YU Z., JIAO P., PAN S., HE D., WU J., YU P.S. ZHANG W. 2023. A Survey of Community Detection Approaches: From Statistical Modeling to Deep Learning, *IEEE Transactions on Knowledge & Data Engineering*, **35**.

[2] BLONDEL, V. D., GUILLAUME, J. L., LAMBIOTTE, R., LEFEBVRE, E. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*.

[3] LANCICHINETTI, A.; FORTUNATO, S.; RADICCHI, F. 2008: Benchmark graphs for testing community detection algorithms. *Physical review E*, **78.4**.

ASSESSING AND IMPROVING DATA QUALITY IN OPEN SPATIAL DATA: A CASE STUDY WITH ANAC DATA

Vincenzo Nardelli[1], Niccolò Salvini[2]

[1] Department of Economics, quantitative methods and business strategies, University Milano Bicocca,
(e-mail: v.nardelli2@campus.unimib.it)

[2] Department of Healthcare Management, (e-mail: niccolo.salvini@unicatt.it)

ABSTRACT: In this paper, we focus on assessing data quality in the context of open anti-corruption data, using data from the National Italian Anti-Corruption Authority (ANAC). The open data movement promotes governments to publish data sets to enhance transparency and accountability, which has been particularly beneficial in combating corruption. Nonetheless, open data is not exempt from challenges, one of which is data quality. We investigate missingness of the data to determine if it is missing at random, not at random, or completely at random. We then present a data quality algorithm, specifically designed for ANAC data, to sanitize errors. To further investigate the phenomena of missingness, we enriched the dataset with socio economic indicators at municipality level coming from official sources (such as ISTAT and Ministry of Economy and Finance, etc.). Finally, we use modelling to determine the factors that contributed to the missingness . An important addition to the classical modeling approaches widely used in literature is to assess if the missingness depends also on the geo-localization of the municipality. This is carried on testing whether it exists an autocorrelation on residuals not explained by classical methods. Our results indicate that addressing missing data with the proposed methodology can lead to more accurate and reliable data for anti-corruption assessments. This study contributes to literature on data quality assessment and provides insights into the challenges and potential solutions to missing data in open data initiatives exploiting spatial statistics techniques.

KEYWORDS: SWORD, data quality, spatial modelling.

VISUALIZING INTERVAL FISHER DISCRIMINANT ANALYSIS RESULTS

M. Rosário Oliveira [1], Diogo Pinheiro[2] and Lina Oliveira[3]

[1] CEMAT and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, (e-mail: rosario.oliveira@tecnico.ulisboa.pt)

[2] Instituto Superior Técnico, Univ. Lisboa, Portugal, (e-mail: diogo.pinheiro.99@tecnico.ulisboa.pt)

[3] CAMGSD and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, Portugal, (e-mail: lina.oliveira@tecnico.ulisboa.pt)

ABSTRACT: In Data Science, entities are usually described by single-valued measurements. Symbolic Data Analysis (SDA) can model more complex data structures such as intervals and histograms that possess internal variability. In this work, we propose an extension of the multi-class Fisher Discriminant Analysis to the interval case based on Mallows' distance and Moore's algebraic structure. Similarly to the conventional case, test observations can be wrongly classified. However, the question is whether the observation is wrongly classified or there exists a labelling switch. Problem may also arise when an observation is atypical. We adress the symbolic data classification problems outline above and use the Mallows' distance adapted to extend classmaps and farness to the SDA setting. Real data is used to illustrate our approach.

KEYWORDS: Symbolic Data Analysis, Classification, Symbolic Fisher Discriminant Analysis, Classmap, Farness.

1 Introduction

Classification is of utmost importance in data science, and the symbolic community is fully aware of that. In a classification problem, the aim is to create a decision rule that assigns a label (or class) to an object (observation) by studying a set of measurements (or variables) characterizing the objects. Conceptually, we can divide the space of the original set of variables into different regions, each associated with one specific label. Sometimes, in the list of variables available, there are a few that do not contribute to the separation of the classes (named irrelevant) or only have repeated information about the objects (called redundant). A common possible way to circumvent these problems is

to project the observations in a space of lower dimension that turns the separation between classes clearer, which in principle leads to better classification performance. Conventional Fisher discriminant analysis uses this strategy, by finding the directions $\alpha \in \mathbb{R}^p$ that best separate the different classes in the projected space: $Z = \alpha^T X$, where $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$, $p \in \mathbb{N}$, is a real-valued random vector with $\mathrm{E}(X|Y = j) = \mu_j \in \mathbb{R}^p$, $\mathrm{Var}(X|Y = j) = \Sigma \in \mathbb{R}^{p \times p}$, for $j = 1, \ldots, g$, and Y represents the class of a given observation, called class-variable. Assuming that within a class the variances of $X|Y = j$, $j = 1, \ldots, g$, are equal, we can compute a pooled sample covariance matrix, S, to estimate Σ, and the Fisher problem can be formulated as the following maximization problems to estimate the sample i-th discriminant vector, $\hat{\alpha}_i$

$$\hat{\alpha}_i = \begin{cases} \underset{\alpha:\ \alpha^T S \alpha = 1}{\arg\max} \dfrac{\alpha^T B \alpha}{\alpha^T W \alpha} \\ \hat{\alpha}_j^T S \alpha = 0, \quad j \in \{1, \ldots, i-1\} \end{cases}, \ i = 1, \ldots, s \leq \min\{g-1, p\},$$

where $W = (n-g)S$, $B = \sum_{j=1}^{g} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T$, $\bar{x}$ is the overall sample mean, $\bar{x}_j$ is the sample mean on the j-th class, n_j is the sample size of the j-th class, and $n = n_1 + \ldots + n_g$ is the total sample size. Moreover, it is known that $T = B + W$, with $T = \sum_{j=1}^{g} \sum_{h=1}^{n_j} (x_{hj} - \bar{x})(x_{hj} - \bar{x})^T$, where x_{hj} represents the observed measurements on the h-th object of the j-th class.

For interval-valued data, the sum of squared total verifies $T = B + W$, and it can be extrapolated using the Mallows' distance instead of the usual Euclidean distance (see Irpino & Verde, 2006), which combined with Moore's definition of linear combination leads to the following maximization problems for interval-valued variables:

$$\alpha_i = \begin{cases} \underset{\alpha:\ \alpha^T S \alpha = 1}{\arg\max} \dfrac{\alpha^T B_C \alpha + \delta |\alpha|^T B_R |\alpha|}{\alpha^T W_C \alpha + \delta |\alpha|^T W_R |\alpha|} \\ \alpha_j^T S \alpha = 0, \quad j \in \{1, \ldots, i-1\}, \end{cases}$$

where $|\alpha| = (|\alpha_1|, ..., |\alpha_p|)^T$, B_l (W_l) is the between (within) sum of square matrix, defined before, based on the centers of the p-dimensional interval-valued observations, if $l = C$, and on its ranges when $l = R$.

Estimating the first $r \leq s$ directions, $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_r\}$, a new observation x_0 is assigned to the k-th class, $k \in \{1, \ldots, g\}$, whenever

$$k = \underset{j \in \{1, \ldots, g\}}{\arg\min} \sum_{t=1}^{r} d_M^2(\hat{\alpha}_t^T x_0, \hat{\alpha}_t^T \bar{x}_j),$$

where $d_M(x,y)$ represents the Mallows' distance between x and y, two p-dimensional interval-valued observations.

To evaluate the performance of the classifier, we split the dataset into the training set, used to estimate the classification rule, and the test set used to independently assess its performance. The test set observations are classified, and the assigned class is compared with the true class to construct the confusion matrix, based on which several global and local measures of performance can be computed.

The classes of the dataset observations are assumed to be mistake free, but with real data, this may not be always true. Moreover, data may contain outlying observations that, even though correctly classified, may reveal atypical patterns when compared with its class or any other class under study. In Raymaekers *et al.*, 2022 and Raymaekers & Rousseeuw, 2022, the authors proposed graphical displays whose goal is to visualize aspects of the classification results to obtain insight into the data, adding interpretability to the results summarized by the confusion matrix. The problem of label switching or atypical observations can be discussed with the help of these plots. In this work, we extend these ideas to the classification problem for interval-valued data. These generalizations rely on the Mallows' distance and we exemplify their relevance and applicability using real examples.

References

IRPINO, ANTONIO, & VERDE, ROSANNA. 2006. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. *Page 185–192 of:* BATAGELJ, V., BOCK, H.-H., FERLIGOJ, A., & ZIBERNA, A. (eds), *Data science and classification, Proc. IFCS'06.* Berlin, Heidelberg: Springer Berlin Heidelberg.

RAYMAEKERS, JAKOB, & ROUSSEEUW, PETER J. 2022. Silhouettes and quasi residual plots for neural nets and tree-based classifiers. *J. Comput. Graph. Stat.*, 1–12.

RAYMAEKERS, JAKOB, ROUSSEEUW, PETER J., & HUBERT, MIA. 2022. Class Maps for Visualizing Classification Results. *Technometrics*, **64**, 151–165.

NONPARAMETRIC LOCAL INFERENCE FOR FUNCTIONAL DATA DEFINED ON MANIFOLD DOMAINS

Niels Lundtorp Olsen[1], Alessia Pini [2] and Simone Vantini[3]

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark , (e-mail: nalo@dtu.dk)

[2] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, (e-mail: alessia.pini@unicatt.it)

[3] MOX - Department of Mathematics, Politecnico di Milano, (e-mail: simone.vantini@polimi.it)

ABSTRACT: We propose a method to test locally functional data whose domain is a Riemaniann manifold. The procedure is based on testing hypotheses on a suitably defined family of balls of the domain, and can be applied to a vast variety of different functional tests. The final result is an adjusted p-value function defined on the same domain as functional data, and controlling the ball-wise error rate.

KEYWORDS: functional data, manifolds, permutation tests, adjusted p-value.

1 Introduction

In functional data analysis (FDA), the object of statistical analysis are typically functions modeled as random elements of a Hilbert space. Inference on functional data is particularly challenging since it deals with elements of infinite dimensional spaces. A currently popular topic in FDA is local inference, i.e., the continuous statistical testing of a null hypothesis along the domain of data. The principal issue in this topic is the infinite amount of tested hypotheses, which can be seen as an extreme case of multiple testing. Local inferential techniques are either based on simultaneous confidence bands (Liebl & Reimherr, 2023), or on the definition of a p-value function, that is a function assigning a p-value at each point of the domain. Methods based on such a p-value function typically adjust p-values for guaranteeing a control of a quantity related with the error rate on the whole domain, that could either be related to the family-wise error rate (e.g., Pini & Vantini, 2017, Abramowicz *et al.*, 2022) or to the false discovery rate (e.g., Lundtorp Olsen *et al.*, 2021). In particular, Pini & Vantini, 2017 introduced the interval-wise testing procedure which performs local inference for functional data defined on an interval domain, where the output

is an adjusted p-value function that controls for type I errors on intervals. The interval-wise testing procedure provides a control of the interval-wise error rate, that is the probability that, if on an interval the null hypothesis is true, at least one part of it is detected as significant.

Most of the current literature focuses on functional data whose domain is an interval of $\mathbb{R}$. The few exceptions considering more complex domains are based on the false discovery rate control (Lundtorp Olsen *et al.*, 2021), or on an asymptotic control of the family-wise error rate (Abramowicz *et al.*, 2022). In this work, instead, we extend the method proposed by Pini & Vantini, 2017 to functional data defined on manifold domains. The resulting method will provide a finite sample control of the ball-wise error rate, which is an extension of the interval-wise error rate to the multidimensional setting.

We extend this idea to a general setting where domain is a Riemannian manifolds. This requires new methodology such as how to define adjustment sets on product manifolds and how to approximate the test statistic when the domain has non-zero curvature. The resulting method will provide a finite sample control of the ball-wise error rate, which is an extension of the interval-wise error rate to the multidimensional setting. This extended abstract describes an overview of the proposed statistical method. More details on the method, its theoretical properties, a simulation and an application to real data can be found in Lundtorp Olsen *et al.*, 2023.

2 Methods

We will assume that the domain of our functional data are *Riemannian manifolds*. In the following, we give a definition of the manifold, as well as the one of ball, that will be of particular importance to define the error control provided by the method.

Definition 1 *A manifold M of finite dimension is a smooth manifold together with a smoothly varying 2-tensor field g on M which is an inner product at each point. The inner product g defines a metric d and a measure μ on M, which we will refer to as the* Riemannian metric *and the* Riemannian measure*, respectively.*

Definition 2 *For a given manifold M with metric d, define the ball of radius ε and center x as*

$$B(x,\varepsilon) = \{y \in M | d(x,y) < \varepsilon\}, \quad x \in M, \varepsilon > 0.$$

Let M be a manifold with metric d. We assume that we have observed n smooth functional data $\xi_1, \ldots, \xi_n$: $M \mapsto \mathbb{R}$. For simplicity of notation, here, functional data are assumed to be observed on a single manifold domain. We refer to Lundtorp Olsen *et al.*, 2023 for a more general version, where the domain can be as well a product of a finite number of manifolds.

Assume that we would like to test at every point $x \in M$, a pointwise null hypothesis $H_0(x)$, against an alternative hypothesis $H_1(x)$. We further assume that hypotheses can be tested by means of a pointwise test statistic $T(x)$, which is stochastically greater under $H_1(x)$ than under $H_0(x)$. Finally, let $p(x)$ denote the unadjusted p-value of the test at point x.

The procedure to define an adjusted p-value function on this setting is based on testing the null and alternative hypothesis on every ball of M of size $\varepsilon \leq r$, with a fixed r (ball-wise testing), and then adjusting the p-values in order to obtain a desired multiplicity control.

Ball-wise testing. Let $B = B(y, \varepsilon)$ be a fixed ball in M. We define the null and alternative hypotheses on the ball as

$$H_0^B : \cap_{x \in B(y,\varepsilon)} H_0(x); \quad H_1^B : \cup_{t \in B(y,\varepsilon)} H_1(x). \tag{1}$$

The hypotheses 1 can be tested with the integral test statistic

$$T^B = \int_B T(x) \mathrm{d}\mu(x) \tag{2}$$

Let p^B be the p-value of the obtained test on ball B. In the ball-wise testing phase, the null and alternative hypotheses H_0^B and H_1^B are tested on every ball $B \in M$ with radius $\varepsilon \leq r$, with a fixed r. The constant r is a parameter of the procedure, and will affect the power and error control of the obtained procedure. We refer to Lundtorp Olsen *et al.*, 2023 for a discussion on the effect of the parameter in the test results.

We here give the general definition of ball-wise hypotheses and p-values. Note that the tests can be performed with any procedure, given that the obtained p-values are exact. In particular, in Lundtorp Olsen *et al.*, 2023 we propose to use permutation tests for testing pointwise and ball-wise hypotheses.

Adjustment. Let $\mathcal{B}$ denote the set of all balls $B \in M$ with radius $\varepsilon \leq r$. The adjusted p-value at point $x \in M$ is defined as

$$\tilde{p}(x) = \sup_{B \in \mathcal{B} : x \in B} p^B. \tag{3}$$

In particular, the null hypothesis $H_0(x)$ is rejected by $\tilde{p}(x)$ at level α only if all null hypotheses on balls $B \in \mathcal{B}$ that contain the point x are also rejected at the same level. This is sufficient to guarantee that the procedure controls the ball-wise error rate Lundtorp Olsen *et al.*, 2023, that is, $\forall \alpha \in (0,1)$:

$$\forall B \in \mathcal{B} : H_0^B \text{ is true, } \mathbb{P}(\exists x \in B : \tilde{p}(x) \leq \alpha) \leq \alpha. \tag{4}$$

References

ABRAMOWICZ, K., PINI, A., SCHELIN, L., SJÖSTEDT DE LUNA, S., STAMM, A., & VANTINI, S. 2022. Domain selection and familywise error rate for functional data: A unified framework. *Biometrics*.

LIEBL, D., & REIMHERR, M. 2023. Fast and fair simultaneous confidence bands for functional parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 04.

LUNDTORP OLSEN, N., PINI, A., & VANTINI, S. 2021. False discovery rate for functional data. *Test*, **30**(3), 784–809.

LUNDTORP OLSEN, N., PINI, A., & VANTINI, S. 2023. *Local inference for functional data on manifold domains using permutation tests*. Tech. rept. arXiv.

PINI, A., & VANTINI, S. 2017. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, **29**(2), 407–424.

CASE-CONTROL VARIATIONAL INFERENCE FOR LARGE SCALE STOCHASTIC BLOCK MODELS

Silvia Pandolfi [1], Francesco Bartolucci [1]

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

ABSTRACT: A scalable variational inference approach for stochastic block models is proposed. The approach is based on a case-control approximation of the likelihood function, which is an unbiased estimator of the full likelihood. Using the case-control likelihood under a variational inference perspective allows us to strongly reduce the computational complexity, making model estimation feasible for large networks. We evaluate the performance of the proposed algorithm using both simulated and real data coming from a Facebook derived social network.

KEYWORDS: clustering, EM algorithm, random graphs, subsampling.

1 Introduction

Stochastic block models (SBMs; e.g. Snijders & Nowicki, 1997) represent a powerful tool for modeling social network data that can discover communities and clusters of nodes according to their social behavior. Under this formulation, the nodes in the network are assumed to belong to a finite number of latent blocks, identified by individual-specific discrete latent variables, with the probability of connection between two nodes only depending on their block membership.

The predominant method of inference for these models is based on a variational approximation of the model log-likelihood (Daudin *et al.*, 2008). However, the complexity of the corresponding estimation algorithm, keeping the number of blocks fixed, is of the order of $O(n^2)$, where n is the number of nodes. This implies that model estimation is computationally intractable for large-scale networks, limiting its use to a narrow range of applications.

Here, following a previous approach (Roy *et al.*, 2019), we propose a case-control approximation of the target function maximized under the variational inference approach, which leads to a strong reduction of the computational complexity, so that the resulting estimation algorithm may be efficiently applied to large networks. The effectiveness of our proposal will be illustrated via simulation and through a real data application.

2 Stochastic block models

Let $\boldsymbol{Y}$ denote an adjacency matrix referred to n nodes and whose generic element, Y_{ij}, is a binary random variable that is equal to 1 if there is an edge between nodes i and j and to 0 otherwise; $\boldsymbol{y}$ and $y_{ij}, i, j = 1, \ldots, n$, are used to denote the realizations of $\boldsymbol{Y}$ and Y_{ij}, respectively. We focus on *binary undirected* networks with no self-loops, leading to a symmetric adjacency matrix with missing values on the main diagonal.

SBMs assume that nodes in the network belong to one out of k distinct unobserved blocks; these are described by means of independent and identically distributed, node-specific, latent variables U_i, $i = 1, \ldots, n$, defined over the discrete support $\{1, \ldots, k\}$ with probabilities $p(U_i = u) = \pi_u$, $u = 1, \ldots, k$.

SBMs also postulate a *local independence assumption* between nodes: conditional on the latent variables U_i and U_j, responses Y_{ij} are assumed to be independent Bernoulli random variables with success probabilities given by $\phi_{uv} = p(Y_{ij} = 1 | U_i = u, U_j = v)$. Therefore, the conditional distribution of Y_{ij} only depends on the block memberships of nodes involved in the relation. Moreover, parameters ϕ_{uv} must satisfy the invariance property with respect to *reflection*, that is, $\phi_{uv} = \phi_{vu}$ for all $u < v$.

2.1 Classical variational inference

Let $\boldsymbol{\theta}$ denote the vector of all model parameters. For parameter estimation, we may rely on the maximization of the following likelihood function:

$$\mathcal{L}(\boldsymbol{\theta}) = p(\boldsymbol{y}) = \sum_{\boldsymbol{u}} p(\boldsymbol{y}|\boldsymbol{u}) p(\boldsymbol{u}), \tag{1}$$

where $\boldsymbol{u} = (u_1, \ldots, u_n)'$ is a realization of the random vector $\boldsymbol{U} = (U_1, \ldots, U_n)'$, and

$$p(\boldsymbol{y}|\boldsymbol{u}) = \sqrt{\prod_{i \leq n} \prod_{j \neq i} p(y_{ij}|u_i, u_j)}, \quad p(\boldsymbol{u}) = \prod_{i \leq n} \pi_{u_i}.$$

As known, the likelihood function in equation (1) involves summation over the configurations of all latent variables in the model, so that the computational burden is prohibitive also when dealing with networks of a very limited size. Moreover, also the posterior expectation of the complete data log-likelihood, which is used within the Expectation-Maximization (EM) algorithm, is intractable. Therefore, a classical solution is to rely on a variational approximation of the EM algorithm (VEM; Daudin *et al.*, 2008), which is based on

the maximization of the following lower-bound of the likelihood function in equation (1):

$$\mathcal{J}(\boldsymbol{\theta}) = \log \mathcal{L}(\boldsymbol{\theta}) - KL[R(\boldsymbol{u}) \,||\, p(\boldsymbol{u}|\boldsymbol{y})], \tag{2}$$

where $p(\boldsymbol{u}|\boldsymbol{y})$ denotes the (intractable) posterior distribution of the latent vector $\boldsymbol{u}$ given the observed adjacency matrix $\boldsymbol{y}$, $R(\boldsymbol{u})$ denotes its approximation, and $KL[\cdot \,||\, \cdot]$ is the Kullback-Leibler divergence between these two distributions. A typical choice for $R(\boldsymbol{u})$ is that based on the conditional independence between the latent variables in the network, given the observed data, implying that $R(\boldsymbol{u}) = \prod_{i \leq n} h(u_i, \boldsymbol{\tau}_i)$, where $h(\cdot, \boldsymbol{\tau}_i)$ denotes a Multinomial probability distribution with parameters 1 and $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{ik})'$. The generic element of $\boldsymbol{\tau}_i$, say τ_{iu}, can be interpreted as an approximation of $p(U_i = u|\boldsymbol{y})$.

Parameter estimates are obtained by alternating two separate steps until convergence of the algorithm. In the variational E-step, $\mathcal{J}(\boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\tau}_i, i = 1, \ldots, n$, with $\boldsymbol{\theta}$ fixed at the values obtained from the previous iteration, under the constraints that these quantities are non-negative and $\sum_u \tau_{iu} = 1$ for all i. In the variational M-step, $\mathcal{J}(\boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$, with the $\boldsymbol{\tau}_i$'s fixed at the values obtained from the E-step.

Besides the several advantages of the variational approximation procedure, the complexity of the iterative algorithm used for deriving parameter estimates, as already mentioned, is of order $O(n^2)$ and this may lead to a excessive computational effort when dealing with large-scale networks.

3 Proposed case-control variational inference

The case-control idea derives from cohort studies where the aim is to compare a group having the outcome of interest ("case") with a control group with regard to one or more characteristics. Usually, the presence of case subjects is relatively rare compared to that of control subjects, and it is impossible or too expensive to select a simple random sample with enough cases to draw conclusions. Accordingly, in a case-control study, all available cases are collected and the corresponding controls are sampled from the corresponding cohort.

In the context of network data, we can view the presence of connections (that is, the 1's) as cases and the absence of connections (the 0's) as controls, and we can rely on this analogy to propose a case-control approximation of the target function in (2). In particular, for every node i, let $\mathcal{A}_i$ denotes the random subset of $\{j : y_{ij} = 0,\ j \neq i\}$, with $n_{i0} = \sum_{j \neq i}(1 - y_{ij})$ being the total number of nodes that are not connected with node i. We also define $\mathcal{B}_i$ as the random subset of $\{j : y_{ij} = 1,\ j \neq i\}$, with $n_{i1} = \sum_{j \neq i} y_{ij}$ being the total number of nodes

connected with i. We may derive the following approximation of $p(\boldsymbol{y}|\boldsymbol{u})$:

$$\tilde{p}(\boldsymbol{y}|\boldsymbol{u}) = \sqrt{\prod_{i\leq n}\left[\left(\prod_{j\in\mathcal{A}_i} p(y_{ij}|u_i,u_j)\right)^{n_{i0}/|\mathcal{A}_i|}\left(\prod_{j\in\mathcal{B}_i} p(y_{ij}|u_i,u_j)\right)^{n_{i1}/|\mathcal{B}_i|}\right]}.$$

Since $\tilde{p}(\boldsymbol{y}|\boldsymbol{u})$ is based on random samples from the 1's and 0's, we get an unbiased estimator of $p(\boldsymbol{y}|\boldsymbol{u})$. The case-control approximate likelihood is then defined as $\tilde{L}(\boldsymbol{\theta}) = \sum_{\boldsymbol{u}} \tilde{p}(\boldsymbol{y}|\boldsymbol{u})p(\boldsymbol{u})$ and the corresponding lower bound may be derived as in equation (2), leading to the approximate target function $\tilde{J}(\boldsymbol{\theta})$. Given the assumption of *a posteriori* independence of the latent variables and denoting by $w_{i0} = n_{i0}/|\mathcal{A}_i|$ and $w_{i1} = n_{i1}/|\mathcal{B}_i|$ the sampling rates, we have

$$\begin{aligned}\tilde{J}(\boldsymbol{\theta}) &= \sum_{\boldsymbol{u}} R(\boldsymbol{u})\log[p(\boldsymbol{u})\tilde{p}(\boldsymbol{y}|\boldsymbol{u})] - \sum_{\boldsymbol{u}} R(\boldsymbol{u})\log R(\boldsymbol{u}) = \sum_{i\leq n}\sum_{u\leq k}\tau_{iu}\log\pi_u \\ &+\frac{1}{2}\sum_{i\leq n}\sum_{u\leq k}\tau_{iu}\left[w_{i0}\sum_{j\in\mathcal{A}_i}\sum_{v}\tau_{jv}\log(1-\phi_{uv}) + w_{i1}\sum_{j\in\mathcal{B}_i}\sum_{v}\tau_{jv}\log\phi_{uv}\right] \\ &-\sum_{i\leq n}\sum_{u\leq k}\tau_{iu}\log\tau_{iu}.\end{aligned}$$

Parameter estimation may be then obtained by means of a modified VEM algorithm that maximizes $\tilde{J}(\boldsymbol{\theta})$. Denoting by $m < n$ the average number of 1's and 0's selected for each node, the complexity of the proposed estimation algorithm reduces to $O(n \times m)$. For large networks that are usually sparse, we can randomly choose a very small subset of 0's, so as to obtain a strong reduction of the computing time. Moreover, alternative sampling schemes based on descriptive network statistics may also be considered in order to increase the efficiency of the algorithm and the accuracy of the estimates.

References

DAUDIN, J-J., PICARD, F., & ROBIN, S. 2008. A mixture model for random graphs. *Statistics and Computing*, **18**, 173–183.

ROY, S., ATCHADÉ, Y., & MICHAILIDIS, G. 2019. Likelihood inference for large scale stochastic blockmodels with covariates based on a divide-and-conquer parallelizable algorithm with communication. *Journal of Computational and Graphical Statistics*, **28**, 609–619.

SNIJDERS, T.A., & NOWICKI, K. 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, **14**, 75–100.

ISSUES WITH SPARSE SPATIAL RANDOM GRAPHS

Francesca Panero [1]

[1] Department of Statistics, London School of Economics and Political Science, (e-mail: f.panero@lse.ac.uk)

ABSTRACT: Spatial networks describe relations among agents that live in a metric space and whose locations affect the probability of connections. Recently, nonparametric Bayesian statistics (BNP) proved itself to be a powerful tool to provide random graph models that mimic real world networks, but no proposals have been made so far to include spatial covariates. I will show how some available models fail in recovering spatial information and conjecture a way to solve the problem.

KEYWORDS: networks, spatial statistics, baysian nonparametrics.

1 Introduction

Relational data can be described by mathematical objects known as graphs, collection of nodes, which represent agents of any nature, connected by edges or links, indicating a relation between those nodes. In applications, a graph is usually called network. Networks describe a plethora of relational phenomena, like transportation, social interactions, email exchanges, protein interactions, internet connections and many more. Network data have been collected extensively in the last decades and have pushed the frontier of research to offer refined models able to fit the complexity of their information.

There are multiple characteristics of real network that researchers try to adhere to when designing a random graph model. The degree of a node is the number of edges departing from it, and the degree distribution of the network is an interesting aspect to study, which in many real examples has been observed to be close to a power-law. Additionally, real networks often display a strictly positive clustering coefficient - defined as number of triangles over triplets - which indicates the presence of transitivity in connections (a friend of a friend will most likely become a friend). Also, the density of edges in many of the gigantic networks we deal with nowadays seems to be very low, meaning that the number of edges does not grow as fast as the number of nodes squared.

The graphon framework received a lot of attention in the last two decades for its possibility of describing node exchangeable graphs, i.e. networks where

a reshuffling of the labels of the nodes does not affect the probability of connections. Being exchangeability a convenient property underpinning many Bayesian models, it comes as no surprise that the graphon model became connected with many Bayesian proposals. The graphon also contains as special cases popular models like the stochastic block model and the latent factor model. Nevertheless, this framework is misspecified for sparse networks, being able to fit only dense or empty ones (see Orbanz & Roy, 2015 for a review). In section 2 I will review the model originally proposed in Caron & Fox, 2017 which uses BNP to overcome the sparsity limitation of graphons and fit some of the properties we observe in real data. This model stimulated an interesting line of research under the name of graphex process. The new proposals, though, fail to describe networks whose edges need a spatial component to be described. In section 3 I will show how Caron & Fox, 2017 fails to represent data that feature a strong spatial component. I will include in the comparison the multidimensional scaling algorithm and show how this spatial algorithm fails to describe such data as well. I will finally conjecture how we can move forward with a spatial network model under the graphex framework.

2 A Bayesian nonparametric model for sparse graphs

The model by Caron & Fox, 2017 defines a network as a Poisson point process on the positive real plane, $Z = \sum_{i,j\geq 1} z_{ij}\delta_{(\theta_i,\theta_j)}$,where z_{ij} is equal to 1 if there is an edge between nodes i, j and 0 otherwise, and $\theta_i \in \mathbb{R}_+$ is the label of the node. The model is heterogeneous, since the probability of connection depends on the node sociability weight $w_i \in \mathbb{R}_+$ (as opposed to homogeneous models with equal probability across all pairs of nodes):

$$\mathbb{P}(z_{ij} = 1|(w_k,\theta_k)_{k\geq 1}) = 1 - e^{-2w_i w_j}. \tag{1}$$

To tune the distribution of w, the authors propose $(\theta_k, w_k)_k$ to be sampled from a Poisson process with intensity $\lambda(d\theta)\rho(dw)$, with λ Lebesgue measure and ρ a Lévy measure. Equivalently, we can describe $W = \sum_{i\geq 1} w_i\delta_{\theta_i}$ as distributed according to a homogeneous completely random measure (CRM). CRMs are a BNP building block, being used as flexible prior distributions over functional spaces (Lijoi & Prünster, 2010). Caron & Fox, 2017 assume ρ to be regularly varying at 0 with exponent $\sigma \in [0,1]$, which intuitively means that ρ behaves similarly to a power function with exponent σ in a neighborhood of 0 (for the formal definition, see Caron *et al.* , 2023). Under this assumption, they prove that the model describes empty, dense or sparse networks (with sparsity level

tuned by σ) and that the degree distribution is a power-law with exponent $1+\sigma$ for high degree nodes. Caron *et al.* , 2023 additionally prove that the clustering coefficient of such model is asymptotically strictly positive.

3 Issues of current models with sparse spatial networks

Spatial networks are networks whose nodes live in a metric space, and their positions affect the probability of connections. An example is the network of airports, where nodes are airports and edges represent flight connections between them. An instance of it is available as the network of flight connections in the United States of America in 2010*. We focus on the continental part of the US, excluding Alaska and Hawaii, for a total of 713 airports and 10^4 connections. The network is sparse with power-law degree distribution. We can easily convince ourselves that connections are determined partly by the size of the airport (a "sociability"), and partly by its location.

We fitted eq. 1 to the dataset in order to estimate the sociability of each airport, using a generalised gamma process as prior for the weights (the set up is as described in Caron & Fox, 2017). Once obtained estimates, we sampled 100 networks from the posterior predictive and we compared the clustering coefficient against its true value. Clustering is usually associated with a strong space component, since spatial models favour connections between nodes that are close (therefore inducing transitivity). The clustering coefficient of the real data is 0.50, while the posterior predictive mean is 0.29 (95% credible interval $[0.25, 0.34]$). The BNP model provides a positive value, but still the true value sits far away from the estimated one, suggesting that sociability is not enough to capture the underlying dynamics of the airport data.

Another possibility to fit such data is to use a multidimensional scaling algorithm, which takes a pairwise similarity matrix between nodes (in our case, the binary adjacency matrix) and computes latent locations for the nodes which minimise a loss function known as strain (Mead, 1992). Applying the algorithm to the dataset and fixing a 2-dimensional latent space, we obtain figure 1. On the left side, longitude is plotted against the two projections, showing that none of them is able to recover the true locations (the results for latitude are similar). The orange dots represent the nodes with highest degree (hubs). On the right, where nodes are shown in the 2-dimensional latent space, we can clearly see that the positions are determined by the degree of the nodes, since the hubs are all projected in a tight central position.

*https://toreopsahl.com/datasets/

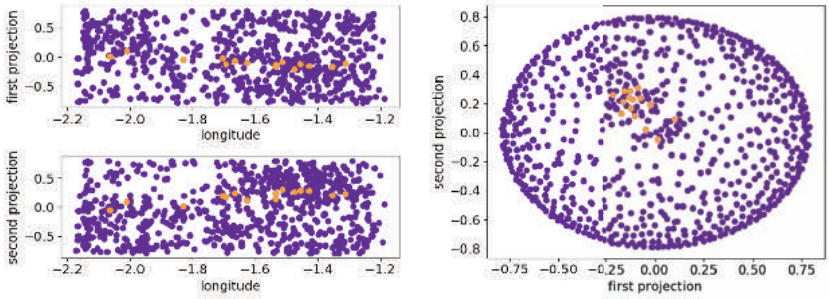


Figure 1. *2-dimensional MDS. On the left, longitude against the two projections, on the right nodes in the latent space. Orange represents nodes with high degree.*

The experiments suggest that a model to describe sparse spatial networks is needed. I conjecture that this could be a modification of eq. 1 with an additional spatial component. The model would inherit the interesting properties of sparsity, power-law degrees and interpretability. This would be beneficial not only for networks with a concrete notion of space, but also for those whose connections can be described by similarity of nodes measured in an abstract latent space (e.g. for qualitative covariates with no notion of distance).

References

CARON, F., & FOX, E. 2017. Sparse Graphs using Exchangeable Random Measures. *Journal of the Royal Statistical Society B*, **79**, 1–44. Part 5.

CARON, F., PANERO, F., & ROUSSEAU, J. 2023. On sparsity and power-law and clustering properties of graphex processes. *Advances in Applied Probability, to appear*.

LIJOI, A., & PRÜNSTER, I. 2010. Models beyond the Dirichlet process. *In:* HJORT, N. L., HOLMES, C., MÜLLER, P., & WALKER, S. G. (eds), *Bayesian Nonparametrics*. Cambridge University Press.

MEAD, AL. 1992. Review of the development of multidimensional scaling methods. *Journal of the Royal Statistical Society: Series D*.

ORBANZ, P., & ROY, D. M. 2015. Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(2), 437–461.

CORPORATE BANKRUPTCY PREDICTION: APPLICATION OF STATISTICAL LEARNING METHODS

Barbara Pawełek[1] and Maria Sadko[1]

[1] Department of Statistics, Krakow University of Economics,
(e-mail: barbara.pawelek@uek.krakow.pl, maria.sadko@uek.krakow.pl)

ABSTRACT: Corporate bankruptcies are an integral part of the functioning of economies. Statistical learning methods are increasingly used in the prediction of corporate bankruptcy. However, there is the problem of choosing an approach, method, procedure or algorithm that would be the most effective from a forecasting point of view. The purpose of the paper is to present the results of the application of statistical learning methods for corporate bankruptcy prediction. The research question was formulated: whether the use of semi-supervised statistical learning methods can improve the effectiveness of corporate bankruptcy prediction. The study used financial data of industrial processing enterprises operating in Poland. Corporate bankruptcy was predicted one, two and three years in advance. Selected methods of unsupervised, supervised and semi-supervised statistical learning were applied.

KEYWORDS: Bankruptcy prediction, statistical learning, unsupervised learning, supervised learning, semi-supervised learning.

References

BALZANO, S., PORZIO, G.C., SALVATORE, R., VISTOCCO, D., & VICHI, M. (Eds.) 2021. *Statistical Learning and Modeling in Data Analysis. Methods and Applications*. Cham: Springer.

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. 2021. *An Introduction to Statistical Learning with Applications in R*. Second Edition, New York: Springer.

PAWEŁEK, B. 2019. Extreme Gradient Boosting Method in the Prediction of Company Bankruptcy. *Statistics in Transition: new series.*, **20**, 155-171.

USING MACHINE LEARNING AND AI IN SCIENCE OF SCIENCE

Daniele Pretolesi[1], Andrea Vian[2] and Annalisa Barla[3]

[1] Austrian Institute of Technology, Vienna, Austria (e-mail: daniele.pretolesi@aic.ac.at)

[2] Department of Architecture and Design, University of Genoa (e-mail: andrea.vian@unige.it)

[3] Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa and Machine Learning Genoa Center, University of Genoa (e-mail: annalisa.barla@unige.it)

ABSTRACT: Complexity has become deeply ingrained in every aspect of our society, and navigating this complexity has become a pressing challenge. Science, in particular, is evolving at an unprecedented rate, constantly pushing the boundaries of human knowledge and understanding. In order to make sense of this rapid scientific evolution, science itself must adapt, employing systems that facilitate the interaction among researchers and that allow to grasp the interconnectedness and evolution of various interdisciplinary fields. In this endeavor, technologies such as natural language processing, network analytics, and machine learning play a pivotal role. These tools provide the essential support needed to analyze vast amounts of scientific data, extract meaningful insights, and uncover hidden patterns.

KEYWORDS: complexity, machine learning, science of science, keyword attribution

1 Introduction

Complexity permeates every facet of our existence, spanning from personal connections to global issues like pandemics and climate change. It is undeniable that comprehending and effectively dealing with complexity has become the paramount challenge of our current era and will continue to be so in the times to come. In scientific research, the intricacy of the subject matter compels researchers to venture beyond their familiar territory and actively pursue collaboration and expertise from scholars in diverse domains. Over the past few years, there has been a rise in scientific collaboration among researchers, leading to intricate interactions involving individuals operating within the same discipline, as well as from different fields of study. This paper is set in the context of the *Science of Science (SciSci)* framework, where the main aim is to

leverage the ever-increasing digital information on scientific production and AI-driven approaches to gain insight into the progress of science, the amount of scientific collaboration between researchers, and the degree of openness (Fortunato *et al.*, 2018).

2 Materials and Methods

We analyze Academic Collaboration Networks (ACNs), which are complex graphs of researchers' scientific output. Each publication in ACNs has important attributes like title, abstract, and keywords, indicating the research topic. By preserving the semantics of collaboration graphs, we connect the academic community and recommend research topics, works, and people. We use advanced technologies like natural language processing (NLP), network analytics (Barabási, 2013), and machine learning (ML)(Hastie *et al.*, 2009; Goodfellow *et al.*, 2016) to attribute missing keywords to publications, which improves our understanding of researchers' scientific interests. This enhanced representation helps us comprehend their scientific endeavors and areas of expertise.

To display the effectiveness of statistical and AI-based methods in this context, we consider the MaLGa dataset, that represents the scientific production of a large interdisciplinary group of scientists in the field of machine learning research, namely the Machine Learning Genoa Center (MaLGa - `https://malga.unige.it`). We collected data of the papers published by the MaLGa faculty members ($n = 14$) during the period 1984–2022 Among a total number of 624 publications, 341 papers are equipped with previously assigned keywords by venues or authors, and 573 papers with abstracts.

With the avilable data, we build a heterogeneous graph considering: (**P**) the *papers* published by MaLGa members together with co-authors, (**Y**) the *year* of paper publication, (**V**) the *venue* in which the paper was published and (**A**) *authors*, i.e. MaLGa members and their co-authors. We consider edges of the type **A**-**P**, **P**-**V**, and **P**-**Y**. The resulting graph is depicted in Fig. 1. The obtained ACN comprises 2007 total nodes, of which 1023 authors, 624 publications, 322 venues, and 38 years. The total amount of edges is 4098, with 2854 **A**-**P** connections, 620 **P**-**V**, and 624 **P**-**Y**.

3 Experimental results

We first use several keyword attribution techniques of increasing complexity to assign missing keywords to publications based on their title and abstract, then we exploit MetaPath2Vec Dong *et al.*, 2017 to represent the graph with

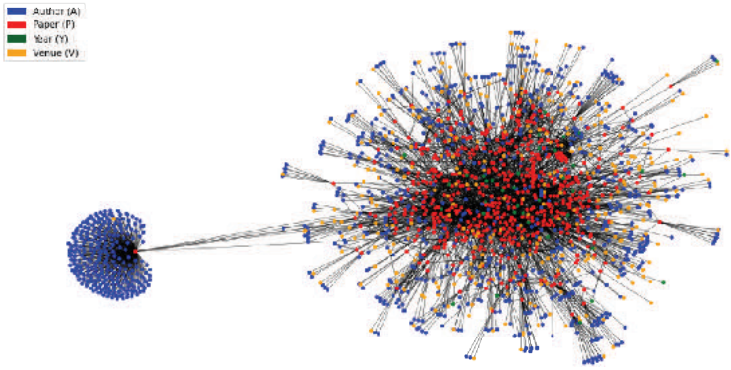


Figure 1. *MaLGa Collaboration as a heterogeneous graph. Blue nodes represent authors* ***A****, Red nodes represent publications* ***P****, green nodes are associated to years* ***Y****, and yellow nodes to venues* ***V****.*

and without keywords. We visualize such embedding on a two dimensional space and assess that this visualization is more informative than the one obtained from a keyword-less graph. We consider several keyword attribution methods: *n*-Grams, RAKE, BERT, SingleRank, TopicRank, MultipartiteRank, and YAKE. We evaluate their performance in terms of Recall, Precision, and F1-score against those ($n = 341$) papers that were associated with benchmark keywords, manually provided by authors or journals. Our results (data not shown) indicate that SingleRank keyword attribution method achieves the best performance for all three metrics considered. To visualise the impact of keywords as additional attributes of nodes, we consider the 14 faculty MaLGa members represented through the embedding with and without keywords and perform a PCA analysis, as shown in Fig. 2.

Adding the keywords into the graph representation improved the quality of the information stored in the network. Indeed, the visualisation clearly indicates that keywords improve the similarity among authors that share the same specific research interests. For example, if we consider M. Santacesaria* and G. S. Alberti†, we note that using embedding that incorporates keywords significantly improve their similarity in terms of the distance measured in the projected space (right panel of Fig. 2, yellow and dark green dots).

*https://scholar.google.com/citations?user=iVlCw_gAAAAJ

†https://scholar.google.com/citations?user=boBf5cgAAAAJ

Figure 2. *PCA projections of MaLGa faculty members represented with embeddings with (right) and without (left) considering their keywords.*

4 Conclusion

In conclusion, our paper suggests how combining graph data structures, graph embeddings, NLP and ML techniques may provide valuable insights on complex topics. This integrated framework offers a holistic perspective by leveraging the strengths of each approach, enabling us to construct statistical models capable of accurately representing the intricate nature of the real world.

References

BARABÁSI, ALBERT-LÁSZLÓ. 2013. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **371**(1987), 20120375.

DONG, YUXIAO, CHAWLA, NITESH V, & SWAMI, ANANTHRAM. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. *Pages 135–144 of: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*.

FORTUNATO, SANTO, BERGSTROM, CARL T, BÖRNER, KATY, EVANS, JAMES A, HELBING, DIRK, MILOJEVIĆ, STAŠA, PETERSEN, ALEXANDER M, RADICCHI, FILIPPO, SINATRA, ROBERTA, UZZI, BRIAN, *et al.* 2018. Science of science. *Science*, **359**(6379), eaao0185.

GOODFELLOW, IAN, BENGIO, YOSHUA, & COURVILLE, AARON. 2016. *Deep learning*. MIT press.

HASTIE, TREVOR, TIBSHIRANI, ROBERT, FRIEDMAN, JEROME H, & FRIEDMAN, JEROME H. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

DISTANCES, ORDERS AND SPACES

Pascal Préa [1,2]

[1] Aix-Marseille Université, CNRS, LIS, Marseille, France
(pascal.prea@lis-lab.fr)

[2] École Centrale de Marseille, Marseille, France

ABSTRACT: A dissimilarity d on a set X is said to be *Robinson* if there exists a total order, said *compatible*, on X such that $x < y < z \implies d(x,z) \geq \max\{d(x,y), d(y,z)\}$. Roughly speaking, d is Robinson if the points of X can be represented on a line *ie.* Robinson dissimilarities generalize line distances.

In this paper, we define k-dimensional Robinson dissimilarities, which generalize the possibility, for a metric set (X,d), to be embedded into a k-dimensional Euclidean space. This generalization is more flexible than the classical embedding and we show that every dissimilarity on an n-set X is $(\log n)$-dimensional Robinson. We give an $O(n^3)$ algorithm which builds such an embedding. This algorithm is based on an incremental algorithm to recognize Robinson dissimilarities.

KEYWORDS: Robinson dissimilarities, embeddings, incremental algorithms.

1 Introduction

Given a finite set set X, a *dissimilarity* on X is a symmetrical function $X \times X \longmapsto \mathbb{R}^+$ such that $\forall x \in X, d(x,x) = 0$ (we say that (X,d) is a dissimilarity space). Dissimilarities generalize distances (a distance is dissimilarity with the triangular inequality).

Given a dissimilarity on a set X, a fundamental problem is to derive "geometrical" properties of X from d, or to characterize dissimilarities from which such properties can be obtained. For instance Robinson dissimilarities (Robinson 1951) correspond to points on a line. These dissimilarities were invented to solve seriation problems in Archeology, but they are now a classical tool for seriation problems in any field. They are also linked with Pyramids (Diday 1986, Durand & Fichet 1988), the standard model with overlapping classes. Moreover, they play an important role ro recognize tractable cases for TSP (Çela & al. 2023).

In this paper, we generalize Robinson dissimilarities to *k(-dimensional)-Robinson* dissimilarities, which represent the fact for X to be embedded into a k-dimensional space. This embedding is less strict than the usual Euclidean

embedding. We show that, if d is a dissimilarity on a set X with $|X| = n$, then d is $(\log n)$-Robinson and we give a $O(n^3)$ algorithm which builds such an embedding. This algorithm is based on an incremental algorithm to recognize Robinson dissimilarities which is presented in the last section.

2 Robinson dissimilarities

A dissimilarity space (X,d) is *Robinson* if there exists a total order, which is said to be *compatible*, on X such that

$$x < y < z \implies d(x,z) \geq \max\{d(x,y), d(y,z)\} \tag{1}$$

Let (X,d) a dissimilarity space and $<$ be an order on X. Notice that $<$ is a compatible order of (X,d) (which is thus a Robinson space) if and only if:

$$x \leq y < z \leq t \implies d(y,z) \leq d(x,t) \tag{2}$$

Given a total order $<$ on X and $x,y,z \in X$, we say that y is *between* x and z for $<$ if $x < y < z$ or $z < y < x$. The set of the elements between x and z is an *interval* for $<$ and we denote it by $[x,z]_<$. Notice that $[x,z]_< = [z,x]_<$.

3 Multidimensional Robinson dissimilarities

Let (X,d) a dissimilarity space and $k \in \mathbb{N}_1$. We say that (X,d) is *k-Robinson* if there exist k orders $<_1, <_2, \ldots, <_k$ such that:

$$\forall x,y,z,t \in X, (\forall 1 \leq i \leq k,\ y,z \in [x,t]_{<_i}) \implies d(y,z) \leq d(x,t)$$

We say that (X,d) is *k-quasi-Robinson* if there exist k orders $<_1, <_2, \ldots, <_k$ such that:

$$\forall x,y,z \in X, (\forall 1 \leq i \leq k,\ y \in [x,z]_{<_i}) \implies d(x,z) \geq \max\{d(x,y), d(y,z)\}$$

If $k = 1$, it is equivalent for a dissimilarity space to be Robinson or 1-quasi-Robinson. For $k \geq 2$, then if (X,d) is k-Robinson, then (X,d) is k-quasi-Robinson, but the converse is false (see Figure 1). Notice in addition that, if (X,d) is k-(quasi-)Robinson, then (X,d) is $k+1$-(quasi-)Robinson. The smallest k such that (X,d) is the *Robinson dimension* of (X,d).

If a metric space (X,d) can be embedded into a $\mathbb{R}^k$, then (X,d) is k-Robinson. But the Robinson dimension of (X,d) is generally smaller. For instance, if $|X| = n$ and d is the constant dissimilarity, then (X,d) is Robinson (its Robinson dimension is 1) although it needs an $n-1$-dimensional Euclidean space to be embedded. Moreover, we have:

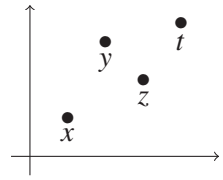


Figure 1. *A set X with four points x,y,z,t. If (X,d) is 2-quasi-Robinson with the two orders represented by the two axis, then no condition is imposed on $d(y,z)$ and we can set $d(y,z) > d(x,t)$. If (X,d) is 2-Robinson (with the same orders), then $d(y,z) \leq d(x,t)$.*

Proposition 1 *The Robinson dimension of a dissimilarity space (X,d) with $|X| = n$ is $\leq \lceil \log_2 \lceil \frac{n}{3} \rceil \rceil + 1$.*

Algorithm 1 returns an approximate value for the Robinson dimension of a dissimilarity space.

Algorithm 1: APPROXIMATE-ROBINSON-DIMENSION

Input: (X,d), a dissimilarity space.
Output: An upper bound on the Robinson dimension of (X,d).
begin
  $X' \leftarrow X$; $k \leftarrow 0$;
  SORT-LINES(X,d) ;
  while $X' \neq \emptyset$ **do**
    $S \leftarrow$ MAXIMAL-ROBINSON-SUBSPACE(X',d) ;
    $X' \leftarrow X' \setminus S$;
    $k \leftarrow k+1$;
  return $\lceil \log_2 k \rceil + 1$;

The function SORT-LINES(X,d), for every $x \in X$, sorts the points of X by increasing values of their distance from x. This function runs in $O(n^2 \log n)$ where $n = |X|$. The function MAXIMAL-ROBINSON-SUBSPACE returns a subset S of X', maximal for inclusion and such that (S,d) is Robinson. This can be easily implemented by a greedy algorithm. We will see in Section 4 that, after SORT-LINES, such a greedy version of MAXIMAL-ROBINSON-SUBSPACE runs in $O(|X'|^2)$. So, as there is at most $n/3$ iterations of the **while** loop, Algorithm 1 runs in $O(n^3)$.

4 An incremental algorithm to recognize Robinson dissimilarities

In order to implement MAXIMAL-ROBINSON-SUBSPACE, we need a function ADD-AND-TEST which takes as entry a dissimilarity space (X,d), a set $S \subset X$ such that (S,d) is Robinson, the PQ-tree $\mathcal{T}_P(S,d)$ and a point $x \in X \setminus S$. A *PQ-tree* (Booth & Lueker 1976) is a data structure which can encode all the compatible orders of a Robinson dissimilarity. ADD-AND-TEST returns the PQ-tree $\mathcal{T}_P(S \cup \{x\},d)$ (If $(S \cup \{x\},d)$ is not Robinson, then $\mathcal{T}_P(S \cup \{x\},d)$ = **none**). The algorithm of ADD-AND-TEST can be sketched as follows:

1. Compute the sets $B_\delta^S := B_\delta(x) \cap S$.
2. Insert the sets B_δ^S into $\mathcal{T}_\mathcal{P}(S,d)$. We get a PQ-tree $\mathcal{T}_\mathcal{P}{}^x(S,d)$.
3. Add the point x to $\mathcal{T}_\mathcal{P}{}^x(S,d)$. We get the PQ-tree $\mathcal{T}_\mathcal{P}(S \cup \{x\},d)$. This will be done in two steps:
   (a) Consider only the points of S the closest from x.
   (b) Consider the other points of S.

Acknoledgements

This work was supported in part by ANR project DISTANCIA (ANR-17-CE40-0015).

References

BOOTH, K.S. & LUEKER,G.S. 1976, Testing for the Consecutive Ones Property, Interval Graphs and Graph Planarity Using PQ-Tree Algorithm, *Journal of Computer and System Sciences* 13, 335–379.

ÇELA, E., DEINEKO, V. AND WOENINGER G.J. 2023, Recognising permuted Demidenko matrices, ArXiv:2302.05191v1.

DIDAY, E. 1986, Orders and overlapping clusters by pyramids in *Multidimensionnal Data Analysis*, J. de Leeuw, W. Heiser, J. Meulman and F. Critchley Eds., 201–234, DSWO.

DURAND, C. & FICHET, B. 1988, One-to-one correspondences in pyramidal representation: an unified approach, in *Classification and Related Methods of Data Analysis*, H.H. Bock Ed., 85–90, North-Holland.

ROBINSON, W.S. 1951, A method for chronologically ordering archeological deposits, *American Antiquity* 16, 293–301.

MODEL-BASED CLUSTERING VIA PARSIMONIOUS MIXTURES OF DIMENSION-WISE SCALED NORMAL MIXTURES

Antonio Punzo[1], Luca Bagnato[2] and Salvatore Daniele Tomarchio[1]

[1] Dipartimento di Economia e Impresa, Università di Catania, (e-mail: antonio.punzo@unict.it, daniele.tomarchio@unict.it)

[2] Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore, (e-mail: luca.bagnato@unicatt.it)

ABSTRACT: Dimension-wise scaled normal mixtures (DSNMs; Punzo & Bagnato, 2022) are a recently defined family of d-variate continuous distributions that generalize the multivariate normal (MN) to allow for 1) a more general central symmetry, and 2) an excess kurtosis that can vary dimension-wise. DSNMs have the further interesting property, shared by the MN distribution too, that no correlation implies independence. These peculiarities are obtained in an MN scale mixture framework by introducing a d-variate mixing random variable with independent and similar components acting separately for each dimension.

We introduce parsimonious finite mixtures of DSNMs for model-based clustering in the presence of symmetric clusters with an amount of excess kurtosis that can vary in each dimension. For illustrative purposes, we describe two members of the DSNM mixture family obtained in the case of mixing random variables being either uniform or shifted exponential; these are examples of mixing distributions that guarantee a closed-form expression for the joint density of the DSNM. For the two described members, we introduce parsimony by putting convenient constraints on the conditional correlation and scale parameters, as well as on the tailedness parameters. This gives rise to a family of 60 interpretable models. Depending on the model under consideration, we describe and use one of two possible variants of the expectation-maximization algorithm to obtain maximum likelihood estimates of the parameters. Finally, we consider simulated and real data to appreciate the advantages of our mixture models over well-established mixtures of symmetric heavy-tailed distributions.

KEYWORDS: Central symmetry, Heavy-tailed distributions, Scale mixtures.

References

PUNZO, A., & BAGNATO, L. 2022. Dimension-wise scaled normal mixtures with application to finance and biometry. *Journal of Multivariate Analysis*, **191**, 105020.

MODEL-BASED SIMULTANEOUS CLASSIFICATION AND REDUCTION FOR THREE-WAY ORDINAL DATA

Monia Ranalli [1], Roberto Rocci[1]

[1] Department of Statistics, Sapienza University, (e-mail: monia.ranalli@uniroma1.it, roberto.rocci@uniroma1.it)

ABSTRACT: A finite mixture model for the unsupervised classification of three-way ordinal data is proposed. Technically, it is a finite mixture of Gaussians observed only through a discretization of its variates. Group specific means and covariances are reparameterized according to parsimonious models. Estimation is carried out through a composite approach to reduce the computational burden.

KEYWORDS: three-way ordinal data, mixture models, composite likelihood, EM algorithm.

1 Introduction

In a cluster analysis context, finite mixtures of Gaussians are frequently used to classify a sample of observations (see for example Hennig *et al.*, 2015), even with complex data structure. This may happens when there are different types of variables or different occasions, i.e. same observations and variables measured at different time points or under different experimental settings. The Gaussian mixture model has been generalized to mixtures of matrix Normal distributions under a frequentist (Viroli, 2011a) and a Bayesian (Viroli, 2011b) framework. The main disadvantage of this model is given by the large number of parameters involved. In the literature, there is a broad consensus in identifying as a possible solution an approach based on performing clustering and dimensionality reduction simultaneously. Indeed, several authors have already proposed such methods (see for example: Rocci & Vichi, 2005, Vichi *et al.*, 2007, Tortora *et al.*, 2016) but only using an optimization approach. In this paper, we focus on three-way ordinal data following a model based approach. We assume that the ordinal variables are variates of a mixture only partially observed through a discretization (Ranalli & Rocci, 2016). This allows us to capture the cluster structure underlying the data, since each component of the mixture corresponds to an underlying group. To reduce the dimensionality, group-specific mean vectors and group-specific covariance matrices are

reparametrized according to parsimonious models that are able to highlight the discrimination power of both variables and occasions while taking into account the three-way structure of the data. The presence of ordinal variables makes the maximum likelihood estimation unfeasible (see for details Ranalli & Rocci, 2016). To overcome the computational issues due to the presence of high dimensional integrals, a composite likelihood (Lindsay, 1988) approach is proposed. The computation of parameter estimates is carried out through an EM-like algorithm.

2 The model

Let $\mathbf{x} = [x_{11}, x_{21}, \ldots, x_{P1}, \ldots, x_{1R}, x_{2R}, \ldots, x_{PR}]'$ be a random vector of P ordinal variables observed at R different occasions. For each ordinal variable we observe $c_p = 1, \ldots, C_p$ categories with $p = 1, \ldots, P$ in each occasion. Following the underlying response variable approach, the observed ordinal variables $\mathbf{x}$ are considered as a discretization of some continuous latent variables $\mathbf{y} = [y_{11}, y_{21}, \ldots, y_{P1}, \ldots, y_{1R}, y_{2R}, \ldots, y_{PR}]'$. The relationship between $\mathbf{x}$ and $\mathbf{y}$ is

$$\gamma_{c_p-1}^{(p)} \leq y_{pr} < \gamma_{c_p}^{(p)} \Leftrightarrow x_{pr} = c_p,$$

where $-\infty = \gamma_0^{(p)} < \gamma_1^{(p)} < \ldots < \gamma_{C_p-1}^{(p)} < \gamma_{C_p}^{(p)} = +\infty$ are non observable thresholds defining the C_p categories and constant over the occasions. We assume that $\mathbf{y}$ follows a heteroscedastic Gaussian mixture model, which is only partially observed,

$$f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi(\mathbf{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{1}$$

where ϕ is the multivariate normal density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, while p_g is the group-specific weight, with $p_g > 0 \ \forall g = 1, \ldots, G$ and $\sum_{g=1}^{G} p_g = 1$. To reduce the number of parameters, the group-specific covariance matrix is modelled as follows (Browne, 1984)

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_{O,g} \otimes \boldsymbol{\Sigma}_{V,g}, \tag{2}$$

where $\otimes$ is the Kronecker product of matrices; while $\boldsymbol{\Sigma}_{O,g}$ and $\boldsymbol{\Sigma}_{V,g}$ represent the group-specific covariance matrices of occasions and variables, respectively. The dimensionality reduction is performed on the group-specific mean vectors following a Tucker2 model (Tucker, 1966). The $G \times (PR)$ matrix collecting the group-specific means is given by

$$\mathbf{M} = (\mathbf{C} \otimes \mathbf{B})\mathbf{N}, \tag{3}$$

where $\mathbf{N}$ collects the scores of the G groups on the Q latent variables under S latent occasions, $\mathbf{B}$ is the loadings matrix that connects the P variables with Q latent variables, $\mathbf{C}$ is the loadings matrix that connects R occasions with the S latent occasions. This trilinear model allows us to project the within-group means, lying into a PR dimensional space, onto a reduced subspace of dimension QS. The number of parameters can be further reduced by observing that $\mathbf{B}$ can be decomposed as follows,

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_U \\ \mathbf{B}_L \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{B}_L \mathbf{B}_U^{-1} \end{bmatrix} \mathbf{B}_U = \widetilde{\mathbf{B}} \mathbf{B}_U$$

where $\mathbf{B}_U$ is assumed to be invertible. The same can be done with $\mathbf{C}$, leading to a more parimonious model for the group-specific mean, that is

$$\begin{aligned} \mathbf{M} = (\mathbf{C} \otimes \mathbf{B})\mathbf{N} &= \left[(\widetilde{\mathbf{C}} \mathbf{C}_U) \otimes (\widetilde{\mathbf{B}} \mathbf{B}_U) \right] \mathbf{N} \\ &= (\widetilde{\mathbf{C}} \otimes \widetilde{\mathbf{B}})(\mathbf{C}_U \otimes \mathbf{B}_U)\mathbf{N} \\ &= (\widetilde{\mathbf{C}} \otimes \widetilde{\mathbf{B}})\widetilde{\mathbf{N}}. \end{aligned}$$

For a i.i.d. random sample of size N, the log-likelihood is given by

$$\ell(\boldsymbol{\theta}) = \sum_{l=1}^{L} n_l \log \left[\sum_{g=1}^{G} p_g \pi(\mathbf{x}_l; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma}) \right]$$

where $\mathbf{x}_l = (c_{11}^{(1)}, \ldots, c_{P1}^{(P)}, \ldots, c_{1R}^{(1)}, \ldots, c_{PR}^{(P)})$ is a particular response pattern with the frequence n_l ($\sum_{l=1}^{L} n_n = N$) and

$$\pi(\mathbf{x}_l; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\gamma}) = \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \cdots \int_{\gamma_{c_{(PR)-1}}^{(P)}}^{\gamma_{c_{PR}}^{(P)}} \phi(\mathbf{y}; \boldsymbol{\mu}_g, \Sigma_g) d\mathbf{y}$$

is its probability in the g-th component of the mixture. This likelihood causes non trivial computational problems due to the presence of multidimensional integrals. To overcome computational issues, we adopt a composite likelihood, based on low-dimensional margins.

Further details will be given in the extended version of the paper along with simulation and real data results to show the effectiveness of the proposal.

References

BROWNE, MICHAEL W. 1984. The decomposition of multitrait-multimethod matrices. *British journal of mathematical and statistical psychology*, **37**(1), 1–21.

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. 2016. Mixture models for ordinal data: a pairwise likelihood approach. *Statistics and Computing*, **26**, 529–547.

RANALLI, MONIA, & ROCCI, ROBERTO. 2017. Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis*, **110**(C), 87–102.

ROCCI, ROBERTO, & VICHI, MAURIZIO. 2005. Three-mode component analysis with crisp or fuzzy partition of units. *Psychometrika*, **70**(4), 715.

TORTORA, CRISTINA, SUMMA, MIREILLE GETTLER, MARINO, MARINA, & PALUMBO, FRANCESCO. 2016. Factor probabilistic distance clustering (FPDC): a new clustering method. *Advances in Data Analysis and Classification*, **10**, 441–464.

TUCKER, LEDYARD R. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**(3), 279–311.

VICHI, MAURIZIO, ROCCI, ROBERTO, & KIERS, HENK AL. 2007. Simultaneous component and clustering models for three-way data: within and between approaches. *Journal of Classification*, **24**(1), 71–98.

VIROLI, CINZIA. 2011a. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, **21**, 511–522.

VIROLI, CINZIA. 2011b. Model based clustering for three-way data structures. *Bayesian Analysis*, **6**(4), 573–602.

The cellwise Minimum Covariance Determinant estimator

Jakob Raymaekers [1] and Peter J. Rousseeuw [2]

[1] Department of Quantitative Economics, Maastricht University, The Netherlands, (e-mail: j.raymaekers@maastrichtuniversity.nl)

[2] Section of Statistics and Data Science, University of Leuven, Belgium, (e-mail: peter@rousseeuw.net)

Abstract: The usual Minimum Covariance Determinant (MCD) estimator of a covariance matrix is robust against casewise outliers. These are cases (that is, rows of the data matrix) that behave differently from the majority of cases, raising suspicion that they might belong to a different population. On the other hand, cellwise outliers are individual cells in the data matrix. When a row contains one or more outlying cells, the other cells in the same row still contain useful information that we wish to preserve. We propose a cellwise robust version of the MCD method, called cellMCD. Its main building blocks are observed likelihood and a sparsity penalty on the number of flagged cellwise outliers. It possesses good breakdown properties. We construct a fast algorithm for cellMCD based on concentration steps (C-steps) that always lower the objective. The method performs well in simulations with cellwise outliers, and has high finite-sample efficiency on clean data. It is illustrated on real data with visualizations of the results.

Keywords: cellwise outliers, covariance matrix, likelihood, missing values, sparsity.

A NEW ACCURATE HEURISTIC ALGORITHM TO SOLVE THE RANK AGGREGATION PROBLEM WITH A LARGE NUMBER OF OBJECTS

Maurizio Romano [1] and Roberta Siciliano[2]

[1] Department of Business and Economics, University of Cagliari, (e-mail: romano.maurizio@unica.it)

[2] Department of Electrical and Information Technology, University of Naples Federico II, (e-mail: roberta@unina.it)

ABSTRACT: The analysis of preference rankings has become an important topic in the general field of data analysis in recent years. The classic meaning of preference rankings understood as orders expressed by a series of judges have been joined by the concept of judges is no longer always understood as human beings, but as resulting from automatic evaluation procedures. This paper provides a particle swarm-based optimization algorithm that provides an accurate solution to the rank aggregation problem, namely producing a ranking that best synthesizes the orderings stated by each judge, when the number of items to be evaluated is large

KEYWORDS: Kemeny problem, tied rankings, heuristics, particle swarm optimization

1 Introduction

The rank aggregation problem is known to be a NP hard problem. For this reason, several heuristic solutions have been proposed over the years. For example:

- Amodio *et al.*, 2016, proposes FAST, an heuristic algorithm based on QUICK that estimates consensus rankings from aggregate preferences. Computational efficiency and accuracy are shown with simulations and real data case studies;
- D'Ambrosio *et al.*, 2017, introduces DECoR, a Differential Evolution algorithm for Consensus Ranking, able to work with full, partial, and incomplete rankings. It outperforms previous proposals in both accuracy and speed while handling large datasets;

- Aledo *et al.*, 2017, considers the Optimal Bucket Order Problem (OBOP) by proposing improvements to the standard greedy algorithm (BPA), resulting in improved accuracy and reduced output variance;
- Aledo *et al.*, 2018, presents $(1+\lambda)$ evolution strategies for solving OBOP, with specific mutation operators and initialization methods. Accuracy improvement with respect to the state-of-the-art algorithm is shown with simulated data;
- Aledo *et al.*, 2021, proposes greedy algorithms based on sort-first and cluster-second strategies to efficiently solve OBOP. Accuracy and scalability improvements of the proposed algorithms with respect to the state-of-the-art algorithm are shown with simulated data;
- Acampora *et al.*, 2021, introduces a memetic algorithm combining genetic algorithms with hill-climbing search for rank aggregation. In particular, results are compared with the DeCoR algorithm (D'Ambrosio *et al.*, 2017).

We follow Kemeny's axiomatic approach (Kemeny, 1959; Kemeny & Snell, 1962), according to which the median (or consensus) ranking is that ranking, or those rankings, that minimize the sum of the distances between a candidate ranking belonging to the universe of rankings and the orderings expressed by a set of judges. Moreover, especially when the number of items to be ranked is large, we assume that all possible tied rankings are allowed either in the data matrix containing the orderings or in the final solution. In other words, we assume that tied rankings are not indifference declarations, but they are 'positive statements of agreement' (Emond & Mason, 2002).

2 Particle swarm optimization for preference rankings

We propose a particle swarm optimization algorithm for the rank aggregation problem (PSORaP). We compared the solutions achieved by the DECoR algorithm (Differential Evolution algorithm for Consensus Ranking, D'Ambrosio *et al.*, 2017) and our PSOPaR on the USA ranks data set (O'Leary Morgan & Morgon, 2010), that contains rankings of the 50 US states with respect to various aspects about the economic and social situation, security, etc., included in the internal repository of the `R` package `ConsRank` (D'Ambrosio, 2021). Table 1 shows the solutions obtained by the DECoR and our PSO algorithm, evaluated through the τ_X rank correlation coefficient (Emond & Mason, 2002). The solutions are really similar (DECoR $\tau_X = 0.2976688$, PSORaP $\tau_X = 0.297449$).

Table 1. *Direct comparison of the consensus generated by DECoR and PSO*

Rank	DECoR	PSORaP	Rank	DECoR	PSORaP
1	California	California	26	Colorado	{Colorado
2	New.York	New.York	27	Connecticut	Minnesota}
3	Florida	Florida	28	Minnesota	Alabama
4	Maryland	Maryland	29	Alabama	{Connecticut
5	Louisiana	Louisiana	30	South.Carolina	South.Carolina}
6	Illinois	New.Mexico	31	Oregon	Oregon
7	New.Mexico	{Illinois	32	Oklahoma	Oklahoma
8	Delaware	Texas}	33	Mississipi	Mississipi
9	Texas	Pennsylvania	34	Arkansas	Arkansas
10	Pennsylvania	Michigan	35	Hawaii	Hawaii
11	Michigan	{Georgia	36	Kentucky	Kentucky
12	Georgia	North.Carolina}	37	{Kansas	{Kansas
13	North.Carolina	New.Jersey	38	Rhode.Island}	Rhode.Island}
14	New.Jersey	{Massachusetts	39	Utah	Utah
15	Massachusetts	Washington}	40	{Iowa	{Iowa
16	Washington	Nevada	41	Nebraska}	Nebraska}
17	Ohio	Delaware	42	Wyoming	Wyoming
18	Virginia	Ohio	43	West.Virginina	West.Virginina
19	Tennessee	{Arizona	44	Idaho	Idaho
20	Nevada	Virginia}	45	Maine	Maine
21	Arizona	Tennessee	46	Montana	Montana
22	Missouri	Missouri	47	New.Hampshire	New.Hampshire
23	Indiana	{Alaska	48	South.Dakota	South.Dakota
24	Alaska	Indiana}	49	Vermont	Vermont
25	Wisconsin	Wisconsin	50	North.Dakota	North.Dakota

The differences between the solutions are mainly that DECoR returns less tied US states in the first part, with Delaware ranked 8 for DECoR and 17 for PSORaP.

3 Concluding remarks

In this paper, a particle swarm optimization heuristic algorithm for the rank aggregation problem has been introduced. A comparison with an already proposed differential evolution algorithm shows that the results are encouraging. A deeper study of the behavior of PSORaP will be carried out in the future to better understand how setting the tuning parameters for improving the performance of the algorithm in terms of the accuracy of the solution.

References

ACAMPORA, GIOVANNI, IORIO, CARMELA, PANDOLFO, GIUSEPPE, SICILIANO, ROBERTA, & VITIELLO, AUTILIA. 2021. A memetic algorithm for solving the rank aggregation problem. *Algorithms as a Basis of Modern Applied Mathematics*, 447–460.

ALEDO, JUAN A, GÁMEZ, JOSÉ A, & ROSETE, ALEJANDRO. 2017. Utopia in the solution of the bucket order problem. *Decision Support Systems*, **97**, 69–80.

ALEDO, JUAN A, GÁMEZ, JOSÉ A, & ROSETE, ALEJANDRO. 2018. Approaching rank aggregation problems by using evolution strategies: the case of the optimal bucket order problem. *European Journal of Operational Research*, **270**(3), 982–998.

ALEDO, JUAN A, GÁMEZ, JOSÉ A, & ROSETE, ALEJANDRO. 2021. A highly scalable algorithm for weak rankings aggregation. *Information Sciences*, **570**, 144–171.

AMODIO, SONIA, D'AMBROSIO, ANTONIO, & SICILIANO, ROBERTA. 2016. Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach. *European Journal of Operational Research*, **249**(2), 667–676.

D'AMBROSIO, ANTONIO. 2021. *ConsRank: Compute the Median Ranking(s) According to the Kemeny's Axiomatic Approach*. R package version 2.1.2.

D'AMBROSIO, ANTONIO, MAZZEO, GIULIO, IORIO, CARMELA, & SICILIANO, ROBERTA. 2017. A differential evolution algorithm for finding the median ranking under the Kemeny axiomatic approach. *Computers & Operations Research*, **82**, 126–138.

EMOND, EDWARD J, & MASON, DAVID W. 2002. A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, **11**(1), 17–28.

KEMENY, J.G. 1959. Mathematics without numbers. *Daedalus*, **88**.

KEMENY, J.G., & SNELL, J.L. 1962. Preference rankings: An axiomatic approach. *Pages 9–23 of:* KEMENY, J.G., & SNELL, J.L. (eds), *Mathematical models in the social sciences*. New York: Blaisdell.

O'LEARY MORGAN, K., & MORGON, S. 2010. *State Rankings 2010: A Statistical view of America; Crime State Ranking 2010: Crime Across America; Health Care State Rankings 2010: Health Care Across America.* CQ Press.

USING ML TECHNIQUES FOR ESTIMATION WITH NON-PROBABILISTIC SURVEY DATA

Jorge Rueda[1], Maria del Mar Rueda [1], Ramón Ferri[1] and Beatriz Cobo[2]

[1] Department of Statistics and O.R., University of Granada, (e-mail: jorgerueda279@correo.ugr.es, mrueda@ugr.es, rferri@ugr.es)

[2] Department of Quantitative Methods for Economics and Business, University of Granada, (beacr@ugr.es)

ABSTRACT: Online surveys, despite their cost and effort advantages, are particularly prone to selection bias due to the differences between target population and potentially covered population. Some techniques have arisen in the last years regarding this issue. Propensity Score Adjustment, kernel weighting, Statistical Matching (or mass imputation), double robust estimation and superpopulation modeling are relevant techniques to mitigate selection bias. These techniques use the sample to train a model capturing the behaviour of a target variable which is to be estimated, or the propensity of the units to participate in the volunteer sample. The modeling step has been usually done with linear regression, but machine learning (ML) algorithms have been pointed out as promising alternatives. In this study we examine the use of these algorithms in the nonprobability survey context, in order to evaluate and compare their performance and adequacy to the problem.

KEYWORDS: survey sampling, non-probability samples, propensity score adjustment, machine learning.

1 Estimation in non probability surveys

The use of web surveys and big data sources for population inference is an active research field in social science and survey research. Such data sources allow to produce statistics cheaper, faster, and on a higher level of detail. However, these data most often lacks a sampling design, population coverage is incomplete and the data-generating mechanism is unknown. No valid inferences can be drawn and new methodologies are needed to evaluate the potential biases and make accurate estimates of the population parameters.

Different inference procedures are proposed in the literature to correct for selection bias induced by non-random selection mechanisms. There are three important approaches: the pseudo-design based inference (or pseudo-randomization),

statistical matching and predictive inference.

Pseudo-randomization and Statistical Matching require, apart from the nonprobability sample, a probability sample to do the adjustments. Propensity score adjustment (PSA) originally developed for balancing groups in nonrandomized clinical trials (Rosenbaum & Rubin, 1983) is the most used method for removing bias in nonprobability surveys (Lee & Valliant, 2009). Statistical Matching was firstly proposed in Rivers, 2007. The difference between both methods is the sample used in the estimators: PSA estimates the propensity of each individual of the nonprobability sample to participate in the survey and then this propensity is used to construct the weights of the estimators, while Statistical Matching adjusts a prediction model using data from the nonprobability sample, applies it in the probability sample to predict their values for the target variable y and uses them in the parametric estimators.

Superpopulation modelling requires data from the complete census of the target population for the covariates used in the adjustment, which is assumed to be a realization (sample) of a superpopulation where the (unknown) target values follow a model. The main idea is to fit a regression model on the target variable with data from the nonprobability sample, and use the model to predict the values of the target variable for each individual in the population. The prediction can be used for estimation using a model-based approach or some alternative versions such as model-assisted and model-calibrated.

Usually the linear regression model is considered for estimation, $E_m(y_i|\mathbf{x}_i) = \mathbf{x}_i^T\beta$, and the predicted values of y_i in the probability sample (in the nonsampled individuals) are used for making estimators in the statistical matching inference (in the predictive inference). Logistic regression is usually used in PSA to predict the propensity (probability of the i-th individual of being included in the sample), $\pi_{vi} = P(I_{vi} = 1|\mathbf{x}_i)$.

Alternatively to the linear regression models, Machine Learning (ML) methods have been proposed for the estimation of the propensities and the nonsampled population values. In situations where additivity and/or linearity do not hold, ML algorithms are more suitable for regression and classification. Some of these algorithms, such as decision trees and related (Random Forests, Gradient Boosting Machines) can also take interactions into account without the need of specifying the terms. The use of some ML algorithms for non probability samples has been studied in the last few years (e.g. Buelens *et al.*, n.d.,

Ferri-García *et al.*, 2021, Castro-Martín *et al.*, 2021. In this work we consider some of the most important ML algorithms that can be used to define different estimators for a non-probability sample.

References

BUELENS, BART, BURGER, JOEP, & VAN DEN BRAKEL, JAN A. Comparing Inference Methods for Non-probability Samples. *International Statistical Review*, **86**(2), 322–343.

CASTRO-MARTÍN, LUIS, RUEDA, MARÍA DEL MAR, FERRI-GARCÍA, RAMÓN, & HERNANDO-TAMAYO, CÉSAR. 2021. On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures. *Mathematics*, **9**(23).

FERRI-GARCÍA, RAMÓN, CASTRO-MARTÍN, LUIS, & DEL MAR RUEDA, MARÍA. 2021. Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling. *Mathematics and Computers in Simulation*, **186**, 19–28. MATCOM Special Issue MACMAS 2019: First International Conference on Mathematical and Computational Modelling, Approximation and Simulation.

LEE, SUNGHEE, & VALLIANT, RICHARD. 2009. Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research*, **37**(3), 319–343.

RIVERS, DOUGLAS. 2007. "Sampling for Web Surveys.".

ROSENBAUM, PAUL, & RUBIN, D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

MULTICLASS CLASSIFICATION OF DISTRIBUTIONAL DATA

Ana Santos [1], Sónia Dias[2], Paula Brito[3] and Paula Amaral[4]

[1] Faculdade de Ciências, Universidade do Porto, Portugal (e-mail: up202103086@fc.up.pt)

[2] ESTG - Instituto Politécnico de Viana do Castelo & LIAAD-INESC TEC, Portugal (e-mail: sdias@estg.ipvc.pt)

[3] Faculdade de Economia, Universidade do Porto & LIAAD-INESC TEC, Portugal (e-mail: mpbrito@fep.up.pt)

[4] Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa & CMA, Portugal (e-mail: paca@fct.unl.pt)

ABSTRACT: In this work, classification of distributional data is addressed, where units are described by histogram-valued variables. The proposed approaches aim at extending the linear discriminant method developed for two-class classification to multiclass classification. This method is then applied to discrimination of network models. The goal is to identify the network model used to generate the networks, considering the distribution of four centrality measures.

KEYWORDS: histogram data, linear discriminant function, Mallows distance, Symbolic Data Analysis.

1 Introduction

The need to analyse complex data makes it necessary to innovate and develop new statistical methods. In the Symbolic Data Analysis (SDA) framework the cells of data arrays may contain finite sets of values/categories, intervals or distributions, representing the variability associated with each unit (Brito, 2014). In Dias *et al.* , 2021, a linear discriminant method for distributional data was proposed. The model aims at obtaining a linear combination of features, now defined by distributions or intervals, that characterize the units and that allows classifying them in different *a priori* groups.

2 Histogram-valued variables

This work focus on histogram-valued variables, a particular type of distributional-valued variables. For each unit i, the observation of this type of variables

is a histogram $X(i) = \{I_{X(i)1}, p_{X(i)1}; I_{X(i)2}, p_{X(i)2}; ...; I_{X(i)m}, p_{X(i)m}\}$, where $I_{X(i)l}$ represents the subinterval l, $p_{X(i)l}$ is the weight associated with the subinterval $I_{X(i)l}$ and $\sum_{l=1}^{m} p_{X(i)l} = 1$. The subinterval $I_{X(i)l}$ may be represented by its bounds or by its center, $c_{X(i)_l}$ and (half)-range, $r_{X(i)_l}$. Within each subinterval, a uniform distribution is assumed. Each realisation of the variable can be, alternatively, represented by the quantile function:

$$\Psi_{X(i)}(t) = \begin{cases} c_{X(i)_1} + \left(\frac{2t}{w_1} - 1\right) r_{Y(i)_1} & \text{if} \quad 0 \le t < w_1 \\ c_{X(i)_2} + \left(\frac{2(t-w_1)}{w_2 - w_1} - 1\right) r_{Y(i)_2} & \text{if} \quad w_1 \le t < w_2 \\ \vdots & \\ c_{X(i)_m} + \left(\frac{2(t-w_{m-1})}{1-w_{m-1}} - 1\right) r_{Y(i)_m} & \text{if} \quad w_{m-1} \le t \le 1 \end{cases} \tag{1}$$

where $w_{i\ell} = \sum_{h=1}^{\ell} p_h$, $\ell \in \{1, \ldots, m\}$, and m is the number of subintervals in $X(i)$.

The empirical quantile functions are the inverse of cumulative distribution functions, which under the uniformity hypothesis are piecewise linear functions with domain $[0,1]$. 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.

The Mallows distance is considered as an adequate measure to evaluate the similarity between distributions. The criterion to be optimized to define linear models is based on this distance. Assuming that the "values" of the histogram-valued variables X and Y are represented by the quantile functions Ψ_X and Ψ_Y, both with m pieces and the same set of weights, $\{p_1, \ldots, p_m\}$, the Mallows distance between them can be written as $D_M(\Psi_X(t), \Psi_Y(t)) = \sqrt{\int_0^1 (\Psi_X(t) - \Psi_Y(t))^2 \, dt}$.

Given a set of n units, we may then compute the *barycentric histogram*, X_b, represented by the quantile function $\Psi_{X_b}(t)$, as the solution of the minimization problem $\min \sum_{i=1}^{n} D_M^2(\Psi_{X(i)}(t), \Psi_{X_b}(t))$. The optimal solution, the *barycentric histogram*, X_b, is a histogram where the centre and half range of each subinterval ℓ is the classical mean, respectively, of the centres and of the half ranges ℓ, of all units i (Irpino & Verde, 2006).

3 Linear Discriminant Analysis

3.1 Linear Discriminant Function

Since the space of quantile functions is a semi-vector space, the definition of linear combination for histogram-valued variables proposed in Dias *et al.* , 2021 uses the quantile function of the observed histograms $\Psi_{X_j(i)}(t)$, together with those of the corresponding symmetric histograms $-\Psi_{X_j(i)}(1-t)$, $j = 1,\ldots,p$. The score of unit i is the quantile function:

$$\Psi_{S(i)}(t) = \sum_{j=1}^{p} a_j \Psi_{X_j(i)}(t) - \sum_{j=1}^{p} b_j \Psi_{X_j(i)}(1-t) \tag{2}$$

with $t \in [0,1]$; $a_j, b_j \geq 0$, $j \in \{1,2,\ldots,p\}$.

The function to optimize in order to obtain the coefficients of the linear discriminant function, $a_j, b_j, j = 1,\ldots,p$, is based on the total inertia decomposition with respect to a barycentric histogram, defined with the Mallows distance. Irpino & Verde, 2006 proved that the total inertia may be decomposed into within and between classes inertia, according to the Huygens theorem. The coefficients of the discriminant function are then obtained by maximizing the ratio of the between to the within classes inertia, subject to non-negativity constraints. This defines a constrained fractional quadratic problem that is non-convex and finding the global optima requires a high computacional effort. Softwares like BARON, that use the Branch and Bound technique, may be used to obtain a good solution. To confirm that the solution is optimal is only possible using conic relaxation techniques (Dias *et al.* , 2021).

3.2 Classification

For the classification of a unit in one of the two groups, the Mallows distance between its score and the score obtained for the barycentric histogram of each class is computed. The observation is then assigned to the closest class (with random assignment in case of equality).

When considering more than two *a priori* classes, there are two ideas that arise:

1. Divide the multi-class classification dataset into several binary classification subproblems. In this case, identifying the best multi-class classifier involves finding the best binary classifiers. In other words, we are extending the already existing binary class classifier. Concerning this ap-

proach, there are two well-known multi-class classification techniques: (a) One-Versus-One (OVO); (b) One-Versus-All (OVA).

2. Define several linear discriminant functions, maximizing the same criterion, under the condition that each new discriminant function must be uncorrelated with all previous ones. This imposes new constraints in addition to the non-negativity of the coefficients. This idea is referred to as Consecutive Linear Discriminant Functions (CLDF). This leads to several score histogram-valued variables with null symbolic linear correlation coefficient.

4 Application - Network Data

The network data was artificially obtained. Fifty six networks were constructed, considering the Erdős-Renyi, Watts-Strogatz and Barabási-Albert models, with parameters carefully chosen. Each network is described by the distribution over the network's nodes of standard graph measures: nodes' degree, betweenness centrality, closeness centrality and eigenvector centrality, as done in Giordano & Brito, 2014. To obtain symbolic data sets aggregations were performed, where the first-level units were the nodes and the higher-level units were the network to which the nodes belong. Therefore, the dataset has 168 units and four histogram-valued variables. The classification goal is to identify the model used to develop each network. The OVA strategy displays the worst performance, OVO performs extremely well, regardless of the model used to produce the networks, and tends to perform better than CLDF.

Acknowledgements: This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

References

BRITO, P. 2014. Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics. *WIREs DMKS*, **4**(4), 281–295.

DIAS, S., BRITO, P., & AMARAL, P. 2021. Discriminant analysis of distributional data via fractional programming. *EJOR*, **294**(1), 206–218.

GIORDANO, G., & BRITO, P. 2014. Social networks as symbolic data. *In: Analysis and Modeling of Complex Data in Behavioral and Social Sciences*.

IRPINO, A., & VERDE, R. 2006. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. *In: Data Science and Classification, Proc. IFCS'06*.

LATENT BAYESIAN CLUSTERING FOR TOPIC MODELLING

Lorenzo Schiavon [1]

[1] Department of Economics, Ca' Foscari University of Venice, (e-mail: lorenzo.schiavon@unive.it)

ABSTRACT: The main objective in topic modelling is uncovering the underlying themes present in a corpus of text data. This process is generally constituted by two phases: (i) identifying the main words associated with each topic; (ii) grouping documents that contain similar sets of words together. In this work, we exploit recent advances in Bayesian factor models to represent the high-dimensional space of the observed words through a set of low-dimensional latent variables, and to jointly cluster the documents according to their distribution over such latent constructs. Groups and underlying constructs are interpreted as document topics and language concepts, respectively, with the number of such dimensions that is not required in advance. We apply the proposed approach to a data set of newspaper headlines.

KEYWORDS: Dirichlet process, infinite factor model,nonparametric Bayes, text data

1 Introduction

Nowadays, the digitalization is making available huge quantities of data, which require, on one hand, suitable automatized procedure to recognize, classify and organize such information and, on the other, allows the development and training of the algorithms to respond to such demand. In particular, it has become widespread in several field of studies and businesses the necessity of powerful tools that emulate human being capacity in extracting and summarizing the information expressed in text data. For instance, in political sciences, the use of automatic methods applied to large corpus of institutional reports and documents can represent a fast methodology to uncover and highlight the topics on which public organizations focus more. Indeed, topic modelling techniques aim to reveal the underlying semantic structures in large collections of documents and cluster them in topics. Such methodologies are based on vector space models that represent each document as a vector. In last decades, several techniques has been considered in topic modelling, including low rank decomposition of the document-term matrix, as non-negative matrix factorization, and probabilistic latent semantic analysis, as Latent Dirichlet Allocation.

Inspired by these approaches and exploiting recent advances in Bayesian nonparametric models, we propose a factor model able to jointly cluster the documents in topics and to recover a distinct set of latent concepts. A Bayesian nonparametric approach allows the number of topics can be inferred along with posterior distribution, as for the dimension of the latent semantic space. In addition, we exploit shrinkage prior to promote sparse structures on the low-rank matrices favouring parsimony and interpretation.

2 Latent mixture model in infinite factorization

We are interested in specifying a model able to provide a parsimonious representation of a document-term matrix along both matrix dimensions: on one hand clustering the documents in few groups, on the other reducing the term-space to a low-dimensional space of latent concepts. In view of this, we rely on the general class of latent factor mixture models (LAMB), proposed by Chandra *et al.* (2020), which is able to combine a dimensional reduction via factorization and a suitable use of Bayesian nonparametric framework to cluster the subjects. In particular, considering y_i the p-variate vector including a binary indication on the presence-absence of p terms in the document i, we adjust the LAMB specification by defining the probit model

$$y_i = \mathbb{1}(y_i^* > 0), \quad y_i^* \sim N_p(\Lambda\eta_i, I_p), \quad \eta_i \sim \sum_{k=1}^{\infty} \pi_k N_H(\mu_k, \Delta_k), \tag{1}$$

where Λ is a $p \times H$ matrix of factor loadings with $H \ll p$. The latent factor scores $\eta_i = (\eta_{i1}, \ldots, \eta_{iH})^\top$ are modelled according to an infinite mixture of Gaussian distributions with $\{\pi_k\}_{k=1}^{\infty}$ following a stick-breaking representation

$$\pi_k = v_k \prod_{l<k}(1 - v_l), \qquad v_l \sim \text{Beta}(1, \alpha). \tag{2}$$

Then, clusters are determined by the membership of η_i to the posterior kernels.

Differently from Infinite Mixture of Factor Analyser (Murphy *et al.* , 2020) models, where observations are clustered over kernels with factorized covariance, LAMB defines the clustering over the low-dimensional space of latent constructs ensuring parsimony in the dimension of cluster-specific parameters. In addition, having a unique loadings matrix shared by the different clusters aids the interpretation of the latent factors as language concepts such that each of them explains the presence or absence of several terms belonging to the same semantic area.

In view of this, we carefully specify the prior of the loadings matrix Λ. We use the cumulative shrinkage process (CUSP) proposed by Legramanti *et al.* (2020), which exploits an over-parameterized model with an infinite number of factors and increasing probability of loadings being shrunk as the column index increases. In particular, we specify a spike and slab construction over the columns of Λ with spike probability mass ϖ_h increasing over the column index according to a stick-breaking construction. To allow for a local behaviour, we follow Schiavon *et al.* (2022) including a Bernoulli local scale $\phi_{jh} \sim \text{Ber}(c_p)$ in the variance of each element λ_{jh}. The mean $c_p \in (0,1)$ is set equal to a small positive offset to guarantee sparsity when p is large. Formally, we assume

$$\lambda_{jh} \sim N(0, \theta_h), \quad \theta_h = \phi_{jh}\rho_h(\vartheta_h - \theta_\infty) + \theta_\infty \tag{3}$$

$$\rho_h \sim \text{Ber}(1 - \varpi_h), \quad \vartheta_h^{-1} \sim \text{Ga}(a_\theta, b_\theta), \tag{4}$$

with θ_∞ a positive constant close to zero. Posterior distribution is approximated via MCMC exploiting an adaptive Gibbs sampling strategy.

3 Latent topic extraction

To illustrate the validity of our approach, we initially apply the model to a set of $n = 213$ newspaper sport headlines published by two newspapers of GEDI* in Autumn 2021. After removing the stopwords, we frame the headlines in a document-term matrix. Considering only the unigram and bigram which recur at least twice in the corpus, we obtain a binary matrix y registering the presence or absence of $p = 522$ distinct terms.

We follow Chandra *et al.* (2020) and Schiavon *et al.* (2022) to set the hyper-parameters. The offset c_p is set equal to the average word frequency in y. After running the MCMC algorithm, we recover meaningful posterior summary of the low-rank matrices Λ and η, we compute the posterior means only after having aligned the posterior samples. The algorithm estimates a nine latent factors model with 19 topics. Each factor can be interpreted as a concept characterized by high loadings in correspondence of terms belonging to a specific semantic area. Figure 1 reports a graph representation of the partial correlation matrix between the terms. Every term j, for $j = 1, \ldots, p$ is coloured according to its characterizing concept—i.e. $\text{argmax}_{h \in \{1,\ldots,9\}} \lambda_{jh}$—while the legend reports the term with the highest loading for every latent concept. As one may expect, different concepts refer to different sports or

*GEDI Gruppo Editoriale S.p.A. is an Italian media conglomerate based in Turin.

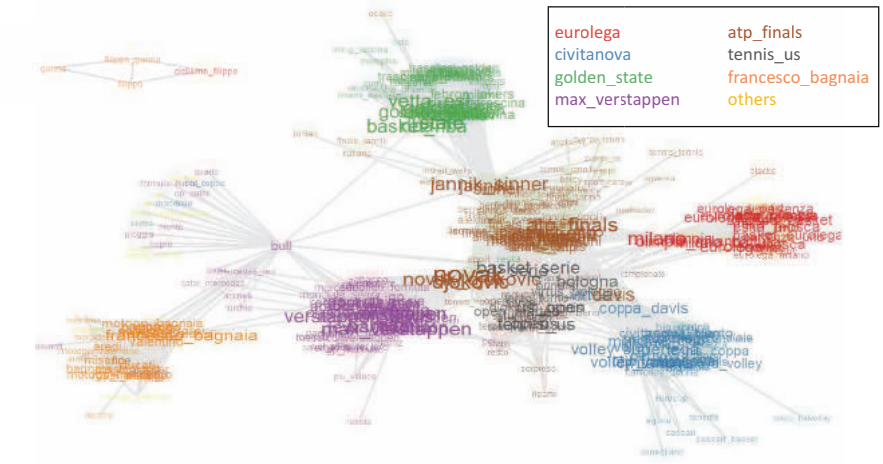


Figure 1. *Graphical representation based on posterior mean of partial correlation matrix. Terms are positioned using a FruchtermanReingold force-direct algorithm and coloured according to the highest loadings.*

competitions. Headlines are defined as weighted combination of such semantic concepts and grouped in topics with similar combination of concepts.

We aim to apply the same approach to documents and reports regarding health and ageing public plans published by Italian regional institutions to uncover the concepts and the themes on which Italian regions are focusing their efforts.

References

CHANDRA, NOIRRIT KIRAN, CANALE, ANTONIO, & DUNSON, DAVID B. 2020. Escaping the curse of dimensionality in Bayesian model based clustering. *arXiv preprint arXiv:2006.02700*.

LEGRAMANTI, SIRIO, DURANTE, DANIELE, & DUNSON, DAVID B. 2020. Bayesian cumulative shrinkage for infinite factorizations. *Biometrika*, **107**(3), 745–752.

MURPHY, KEEFE, VIROLI, CINZIA, & GORMLEY, ISOBEL CLAIRE. 2020. Infinite Mixtures of Infinite Factor Analysers. *Bayesian Analysis*, **15**(3), 937 – 963.

SCHIAVON, LORENZO, CANALE, ANTONIO, & DUNSON, DAVID B. 2022. Generalized infinite factorization models. *Biometrika*, **109**(3), 817–835.

A NOVEL MULTI-VIEW ENSEMBLE CLUSTERING FRAMEWORK FOR CANCER SUBTYPE DISCOVERY

Michael G. Schimek[1], Bastian Pfeifer[2] and Marcus D. Bloice[3]

[1] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria, (e-mail: michael.schimek@medunigraz.at)

[2] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria, (e-mail: bastian.pfeifer@medunigraz.at)

[3] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria, (e-mail: marcus.bloice@medunigraz.at)

ABSTRACT: Multi-view clustering methods are essential for the stratification of patients into sub-groups of similar molecular characteristics. Recently, a wide range of methods has been developed for this purpose. However, due to the high diversity of cancer-related data, a single method may not perform sufficiently well in all instances. We present a multi-view hierarchical ensemble clustering framework of methods. We apply and validate it on real-world multi-view cancer patient data. Our approach outperforms the current state-of-the-art in all but one case. It is integrated into our Python package *Pyrea* [https://github.com/mdbloice/Pyrea].

KEYWORDS: multi-view clustering, ensemble clustering, hierarchical clustering, multi-omics, disease subtyping

1 Introduction

Multi-view data contain information relevant for the identification of patterns or clusters that allow us to specify groups of subjects or objects. This presentation is based on (Pfeifer *et al.*, 2023) with a focus on patients for which we have bio-medical and/or clinical observations describing their characteristics obtained from various diagnostic procedures or different molecular technologies. The different types of subject characteristics constitute views related to the patients of interest. Integrative clustering of these views facilitates the detection of patient groups, with the advantage of improved clinical diagnostic and treatment schemes.

Simple integration of single view clustering results is not appropriate for the diversity and complexity of available medical information. Even state-of-the-art multi-view approaches have their limitations, although ensemble clustering has the potential to overcome some of them (Alqurashi & Wang, 2019).

Data views can stem from highly heterogeneous input sources. Therefore, each view needs to be clustered with the most adequate strategy. Multi-view clustering methods are widely applied within the bio-medical domain, where often molecular data are retrieved from different biological layers for the same set of patients. Those clusters inferred from these multi-omics observations facilitate the stratification of cancer patients into sub-groups, providing a useful tool towards precision medicine.

There are two basic types of a multi-view clustering integration, one horizontal and the other vertical (Richardson *et al.*, 2016). Horizontal integration is the aggregation of homogeneous data views, while vertical integration entails the joint analysis of heterogeneous data views from the same group of patients. When data are highly diverse with respect to their probability distributions, problems can arise in vertical integration. Simple data concatenation and the application of single-view methods are most likely to produce biased results.

Clustering ensembles and multi-view clustering methods should provide more robust and accurate clustering results compared to an individual clustering algorithm. A wide range of multi-view clustering methods has been developed, for instance (Xue *et al.*, 2019), (Liu *et al.*, 2021), and (Yang *et al.*, 2022). Other recent approaches, e. g. (Rappoport & Shamir, 2019), (John *et al.*, 2020), and (Pfeifer & Schimek, 2021), have specialised in biomedical applications such as disease subtype detection. However, only a few contributions have investigated the possibility of combining the strengths of both ensemble clustering and multi-view clustering to further improve consistency and accuracy. Here, in contrast to the above-mentioned as well as many other methods, we aim at a generic theoretical and practical framework to enhance flexible ensemble-based multi-view clustering. Our framework is flexible with regard to those clustering techniques that are most suitable for the considered data. Furthermore, the framework allows to construct arbitrarily complex ensemble architectures.

2 The ensemble architecture and proposed methodology

Each view $V \in \mathbb{R}^{n \times p}$ is associated with a specific clustering method c, where n is the number of samples and p is the number of predictors. In total let us have N data views. An ensemble, called $\mathcal{E}$, can be modelled using a set of views $\mathcal{V}$ and an associated fusion algorithm f. Let us have $\mathcal{V} \leftarrow\!\shortmid \{(V \in \mathbb{R}^{n \times p}, c)\}$, $\mathcal{E}(\mathcal{V}, f) \mapsto \widetilde{V} \in \mathbb{R}^{n \times n}$, and $\mathcal{V} \leftarrow\!\shortmid \{(\widetilde{V} \in \mathbb{R}^{n \times n}, c)\}$. From these equations we can see that a specified ensemble $\mathcal{E}$ creates a view $\widetilde{V} \in \mathbb{R}^{n \times n}$ which again can

be used to specify $\mathcal{V}$, including an associated clustering algorithm c. With this concept it is possible to stack layer-wise views and ensembles into arbitrarily complex ensemble architectures. It should be noted, however, that the resulting view of a specified ensemble $\mathcal{E}$ forms an affinity matrix of dimension $n \times n$, and thus only those clustering methods which are compatible with an affinity or distance matrix as input are applicable. The data views are clustered with up to N different hierarchical clustering methods $hc_1, hc_2, \ldots, hc_N$, where N is the number of views. The best combination of clustering methods is inferred by a genetic algorithm, where the silhouette coefficient is adopted as a fitness function. For technical details see (Pfeifer & Schimek, 2021). The Parea$_{hc}$ ensemble approach comprises two different strategies: Parea$^1_{hc}$ is limited to the application of two selected hierarchical clustering methods, while Parea$^2_{hc}$ allows for a variation of hierarchical clustering methods in the data fusion process. Based on machine learning benchmark data sets, a comparison with state-of-the-art methods, such as multi-view spectral clustering and multi-view k-means clustering, was carried out in support of the described approach.

3 Multi-omics clustering for disease subtype discovery

We applied our methodology to a set of real patient data, often used as benchmark data (Rappoport & Shamir, 2018), of seven different cancer types, namely glioblastoma multiforme (GBM), kidney renal clear cell carcinoma (KIRC), liver hepatocellular carcinoma (LIHC), skin cutaneous melanoma (SKCM), ovarian serous cystadenocarcinoma (OV), sarcoma (SARC), and acute myeloid leukemia (AML), aiming at the externally known survival outcome. The Parea$_{hc}$ ensemble approach was studied on multi-omics data, including gene expression (mRNA), DNA methylation, and micro-RNA. Parea$_{hc}$ was compared with SNF (Wang *et al.*, 2014), NEMO (Rappoport & Shamir, 2019), HCfused (Pfeifer & Schimek, 2021), and PINSplus (Nguyen *et al.*, 2019). It is important to mention that the cancer patients were exclusively clustered based on their genomic footprints.

The survival data of all patients were used for the validation of the obtained patient clusters. For the quantification of differences between the studied methods, the Cox log-rank test was applied. The obtained p-values are displayed in Table 1. Our Parea$_{hc}$ ensembles outperform the alternative approaches in almost all cases. SKCM is the only cancer type for which HCfused achieved a superior result. Notably, the spectral-based clustering methods NEMO and SNF performed poorly.

Table 1. *Survival analysis of TCGA cancer group clusters*

Cancer type	Sample size	SNF	PINSplus	NEMO	HCfused	Parea$^1_{hc}$	Parea$^2_{hc}$
GBM	538	0.1304	0.2223	**0.0347**	0.0997	**0.0447**	**0.0347**
KIRC	606	0.3962	0.4005	0.3464	0.0561	**0.0137**	**0.0400**
LIHC	423	0.5357	0.6731	0.4354	0.2062	**0.0334**	**0.0436**
SKCM	473	0.5153	0.3956	0.4565	0.0699	0.1677	0.1629
OV	307	0.4042	0.5300	0.3593	0.2594	0.1685	0.2870
SARC	265	0.1622	0.2024	0.0979	**0.0408**	**0.0076**	**0.0109**
AML	173	0.0604	0.1973	**0.0440**	0.1148	**0.0167**	0.0502

Results based on (Pfeifer *et al.*, 2023): Median p-values of the Cox log-rank test. Significant results ($\alpha = 0.05$) for the separation of patient cluster survival curves in **bold**.

References

ALQURASHI, TAHANI, & WANG, WENJIA. 2019. Clustering ensemble method. *International Journal of Machine Learning and Cybernetics*, **10**(6), 1227–1246.

JOHN, CHRISTOPHER R, WATSON, DAVID, BARNES, MICHAEL R, PITZALIS, COSTANTINO, & LEWIS, MYLES J. 2020. Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, **36**(4), 1159–1166.

LIU, JIANLUN, TENG, SHAOHUA, FEI, LUNKE, ZHANG, WEI, FANG, XIAOZHAO, ZHANG, ZHUXIU, & WU, NAIQI. 2021. A novel consensus learning approach to incomplete multi-view clustering. *Pattern Recognition*, **115**, 107890.

NGUYEN, HUNG, SHRESTHA, SANGAM, DRAGHICI, SORIN, & NGUYEN, TIN. 2019. PINSplus: a tool for tumor subtype discovery in integrated genomic data. *Bioinformatics*, **35**(16), 2843–2846.

PFEIFER, BASTIAN, & SCHIMEK, MICHAEL G. 2021. A hierarchical clustering and data fusion approach for disease subtype discovery. *Journal of Biomedical Informatics*, **113**, 103636.

PFEIFER, BASTIAN, BLOICE, MARCUS, D., & SCHIMEK, MICHAEL G. 2023. Parea: Multi-view ensemble clustering for cancer subtype discovery. *Journal of Biomedical Informatics*, **143**, 104406.

RAPPOPORT, NIMROD, & SHAMIR, RON. 2018. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, **46**(20), 10546–10562.

RAPPOPORT, NIMROD, & SHAMIR, RON. 2019. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, **35**(18), 3348–3356.

RICHARDSON, SYLVIA, TSENG, GEORGE C, & SUN, WEI. 2016. Statistical methods in integrative genomics. *Annual Review of Statistics and its Application*, **3**, 181–209.

WANG, BO, MEZLINI, AZIZ M, DEMIR, FEYYAZ, FIUME, MARC, TU, ZHUOWEN, BRUDNO, MICHAEL, HAIBE-KAINS, BENJAMIN, & GOLDENBERG, ANNA. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, **11**(3), 333–337.

XUE, ZHE, DU, JUNPING, DU, DAWEI, & LYU, SIWEI. 2019. Deep low-rank subspace ensemble for multi-view clustering. *Information Sciences*, **482**, 210–227.

YANG, MOUXING, LI, YUNFAN, HU, PENG, BAI, JINFENG, LV, JIANCHENG, & PENG, XI. 2022. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**(1), 1055–1069.

REDUCING SELECTION BIAS IN NON-PROBABILITY SAMPLE BY SMALL AREA ESTIMATION

Francesco Schirripa Spagnolo [1], Gaia Bertarelli[2], Nicola Salvati [1], Donato Summa [3], Monica Scannapieco [3], Stefano Marchetti [1] and Monica Pratesi [13]

[1] Department of Economics and Management, University of Pisa, (e-mail: francesco.schirripa@unipi.it, nicola.salvati2@unipi.it, stefano.marchetti@unipi.it, monica.pratesi@unipi.it)

[2] Department of Economics, Ca' Foscari University of Venice (e-mail: gaia.bertarelli@unive.it)

[3] ISTAT (e-mail: summa@istat.it, scannapieco@istat.it, pratesi@istat.it)

ABSTRACT: Nowadays, the availability of a huge amount of data produced by a wide range of new technologies is increasing. However, data obtainable from these sources are often the result of a non-probability sampling process. We propose a method to reduce the selection bias associated with the big data in the context of Small Area Estimation. Our approach is based on data integration and it combines a big data sample and a probability sample. Real data examples are considered in the context of Italian enterprises sensitiveness towards Sustainable Development Goals and e-commerce.

KEYWORDS: official statistics, big data, data integration, SDGs, e-commerce.

1 Introduction

For many decades probability surveys have been the standard for producing official statistics. However, the decline in response rates in probability surveys associated with the increasing cost of data collection have become big issues for producing official statistics. Due to technological innovations, over the past decade, there has been an unprecedented increase in the volume of "new" data, called *big data*, which are often the results of non-probability sampling processes but, at the same time, they offer very rich data sets. Anyway the "nature" itself of the data, as collected without a probability scheme, opens the door to possible selection bias, even at domain level.
Although, there is a trend to modernize official statistics through a more extensive use of big data, making reliable inferences from a non-probability sample alone is very challenging and a naive use of these data can lead to

biased estimates as affected by selection bias and measurement error. The Italian National Statistical Institute has a strategic program of investments on the use of these new data sources to complement and enrich official statistics. In this context a roadmap document, named "Roadmap for Trusted Smart Statistics"(RTSS), has been released. This work must be laid in the methodological action of the RTSS related to quality improvement by reducing non-representativeness of Big Data sources at survey unplanned domain level.

2 Notation

We consider a population U of size N divided into m non-overlapping subsets U_i of size N_i, $i = 1,\ldots,m$. Let y_{ij} denote the value of the target variable for the unit j belonging to the area i. A non-probability sample B is available for the target population, with $B \subset U$. We assume that the non-probability sample is available in each area of interest: B_i is the non-probability sample in the area i, $B_i \subset U_i$. We denote the inclusion indicator in B_i as δ_{ij}; in other words, $\delta_{ij} = 1$ if $j \in B_i$, $\delta_{ij} = 0$ otherwise; therefore $N_{B_i} = \sum_{j=1}^{N_i} \delta_{ij}$. The study variable y_{ij} is observed only when $\delta_{ij} = 1$. The non-probability sample contains other auxiliary variables, denoted by $\mathbf{x}$.

A survey data of size n, denoted by A, is also available; $A_i \in U_i$ drawn randomly. The survey data do not contain the variable of interest but contain only the auxiliary variables $\mathbf{x}$. The area-specific samples A_i are available in each area, but the number of sample units in each area, $n_i > 0$, is limited. Therefore, the areas of interest can be denoted as "small areas". In general, an area is regarded as "small" if the domain-specific sample size is not large enough to obtain direct estimates with acceptable statistical significance. In these cases, SAE techniques need to be employed.

In summary, the available data can be denoted by $\{(y_{ij}, x_{ij}), i \in B\}$ and $\{(x_{ij}), i \in A\}$, and the quantities of interest are the area means $\bar{Y}_i = N_i^{-1} \sum_{j \in U_i} y_{ij}, i = 1,\ldots,m$. By using B we can estimate $\bar{Y}_i$ by:

$$\bar{Y}_{B_i} = N_{B_i}^{-1} \sum_{j \in B_i} y_{ij},$$

where $N_{B_i} = \sum_{j=1}^{N_i} \delta_{ij}$ and y_{ij} is the jth observation in the area i. Because of the selection bias and the measurement error, the sample mean $\bar{Y}_{B_i}$ from the non-probability sample is biased, and it does not represent the target population (Kim & Wang, 2019). Therefore, we propose a techniques in order to make valid inference from big data sources when the aim is to provide reliable estimates at small area level.

3 Reducing selection bias in big data: a data integration approach using SAE methods

We consider a data integration method for combining probability and non-probability samples in order to reduce the bias which is assisted by unit level small area model, following the approach of Kim and Wan (2019). We consider the case in which the survey data and the big data are available in each small area of interest. We also assume that the selection mechanism for the big data is non-informative :

$$P(\delta_{ij} = 1|\mathbf{x}_{ij}, y_{ij}; u_i) = P(\delta_{ij} = 1|\mathbf{x}_{ij}; u_i)$$

where u_i is an area-specific random effect characterizing the between-area differences in the distribution of y_{ij} given the covariates $\mathbf{x}_{ij}$.

Moreover, we can observe δ_{ij}, the big data sample inclusion indicator, from the sample A. We can use the data $\{(\delta_{ij}, \mathbf{x}_{ij})\} \in A_i$ to fit a model for the for the propensity scores $P(\delta_{ij} = 1|\mathbf{x}_{ij}) = p(\mathbf{x}, \lambda)$ in sample B based on the missing at random. Usually, a logistic regression model for the binary variable δ_{ij} can be used in order to obtain estimators $\hat{p}_{ij}$ in sample B.

In order to take into account the hierarchical structure of the data, we consider the following generalized linear random intercept model for the propensity scores:

$$\hat{p}_{ij}(\hat{\lambda}, \hat{u}_i) = g^{-1}(\mathbf{x}_{ij}^T \hat{\lambda} + \hat{u}_i),$$

where $g(\cdot)$ is a logit link function; $\hat{\lambda}$ and $\hat{u}_i$ are the ML estimates of λ and u_i.

To develop our estimator we suppose that the following working population model holds for sample B:

$$E[y_{ij}|\mathbf{x}_{ij}, \gamma_i] = \mu_{ij} = h^{-1}\left(\mathbf{x}_{ij}^T \beta + \gamma_i\right), \tag{1}$$

where $h(\cdot)$ is the link function, assumed to be known and invertible, γ_i is the area-specific random effect for area i characterizing the between-area differences in the distribution of y_{ij} given the covariates $\mathbf{x}_{ij}$. It should be noted that the covariates used here could be different from those used to fit the propensity model. Model in equation (1) includes three important special cases: the linear model obtained with $h(\cdot)$ equal to the identity function and y_{ij} is a continuous variable; logistic generalized linear random intercept model, where $h(\cdot)$ is the logistic link function and the outcome variable is binomial; the Poisson-log generalized linear random intercept model where $h(\cdot)$ is the log link function

and the individual y_{ij} values are taken to be independent Poisson random variable. Using data from the big data sample B, assuming the model is correctly specified, we obtain an estimator of $\hat{\beta}$ which is consistent for β (Rao, 2021). Then a doubly robust (DR) estimator of the mean is given by:

$$\hat{\theta}_{i;DR}^{EBLUP} = \frac{1}{N_i}\left\{\sum_{j\in B_i}\frac{1}{\hat{p}_{ij}(\hat{\lambda},\hat{u}_i)}(y_{ij}-\hat{\mu}_{ij}) + \frac{N_i}{n_i}\sum_{j\in A_i}\hat{\mu}_{ij}\right\}, \tag{2}$$

where $\hat{\mu}_{ij} = h^{-1}\left(\mathbf{x}_{ij}\hat{\beta}+\hat{\gamma}_i\right)$ and $\hat{\beta}$ and $\hat{\gamma}_i$ are respectively the estimated regression coefficients and the random effects based on the big data sample.

The estimator in Eq. (2) is DR in the sense that it is consistent if both the model for propensity scores and the model for the study variable are correctly specified (Kim & Wang, 2019, Rao, 2021).

4 Real data examples

The proposed methodology has been applied to estimate the proportion of enterprises sensitive to Sustainable Development Goals (SDGs) of the 2030 Agenda at the provincial level in Italy. The Big Data sample is represented by the enterprises' websites accessed due to a web scraping procedure. The probabilistic sample dataset is a sub-sample of the survey "*Situazione e prospettive delle imprese nell'emergenza sanitaria Covid-19"* (2020). The target variable is a binary indicator computed for each enterprise and represents if the enterprise is sensitive or not to SDGs. This indicator has been computed through machine learning methods by analyzing the big data sample and looking for a set of pre-defined SDGs-related words on each website. Furthermore, an application related to the diffusion of e-commerce in Italian companies, using the same data of the application on sustainability, will be considered.

References

KIM, JAE KWANG, & WANG, ZHONGLEI. 2019. Sampling techniques for big data analysis. *International Statistical Review*, **87**, S177–S191.

RAO, JNK. 2021. On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, **83**, 242–272.

SPARSE AND ROBUST ESTIMATORS FOR OUTLIER DETECTION IN DISTRIBUTIONAL DATA

Pedro Duarte Silva [1], Peter Filzmoser [2] and Paula Brito [3]

[1] Católica Porto Business School & CEGE, Universidade Católica Portuguesa, Porto, Portugal, (e-mail: psilva@ucp.pt)

[2] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria, (e-mail: peter.filzmoser@tuwien.ac.at)

[3] Faculdade de Economia, Universidade do Porto & LIAAD-INESC TEC, Porto, Portugal, (e-mail: mpbrito@fep.up.pt)

ABSTRACT: The classical data representation model is too restrictive when the data to be analysed are not real numbers but comprise variability. In this talk, we are interested in numerical distributional data, where units are described by histogram or interval-valued variables. We consider parametric probabilistic models, which are based on the representation of each distribution by a location measure and interquantile ranges. A multivariate outlier detection method is proposed that makes use of restricted configurations for the covariance matrix, and is based on a sparse robust estimator of its inverse. The computations rely on an efficient adaptation of the graphical lasso algorithm. A simulation study puts in evidence the usefulness of the robust estimates for outlier detection.

KEYWORDS: outliers, robust statistics, distributional data, Mahalanobis distance, graphical lasso

1 Introduction

Multivariate datasets often include atypical data points known as *outliers*, i.e. points that deviate from the main pattern. Outlier detection is important because outlying data points may reveal nonconforming phenomena and the results of usual multivariate methods can be heavily influenced by them.

In this paper we address the problem of outlier detection in multivariate distributional data. Distributional data may result from the aggregation of large amounts of open/collected/generated data, or may be directly available in a structured or unstructured form, describing the variability of some features. In recent years, different approaches have been investigated and methods proposed for the analysis of such data. However, most existing methods rely on non-parametric descriptive approaches.

A common approach for multivariate outlier detection measures outlyingness by Mahalanobis distances. Given a sample of n observations, a point i is considered an outlier if its distance $D^2_{\hat{\mu},\hat{\Theta}}(i)$ from an appropriate mean estimate, $\hat{\mu}$, is above a relevant threshold. Here, $\hat{\Theta}$ is an estimate of the precision matrix, $\Theta = \Sigma^{-1}$, and Σ denotes the population covariance. However, if $\hat{\mu}$ and $\hat{\Theta}$ are chosen to be the classical sample mean vector and inverse covariance matrix, S^{-1}, this procedure is not reliable, as $D^2_{\hat{\mu},\hat{\Theta}}(i)$ may be strongly affected by atypical observations. Furthermore, S^{-1} has a large sample variability when its dimension, d, is close to n, and it is is not even computable when $d > n$. To address these issues Öllerer and Croux (Öllerer & Croux, 2015), proposed sparse precision matrix estimators based on the GLASSO L_1-penalized log-likelihood function (Friedman *et al.* , 2008).

In this paper we address the problem of outlier detection in distributional data, combining Öllerer and Croux estimators with a parametric modelling of distributional data, along the lines of Brito & Duarte Silva, 2012, and Duarte Silva *et al.* , 2018.

2 Distributional Variables

Let $S = \{s_1, \ldots, s_n\}$, be the set of n units under analysis. We consider that for each unit, the descriptive variables are (in general) not constant, but present variability.

We represent the "values" of a numerical distributional variable by an ordered vector of quantiles, always including the minimum and the maximum. Formally, a numerical distributional variable is defined by an application

$$\begin{gathered} Y : S \to T \\ s_i \to Y(s_i) = (Min_i, \psi_{1i}, \ldots, \psi_{qi}, Max_i) \end{gathered}$$

Let $Y_1, \ldots, Y_p$ be the p numerical distributional variables, defined on S. Here we assume that all variables are represented by the same set of $q+2$ quantiles, and that $Min_{ij} < \psi_{1ij} < \ldots < \psi_{qij} < Max_{ij}, 1 \leq i \leq n, 1 \leq j \leq p$ (strict inequalities).

The model consists in representing $Y_j(s_i)$ by

- a central statistic C_{ij}, typically the Median Med_{ij} or the MidPoint $\frac{Max_{ij}+Min_{ij}}{2}$
- the $[Min, \psi_1[$ range: $R_{1ij} = \psi_{1ij} - Min_{ij}$
- the $[\psi_1, \psi_2[$ range: $R_{2ij} = \psi_{2ij} - \psi_{1ij}$
- ...

- the $[\psi_q, Max[$ range: $R_{mij} = Max_{ij} - \psi_{qij}$

Typical cases consist in using the median, or else the midpoint, as central statistics, and quartiles, or other equally-spaced quantiles.

The proposed model consists in assuming that the joint distribution of the central statistic C and the logarithms of the ranges $R^*_\ell, \ell = 1, \ldots, m$, is Gaussian:

$(C, R^*_1, \ldots, R^*_m) \sim N_{(m+1)p}(\mu, \Sigma)$

In the most general formulation (configuration 1) we allow for non-zero correlations among all central statistics and log-ranges; for distributional variables there are however other cases of interest: the distributional-valued variables Y_j are non-correlated, but for each variable, the central statistic and all its log-ranges may be correlated among themselves (configuration 2); central statistics (respectively, log-ranges) of different variables may be correlated, but no correlation between central statistics and log-ranges is allowed (configuration 3); central statistics (respectively, each log-range) of different variables may be correlated, but no correlation between central statistics and log-ranges or between non-corresponding log-ranges is allowed (configuration 4); and, finally, all central statistics and log-ranges are non-correlated (configuration 5).

3 Outlier Detection of Distributional Data

Let $X_i = \left[C_i^t, R_{1i}^{*\,t}, \ldots, R_{mi}^{*\,t}\right]^t$ be the $d = (m+1)p$ dimensional column vector comprising all central statistics and log-ranges for s_i, $i = 1, \ldots, n$.

The identification of outliers is based on robust Mahalanobis distances, $D^2_{\hat{\mu},\hat{\Theta}}(i) = (x_i - \hat{\mu})^t \hat{\Theta} (x_i - \hat{\mu})$ from each data point to a robust location vector, $\hat{\mu}$, which are then compared with the 97.5% quantile of a chi-squared distribution with d-degrees of freedom. In our approach we choose as location vector, the L_1 median (Fritz *et al.* , 2012), which has a break-down point of 0.5 and, given our Gaussian assumption, is a robust estimator of μ.

Following Öllerer and Croux (2015) we estimate $\Theta = \Sigma^{-1}$ by

$$\hat{\Theta} = argmax_{\Theta \in \vartheta} \log\det(\Theta) - tr(\hat{\Sigma}\Theta) - \rho \sum_{j,k=1}^{d} |(\Theta)_{jk}| \qquad (1)$$

where $\vartheta := \{\Theta \in \mathbb{R}^{d \times d} : \Theta \succ 0\}$ is the space of d-dimensional positive-definite matrices, $\hat{\Sigma}$ is a robust covariance estimate, and ρ a regularization parameter.

For each covariance configuration, we set the null elements of Σ to zero in its initial $\hat{\Sigma}$ estimate, and for the remaining elements we use the formula

$$\hat{\Sigma}_{j,k} = scale(X^j)\, scale(X^k)\, r(X^j, X^k) \tag{2}$$

where X^j, X^k are the j^{th} and k^{th} columns of X, $scale(X^j), scale(X^k)$ are robust scale estimators (see Rousseeuw & Croux, 1993), and $r(X^j, X^k)$ is the Gaussian rank correlation (Boudt *et al.* , 2012) between X^j and X^k.

The above procedure was evaluated in a controlled simulation experiment that showed promising results for the proposed approach.

References

BOUDT, KRIS, CORNELISSEN, JONATHAN, & CROUX, CHRISTOPHE. 2012. The Gaussian rank correlation estimator: robustness properties. *Statistics and Computing*, **22**, 471–483.

BRITO, PAULA, & DUARTE SILVA, A. PEDRO. 2012. Modelling Interval Data with Normal and Skew-Normal distributions. *Journal of Applied Statistics*, **39**(1), 3–20.

DUARTE SILVA, A. PEDRO, FILZMOSER, P., & BRITO, PAULA. 2018. Outlier detection in Interval Data. *Advances in Data Analysis and Classification*, **12**(3), 785–822.

FRIEDMAN, JEROME, HASTIE, TREVOR, & TIBSHIRANI, ROBERT. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.

FRITZ, HEINRICH, FILZMOSER, PETER, & CROUX, CHRISTOPHE. 2012. A comparison of algorithms for the multivariate L1-median. *Computational Statistics*, **27**, 393–410.

ÖLLERER, VIKTORIA, & CROUX, CHRISTOPHE. 2015. Robust high-dimensional precision matrix estimation. *Pages 325–350 of: Modern Nonparametric, Robust and Multivariate Methods*. Springer.

ROUSSEEUW, PETER J, & CROUX, CHRISTOPHE. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, **88**(424), 1273–1283.

CLUSTERING GENES SPATIAL EXPRESSION PROFILES WITH THE AID OF EXTERNAL BIOLOGICAL KNOWLEDGE

Andrea Sottosanti [1], Sara Agavni' Castiglioni [2], Stefania Pirrotta [3] , Enrica Calura[3], and Davide Risso [2]

[1] Department of Medicine, University of Padova,
(e-mail: andrea.sottosanti@unipd.it)

[2] Department of Statistical Sciences, University of Padova,

[3] Department of Biology, University of Padova

ABSTRACT: In the analysis of spatial transcriptomic experiments, the recently proposed SpaRTaCo model (Sottosanti & Risso, 2022) allows for the simultaneous clustering of genes and cells of a tissue sample, providing interesting insights abouve the underlying biological processes. In this work, we discuss how to integrate external knowledge such as manual cell-type annotations to inform gene clustering, with the by-product of substantially reducing the computational burden.

KEYWORDS: clustering, genomics, spatial statistics, spatial transcriptomics.

1 Introduction

Spatial transcriptomics is an innovative class of sequencing technologies, capable of providing the expression levels of thousands of genes in a tissue sample while retaining the spatial conformation of the analyzed tissue. With the aid of additional spatial information, researchers can better understand the complex biological processes that depend on the cellular organization of the tissue. New insights come from the discovery of *spatially expressed* (s.e.) genes, i.e., genes that exhibit specific patterns of variation in space (Svensson *et al.*, 2018).

Recently, we proposed SpaRTaCo (Sottosanti & Risso, 2022), a co-clustering model for spatial transcriptomic experiments, which has shown to be capable of determining s.e. genes active only in specific areas of a sample, providing insights that could not be achieved by competing methods in the literature. Clearly, it represents a useful tool for spatial transcriptomic data analysis; nevertheless, its estimation process is highly computationally demanding.

Here, we propose a modification of the original SpaRTaCo formulation that integrates external biological knowledge to speed up the computation. In

fact, spatial experiments often come with a manual annotation of the cellular composition of a sample made by a pathologist, providing a relevant source of information that can be integrated into the inferential process. Furthermore, we propose to estimate SpaRTaCo with a penalized maximum likelihood approach to prevent the model from capturing spurious spatial correlation, retaining relevant patterns only. We conclude with the analysis of a prostate cancer tissue sample analyzed with a recent spatial transcriptomic technology.

2 The semi-supervised SpaRTaCo with L_1 and L_2 penalizations

Let $\mathbf{X}$ be the $n \times p$ matrix of a spatial experiment having the expression of n genes measured over p spots, whose spatial locations are known. SpaRTaCo assumes the existence of K gene clusters and R spot clusters, inducing a partition of the experiment matrix into $K \times R$ blocks. Thus, the kr-th block has dimension $\dim(\mathbf{X}^{kr}) = n_k \times \ p_r$, and $\mathbf{X} = \left(\mathbf{X}^{kr}\right)$, with $k = 1, \ldots, K$ and $r = 1, \ldots, R$. The expression of the i-th gene with the kr-th block distributes as

$$\mathbf{x}_{i\cdot}^{kr} | \sigma^2_{kr,i} \sim N_{p_r}\left(\mu_{kr}\mathbf{1}_{p_r}, \sigma^2_{kr,i}\boldsymbol{\Delta}_{kr}\right), \qquad \sigma^2_{kr,i} \sim I\mathcal{G}(\alpha_{kr}, \beta_{kr}) \tag{1}$$

where μ_{kr} is a mean parameter, $\mathbf{1}_{p_r}$ is a vector of ones of length p_r, $\sigma^2_{kr,i}$ is a gene-specific variance, and $\boldsymbol{\Delta}_{kr}$ is the covariance matrix of the spots with form

$$\boldsymbol{\Delta}_{kr} = \tau_{kr}\mathcal{K}(\mathbf{S}_r; \phi_r) + \xi_{kr}\mathbb{1}_{p_r}. \tag{2}$$

Notice that $\boldsymbol{\Delta}_{kr}$ is expressed as a linear combination of two matrix terms: the first is a kernel matrix with isotropic spatial covariance function $k(\cdot; \phi_r)$ that models the gene expression correlation across the spots of cluster r (with spatial coordinates $\mathbf{S}^r = (s_j)$), the second is an identity matrix of size p_r. The parameters τ_{kr} and ξ_{kr} quantify the amount of spatial variation and residual intra-block variability, respectively. Moreover, the quantity τ_{kr}/ξ_{kr} can be used to measure the amount of spatial variability compared to the residual variability, and for this reason it is called *spatial signal-to-noise ratio*. Last, the parameter $\sigma^2_{kr,i}$ in (1) is used to model the variance specific of gene i in the kr-th to account for the possible dependence across genes in the same cluster.

Even though SpaRTaCo is designed for clustering both rows (genes) and columns (cells) of $\mathbf{X}$, when a manual annotation of the tissue image is available, we can include it in the model in place of the column clustering labels to inform the inferential process. In addition, to improve the stability of parameter estimation, we can estimate the model with a penalized maximum likelihood approach. A *lasso* penalty on the parameters τ_{kr} discourages the

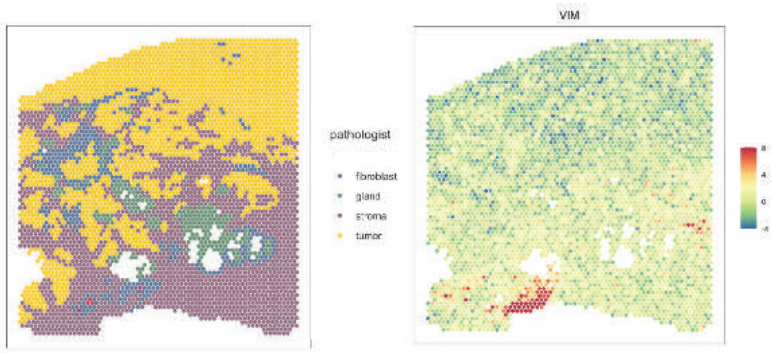


Figure 1: Left: human prostate tissue diagnosed with adenocarcinoma. Spot colours denote Dr. Esposito's annotation (red spots are not considered as they appear only 5 times). Right: spatial distribution of gene VIM.

model from capturing spurious correlation when no spatial effect is present, while a *ridge* penalty regularizes the mean parameters μ_{kr} since zero values do not have a clear biological meaning. The estimates of the model parameters $\Theta = \cup_r\{\cup_k(\mu_{kr}, \tau_{kr}, \alpha_{kr}, \beta_{kr}), \phi_r\}$ and of the clustering labels $\mathcal{Z}$ are obtained maximizing

$$\log \mathcal{L}(\Theta, \mathcal{Z}|\mathbf{X}, \mathcal{W}) - \lambda_\tau \sum_{k=1}^{K}\sum_{r=1}^{R} |\tau_{kr}| - \lambda_\mu \sum_{k=1}^{K}\sum_{r=1}^{R} \mu_{kr}^2, \tag{3}$$

where $\mathcal{W}$ is the vector containing the spot clustering labels that come with the data, $\log \mathcal{L}(\Theta, \mathcal{Z}|\mathbf{X}, \mathcal{W})$ is the classification log-likelihood, and λ_τ and λ_μ are the penalization terms associated to the τ and μ parameters, respectively. Simulation studies not reported here showed that $\lambda_\mu = 1.5$ and $\lambda_\tau = 0.3$ guarantee robust parameter estimates and prevent the model from capturing spurious spatial correlation. Notice that the parameters ξ_{kr} are not estimated, but are fixed a priori, for identifiability reasons. An exact solution to the maximization of (3) can be obtained using a classification EM algorithm.

3 Application to human prostate cancer data

We analyze a human prostate tissue diagnosed with adenocarcinoma processed with 10X-Visium platform (Righelli *et al.*, 2022). The slide was manually annotated by the pathologist Dr. Esposito (Veneto Oncology Institute, Italy), by analyzing microscope images that consider the cytoarchitecture of

the cells, i.e., the spatial organization and arrangement of cells within the tissue. Based on these characteristics, the tissue was divided into four macro categories: fibroblasts, glands, stroma, and tumour (Figure 1, left). After preliminary gene filtering and count normalization (Townes *et al.*, 2019), the final dataset had 1000 genes measured over 4366 locations (spots).

We estimated the semi-supervised SpaRTaCo using $K \in \{1,\dots,9\}$ and, after evaluating the *integrated complete log-likelihood* criterion and the clustering uncertainties (Sottosanti & Risso, 2022, Section 3.3 and 3.4), we selected the model with $K = 5$ gene clusters. The first two clusters that the model identifies have a substantial spatial variability in all tissue areas ($\hat{\tau}_{kr}/\hat{\xi}_{kr} > 1.5$, for $k = 1,2$ and $\forall r$) and particularly pronounced in the tumour area ($\hat{\tau}_{14}/\hat{\xi}_{14} = 7.12$, $\hat{\tau}_{24}/\hat{\xi}_{24} = 2.46$). In comparison, the remaining three gene clusters have moderate or absent spatial variability throughout the tissue and show substantial differences only at the mean level.

Thanks to gene-specific variance parameters $\sigma^2_{kr,i}$, we can provide a list of the most variable genes in every tissue area. As an example, the gene VIM appeared among the 20 most variable genes in the stromal region (Figure 1, right). VIM is a cancer growth promoter gene, and therefore, from this observed expression pattern, it can provide helpful information about the nature of the tumour and be the starting point for biological investigations. Alternative algorithms for selecting highly variable genes (e.g., Townes *et al.*, 2019) do not include VIM among the top 80 most informative genes, showing the importance of accounting for the spatial variability of the data in the analysis.

References

RIGHELLI, DARIO, WEBER, LUKAS M, CROWELL, HELENA L, & PARDO, BRENDA, ET AL. 2022. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. *Bioinformatics*, **38**(11), 3128–3131.

SOTTOSANTI, ANDREA, & RISSO, DAVIDE. 2022. Co-clustering of Spatially Resolved Transcriptomic Data. *The Annals of Applied Statistics*. In press.

SVENSSON, V., TEICHMANN, S., & STEGLE, O. 2018. *SpatialDE: identification of spatially variable genes | Nature Methods*.

TOWNES, F. WILLIAM, HICKS, STEPHANIE C., ARYEE, MARTIN J., & IRIZARRY, RAFAEL A. 2019. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, **20**(1), 295.

STRUCTURAL EQUATION MODELING WITH LATENT/EMERGENT VARIABLES: RGCCAc

Arthur Tenenhaus[1], Michel Tenenhaus[2] and Theo Dijkstra

[1] Universite Paris-Saclay, CentraleSupelec, Laboratoire des Signaux et Systèmes
(e-mail: arthur.tenenhaus@centralesupelec.fr)

[2] HEC Paris
(e-mail: tenenhaus@hec.fr)

ABSTRACT: We present how to use Regularized Generalized Canonical Correlation Analysis (RGCCA) in structural equation modeling with latent and/or emergent variables. This new approach, named consistent RGCCAc (RGCCAc), produces consistent and asymptotically normal estimators of the parameters. RGCCAc relies on a well-grounded optimization problem and the global convergence of the algorithm used to solve this problem is guaranteed. RGCCAc contains composite models as special case, keeps the robustness and simplicity of PLSc and cSEM and corrects their shortcomings. RGCCAc, cSEM and Maximum Likelihood (ML) based-approach are evaluated in a Monte Carlo simulation and on a case study and produce similar results.

KEYWORDS: Structural Equation Modeling, RGCCA, Consistent PLS, Composite models, composite-based SEM.

ON SOME PROPERTIES OF RECONSTRUCTED TRAJECTORIES FROM SPARSE LONGITUDINAL DATA

Yoshikazu Terada [1]

[1] Graduate School of Engineering Science, Osaka University / RIKEN AIP, (e-mail: terada.yoshikazu.es@osaka-u.ac.jp)

ABSTRACT: In sparse longitudinal data, we only have a few measurements on irregularly spaced time points from a hidden continuous stochastic process (trajectory) for each subject. The prediction of individual trajectories is sometimes useful for functional data analysis of such data, and the properties of the reconstructed trajectories play important roles in theoretical analysis. When we have measurements on a dense grid of time points for each subject, we can reconstruct the individual trajectories independently. However, for sparse longitudinal data, we often use the reconstruction method (Yao *et al.* , 2005) based on functional principal component analysis (FPCA). In this case, the predicted trajectories are not independent. In this complicated situation, we demonstrate some fundamental properties of the individual trajectories reconstructed by FPCA.

KEYWORDS: functional data analysis, weak convergence.

References

YAO, FANG, MÜLLER, HANS-GEORG, & WANG, JANE-LING. 2005. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, **100**(470), 577–590.

CLUSTERPATH GAUSSIAN GRAPHICAL MODELING

Daniel J.W. Touw[1], Patrick J.F. Groenen[1], Ines Wilms[2], and Andreas Alfons[1]

[1] Erasmus School of Economics, Erasmus University Rotterdam, (e-mail: touw@ese.eur.nl, groenen@ese.eur.nl, alfons@ese.eur.nl)

[2] Department of Quantitative Economics, Maastricht University, (e-mail: i.wilms@maastrichtuniversity.nl)

ABSTRACT: Gaussian graphical models (GGMs) serve as a means of summarizing conditional dependencies among a set of p variables. Such models are structured as networks, in which nodes represent individual variables and edges denote the presence of conditional dependence between two variables. Estimating GGMs in cases where the sample size n is smaller than the number of variables ($n < p$) can present a challenge. To address this issue, existing estimation methods frequently rely on applying regularization techniques to the edges within the network, with the aim of obtaining a sparse network where many variables are represented as conditionally independent (see, e.g., Cai *et al.*, 2011; Friedman *et al.*, 2008; Meinshausen & Bühlmann, 2006; Peng *et al.*, 2009; Rothman *et al.*, 2008; Yuan, 2010).

Nevertheless, relying solely on edge sparsity does have limitations. First, when the number of variables is substantially larger than the sample size ($n \ll p$), the conditional dependencies between variables may become too weak to detect (Eisenach *et al.*, 2020). Second, sparse GGMs that include many variables can still contain a substantial number of edges, making interpretation difficult (Grechkin *et al.*, 2015). Last, real-world networks often exhibit more complex structures than mere edge sparsity (Heinävaara *et al.*, 2016; Hosseini & Lee, 2016).

To overcome these challenges, node aggregation has emerged as a means to perform dimension reduction in GGMs (see, e.g., Hosseini & Lee, 2016; Pircalabelu & Claeskens, 2020; Tarzanagh & Michailidis, 2018; Wilms & Bien, 2022). For example, instead of estimating the conditional dependencies between all observed variables, one may be interested in identifying the dependencies among a smaller number of clusters that share the same behavior. To achieve this, we propose the clusterpath GGM (CGGM), a model-based convex clustering Gaussian graphical model that automatically clusters groups of variables by means of the penalty structure used in the convex clustering literature (Hocking *et al.*, 2011; Lindsten *et al.*, 2011; Pelckmans *et al.*, 2005).

KEYWORDS: convex clustering, dimension reduction, graphical modeling, regulatization

References

CAI, T., LIU, W., & LUO, X. 2011. A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, **106**(494), 594–607.

EISENACH, C., BUNEA, F., NING, Y., & DINICU, C. 2020. High-Dimensional Inference for Cluster-Based Graphical Models. *Journal of Machine Learning Research*, **21**(53), 1–55.

FRIEDMAN, J., HASTIE, T., & TIBSHIRANI, R. 2008. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, **9**(3), 432–441.

GRECHKIN, M., FAZEL, M., WITTEN, D., & LEE, S.-I. 2015. Pathway Graphical Lasso. *In: Proceedings of the AAAI conference on artificial intelligence*, vol. 29.

HEINÄVAARA, O., LEPPÄ-AHO, J., CORANDER, J., & HONKELA, A. 2016. On the Inconsistency of ℓ_1-penalised Sparse Precision Matrix Estimation. *BMC bioinformatics*, **17**(16), 99–107.

HOCKING, T.D., JOULIN, A., BACH, D., & VERT, J.-P. 2011. Clusterpath: An Algorithm for Clustering Using Convex Fusion Penalties. *In: The 28th International Conference on Machine Learning*.

HOSSEINI, M.J., & LEE, S.-I. 2016. Learning Sparse Gaussian Graphical Models with Overlapping Blocks. *Advances in Neural Information Processing Systems*, **29**.

LINDSTEN, F., OHLSSON, H., & LJUNG, L. 2011. *Just Relax and Come Clustering!: A Convexification of K-Means Clustering*. Tech. rept. Department of Electrical Engineering, Linköping University, Linköping, Sweden.

MEINSHAUSEN, N., & BÜHLMANN, P. 2006. High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, **34**(3), 1436–1462.

PELCKMANS, K., DE BRABANTER, J., SUYKENS, J.A.K., & DE MOOR, B. 2005. Convex Clustering Shrinkage. *In: PASCAL Workshop on Statistics and Optimization of Clustering Workshop*.

PENG, J., WANG, P., ZHOU, N., & ZHU, J. 2009. Partial Correlation Estimation by Joint Sparse Regression Models. *Journal of the American Statistical Association*, **104**(486), 735–746.

PIRCALABELU, E., & CLAESKENS, G. 2020. Community-Based Group Graphical Lasso. *Journal of Machine Learning Research*, **21**(1), 2406–2437.

ROTHMAN, A.J., BICKEL, P.J., LEVINA, E., & ZHU, J. 2008. Sparse Permutation Invariant Covariance Estimation. *Electronic Journal of Statistics*,

2, 494–515.

TARZANAGH, D.A., & MICHAILIDIS, G. 2018. Estimation of Graphical Models through Structured Norm Minimization. *Journal of Machine Learning Research*, **18**(1), 1–48.

WILMS, I., & BIEN, J. 2022. Tree-based Node Aggregation in Sparse Graphical Models. *Journal of Machine Learning Research*, **23**(243), 1–36.

YUAN, M. 2010. High Dimensional Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research*, **11**(79), 2261–2286.

HOUSING POVERTY IN EUROPE. MULTIDIMENSIONAL ANALYSIS

Paweł Ulman[1], Małgorzata Ćwiek[1] and Maria Sadko[3]

[1] Department of Statistics, Cracow University of Economics,
(e-mail: pawel.ulman@uek.krakow.pl)

[2] Department of Statistics, Cracow University of Economics,
(e-mail: malgorzata.cwiek@uek.krakow.pl)

[3] Department of Statistics, Cracow University of Economics,
(e-mail: maria.sadko@uek.krakow.pl)

ABSTRACT: The aim of the study is to determine the scale of housing poverty in Europe. To describe poor housing quality a multi-dimensional tool based on the the Integrated Fuzzy and Relative (IFR) methodology was used. This approach allows to include a large number of variables in the analysis and solves the problem of correlation between variables by assigning weights, rather than limiting the number of variables as before. The risk evaluation of bad housing situation is based on micro-data from the EU statistics on income and living conditions (EU-SILC) for 2021. The study took into account 13 variables describing the technical characteristics of the apartment, the surroundings of the place of residence and the economic conditions of housing maintenance. This set includes both quantitative and qualitative variables. The analysis was conducted for European 32 countries. The multidimensional approach adopted in this study captures the diversity of housing poverty risk both in selected areas of assessment, as well as in general, which is impossible to achieve using traditional housing deprivation measures.

KEYWORDS: housing poverty; housing in Europe; Integrated Fuzzy and Relative; fuzzy set.

References

BETTI, G., & VERMA, V. 2008. Fuzzy Measures of the Incidence of Relative Poverty and Deprivation: A Multi-dimensional Perspective. *Statistical Methods & Applications*, **17**(2), 225-250.

Housing Europe The State of Housing in the EU2017. 2017. A Housing Europe Review, Brussels.

ULMAN, P., & ĆWIEK, M. 2021. Measuring housing poverty in Poland: a multidimensional analysis. *Housing Studies*, **36**:8, 1212-1230, DOI: 10.1080/02673037.2020.1759515.

EFFICIENCY AND ROBUSTNESS IN SUPERVISED LEARNING

Anand Vidyashankar[1], Fengnan Deng[1], Giacomo Francisci[1], and Xiaoran Jiang[1]

[1] Department of Statistics, George Mason University, College of Engineering and Computing, Faitfax, VA 22030, (e-mail: avidyash@gmu.edu, fdeng2@gmu.edu, gfranci@gmu.edu, xjiang21@gmu.edu)

ABSTRACT: In recent years, there has been an increasing interest in building machine-learning systems that perform adequately when the training and test data differ. In the context of supervised learning, this problem has been addressed within the distributionally robust framework wherein the ambiguity set for the test distributions is allowed to vary within a neighborhood of the training distribution. While such methods are useful, the tradeoff between statistical efficiency and robustness remains unclear. Focusing on the out-of-distribution generalization problem, in this presentation, we describe a precise notion of statistical efficiency and relate the loss of efficiency to the gain in robustness in these contexts. We illustrate our ideas with examples from label shift estimation arising in diagnostic problems, privacy and utility in healthcare, and generalized adversarial networks.

OPTIMAL AND ROBUST COMBINATION OF FORECASTS VIA CONSTRAINED OPTIMIZATION AND SHRINKAGE

Frédéric Vrins [1]

[1] LIDAM/LFIN – UCLouvain, Belgium.

ABSTRACT: We introduce various methods that combine forecasts using constrained optimization with penalty. A non-negativity constraint is imposed on the weights, and several penalties are considered, taking the form of a divergence from a reference combination scheme. In contrast with most of the existing approaches, our framework performs forecast selection and combination in one step, allowing for potentially sparse combining schemes. Moreover, by exploiting the analogy between forecasts combination and portfolio optimization, we provide the analytical expression of the optimal penalty strength when penalizing with the L2-divergence from the equally-weighted scheme. An extensive simulation study and two empirical applications allow us to investigate the impact of the divergence function, the reference scheme, and the non-negativity constraint on the predictive performance. Our results suggest that the proposed models outperform those considered in previous studies.

KEYWORDS: Combination of forecasts, optimization, shrinkage

DIF ANALYSIS WITH UNKNOWN GROUPS AND ANCHOR ITEMS

Gabriel Wallin[1], Yunxiao Chen[1] and Irini Moustaki[1]

1 Department of Statistics, London School of Economics and Political Science (e-mail: g.a.wallin@lse.ac.uk, Y.Chen186@lse.ac.uk, i.moustaki@lse.ac.uk)

ABSTRACT: Measurement invariance across items is key to the validity of instruments like a survey questionnaire or an educational test.

KEYWORDS: latent classes, measurement invariance, EM algorithm.

Differential item functioning (DIF) analysis is typically conducted to assess measurement invariance at the item level. Traditional DIF analysis methods require knowing the comparison groups (reference and focal groups) and anchor items (a subset of DIF-free items) (see e.g Millsap, 2011). Such prior knowledge may not always be available, and psychometric methods have been proposed for DIF analysis when one piece of information is unknown.

The paper proposes a method for the case when both the anchor items and the groups are unknown. The proposed framework combines ideas of mixture IRT modeling for latent DIF analysis and regularised estimation for manifest DIF analysis with unknown anchor items. More specifically, the unknown groups are modelled by latent classes, and the DIF effects are characterised by item-specific DIF parameters. An L_1-regularised marginal maximum likelihood estimator is proposed, assuming that the number of DIF items is relatively small. This estimator penalises the DIF parameters by a Lasso regularisation term, so that the DIF items can be selected by the non-zero pattern of the estimated DIF parameters. Computing the L_1-regularised estimator involves solving a non-smooth optimisation problem. The proposed method simultaneously identifies the latent classes and the DIF items. A computationally efficient Expectation-Maximisation (EM) algorithm is developed to solve the non-smooth optimisation problem for the regularised estimator.

References

MILLSAP, R. E. 2011. *Statistical Approaches to Measurement Invariance*. Routledge, New York.

CONSTRAINT-BASED ATTRACTOR SEARCH IN BOOLEAN NETWORKS USING QUANTUM COMPUTING

Felix M. Weidner [1] , Mirko Rossini[2] , Joachim Ankerhold[2], and Hans A. Kestler[1]

[1] Institute of Medical Systems Biology, University of Ulm, 89081 Ulm, Germany, (e-mail: felix.weidner@uni-ulm.de, hans.kestler@uni-ulm.de)

[2] Institute of Complex Quantum Systems, University of Ulm, 89081 Ulm, Germany, (e-mail: mirko.rossini@uni-ulm.de, joachim.ankerhold@uni-ulm.de)

ABSTRACT: Quantum information offers a new computing paradigm which may yield further increases in computational resources beyond the limits of miniaturisation for Moore's law. Various quantum algorithms such as Grover's search algorithm also offer improvements in complexity when compared to their classical counterparts. In particular, quantum computing can be applied to the context of Boolean network analysis in systems biology. We demonstrate a new quantum algorithm which uses a modified Grover operator to identify attractor states in the dynamics of Boolean networks. This procedure is based on the iterated addition of constraints for previously identified attractors, thus restraining the size of the remaining state space that has to be searched.

KEYWORDS: quantum computing, quantum information, systems biology, network theory

1 Introduction

Boolean networks (BNs) are simple mathematical dynamic models describing gene regulatory interactions (Kauffman, 1969, Schwab *et al.*, 2020). Network components are represented as expressed (1) or not expressed (0). Logical rules combining network components via the operators AND, OR, and NOT describe the system's interactions. Repeated evaluation of these update rules generates complex dynamics, leading to stable states called attractors. Knowledge of attractors is of interest in the biological context as these states correspond to cellular phenotypes (Huang *et al.*, 2005). The dynamic state space of BNs grows exponentially with the number of components, limiting analyses of larger systems. Motivated by the limits of traditional processors and the search for alternative hardware, we make use of gate-based, universal quantum computers which can perform calculations in an exponentially growing state

space using a linearly increasing number of qubits. Thus, quantum hardware can capture the complexity of the model's dynamics. In our previous work we implemented BNs on quantum computers, highlighting how quantum algorithms can be used to obtain information about network dynamics (Weidner *et al.*, 2023a, Weidner *et al.*, 2023b).

2 Results

Building on this, we now propose a quantum circuit performing a search through the entire state space based on Grover's algorithm (Grover, 1997, Liu & Ouyang, 2013), aiming to identify the entire set of attractors. Grover's algorithm works by iteratively amplifying the weight associated with a specified subset of states, followed by a probabilistic readout of a single possible solution. In contrast, we invert the roles of solution and non-solution states, resulting in a suppression of previously identified attractors. After every readout leading to the detection of a new attractor, this state is then added as a constraint to the quantum circuit in all further runs. In this manner, the search space can be restricted to assign increased weight to states which lead to novel attractors without requiring previous knowledge of the distribution or number of attractors in the system. This also allows for the detection of small attractors which may be difficult to find using a classical random sampling of states. Such attractors may nevertheless be biologically interesting, e.g. in the early detection of rare but drug-resistant phenotypes in a network modeling cancer.

We demonstrate this algorithm on a small biologically motivated Boolean network with attractors of different sizes. We analyse its performance when accounting for the noise present in real quantum processors and quantify the improvement that can be gained from error mitigation techniques. Furthermore, we are investigating the possibility of implementing a constraint-based attractor search using quantum annealers rather than gate-based processors, due to the increased number of qubits available on such devices.

References

GROVER, LOV K. 1997. Quantum Mechanics Helps in Searching for a Needle in a Haystack. *Physical Review Letters*, **79**(2), 325.

HUANG, SUI, EICHLER, GABRIEL, BAR-YAM, YANEER, & INGBER, DONALD E. 2005. Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. *Physical Review Letters*, **94**(12), 128701.

KAUFFMAN, STUART A. 1969. Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets. *Journal of Theoretical Biology*, **22**(3), 437–467.

LIU, YANG, & OUYANG, XIAOPING. 2013. A quantum algorithm that deletes marked states from an arbitrary database. *Chinese Science Bulletin*, **58**(19), 2329–2333.

SCHWAB, JULIAN D, KÜHLWEIN, SILKE D, IKONOMI, NENSI, KÜHL, MICHAEL, & KESTLER, HANS A. 2020. Concepts in Boolean network modeling: What do they all mean? *Computational and Structural Biotechnology Journal*, **18**, 571–582.

WEIDNER, FELIX M, SCHWAB, JULIAN D, WÖLK, SABINE, RUPPRECHT, FELIX, IKONOMI, NENSI, WERLE, SILKE D, HOFFMANN, STEVE, KÜHL, MICHAEL, & KESTLER, HANS A. 2023a. Leveraging quantum computing for dynamic analyses of logical networks in systems biology. *Patterns*, **4**(3), 100705.

WEIDNER, FELIX M, ROSSINI, MIRKO, ANKERHOLD, JOACHIM, & KESTLER, HANS A. 2023b. A protocol for the use of cloud-based quantum computers for logical network analysis of biological systems. *STAR Protocols (accepted)*.

CLUSTERING FOR SPARSELY SAMPLED LONGITUDINAL DATA BASED ON BASIS EXPANSIONS

Michio Yamamoto[1] and Yoshikazu Terada[2]

[1] Graduate School of Human Sciences, Osaka University / RIKEN AIP, (e-mail: yamamoto.michio.hus@osaka-u.ac.jp)

[2] Graduate School of Engineering Science, Osaka University / RIKEN AIP, (e-mail: terada.yoshikazu.es@osaka-u.ac.jp)

ABSTRACT: In longitudinal data, the observations often occur at different time points for each subject. In such a case, ordinary clustering algorithms like K-means clustering cannot be applied directly. Instead, one may apply a smoothing technique to get individual continuous trajectories, followed by finding groups among the trajectories using some clustering algorithm. However, this is inappropriate when each subject's data are observed at only a few time points. Thus, we develop a new clustering algorithm for sparsely sampled longitudinal data, which can be considered a natural extension of the K-means clustering. We show the consistency of the proposed estimator under mild regularity conditions. We also evaluate its performance through simulation studies and data applications.

KEYWORDS: clustering, longitudinal data, functional data analysis

TWO EXTENSIONS OF EXTENDED REDUNDANCY ANALYSIS FOR EXPLORATORY DATA ANALYSIS

Naoto Yamashita[1]

[1] Department of Sociology, Kansai University,
(e-mail: nyam@kansai-u.ac.jp)

ABSTRACT: A multivariate analysis procedure called Extended Redundancy Analysis (ERA) regresses dependent variable(s) on component scores that are determined as weighted sums of independent variables. ERA requires knowledge of group structures of the independent variables, which is often not available in real-world problems. This research proposes a new exploratory variant of ERA called Exploratory ERA (ExERA) is proposed. ExERA does not require the group structure but instead estimates the optimal structure using the dataset. ExERA can also be divided into two different procedures according to their methodology, ExERA-Sp and ExERA-R. ExERA-Sp estimates the group structure of independent variables by sparsely estimating the weight matrix under the constraint that the weight matrix has a perfect cluster structure. ExERA-R approximates a similar structure obtained using ExERA-Sp and obliquely rotating the weight matrix. Numerical simulations and a real data example were used to investigate how well the two approaches performed and to demonstrate the validity of the proposed procedures for exploratory data analysis.

KEYWORDS: Redundancy analysis, oblique rotation, sparse estimation, exploratory data analysis

ULTRAMETRIC GAUSSIAN MIXTURE MODELS WITH PARSIMONIOUS STRUCTURES

Giorgia Zaccaria [1]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, (e-mail: giorgia.zaccaria@unimib.it)

ABSTRACT: Multidimensional phenomena are usually characterized by nested latent dimensions associated, in turn, with observed variables. These phenomena, for instance, poverty, well-being, and sustainable development, can often differ across countries, or cities within countries, in terms of dimensions, other than in their relationships to each other, on the one hand, and their importance in the definition of the general concept, on the other hand. This paper discusses several parsimonious structures of the covariance matrix reconstructing relationships among variables which can be implemented in Gaussian mixture models to study complex phenomena in heterogeneous populations.

KEYWORDS: ultrametricity, Gaussian mixture models, parsimony, hierarchical structures

1 Introduction

Nested latent dimensions associated with observed variables usually characterize multidimensional phenomena. The hierarchical structure underlying them is composed of *specific* and *higher-order* dimensions; therefore, they give rise to a hierarchy of latent concepts, whose root is represented by the general one. These phenomena concern several fields such as economy, sustainability, health, but also differ in their definition across countries. To reconstruct hierarchical relationships among variables in heterogeneous populations, Cavicchia *et al.*, 2022, introduced a Gaussian mixture model with a specific hierarchical structure of the component covariance matrix. The latter corresponds to an extended ultrametric covariance matrix, whose main property is to be one-to-one-associated with a hierarchy of latent concepts. Differently from the mixture of factor analyzers model (McLachlan *et al.*, 2003), where a factorial structure in uncorrelated factors is identified, the methodology proposed by Cavicchia *et al.*, 2022, is able to detect correlated latent concepts, each one associated with a group of observed variables, and to delve deeper into their relationships.

Notwithstanding the general formulation of an extended ultrametric covariance structure is useful to study hierarchies composed of their maximum number of internal nodes, i.e., the number of the specific dimensions and their aggregations in pairs, more parsimonious structures can be considered. In this paper, different configurations of the extended ultrametric covariance structure are discussed, as well as their properties and main features (Section 2). Final considerations conclude the paper in Section 3.

2 Ultrametric Gaussian mixture models: parsimonious structures

Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ be a random sample of size n, where $\boldsymbol{x}_i\,(i = 1, \ldots, n)$ takes value in $\mathcal{R}^p$. Suppose that $\boldsymbol{x}_i$ follows a finite mixture of G Gaussian distributions, whose pdf is given by

$$f(\boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g \phi\big(\boldsymbol{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\big), \tag{1}$$

where $\pi_1, \ldots, \pi_G$ are positive weights (mixing proportions of the mixture) such that $\sum_{g=1}^{G} \pi_g = 1$, $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, $g = 1, \ldots, G$, are the mean vectors and the component covariance matrices of the multivariate Gaussian distributions $\phi(\cdot|\cdot)$. In the Ultrametric Gaussian Mixture model, the covariance matrix of the gth component of the mixture is parameterized as

$$\begin{aligned}
\boldsymbol{\Sigma}_g = &\left(\boldsymbol{V}_g \begin{bmatrix} {}_{V_g}\sigma_{11} & 0 & \ldots & 0 \\ 0 & {}_{V_g}\sigma_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & {}_{V_g}\sigma_{QQ} \end{bmatrix} \boldsymbol{V}_g' \right) \odot \boldsymbol{I}_p \\
&+ \left(\boldsymbol{V}_g \begin{bmatrix} {}_{W_g}\sigma_{11} & 0 & \ldots & 0 \\ 0 & {}_{W_g}\sigma_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & {}_{W_g}\sigma_{QQ} \end{bmatrix} \boldsymbol{V}_g' \right) \odot (\mathbf{1}_p \mathbf{1}_p' - \boldsymbol{I}_p) \\
&+ \left(\boldsymbol{V}_g \begin{bmatrix} 0 & {}_{B_g}\sigma_{12} & \ldots & {}_{B_g}\sigma_{1Q} \\ {}_{B_g}\sigma_{12} & 0 & \ldots & {}_{B_g}\sigma_{2Q} \\ \ldots & \ldots & \ldots & \ldots \\ {}_{B_g}\sigma_{1Q} & {}_{B_g}\sigma_{2Q} & \ldots & {}_{B_g}\sigma_{QQ} \end{bmatrix} \boldsymbol{V}_g' \right).
\end{aligned} \tag{2}$$

Each addend of Eq. (2) depends on the matrix $\boldsymbol{V}_g$, which represents the membership matrix determining the partition of the variable space into $Q < p$ groups, and on one of the three parameters characterizing the variable groups.

The first addend corresponds to the diagonal elements of $\boldsymbol{\Sigma}_g$, where ${}_{V_g}\sigma_{11}$, $\ldots, {}_{V_g}\sigma_{QQ}$ are the variances of the Q groups in $\boldsymbol{V}_g$; whereas, the off-diagonal elements of $\boldsymbol{\Sigma}_g$ are defined by ${}_{W_g}\sigma_{qq}$ and ${}_{B_g}\sigma_{qh}$, $q, h = 1, \ldots, Q, h \neq q$, in the second and third addend of Eq. (2), respectively. The latter represent the covariances within and between the Q groups. Specific constraints on these parameters let the extend ultrametric covariance matrix in Eq. (2) be one-to-one associated with a hierarchy. Specifically, an ordering exists among ${}_{V_g}\sigma_{qq}, {}_{W_g}\sigma_{qq}$ and ${}_{B_g}\sigma_{qh}$ so that the group variance is greater than the covariance within the group, which in turn is not lower than the maximum covariance between the groups.

Even if suitable in different situations to represent the hierarchical relationships among variables, the parameterization in Eq. (2) can be further constrained to obtain more parsimonious structures. By setting the membership matrix $\boldsymbol{V}_g$ to be the same across mixture components, the other three sets of parameters can be fixed or left free to vary across them. Therefore, the latter structures pinpoint specific dimensions that are equal across the subpopulations of the mixture while their aggregations, thus higher-order dimensions, can differ across them. We can delve into an example of these hierarchical configurations by considering well-being. OECD identifies eleven key dimensions for measuring it throughout the countries*. Nonetheless, despite sharing the same specific dimensions, the characterization of this complex phenomenon can vary across countries. For instance, the education level is more associated with the possibility of having a better job in less developed economies and more related to a higher civic engagement in more developed economies.

In both cases in which the specific dimensions are equal or not across components, they can be aggregated altogether at the same level, i.e., a unique value occurs in the matrix of the covariances between groups. This structure gives rise to a second-order hierarchy, studied by Cavicchia & Vichi, 2022, in the factor analysis framework. An interesting case that arises from this configuration corresponds to a formative model (Bollen, 2001), where the unique value ${}_B\sigma$ – depending or not on g – equals zero. Indeed, in this hierarchical structure, the specific dimensions result to be uncorrelated and, thus, formed the general concept as unique and not interchangeable part of it. Several examples of formative concepts exist in the literature, such as human development, which is measured by three specific dimensions, i.e., long and healthy life, education, and decent standards of living, usually uncorrelated to each other.

*https://www.oecd.org/wise/measuring-well-being-and-progress.htm

3 Conclusions

When studying multidimensional phenomena, the hierarchical structures of latent dimensions underlying them have to be analyzed to build an index for their measurement. To this aim, Cavicchia *et al.*, 2022 proposed an ultrametric Gaussian mixture model which is able to delve into hierarchical relationships among latent dimensions, on the one hand, and to study different characterization of concepts in heterogeneous populations, on the other hand. In this paper, several parsimonious structures of the component covariance matrices are discussed together with the analysis of their corresponding hierarchies.

References

BOLLEN, K. A. 2001. Indicator: Methodology. *Pages 7282–7287 of:* SMELSER, N. J., & BALTES, P. B. (eds), *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier Science, Oxford.

CAVICCHIA, C., & VICHI, M. 2022. Second-order disjoint factor analysis. *Psychometrika*, **87**, 289–309.

CAVICCHIA, C., VICHI, M., & ZACCARIA, G. 2022. Gaussian Mixture Model with an extended ultrametric covariance structure. *Advances in Data Analysis and Classification*, **16**, 399–427.

MCLACHLAN, G. J., PEEL, D., & BEAN, R. 2003. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, **41**(3), 379–388.

USING RETAIL TRANSACTIONS FOR CONSUMER PRICE INDEX AND EXPENDITURE STATISTICS

Li-Chun Zhang [1] [2]

[1] Statistisk sentralbyrå, Norway

[2] University of Southampton, (e-mail: L.Zhang@soton.ac.uk)

ABSTRACT: Scanner data arising from retail transactions have replaced survey of food price observations for the consumer price index (CPI) for more than a decade. The same data source can provide the expenditure weights needed for the CPI as well, when combined with population data using secure linkage and processing techniques that protect confidentiality. This would alleviate the most burdensome part of diary collection for the Consumer Expenditure Survey that collects expenditure data from households. Due to the sheer amount of transactions, automatic classification of the consumption subclasses of the goods requires natural language processing techniques, as long as there does not exist a catalogue that covers all the goods. Statistical theories pertaining to these big-data expenditure weights and classification are discussed.

KEYWORDS: audit sampling inference, evaluation coverage, entity resolution, maximum entropy classification, entity forest.

1 Big-data proxy expenditure weights

In some countries, scanner data arising from retail transactions have replaced survey of food price observations for the consumer price index (CPI) for more than a decade. The data are typically available on a weekly basis, in the form of unit value (average) price for each consumption goods or *item*. Scanner data constitute a promising source of price data for CPI, which is being expanded to other consumption subclasses such as clothes, electronics. For the price index methodology based on scanner data, we refer to the website of Ottawa Group.

Provided one can connect the transaction items of different consumer subpopulations, it is possible to calculate *proxy* CPI expenditure weights for any specific subpopulation. Unlike the survey-based weights, these proxy weights can be considered to have virtually zero sampling variance for practical purposes because of the sheer amount of data that can be made available. But they are generally biased due to a number of errors that are unavoidable in reality. In particular, these include coverage errors caused by the discrepancy between

the available transactions and the entire consumption of the population, and selection errors from the available transactions because, for various technical reasons, one is not able to code and classify all the items.

In such a situation, where bias completely dominates variance, modelling the intrinsic variability of the proxy weights would be fruitless, as long as it cannot capture the bias. Additional observations of expenditure are necessary to investigate the extent to which the proxy weights may be biased. Zhang (2021) propose and develop audit sampling inference for big-data statistics, which consists of the following elements:

I. clarify the *validity condition* for unbiased big-data statistics,
II. derive *tests* for the unbiasedness of big-data statistics,
III. *measure* the accuracy of the big-data statistics.

The theory of audit sampling inference is applied to the Norwegian data in following setup. First, fully anonymised food expenditure data are obtained for a single weekday in September of 2016, based on extractions provided by the largest debit card payment service and some of the largest supermarket chains. The proxy expenditure weights are calculated from 0.8 million transactions, broken down to four groups according to the age of the cardholder.

Next, take the Consumer Expenditure Survey 2012 as the audit sample, where the survey-based expenditure weights are treated as unbiased estimates of the true CPI food weights for 2012. Setting aside the coverage and selection errors of the available transaction data with respect to all household purchases, the proxy CPI weights do not refer to exactly the same subpopulations as those identified in the survey, because the transaction data refer to a different time point and the proxy weights are broken down by the age of the cardholder instead of the age of the household head. In other words, these proxy weights are *necessarily* biased for the true CPI weights in 2012.

Applying the tests developed for this setup, one is unable to reject the null hypotheses that the proxy-weights CPI for the different subpopulations are unbiased, despite the high power of the tests. However, since one can be certain that the proxy weights are not exactly equal to the true weights, it is sensible to treat these non-rejection results as indications for the usefulness of the resulting big-data CPI, and accuracy measures are still necessary.

Mean squared error (MSE) is a common choice of accuracy measure where bias is known to exist. However, MSE estimation can easily produce negative (hence unusable) results, where the audit sampling variance is large compared to the bias of the big-data statistic. It is unattractive to simply increase the

audit sample size in such situations, which means audit sampling would be more costly in a relatively favourable setting for adopting big-data statistics.

Zhang (2021) proposes and develops *evaluation coverage* as a novel accuracy measure for any big-data statistic, which is generally applicable based on audit sampling and overcomes the problem of limited audit sample size. Whereas the estimation of MSE runs into troubles in the said application, the evaluation coverage provides meaningful results. Indeed, to reach the same evaluation coverage of the proxy-weights CPI, the survey sample size would need to be increased approximately by a factor of 80 in some cases, which is unrealistic in practice.

In short, by the proposed approach of audit sampling inference, one can conclude from the study that proxy CPI weights derived from the transaction data can replace the relevant diary component that is the most burdensome part of the traditional expenditure survey.

2 Classification based on text

The consumption items are classified into subclasses called COICOP groups. Automatic classification of COICOP groups of the large amount of transaction items requires natural language processing techniques, as long as there does not exist a COICOP-catalogue that covers all the items.

Denote by $i = 1, ..., N$ the *items* to be classified. Let $U = \{1, ..., N\}$. Denote by $y = 1, ..., K$ the *groups* to which items are classified. Let $\Gamma = \{1, ..., K\}$. Denote by x any *term* that can be used in item description, e.g. jasminris, toalettpapir, etc. Denote by $\boldsymbol{x}$ the collection of terms in *item description*, possibly in vector-representation , e.g. $\boldsymbol{x} = \{\text{coop, jasminris}\} = (1, 1, 0, 0, ..., 0)_{1\times p}$. Denote by Ω the *corpus* of item description, i.e. $\Omega = \{\boldsymbol{x}_i : i \in U\}$.

For each $i \in U$, let y_i be its group classification. Group classification can be viewed as an entity resolution problem, where Γ are the (known) entities and U the records. The records U_y, $U_y = \{i \in U : y_i = y\}$, are considered to be matched (to each other) via co-reference to the entity y. The *resolution* we seek is the partition $U = \bigcup_{y\in\Gamma} U_y$, denoted by $\mathbb{C} = \{U_y : y \in \Gamma\}$.

One can distinguish generally the *discriminative* or *generative* machine learning approach to classification or entity resolution problems. By the discriminative approach, classification of y_i for any $i \in U$ is based on

$$f(y|\boldsymbol{x}; \Omega) = \Pr(y_i = y \mid \boldsymbol{x}_i = \boldsymbol{x}; \Omega)$$

where the different terms in an item description $\boldsymbol{x}$ are used as distinct features

for $f(y|\boldsymbol{x};\Omega)$. Let $f_U(y|\boldsymbol{x};\Omega)$ be the model function given the corpus Ω and the true resolution $\{U_y : y \in \Gamma\}$. As long as there exists any term x, e.g. $x =$ ekstra, which appears in multiple item descriptions not all belonging to the same group, classification of any $\boldsymbol{x}_i$ that contains this x-term may be incorrect by the *discriminative classifier*

$$y_i = \arg\max_{y\in\Gamma} f_U(y|\boldsymbol{x}_i;\Omega)$$

By the *generative approach*, one would focus on the model function

$$f(\boldsymbol{x}|y;\Omega) = \Pr(\boldsymbol{x}_i = \boldsymbol{x} \mid y_i = y;\Omega)$$

Let $f_U(\boldsymbol{x}|y;\Omega)$ be the model function given the corpus Ω and the true resolution $\{U_y : y \in \Gamma\}$. The corpus Ω is *free of entity-duplication* provided, for any $\boldsymbol{x} \in \Omega$,

$$\sum_{y\in\Gamma} \mathbb{I}\big(f(\boldsymbol{x}|y;\Omega) > 0\big) \equiv 1$$

This *admissibility* condition is in fact necessary for any well-defined mapping from the item descriptions Ω to the groups Γ. Notice that it allows for multiple items with the same $\boldsymbol{x}$ as long as they all belong to the same group. Given any admissible corpus Ω, classification of any $i \in U$ based on $\boldsymbol{x}_i$ would always be correct by the *generative classifier*

$$y_i = \arg\max_{y\in\Gamma} f_U(\boldsymbol{x}_i|y;\Omega)$$

even if there are terms belonging to item descriptions in different groups.

Thus, perfect classification is conceptually possible and easily achievable only by corpus engineering under the generative approach. We develop a generative approach to item classification based on text descriptions. Entity resolution and maximum entropy classification are adopted as the formal framework. In situations where only a subset of all the items have known classifications, we develop supervised learning of an *entity forest* model and associated classification method (based on item descriptions) for the rest of the items.

References

Zhang, L.-C. (2021). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big-data statistics. *Journal of the Royal Statistical Society, Series A*, **184**, 571-588.

Contributed Papers

PROPENSITY TOWARDS MASTER'S DEGREE: CHOICES OF NORTHERN STUDENTS AFTER BAS?

Alfonzetti Giuseppe [1], Grassetti Luca[1] and Rizzi Laura[1]

[1] Department of Economics and Statistics, University of Udine, (e-mail: giuseppe.alfonzetti@uniud.it, luca.grassetti@uniud.it, laura.rizzi@uniud.it)

ABSTRACT: The study aims to explore northern students' choices after Bachelor's degree, focusing on which individual and contextual factors affect the likelihood to continue studying at MAs. The study is population-based, and the used dataset is extracted from the Italian Ministry of University's administrative databases. Students' characteristics are used to study the probability of enrolling in a Master's degree by generalized linear mixed models. Model estimation results can be used to predict the probability of continuing the studies for students at first enrolment and update them during their studies. From the university's point of view, this can represent an essential tool for monitoring the students' careers.

KEYWORDS: Enrollment at MAs, GLMM models, Northern students' mobility', Spatial and contextual effects.

1 Introduction and aims

In the last decades, the literature on educational mobility at national (Barriolu-engo & Flisi, 2017) and international (Van Bouwel & Veugelers, 2013) levels has grown in importance. In the Italian context, much interest has been set on the South–to–North flows at first university enrollment (see, for instance, Attanasio *et al.*, 2020). Furthermore, the multi-cycle organization, based on a 3-years first-cycle degree (Bachelor's degree) and a 2-years second-cycle degree (Master's degree), offers the opportunity to examine further aspects of the students' training paths and study the transition between consecutive levels of the academic studies (Mollica & Petrella, 2017). This study aims to disentangle individual and contextual factors' role in the northern Italian students' behavior after the first level qualification. Besides the classical predictors of academic outcomes, particular attention is devoted to aspects of students' paths, trying to answer the following questions: Do the context of origin and the university of bachelor degree affect the choice of transition? Do stayers and movers at first-

level careers show a different propensity to enroll on a master's programme? Which is the trend of Northern students' enrollment in Master programmes?

The study is structured as follows. Section 2 is devoted to the data and methods description. Model results and discussion are reported in Section 3.

2 Data and Methods

The study uses a cohort-based dataset collected using the Italian Ministry of University's administrative databases (Mobysu.it, 2016, update 2022). The analysis regards the cohorts of students who reached the Bachelor's degree in the academic years 2012/13 - 2016/17, which allow observing the Bachelor (BA) to Master (MS) transition. Therefore, we focus on BA students who attended a high school in northern Italian regions, and we exclude students enrolled in medicine, veterinary, or other 5-year courses. Furthermore, students enrolled in health professions and engineering courses are excluded from the analysis because of their extremely low and high enrolling rates, respectively.

The first step in data analysis is based on simple descriptive statistics. The selected database considers students enrolled in 80 universities during the Bachelor's course (mostly in northern universities $\sim 98.2\%$). The number of Northern students enrolled at MS degrees increased by 76.4% in the period, with great heterogeneity across high school regions (from 45% of Trentino Alto Adige (TAA) to 100% of Liguria). Considering the Northern regions of BA, the relative increase in MS students ranges from 58% of TAA to 111% of Veneto. However, this increase in MS students is mainly due to the growth of students entering the university's first-level programmes. As a result, the transition rates from BA to MS decreased in the period. Figure 1 details the transition rates for students who obtained the BA degree in 2016/17, distinguished by individual and contextual factors. Flows between categories visually highlight the rates of MS enrolling students belonging to the two specific categories of adjacent factors. At the same time, the associated white background labels refer to the enrolling rates conditioned on the flow. Blue background labels, instead, report, from top to bottom row: the category name, its proportion to the whole population and the marginal enrolling rate in that specific category. For example, it highlights the effect of the interaction between the field of study and degree mark, showing that scientific and economic-related graduates display higher enrolling rates, 83.7% and 59.7%, respectively, compared to the average in the mark range (100,109], 66.4%. Briefly, the highest rates of transition in the 16/17 cohort are registered for movers males (62.7%), with BA degree in fields "math-bio" or "other" (75.1% and 77.3%, respectively), with

BA degree mark $>$ 109 and a regular duration of studies at BA.

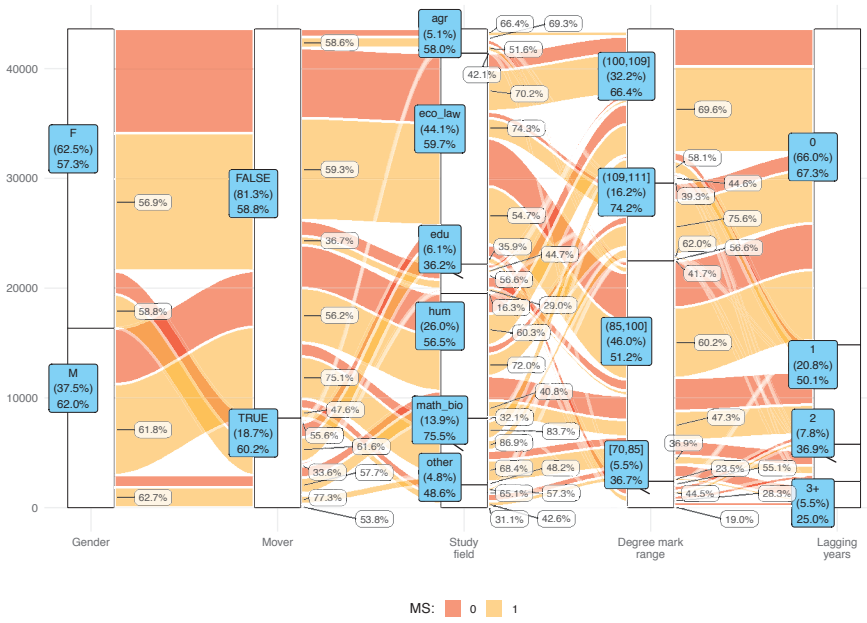


Figure 1. *Rates of transition to MS of students with BA degree in a.a. 2016/17 by Gender, mobility at BA level, the field of study at BA, degree mark at BA, duration of studies at BA.*

To in-depth study this phenomenon, we adopted a model-based approach. In particular, we compare several possible GLMM configurations with random intercept components on a 20-fold cross-validation run on the training data (70% of the available observations). The final model configuration is chosen by monitoring its goodness of fit, via AIC and BIC and its predictive performance, via AUC, on the 20 folds. The models are fitted in `R` with the `glmmTMB` package (Brooks *et al.*, 2017), which allows for the integration of random effects through Laplace approximation.

3 Model results and comments

The chosen model exhibits a cross-classified random intercept, with random components accounting for the high-school municipality and the BA university. Finally, the model is refitted on the whole training partition, obtaining a

predictive AUC on the test sample of the 73%, in line with the cross-validation results. Detailed model estimates are omitted for space reasons.

From an interpretation point of view, the model estimation represents a tool for monitoring the students' careers and predicting their transition behavior. The type of high school attended and the BA study field, as well as their interactions with the corresponding degree marks, emerge as very informative for the MS enrollment choice. Students from scientific fields and students with classic and scientific high school backgrounds are the most likely to further their university education. Furthermore, many individual-specific characteristics play a crucial role in explaining the probability of enrollment, among which the distance from the high-school municipality, as well as the age at graduation and the numbers of years enrolled, play a detrimental role. Finally, the model highlights a significant gender gap, with male students more likely to enrol on an MS. The temporal dimension of the phenomenon enters the model with a set of binary predictors encoding the academic years of reference. The estimates show a sharp drop in the enrollment probability from 2012/2013 to 2014/2015, followed by a softer decrease till 2016/2017.

References

ATTANASIO, M, *et al.* 2020. Chi rimane e chi se ne va? Un'analisi statistica della mobilità universitaria dal Mezzogiorno d'Italia. *In: " Verso Nord. Le nuove e vecchie rotte delle migrazioni universitarie"*. IT.

BARRIOLUENGO, M.S., & FLISI, S. 2017. Student mobility in tertiary education: institutional factors and regional attractiveness. *Publications Office of the European Union, Luxembourg. EUR*, **28867**.

BROOKS, M. E., *et al.* 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, **9**(2), 378–400.

MOBYSU.IT, DATABASE. 2016. *Database Mobysu.it degli studi universitari in Italia.* In Protocollo di ricerca MIUR-Università degli Studi di Cagliari, Palermo, Siena, Torino, Sassari, Firenze e Napoli Federico II. Fonte dei dati: ANS-MIUR/CINECA.

MOLLICA, C., & PETRELLA, L. 2017. Bayesian binary quantile regression for the analysis of Bachelor-to-Master transition. *Journal of Applied Statistics*, **44**(15), 2791–2812.

VAN BOUWEL, L., & VEUGELERS, R. 2013. The determinants of student mobility in Europe: The quality dimension. *European Journal of Higher Education*, **3**(2), 172–190.

CLASSIFYING NORTHERN ITALIAN STUDENTS IN THEIR TRANSITION TO MASTER DEGREE

Alfonzetti Giuseppe [1], Grassetti Luca[1] and Rizzi Laura[1]

[1] Department of Economics and Statistics, University of Udine, (e-mail: giuseppe.alfonzetti@uniud.it, luca.grassetti@uniud.it, laura.rizzi@uniud.it)

ABSTRACT: The university students' behaviour represents a relevant field of study from the management point of view. Given the availability of large administrative data on students' careers, the chance to discover students' profiles in terms of behavioural patterns could be interesting. However, the identification of students' clusters that are informative, feasible and robust at the same time could be complex. The present work aims to define a feasible student clusterisation, adopting an empirical algorithm to treat mixed data and large sample sizes and borrow the syncytial clustering idea developed in the machine learning framework. The proposal is a generalisation of the original algorithm to mixed data cases. Finally, the importance of finding a prototype of students' behaviours is discussed.

KEYWORDS: Two-stage clustering, Hierarchical clustering, Partitioning clustering, Student profiling, Students careers

1 Introduction and aims

A relevant task in education is analysing students' behaviour during their careers. In particular, finding some structured patterns helps defining actions to optimise the supply and organisation of second-level university (Masters) courses and, more in general, of the third-level educational system. Analysing the identified patterns makes it possible to point out both opportunities and shortages in the university education provision.

The availability of individual-level administrative data and their integration with contextual information on the university students (such as the secondary school track) can be considered fundamental for the development of a detailed data mining process able to extract the relevant signals from the humongous set of available data. In particular, the adoption of feasible and robust clustering behavioural has a crucial role in the detection of students' prototype characteristics that can be relevant in providing better services and management.

The dataset considered in the present study comes from a population database regarding Italian university students. We decide to focus on the subset of students from the North of Italy. Even in this restricted framework, the size of the dataset is very large (close to 400,000). Clustering under these settings is not straightforward. The classical hierarchical clustering is unfeasible given the size of the distance matrix, and the partitive solutions (k-means and other machine learning algorithms) are typically difficult to manage. For instance, determining the optimal number of groups or developing a diagnostic for the obtained solution is complex and time-consuming.

In the present work, we propose a solution which inherits the two-stage clustering idea of alternating a hierarchical and a partitive algorithm to reach a more interpretable solution. The typical two-stage clustering algorithm involves a hierarchical clustering first and a partitive approach then, which results unfeasible in large dataset settings. Our approach, instead, connects to the syncytial algorithms outlined in Peterson *et al.* (2018) and Almodóvar-Rivera & Maitra (2020), where the output of a partitive algorithm is used as input for a second agglomerative step.

The main contribution of the present work is the definition of a practical tool for students' prototype recognition based on a syncytial clustering algorithm accommodating mixed data types. The proposal provides enhanced interpretability compared with the classical unsupervised solutions usually adopted in large dataset frameworks.

The structure of the paper is as follows. First, Section 2 details the proposed methodology and presents the data. Then, Section 3 summarizes the results of the empirical analysis.

2 Data and Methods

The characteristics of the Northern Italian students are collected from the Italian Ministry of University's administrative databases (Mobysu.it, 2016, update 2022). In this preliminary analysis, variables considered in the clustering procedures are only some available measures of students' career performance. in particular, the analysis involves a set of dummy variables identifying: Italian students, private secondary schools, and public universities. In addition, factors for the kind of secondary school attended and the gender of students are also included. Finally, some quantitative variables are introduced, including students' age at bachelor's degree, bachelor's course duration, diploma and bachelor's degree marks, years between diploma and bachelor's degree enrolment, and the distance between the first-level university and secondary school

municipalities.

As anticipated, the proposed methodology can be framed as a syncytial clustering algorithm (Peterson *et al.*, 2018; Almodóvar-Rivera & Maitra, 2020). Furthermore, due to the mixed nature of our dataset, where many dummy variables and factors are observed along with some numerical measures, the two steps accommodate algorithms suitable to deal with mixed data. Specifically, the first step implements a k-prototypes clustering (Huang, 1998; Szepannek, 2018), while the second one is a hierarchical clustering procedure based on Gower's distance (Gower, 1971; Maechler *et al.*, 2022). It is worth stressing that the proposed method enjoys easy identification of the optimal clustering solution, along with enhanced interpretability and a robust cluster selection.

3 The empirical analysis

In this section, we analyse a specific clustering solution focusing on students' prototyping and interpret the obtained results.

The optimal number of clusters selected by the procedure is four, as the dendrogram in Figure 1 suggests. Figure 1 also shows the flow of the students through their career characteristics represented for the different clusters. First, Gender is used to describe the population, and then the student patterns are observed over the schooling period. The choice of the kind of school is the first step (for the sake of readability, we reduced the factor to a dummy variable identifying the Liceo secondary school). A fundamental variable affecting the students' career is the diploma mark which is here classified into three levels (≤ 80, $81-90$, and $91+$). The choice of the University is also linked to the opportunity to move from home. The distance variable is a categorized version of the Euclidean distance between the first-level university and secondary school municipalities (0, $1-50$, $51-100$, $101+$ km). The Degree Age is a categorical variable collecting the regular, young, and Late students defined based on their age at bachelor's degree (≤ 22, $23-25$, and $26+$). In the plot, the final collected aspect before the master's degree choice is defined here by Degree Mark, a factor presenting low, mid-low, mid-high, and high levels (≤ 90, $91-100$, $101-110$, *summacumlaude*). All these characteristics flow into the choice of continuing the career or dropping from the university system.

We finally link the students' prototypes to the dropout phenomenon as an example of university outcomes. However, while the clustering approach allows to point out some prototypes of students on the basis of their high-school and first-level university tracks, the identified groups are not connected with

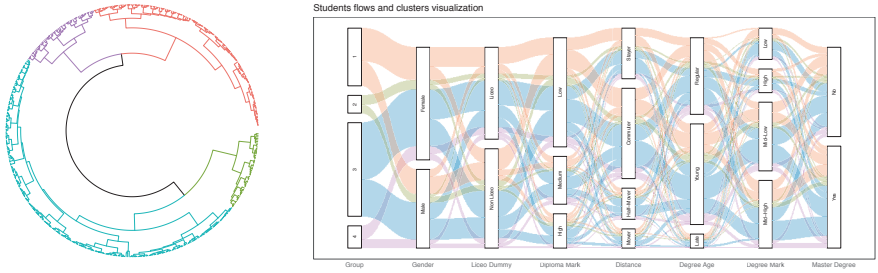


Figure 1. *Behavioural flow of students clustered in the four groups discussed in the text: From birth to master's degree enrolment.*

the propensity to enrol on a master's program. Other factors, such as the field of study, family background and social-economic aspects, may play a relevant role in the choice of transition.

References

ALMODÓVAR-RIVERA, I.A., & MAITRA, R. 2020. Kernel-estimated nonparametric overlap-based syncytial clustering. *The Journal of Machine Learning Research*, **21**(1), 4808–4861.

GOWER, J.C. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

HUANG, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, **2**(3), 283–304.

MAECHLER, M., *et al.* 2022. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.4.

MOBYSU.IT, DATABASE. 2016. *Database Mobysu.it degli studi universitari in Italia.* In Protocollo di ricerca MIUR-Università degli Studi di Cagliari, Palermo, Siena, Torino, Sassari, Firenze e Napoli Federico II. Fonte dei dati: ANS-MIUR/CINECA.

PETERSON, A.D., GHOSH, A.P., & MAITRA, R. 2018. Merging K-means with hierarchical clustering for identifying general-shaped groups. *Stat*, **7**(1), e172.

SZEPANNEK, G. 2018. clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *The R Journal*, **10**(2), 200–208.

CUSTOMER SATISFACTION TROUGH TIME: STRUCTURED TIME SERIES FROM SENTIMENT ANALYSIS OF TRIP ADVISOR DATA

Rosa Arboretti[1], Elena Barzizza[1], Nicolò Biasetton[1], Marta Disegna[1]

[1] Department of Management and Engineering, University of Padova

(e-mail: rosa.arboretti@unipd.it; elena.barzizza@phd.unipd.it; nicolo.biasetton@phd.unipd.it; marta.disegna@unipd.it;)

ABSTRACT:
Nowadays, online reviews (User Generated Content – UGC) written by people on review web platforms, e-commerce website etc. offer the opportunity for business to deeply analyse Customer Satisfaction (CS) with products or services. The availability of such free textual data allows practitioners to study customer behaviour exploiting huge amounts of data. To help processing and analysing such data, sentiment analysis and emotion analysis have been proposed and extensively adopted in literature. Sentiment analysis represents the process of automatic identification and categorization of opinions expressed in a piece of text, especially focusing in determining whether user's attitude towards a particular item (i.e. topic, product or service) is positive, neutral or negative. On the other hand, emotional analysis aims to find the hidden emotions behind texts.
With the aim to provide businesses with insights into trends concerning their product or services, advanced methods, including text mining and sentiment analysis, have been used to transform the unstructured social media data into structured data series. In the recent literature, researchers have devoted efforts to obtain structured time series from texts and images.
After a comprehensive literature review, an application of sentiment analysis for time series data is presented. Through webscraping we extract reviews on Tripadvisor activities concerning movie-set tourism. Being some scenes of "The Lord of the Rings" (LOTR) and "the Hobbit" movies shot in Hinuera, New Zealand, we scraped all comments reviews of LOTR-related activities available there. Linking the sentiment extracted from the reviews to the date in which the reviews have been written allows to obtain customer satisfaction time-series that reflect the trend in the customers' opinion concerning the product/service. Further insights concerning the level of appreciation toward different aspects can be obtained by relating the reviews to the topics they deal with, using LDA topic modelling to extract such topic information for each review. Concluding, the main issues related to sentiment analysis for time series are highlighted offering some suggestions and recommendations for future analysis.
KEYWORDS: sentiment analysis, time series, customer satisfaction, textual reviews, LDA topic modelling

A FLEXIBLE TOPIC MODEL

Ascari Roberto [1] and Giampino Alice [1]

[1] Department of Economics, Management, and Statistics, Piazza dell'Ateneo Nuovo, 1, Milan, University of Milano-Bicocca, Italy, (e-mail: roberto.ascari@unimib.it, a.giampino@campus.unimib.it)

ABSTRACT: In the last two decades, text modeling techniques have been used for various applications, including the analysis of topics in different text documents, where the aim is to provide a document representation in terms of topic distribution. This work aims to show some results on a generalization of the popular latent Dirichlet allocation model, with a particular focus on the clustering of text documents.

KEYWORDS: Dirichlet, latent variable, MCMC, mixture model, textual data.

1 Introduction

Let us consider a collection $\mathcal{C}$ of D text documents, commonly referred to as a "corpus". The d-th document can be thought of as a sequence $(w_{d,1},\ldots,w_{d,N_d})^\intercal$ of N_d words (i.e., $w_{d,n}$ represents the n-th word in the d-th document, $d = 1,\ldots,D$ and $n = 1,\ldots,N_d$). The set $\mathcal{V}$ of the V unique words appearing in the corpus represents a "vocabulary".

Topic modeling techniques assume that each word in a document is generated according to one among T possible topics. As a consequence, the d-th document can be represented through a vector $\boldsymbol{\theta}_d = (\theta_{d,1},\ldots,\theta_{d,T})^\intercal$, where $\theta_{d,t}$ represents the proportion of words in document d generated from topic t. Clearly, $\boldsymbol{\theta}_d$ belongs to the T-part simplex $\mathcal{S}^T = \{\boldsymbol{\theta} : \theta_t > 0, \sum_{t=1}^T \theta_t = 1\}$. Similarly, each topic is represented as a discrete probability distribution $\boldsymbol{\phi}_t$ over the vocabulary $\mathcal{V}$, $t = 1,\ldots,T$, thus $\boldsymbol{\phi}_t \in \mathcal{S}^V$. The most popular topic model is the latent Dirichlet allocation (LDA), introduced by Blei *et al.*, 2003, which supposes both the vectors $\boldsymbol{\theta}_d$ and ϕ_t following a Dirichlet distribution on $\mathcal{S}^T$ and $\mathcal{S}^V$, respectively. Thus,

$$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha}),\ \ \boldsymbol{\alpha} \in \mathbb{R}_+^T \quad \text{and} \quad \boldsymbol{\phi}_t \sim \text{Dir}(\boldsymbol{\beta}),\ \ \boldsymbol{\beta} \in \mathbb{R}_+^V.$$

Despite its popularity, the LDA suffers from the poor parameterization that the Dirichlet deserves for its covariance matrix. Then, the development of a more flexible technique seems to be a relevant issue.

2 The flexible LDA

In this section, we introduce a generalization of the LDA, namely the flexible LDA (FLDA). This model arises by assuming a flexible Dirichlet distribution (FD, Migliorati *et al.*, 2017) for each $\boldsymbol{\theta}_d$. The FD is a (structured) finite mixture model with Dirichlet components:

$$\mathrm{FD}(\boldsymbol{\theta};\boldsymbol{\alpha},\tau,\mathbf{p}) = \sum_{t=1}^{T} p_t \mathrm{Dir}(\boldsymbol{\theta};\boldsymbol{\alpha}+\tau\cdot\mathbf{e}_t),$$

where $\mathbf{p} \in \mathcal{S}^T$, $\tau > 0$, and $\mathbf{e}_t$ is the null vector with the t-th element equal to 1. The additional parameters introduced by the mixture structure of the FD allow for a more flexible modelization of the covariance matrix, thus overcoming some limitations of the Dirichlet. It is noteworthy to mention that the FD includes the Dirichlet distribution as a special case if $\tau = 1$ and $p_t = \alpha_t/\alpha^+$ for $t = 1,\ldots,T$, hence the FLDA model includes the LDA. The FD possesses several statistical properties, among which is the conjugacy to the multinomial scheme. Thus, if $\boldsymbol{\theta}_d \sim \mathrm{FD}(\boldsymbol{\alpha},\tau,\mathbf{p})$, then $\boldsymbol{\theta}_d$ given the corpus (i.e., the observed data) follows an FD distribution with updated parameters $\boldsymbol{\alpha}^*,\tau^*$, and $\mathbf{p}^*$.

To obtain estimates for the FLDA parameters $\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_D\}$ and $\{\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_T\}$, we implement a collapsed Gibbs sampling (CGS), extending the approach proposed by Griffiths & Steyvers, 2004. The main difference with respect to a standard Gibbs sampling is that full conditionals are computed by marginalizing some parameters out. The estimates of the dropped parameters are computed by means of the conjugacy properties. To implement a CGS, we introduce a set of latent (i.e., unobservable) random variables $Z_{d,n}$ representing the topic label of the n-th word in the d-th document, $n = 1,\ldots,N_d$, $d = 1,\ldots,D$.

It is possible to show that the full conditionals, namely the probability that $\{Z_{d,n} = t\}$ (i.e., the word is assigned to topic t) given all the other topic assignments $\mathbf{z}_{-(d,n)}$, take the following form

$$p(Z_{d,n} = t|\mathbf{z}_{-(d,n)},\mathcal{C},\boldsymbol{\alpha},\tau,\mathbf{p},\boldsymbol{\beta}) \propto$$
$$\propto \frac{\left(\alpha_t + c^-_{t,d,\cdot}\right)\left(\beta_{v_{d,n}} + c^-_{t,\cdot,w_{d,n}}\right)}{\left(\beta^+ + c^-_{t,\cdot,\cdot}\right)} \cdot \left\{\sum_{h=1}^{T} p^*_{d,h} + p^*_{d,t}\left(\frac{\tau_t}{\alpha_t + c^-_{t,d,\cdot}}\right)\right\},$$

$t = 1,\ldots,T$, where $p^*_{d,t} = p_t \dfrac{(\alpha_t+\tau)^{[c_{t,d,\cdot}]}}{(\alpha_t)^{[c_{t,d,\cdot}]}}$, $x^{[n]} = x(x+1)\cdot\cdots\cdot(x+n-1)$ denotes the rising factorial function, and $w_{d,n} \in \mathcal{V}$ indicates which term of the

vocabulary is associated with the n-th word in document d. Additionally, we define the quantities $c_{t,d,\cdot}$, $c_{t,\cdot,w}$, and $c_{t,d,\cdot}$ as summation over the proper index of the counts $c_{t,d,v} = \sum_{n=1}^{N_d} \mathbb{I}(z_{d,n} = t,\ w_{d,n} = v)$, the latter representing the number of times that word v is assigned to topic t in document d. Having the full conditionals, the CGS algorithm can be summarized by the following steps:

1. Initialize the vector $\mathbf{z}$ (randomly) and compute the counts $c^{(0)}_{t,d,v}$;
2. For $b = 1, \ldots, B$:
   - For each word in the corpus:
     - sample a new topic $z^{(b)}_{d,n}$ for $w_{d,n}$ from $p(z)$;
     - update the counts $c^{(b)}_{t,d,v}$.
   - Use $\mathbf{z}^{(b)}$ to compute the estimates $\hat{\boldsymbol{\theta}}^{(b)}_d$ and $\hat{\boldsymbol{\phi}}^{(b)}_t$.

By having a sample of size B for the topic labels, namely $\mathbf{z}^{(b)}$, $b = 1, \ldots, B$, and relying on the conjugacy properties, we can estimate $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_t$ as the mean of an FD and Dirichlet distributions with updated parameters, that is

$$\hat{\boldsymbol{\theta}}^{(b)}_d = \frac{\boldsymbol{\alpha} + \mathbf{c}^{(b)}_d + \tau \mathbf{p}^{*(b)}_d / p^{(b)}_+}{\alpha^+ + \tau + N_d} \quad \text{and} \quad \hat{\boldsymbol{\phi}}^{(b)}_t = \frac{\boldsymbol{\beta} + \mathbf{c}^{(b)}_t}{\beta^+ + c^{(b)}_{t,\cdot,\cdot}},$$

where $\mathbf{c}^{(b)}_d = (c^{(b)}_{1,d,\cdot}, \ldots, c^{(b)}_{T,d,\cdot})^\intercal$ and $\mathbf{c}^{(b)}_t = (c^{(b)}_{t,\cdot,1}, \ldots, c^{(b)}_{t,\cdot,V})^\intercal$.

3 Application: The Great Library Heist

During the night, a vandal broke into their professor's study and tore three books into single chapters. The single chapters are not labeled, so the professor is not able to cluster them so to restore the original books. In the following, we consider the $D = 166$ chapters as documents forming the corpus. We will consider $T = 3$ latent topics, each of them hopefully representing one of the destroyed books. Words in the corpus $\mathcal{C}$ compose a vocabulary $\mathcal{V}$ of $V = 16531$ unique terms. We run both the LDA and the FLDA models for $B = 5000$ iterations. Figure 1 displays the topic proportions $\boldsymbol{\theta}_d$ for all the documents, by conditioning on the true topic (i.e., the original book). We can note that both the LDA and FLDA models represent chapters from "Great Expectations" as mainly composed of terms arising from topic 1. The FLDA, thanks to the flexible covariance matrix of the FD, improves the LDA performance by providing more concentrated $\boldsymbol{\theta}_d$'s towards 0 or 1 than the LDA. Similar conclusions hold

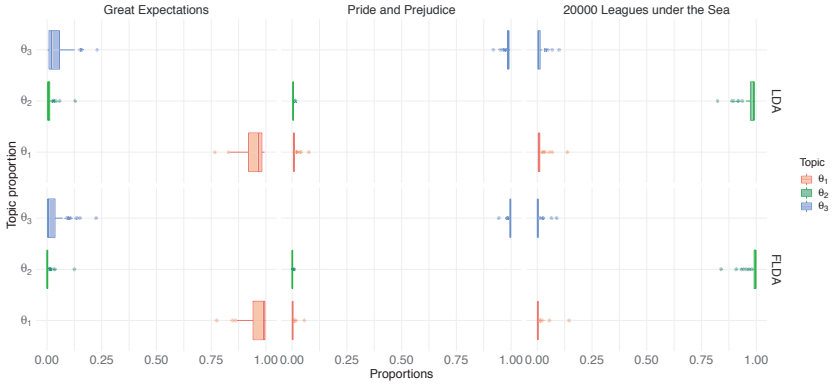


Figure 1. *Boxplots of the elements of* $\boldsymbol{\theta}_d$ *estimates by the LDA (upper panels) and the FLDA (bottom panels) conditioning on the true topic (i.e., the original book).*

Figure 2. *Word clouds representing the 20 most probable words for each topic detected by the FLDA.*

true for chapters from "20000 Leagues Under the Sea" and "Pride and Prejudice", being characterized by high proportions of words from topics 2 and 3, respectively. Topics generated by the FLDA are represented by illustrating the 20 most probable words (Figure 2).

References

BLEI, D.M., NG, A.Y., & JORDAN, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

GRIFFITHS, THOMAS L., & STEYVERS, MARK. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(SUPPL. 1), 5228 – 5235.

MIGLIORATI, S., ONGARO, A., & MONTI, G. S. 2017. A structured Dirichlet mixture model for compositional data: inferential and applicative issues. *Statistics and Computing*, **27**(4), 963–983.

EXPLAINABLE MACHINE LEARNING FOR LENDING DEFAULT CLASSIFICATION

Golnoosh Babaei [1], Paolo Pagnottoni[2] and Thanh Thuy Do[3]

[1] Department of Engineering, University of Pavia, Pavia, Italy, (e-mail: golnoosh.babaei01@universitadipavia.it)

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

[3] Department of Economics, University of Insubria, Varese, Italy, (e-mail: ttdo@studenti.uninsubria.it)

ABSTRACT: Machine Learning (ML) models are often used to support classification decision-making, such as in peer-to-peer lending. However, they usually lack interpretable explanations. While Shapley values and the computationally efficient variant Kernel SHAP may be employed for this aim, the latter makes the assumption that the features are independent. We explain classifiers through a Kernel SHAP method able to handle dependent features in the context of credit risk management for peer-to-peer lending. We demonstrate the effectiveness of our method by considering linear and non-linear models with varying degrees of feature dependence, showing that our approach yields credible estimates of true Shapley values across model and dependence specifications.

KEYWORDS: feature dependence; Shapley values; machine learning; explainability.

1 Introduction

Obermeyer & Emanuel, 2016 pointed out that ML model interpretability enhances medical, healthcare, credit scoring, and fraud detection. Explaining complex ML model predictions is a challenging task, and the model's explanation is crucial for both reliability of the estimates and for fairness and compliance with respect to General Data Protection Regulation compliance. Peer-to-peer lending requires creditworthiness, namely transparent and trustworthy explanations to build trust and help lenders and borrowers make well-informed choices. Credit risk analysis determines peer-to-peer lending rates and creditworthiness, and lenders may distrust complicated ML model predictions. Explainable Artificial Intelligence (XAI) improves classification accuracy, model transparency and interpretability via the concept of game-theoretic

Shapley values. Recent model-agnostic explanation methods simplify understanding of how each predictor affects the prediction; in particular, Aas *et al.*, 2021 expand Kernel SHAP to address interdependent characteristics. We exploit such formulation of Kernel SHAP to build predictive classification ML models and relative model explanations for interpretable peer-to-peer credit risk management. We test our proposal on three predictive ML models, i.e. logistic regression, GAMs, XGBoost, and four structures for modelling feature dependence, i.e. the independent case, Gaussian, empirical distribution and copula. This study reveals that linear and non-linear models with variable feature dependencies give consistent and reliable Shapley value estimates. This enhances the understanding of the drivers of peer-to-peer lending credit risk and outlines best practices for its management via machine learning classification techniques.

2 Kernel SHAP for dependent features

Kernel SHAP computes feature importance using weighted linear regression and local linear regression coefficients. In classical machine learning, a predictive model, $f(x)$, is trained using a training set of size n_{train} comprised of sets y $\left\{y^i, x^i\right\}_{i=1,\ldots,n_{train}}$ where $\quad j = 1,\ldots,n_{train}$. This model attempts to closely approximate the response value y. To explain the prediction $f(x^*)$ for a particular feature vector $x = x^*$, the Kernel SHAP technique only uses the independence assumption $p(x_{\bar{S}} \mid x_S) = p(x_{\bar{S}})$ - see Aas *et al.*, 2021.

We examine how the three different ways of accounting for dependence structures in the features increase ML credit risk model accuracy and feature explainability compared to independence.

2.1 Multivariate Gaussian distribution

Given that the feature vector x is obtained from a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ, then the conditional distribution $p\left(x_S \mid x_S = x_S^*\right)$ is also multivariate Gaussian. By expressing $p(x)$ in terms of $p(x) = p(x_S, x_S) = \mathrm{N}_M(\mu, \Sigma)$ with $\mu = (\mu_S, \mu_S)^\top$ and

$$\Sigma = \begin{bmatrix} \Sigma_{SS} & \Sigma_{S\bar{S}} \\ \Sigma_{\bar{S}S} & \Sigma_{\bar{S}\bar{S}} \end{bmatrix}$$

gives $p\left(x_S|_S = x_S^*\right) = \mathrm{N}_{|\bar{S}}\left(\mu_{\bar{S}|S}, \Sigma_{\bar{S}|S}\right)$, with

$$\mu_{\bar{S}|S} = \mu_{\bar{S}} + \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\left(x_S^* - \mu_S\right)$$

and

$$\Sigma_{\bar{S}|S} = \Sigma_{\bar{S}\bar{S}} - \Sigma_{\bar{S}S}\Sigma_{SS}^{-1}\Sigma_{S\bar{S}}$$

2.2 Gaussian Copula

A d-dimensional copula is a multivariate distribution, C, characterized by uniformly distributed marginal probabilities $\mathrm{U}(0,1)$ over the unit interval of $[0,1]$. Sklar's theorem states that for each multivariate distribution F with univariate distributions $F_1, F_2, \ldots, F_d$ can be written as

$$F(x_1, \ldots, x_d) = C(F_1(x_1), F_2(x_2), \ldots, F_d(x_d)),$$

for some appropriate d-dimensional copula C. In fact, the copula from (12) has the expression

$$C(u_1, \ldots, u_d) = F\left(F_1^{-1}(u_1), F_2^{-1}(u_2), \ldots, F_d^{-1}(u_d)\right)$$

where the F_j^{-1} s are the inverse distribution functions of the marginals. Assuming a Gaussian copula, the following methodology can be employed to generate samples from $p\left(x_S \mid x_S = x_S^*\right)$.

2.3 Empirical conditional distribution

We propose a non-parametric method if x's dependence structure and marginal distributions depart from the Gaussian. The kernel estimator, a classical non-parametric density estimation method, has been modified and improved over the decades. The kernel estimator is impeded by the curse of dimensionality, which rapidly restricts its applicability in multivariate problems. Additionally, the non-parametric estimation of conditional densities is limited to a small number of techniques, particularly when either x_S or $x_{\overline{S}}$ is not one-dimensional. Ultimately, most kernel estimation methods generate a non-parametric density estimate, however, samples from the estimated distribution must be produced. Consequently, we have formulated an empirical conditional method to approximately sample from $p\left(x_{\overline{S}} \mid x_S^*\right)$.

3 Empirical Findings

We compare accuracy and prediction explanations from different ML models and feature dependence settings on four predictive models using the suggested technique. Logistic regression and three more complex predictive models—GAMs, RF, and XGBoost—are chosen. Lending Club (LC) has 2260701

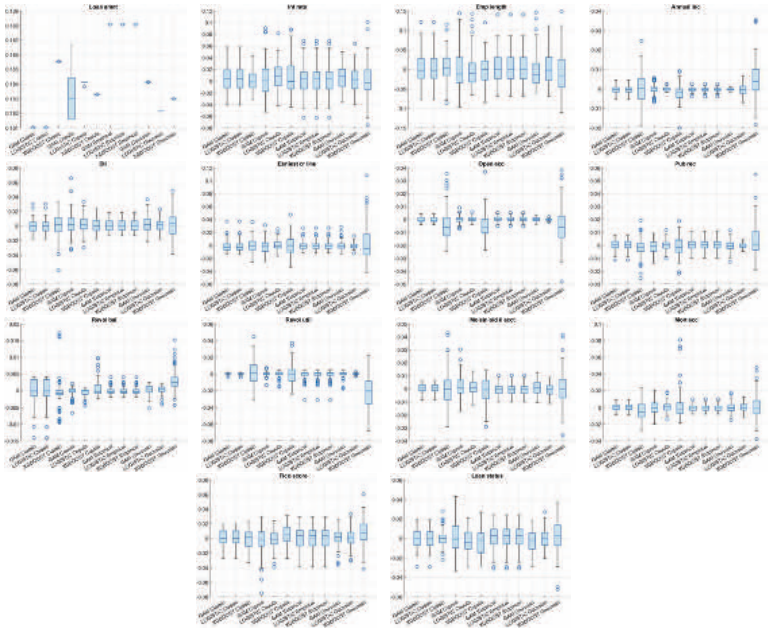


Figure 1. *Distribution of Shapley values from random subsampling for each variable, model and feature dependence structure.*

observations on individual borrowers and their requested loans from 2007 to the fourth quarter of 2018. In this study, we preprocess data and keep 14 variables to analyze the impact of dependencies on the explanations produced by the different ML models. We perform test data random sub-sampling, which provides Shapley values for each of the $n = 100$ iterations. Results are contained in Figure 1. The figure shows that Shapley value estimates are very consistent across model specifications, and that loan amount is the variable fostering the discriminatory power of all the classification models employed.

References

AAS, KJERSTI, JULLUM, MARTIN, & LØLAND, ANDERS. 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, **298**, 103502.

OBERMEYER, ZIAD, & EMANUEL, EZEKIEL J. 2016. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, **375**(13), 1216.

A MULTIVARIATE PERMUTATION TEST FOR ASSOCIATION

Barzizza E.[1], Ceccato R. [1], Harrar S. [2], Pesarin F.[3] and Salmaso L.[1]

[1] Department of Management Engineering, University of Padova, (e-mail: elena.barzizza@phd.unipd.it, riccardo.ceccato.1@unipd.it, luigi.salmaso@unipd.it)

[2] Department of Statistics, University of Kentucky, (e-mail: solomon.harrar@uky.edu)

[3] Department of Statistical Sciences, University of Padova, (e-mail: pesarin@stat.unipd.it)

ABSTRACT: Testing the association between multivariate responses and predictors is an important problem in statistics. A huge number of parametric and non-parametric solutions have been provided in the literature for the univariate case while for multivariate settings only few solutions were introduced. There was a recent attempt for non-parametric test based on high dimensional MANOVA but for the case of a single predictor. In this study we propose a generalization to a multiple predictor situation via the NonParametric combination methodology (NPC), a permutation-based solution which needs only very mild assumptions and affords substantial flexibility in the choice of the test statistic. We evaluate and compare the performances of the proposed procedures in a simulation study.

A COMPETING RISK ANALYSIS OF ACADEMIC CAREERS WITH STUDENTS' ABILITY AND SPEED AS PREDICTORS

Michela Battauz [1]

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

ABSTRACT: A competing risk model in discrete time is employed to analyze the outcomes of students' academic careers, which are degree attainment, drop out or transfer to another course. As covariates, besides using the variables available from the administrative database, we consider also the performance of the students in terms of ability and speed, which are predicted from an IRT model applied to the grades obtained in the exams. An application shows that these variables are good predictors of the outcomes.

KEYWORDS: academic performance, latent variables, survival analysis.

1 Introduction

Competing risks models in discrete time (Tutz & Schmid, 2016) are particularly suitable for the analysis of the students' careers at university since they consider all the possible events that can occur in time (see Scott & Kennedy, 2005, Clerici *et al.* , 2015). Such events are degree attainment, drop out or transfer to another course. The novelty of our proposal is given by the predictors included in the model. In fact, we use an Item Response Theory (IRT) model for the grades obtained by the university students and the time needed to pass the exams that accounts for two sources of censoring: dropout and lack of grades for non-passed exams. Using this model we predict two latent variables for each student, which can be interpreted as ability and speed. More details on this model can be found in Battauz (2023). They are then used as predictors in the competing risk model together with other observed covariates.

2 Models and Methods

To extend the analysis of time-to-event data from the case of one possible event, as usually done in survival analysis, to the case of multiple events, it is

necessary to define a hazard function for each target event

$$\lambda_r(t|\mathbf{x}) = P(T = t, R = r|T \geq t, \mathbf{x}), \tag{1}$$

where $R \in 1,\ldots,m$ denotes the event, $T \in 1,\ldots,t_{max}$ the time, and $\mathbf{x}$ a set of covariates. It is possible to show that the survival function is given by

$$S(t|\mathbf{x}) = P(T \geq t|\mathbf{x}) = \prod_{i=1}^{t}(1 - \lambda(i|\mathbf{x})) \tag{2}$$

and the event probability results

$$P(T = t, R = r|\mathbf{x}) = \lambda_r(t|\mathbf{x})S(t-1|\mathbf{x}). \tag{3}$$

In discrete time, the hazard function is frequently modelled using the multinomial model

$$\lambda_r(t|\mathbf{x}) = \frac{\exp(\beta_{0tr} + \mathbf{x}^\top \beta_r)}{1 + \sum_{i=1}^{m} \exp(\beta_{0ti} + \mathbf{x}^\top \beta_i)}. \tag{4}$$

Once the parameters have been estimated, it is possible to compute the event probabilities for a vector of covariate values $\mathbf{x}$ by means of Equations (2) and (3). The estimation was performed by maximum likelihood.

3 Application

The model was applied to the 2017 cohort of students enrolled in Business or Economics at the University of Udine, composed of 353 people at baseline. These two bachelor's degrees share many courses, especially in the first and second year, thus permitting us to fit the IRT model with censoring to the grades obtained by these students together. From the IRT model we predict two latent variables for each student: the ability (θ) and the speed (τ). The predicted values are then standardized and used as covariates in the competing risk model together with other variables available from the administrative database, which are age at enrolment, grade obtained from high school, type of high school, gender and residence. Table 1 reports the estimates of the coefficients of the variables selected on the basis of statistical significance. Hence, the only variable still significant when θ and τ are included in the model is age, with younger students having higher probabilities of attaining the degree and lower probabilities of dropping out. Both ability and speed have a positive effect on the probability of attaining the degree and also a positive joint effect. Figure 1 shows the cumulative predicted probabilities for different values

of θ and τ. The age was fixed at 19, the most common value. It is apparent the important effect that these variables have on the outcome. Students with moderately high values of both ability and speed present a high probability of attaining the degree. Such probability results definitely lower for students with low values of ability and moderately high speed, while having low values of speed and moderately high ability affects the time of degree attainment with a less severe impact on the probability of obtaining this outcome. Finally, the case of low levels on both the latent variables determines very low probabilities of attaining the degree.

Table 1. *Estimates of the coefficients of the model (coef.) and their standard errors (s.e.).*

	Degree		Drop Out		Transfer	
Variable	coef.	s.e.	coef.	s.e.	coef.	s.e.
time 1	-26.13	0.00	-1.79	0.21	-1.28	0.20
time 2	-9.93	1.45	-3.91	0.38	-3.48	0.41
time 3	-0.35	0.24	-3.37	0.39	-4.74	1.02
time 4	0.71	0.34	-2.94	0.43	-13.01	86.69
time 5	-0.80	0.74	-1.30	0.42	-13.00	133.39
θ	1.87	0.32	-1.77	0.24	-0.52	0.23
τ	1.84	0.27	-0.63	0.19	-0.27	0.23
age at enrollment - 19	-0.20	0.12	0.06	0.03	-0.00	0.04
$\theta \times \tau$	0.55	0.39	-0.62	0.20	-0.15	0.22

4 Conclusion and ongoing work

This paper shows that students' ability and speed, measured through the grades obtained in the exams, play an important role in their career. These two variables are predicted using an IRT model with censoring, and hence the predicted values can be considered as measures with error of the latent variables. Since error-in-variables introduces bias in parameter estimation, in future research, we aim at jointly modelling the grades and the events of students' career so that the measurement error is properly taken into account in the model.

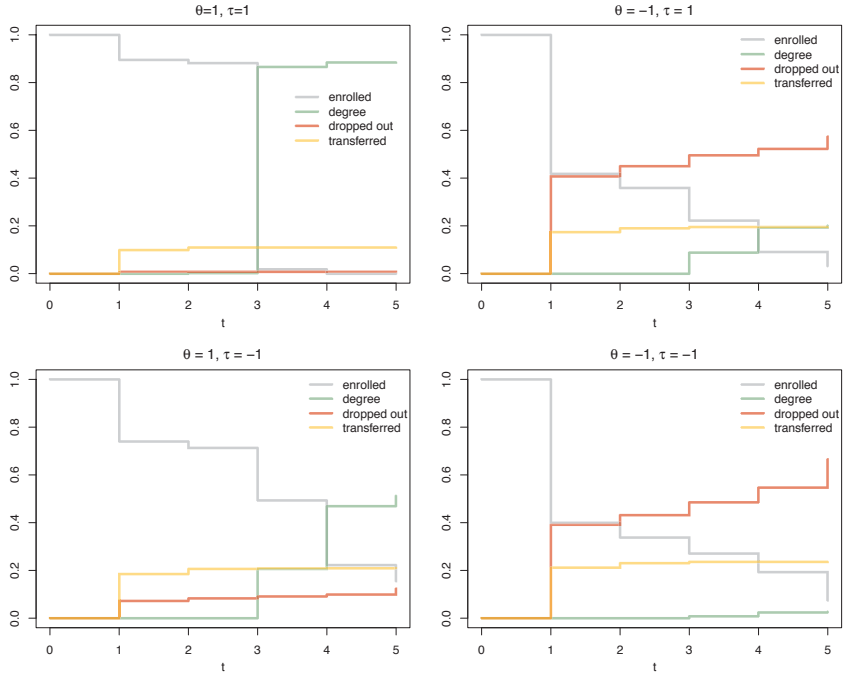


Figure 1. *Cumulative event probabilities obtained from the model for different values of* θ *and* τ.

References

BATTAUZ, M. 2023. Analysis of university grades: An IRT model for responses and response times with censoring. *Submitted to the Conference SIS 2023 - Statistical Learning, Sustainability and Impact Evaluation*.

CLERICI, R., GIRALDO, A., & MEGGIOLARO, S. 2015. The determinants of academic outcomes in a competing risks approach: Evidence from Italy. *Studies in Higher Education*, **40**, 1535–1549.

SCOTT, M. A., & KENNEDY, B. B. 2005. Pitfalls in Pathways: Some Perspectives on Competing Risks Event History Analysis in Education Research. *Journal of Educational and Behavioral Statistics*, **30**, 413–442.

TUTZ, G., & SCHMID, M. 2016. *Modeling Discrete Time-to-Event Data*. New York: Springer.

BAYESIAN ANALYSIS FOR A GRAPHICAL T-MODEL

A. Bekker[1,3], J.T. Ferreira[1,3], J. Pillay[1,3] , and M. Arashi [2]

[1] Department of Statistics, University of Pretoria, South Africa, (e-mails: andriette.bekker@up.ac.za, u18067434@tuks.co.za johan.ferreira@up.ac.za)

[2] Department of Statistics, Ferdowsi University of Mashhad, Iran (e-mail: arashi@um.ac.ir)

[3] Centre of Excellence in Mathematical and Statistical Sciences, Johannesburg, South Africa

ABSTRACT: Modelling noisy data in a network context remains an unavoidable obstacle; fortunately, random matrix theory may comprehensively describe network environments effectively. Thus it necessitates the probabilistic characterisation of these networks (and accompanying noisy data) using matrix variate models. Denoising network data using a Bayes approach is not common in surveyed literature. Thus we briefly introduce a new matrix-variate t model in a prior sense for the noise process following the Gaussian graphical network, for the cases when the assumption of normality is violated in the model and cases when Gaussian distributions is no longer sufficient to explain variation in the data. We investigate the performance of this matrix-variate t distribution applied to a network setting within a Bayesian context. Calculation and approximation of the resulting posterior are of interest to assess the considered model's network centrality measures, which is illustrated using real-life stock price data.

KEYWORDS: adjacency matrix, Bayesian estimation, Gaussian graphical model, matrix-variate t, stock price data.

1 Introduction

Let G_t be a sequence of directed networks for $t = 1, \ldots, T$ for $T \in \mathbb{N}$. Assume that the number of nodes do not change with respect to t, but the number of edges can. Assume that each of the nodes bears a stationary time series of variables that estimates a sequence of networks G_t at time t. Then an adjacency matrix is estimated for G_t at each time index t, say $\boldsymbol{Y}_t$. A stationary time series implies that network structure itself at time t is nothing more than a deviation from an underlying adjacency matrix $\boldsymbol{B}$ independent of time t. In other words, the true graphical network structure is stationary. $\boldsymbol{Y}_t$ is thus viewed as 'noisy

copy' of $\boldsymbol{B}$ given by:

$$\boldsymbol{Y}_t = \boldsymbol{B} + \boldsymbol{E}_t \quad \text{for } t = 1, \ldots, T. \tag{1}$$

$\boldsymbol{E}_t : n \times n$ is a random error term, independent and identically distributed for all $t = 1, \ldots, T$. The matrix-variate Gaussian distribution is fundamental for inference, but is sometimes inadequate for modelling populations where the matrix variate-t distribution may be a better fit. There is extensive literature around a multivariate Gaussian distribution of errors. Articles that date back as early as the classical linear models (Arnold, 1979) to relatively recent ones on engineering processes (Amiri *et al.*, 2018), with recent contributions including the work by Billio *et al.*, 2021. Instead, a t distribution seems a suitable choice to characterise error. Thus, consider $\boldsymbol{E}_t$ as matrix-variate t distributed with corresponding probability density function (pdf), then $\boldsymbol{E}_t \sim t_{n,n}(\boldsymbol{0}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ and the pdf of $\boldsymbol{E}_t$ is given by,

$$f(\boldsymbol{E}_t|\nu, \boldsymbol{0}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \frac{\Gamma_n(\frac{\nu+2n-1}{2})}{\pi^{(\frac{n^2}{2})}\Gamma_n(\frac{\nu+n-1}{2})} |\boldsymbol{\Sigma}_1|^{-\frac{n}{2}} |\boldsymbol{\Sigma}_2|^{-\frac{n}{2}} |\boldsymbol{I}_n + \boldsymbol{\Sigma}_1^{-1} \boldsymbol{E}_t \boldsymbol{\Sigma}_2^{-1} \boldsymbol{E}_t'|^{-\frac{\nu+2n-1}{2}}, \tag{2}$$

where $\Gamma_n(\cdot)$ is the multivariate gamma function. By the linearity property of a matrix-variate t distribution, (1) implies that $\boldsymbol{Y}_t \sim t_{n,n}(\nu, \boldsymbol{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ and is consequently called the matrix-variate t model. Since $\boldsymbol{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ and ν are unknown, they must be estimated. Bayesian methodology for estimating the unknown parameters is followed and implementing the matrix-variate Gamma and inverse matrix-variate Gamma as priors for $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ respectively, and a new graphical t-model as a result. Applying the methodology reveals a clear discrepancy between estimates from raw data and the Bayesian approach, which highlights the misleading impact that noise in data has and how it may lead to more grave consequences for any analysis built upon said noise.

2 A new graphical t-model construction

Assume that the prior density functions are mutually independent. The joint pdf $\pi(\boldsymbol{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \gamma, \nu)$ is then proportional to :

$$\gamma^{a_\gamma - 1} \nu^{a_\nu - 1} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\mathrm{vec}(\boldsymbol{B})' (\boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_1)^{-1} \mathrm{vec}(\boldsymbol{B}) \right) - \frac{\gamma}{b_\gamma} - \frac{\nu}{b_\nu} \right] f(\boldsymbol{\Sigma}_1|\gamma) f(\boldsymbol{\Sigma}_2|\gamma), \tag{3}$$

where $f(\boldsymbol{\Sigma}_1|\gamma), f(\boldsymbol{\Sigma}_2|\gamma)$ are some conditional prior pdfs of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. It follows $\boldsymbol{B}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \gamma, \nu$ has likelihood function equal to:

$$\prod_{t=1}^{\mathrm{T}} \frac{\Gamma_n\left(\frac{\nu+2n-1}{2}\right)}{\pi^{\left(\frac{n^2}{2}\right)}\Gamma_n\left(\frac{\nu+n-1}{2}\right)} |\boldsymbol{\Sigma}_1|^{-\frac{n}{2}} |\boldsymbol{\Sigma}_2|^{-\frac{n}{2}} |\boldsymbol{I}_n + \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{Y}_t - \boldsymbol{B})\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{Y}_t - \boldsymbol{B})'|^{-\frac{\nu+2n-1}{2}}. \tag{4}$$

From (3) and (4) the posterior pdf follows as

$$\begin{aligned}
&\prod_{t=1}^{\mathrm{T}} \frac{\Gamma_n\left(\frac{\nu+2n-1}{2}\right)}{\pi^{\left(\frac{n^2}{2}\right)}\Gamma_n\left(\frac{\nu+n-1}{2}\right)} |\boldsymbol{\Sigma}_1|^{-\frac{n}{2}} |\boldsymbol{\Sigma}_2|^{-\frac{n}{2}} |\boldsymbol{I}_n + \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{Y}_t - \boldsymbol{B})\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{Y}_t - \boldsymbol{B})'|^{-\frac{\nu+2n-1}{2}} \\
&\times \gamma^{a_\gamma - 1} \nu^{a_\nu - 1} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\mathrm{vec}(\boldsymbol{B})'(\boldsymbol{\Omega}_2 \otimes \boldsymbol{\Omega}_1)^{-1}\mathrm{vec}(\boldsymbol{B})\right) - \frac{\gamma}{b_\gamma} - \frac{\nu}{b_\nu}\right] \\
&\times f(\boldsymbol{\Sigma}_1|\gamma) f(\boldsymbol{\Sigma}_2|\gamma).
\end{aligned} \tag{5}$$

For this paper, consider the matrix-variate and inverse matrix-variate Gamma distributions, i.e., $\boldsymbol{\Sigma}_1 \sim MG_n(\delta_1, \beta, (\gamma\boldsymbol{\Phi}_1)^{-1})$ and $\boldsymbol{\Sigma}_2 \sim IMG_n(\delta_2, \beta, (\gamma\boldsymbol{\Phi}_2)^{-1})$ as priors. Notice that the scalar shape parameter β can be fixed, or have a prior imposed on it also. Either way the estimation procedure unaffected. As is usual with Bayesian estimation, an observed matrix $\boldsymbol{B}_i$ from the posterior distribution is an estimate of the true adjacency matrix $\boldsymbol{B}$ - thus, the average of a sample estimates $\boldsymbol{B}$. To simulated a sample, the Gibbs sampling algorithm is used.

3 Application and evaluation

The methodology is applied to the weekly stock prices of 70 European firms, resulting in 105 observations. Granger causality hypothesis tests are applied pairwise for week t. The resulting test statistics belong in a matrix that is an observed $\boldsymbol{Y}_t$ [*]. We employ well-known centrality measures, such as a graph's degree, closeness, eigen centrality, and betweenness, to evaluate a matrix variate estimator. These measures are univariate scores that measure a node's influence in a graph.

The results from the application are shown in Figure 1[†]. It is observed that there are clear discrepancies between the different estimators, with particular

[*]Data provided by Prof. M. Billio, University of Venice, Italy.

[†]The simulations were run on MATLAB R2022b on University of Pretoria server with 501Gb of RAM and 48 cores. Runtime for simulations was 16h excluding time to compute Granger causality test statistics.

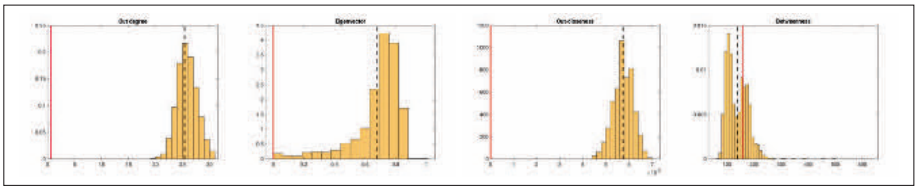


Figure 1: Estimated centrality measures: The solid red line and dashed black lines represent the averages of the raw data, and methodology respectively.

attention to the out-degree, out-closeness, and eigencentrality. The raw data seems to underestimate the centrality measures. In other words, the discrepancies highlight how noise left in data may jeopardise the validity and reliability of analysis built on data.

4 Acknowledgements

This work was based upon research supported in part by the National Research Foundation (NRF) of South Africa (SA), grant RA201125576565 & RA211204653274, nr 145681 & 151035; NRF ref. SRUG2204203865, the Centre of Excellence in Mathematical and Statistical Sciences, based at the University of the Witwatersrand (SA). The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

References

AMIRI, AMIRHOSSEIN, GHASHGHAEI, REZA, & MALEKI, MOHAMMAD REZA. 2018. On the effect of measurement errors in simultaneous monitoring of mean vector and covariance matrix of multivariate processes. *Transactions of the Institute of Measurement and Control*, **40**(1), 318–330.

ARNOLD, STEVEN F. 1979. Linear models with exchangeably distributed errors. *Journal of the American statistical Association*, **74**(365), 194–199.

BILLIO, MONICA, CASARIN, ROBERTO, COSTOLA, MICHELE, & IACOPINI, MATTEO. 2021. A matrix-variate t model for networks. *Frontiers in Artificial Intelligence*, 49.

MODELLING SOCCER PLAYERS FIELD POSITION VIA MIXTURE OF GAUSSIANS WITH FLEXIBLE WEIGHTS

Marco Berrettini [1], Giuliano Galimberti [1], Thomas Brendan Murphy [2] 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)

[2] School of Mathematics and Statistics, University College Dublin,
(e-mail: brendan.murphy@ucd.ie)

ABSTRACT: An empirical analysis on players' position on the field throughout a soccer match is presented. For this purpose, a Bayesian mixture of experts model is defined, allowing for flexible specification of concomitant covariates on the component weights as smooth functions represented by cubic splines.

KEYWORDS: mixtures of experts models, Gibbs sampling, Bayesian P-splines

1 Introduction

Pettersen et al. (2014) present a dataset of body-sensor traces and corresponding videos from three professional soccer games captured in late 2013 at the Alfheim Stadium in Tromsø, Norway. Tromsø - Stromsogodset is selected for this study, since it is the only one which is valid for the national competition. This game was played on November 3rd, 2013, and it ended with no scores. Player data, including field position, are sampled at 20 Hz using the ZXY Sport Tracking system.
The aim of this analysis is to study how a player's position is affected by a teammate's one and possibily identify a finite number of different phases of the game. Obviously, this relationship depends on many factors, such as the two player's role and which area of the field they are supposed to cover. For this reason, this study focuses on a couple of players playing close to each other.

2 Model specification

The study concentrates on the player covering the right full-back position, identified with tag 9, and assuming that his longitude and latitude (y_1 and y_2,

respectively) can reasonably be approximated by a bivariate Gaussian distribution. Then, the two-dimensional location of the centre-back playing closer to him, Player 13, are taken as concomitant covariates (x_1, x_2). Let $\mathbf{c}$ be a vector of latent variables such that, for each time i, $c_i = g$ if i belongs to cluster g. Conditioning on $\mathbf{c}_i$ and $\mathbf{x}_i$, it is assumed that $\mathbf{y}_i$ follows a Gaussian distribution with vector of means μ_{c_i} and positive definite covariance matrix Σ_{c_i}. Hence, the conditional density of $\mathbf{y}_i$ given $\mathbf{x}_i$ can be written as the following mixture of bivariate Gaussians:

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{g=1}^{G} \pi_g(\mathbf{x}_i) f_{MVN_2}(\mu_g, \Sigma_g), \tag{1}$$

with $f_{MVN_2}(\mu_g, \Sigma_g)$ being the density of a bivariate Gaussian distribution and component weights $\pi_g(\mathbf{x}_i) = \Pr(c_i = g|\mathbf{x}_i) > 0$, so that $\sum_{g=1}^{G} \pi_g(x_i) = 1$, for $i = 1, \ldots, 501$ and $g = 1, 2, \ldots, G$. To allow for flexible specification of such probabilities, a similar methodology to that proposed by Berrettini et al. (2021) for latent class models is adapted to the continuous case. More specifically, prior probabilities are expressed as smooth functions of the covariates represented through Bayesian P-splines (Lang & Brezger, 2004), and estimation is carried out following the data augmentation scheme suggested by Früwirth-Schnatter et al. (2012). Regarding the parameters of the component conditional distributions of the mixture, Gaussian and inverse Wishart priors are respectively assigned to μ_g and Σ_g, as in Marin et al. (2005). The resulting MCMC algorithm does not require any Metropolis-Hastings step.

3 Soccer player positions data

To carry out the analysis, some assumptions are made. In particular, the observations are assumed to be independent across time: to make this assumption more realistic, the data are thinned out to 501 observations over more than 90 minutes of play, leading to a distance of approximately 10 seconds between each pair of consecutive observations. Since between the first and the second half of the game the direction of play changes, preparing this dataset requires a 180° rotation of the locations observed during the second half. The two dimensions of the location of the centre-backs, x_1 and x_2, representing the long and short side of the field, respectively, are assumed to have an additive effect on the log-odds of the component weights. For the analysis, the algorithm is run for fixed G ranging from 1 to 6. The results produced by the best models, in terms of AICM, are selected. Observations are allocated into the G components using the maximum-a-posteriori rule.

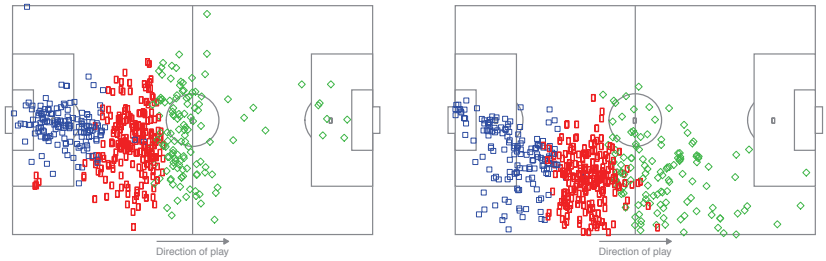


Figure 1. *Locations of Player 13 (left plot) and Player 9 (right plot). Different colors and dot symbols correspond to different clusters.*

4 Results

The best model according to AICM has $G = 3$ components. Figure 1 shows the locations of the two players during the game, allocated according to the 3-component ME model. The clusters does not seem well separated. Indeed, without considering the position of Player 13, the best finite mixture of Gaussians with constant component weights suggests the presence of a single component. These clusters may be interpreted as phases of the game: in particular, the blue dots identify the defensive phase, the green triangles the offensive one, while the red square indicate an intermediate phase. The intermediated phase, originally associated to the first component (in red), is taken as the reference to define the log-odds of mixture weights. The splines' coefficients are transformed accordingly, and, due to space limitations, only the estimated effect of the location of Player 13 on the probability of the defensive phase of Player 9 is reported in Figure 2. The clusters differ mainly with respect to the long side (x_1) of the field, while the location on the short side seems to be less impactful. Lower values of the longitude for Player 13 seem to lead to a higher probability that Player 9 is in the defensive phase, implying him covering the backfield too. This probability drops as x_1 grows, increasing the odds of the offensive phase, characterized by a higher longitude and variability. A huge amount of variability of the estimated effects can be noticed in the plots, especially when the functions reach large absolute values that correspond to 0 or 1 on the scale of the probability. This might be also due to the fact that the locations of the players are not uniformly distributed along the field. It is worth mentioning that this uneven distribution of the observations seems coherent with the specific roles of the two players considered in this analysis.

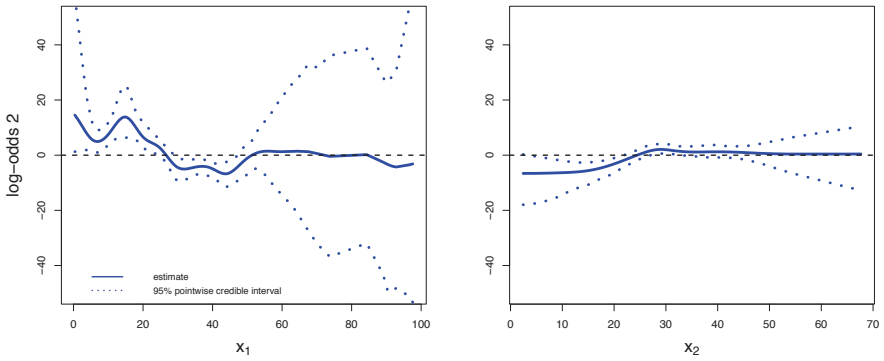


Figure 2. *Estimated effect (and 95% pointwise credible interval) of the location of Player 13,* (x_1, x_2) *on the log-odds of the mixture weights, for Cluster 2 .*

References

BERRETTINI, MARCO, GALIMBERTI, GIULIANO, RANCIATI, SAVERIO, & MURPHY, THOMAS BRENDAN. 2021. Flexible Bayesian modelling of concomitant covariate effects in mixture models. *arXiv preprint arXiv:2105.12852*.

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.

LANG, STEFAN, & BREZGER, ANDREAS. 2004. Bayesian P-splines. *Journal of computational and graphical statistics*, **13**(1), 183–212.

MARIN, JEAN-MICHEL, MENGERSEN, KERRIE, & ROBERT, CHRISTIAN P. 2005. Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, **25**, 459–507.

PETTERSEN, SVEIN ARNE, JOHANSEN, DAG, JOHANSEN, HÅVARD, BERG-JOHANSEN, VEGARD, GADDAM, VAMSIDHAR REDDY, MORTENSEN, ASGEIR, LANGSETH, RAGNAR, GRIWODZ, CARSTEN, STENSLAND, HÅKON KVALE, & HALVORSEN, PÅL. 2014. Soccer video and player position dataset. *Pages 18–23 of: Proceedings of the 5th ACM Multimedia Systems Conference*.

TOURISM AS SUPPORT IN ECONOMIC DEVELOPMENT OF INNER AREAS: A MULTI-SOURCES APPROACH

Antonella Bianchino[1], Daniela Fusco[1], Paola Giordano[1], Maria Antonietta Liguori[1], Maria Carmina Palma[2] and Donato Summa[1]

[1] Istat, Italian National Institute of Statistics, (e-mail: bianchin@istat.it, dafusco@istat.it, pgiordano@istat.it, liguori@istat.it, donato.summa@istat.it)

[2] Università degli studi di Napoli Federico II, Department of political science, (e-mail: melania.palma11@gmail.com)

ABSTRACT: In the 2014-2020 programming period, Italy has put in place a new integrated policy called the National Strategy for Inner Areas (NSIA). This policy aims to contribute to the country's economic and social recovery. Rural tourism is an opportunity for the economic growth of these areas. The aim of the work is to represent the complexity of rural tourism in Inner Areas (IAs), mainly Peripheral and Ultra-peripheral ones, compared with Urban Poles and "Belt" municipalities, using a multisource approach. A pilot study was carried out for the Campania region.

KEYWORDS: Tourism, Inner Areas, Economical grow, Multi-source approach, Indicators

1 Introduction and methodology

According to NSIA, in 2020, the 7,903 Italian municipalities are classified in 7 categories, from Urban Pole (182) to Ultra-peripheral ones (382). IAs include mainly Peripheral and Ultra-peripheral areas (total 1,906), fragile territories with a far "distance" from essential services. In the IAs, the agricultural, pastoral and forestry sectors play a central role as opportunities for economic growth and for the value of care and environmental prevention (Lucatelli, Storti, 2019).

On the other hand, rural tourism can be an effective means of providing socio-economic opportunities to rural communities. It can also help to increase the attractiveness and vitality of rural areas, mitigate demographic challenges, reduce migration and promote a range of local resources and traditions, while retaining the essence of rural life (UNWTO, 2020).

The aim of the work is to represent the complexity of rural tourism in IAs, compared with Urban Poles and "Belt" municipalities, analysing main components and driving forces, by using a multisource approach. The study identifies useful indicators for the evaluation of these phenomena by exploiting the opportunity given by using Big data, open data and traditional sources.

There were identified 15 basic indicators declined in three Pillars:

- Infrastructural density and touristic fluxes;
- Economical impact of touristic sector;

- Agricultural sector support.

For the construction of indicators were used 7 data sources: survey and administrative sources (Continuous Population Census, Survey of Museums and similar institutions, Capacity of Collective tourist accommodation establishments, Statistical Atlas of the Municipalities), Statistical registers (Statistical register of active enterprises - ASIA, Frame SBS) and Big Data on Agritourisms.

To improve data quality (e.g. timeliness, accuracy, punctuality) and to increase the amount of related information, the use of web scraping techniques has been proposed in this work. In this work, web data acquisition focuses on specific web scraping. Customized software programs have been developed to extract information from the website http:///www.agriturismoitalia.gov.it, which is the official website for authorized Italian Agritourisms, of which about 25,000 units.

To complete the study, we analyse the potential of municipalities defined as non-tourism (Istat, 2022) by integrating data with information on the presence of sites of tourist interest using open data.

Based on the Law 77, 2020 (G.U. N. 25/L Law 17 July 2020, n. 77), Istat classifies the Italian municipalities in "main tourist category", i.e. the potential tourist vocation of the municipality identified mainly on the basis of geographical (proximity to the sea, altitude, etc.) and anthropic (large urban municipalities) characteristics. There are 1,704 (21.5%) municipalities considered non-tourism, i.e. where there are no accommodation facilities and/or where tourist flows are absent.

Using the open data available on the website of the members of the National Statistical System (Sistan), we analyse the presence of Natura 2000 areas (DGE, 2023), monuments, local festivals and other attractions in the non-tourism municipalities.

2 Some results and Final remarks

The pilot analysis was carried out for Campania region. The following areas have been joined for the calculation of indicators: Poles/Inter-municipal Poles, Belt (260 municipalities), Intermediate (126 municipalities), Peripheral, Ultraperipheral (165 municipalities).

The potential of accommodation facilities is evidently very high in Peripheral and Ultraperipheral Areas. The exception is Visitor pressure on museum and similar institutions that, as expected, is higher in the Poles. Even the economy of the tourism sector confirms the potential of Peripheral and Ultraperipheral Areas.

The used source of big data, by its nature, excludes a part of the agritourisms, in particular those not in possession of the certification given by Ministry of Agriculture.

In fact, thanks to the “Agriturismo Italia” brand by Ministry of Agriculture, tourists and professional operators can easily distinguish officially accredited businesses. This distinction is very important, especially for the international market

where the Italian agritourism reality is not always perfectly known and the various operators could easily be disoriented by other forms of hospitality, equally present in the rural area.

Therefore, it cannot be excluded that a part of the agritourisms is missing, even if they are located in the Peripheral and Ultraperipheral Areas, but not in possession of the official certification. This is caused by uneven territorial marketing throughout the territory and the lack of associations among the structures.

Finally, some Municipalities considerate non-tourism has a certain number of attractions, underling the potential attractiveness of these areas.

Switching from a single source to multi-source statistics seems therefore the way to go. However, this transition is not easy. Multi-source statistics present new problems that need to be overcome before the resulting output quality is sufficiently high and these statistics can be efficiently produced. What complicates the production of multi-source statistics is that the supporting data are available in many different varieties as data sets can be combined in many different ways (de Waal *et al*., 2019).

References

DE WAAL T., VAN DELDEN A., SCHOLTUS S. 2019: Multi-source statistics: Basic situations and methods. International Statistical Review, **88**(1), 203-228.

DIRECTORATE-GENERAL FOR ENVIRONMENT, 2023. Guidelines for Defining, Mapping, Monitoring and Strictly Protecting EU Primary and Old-Growth Forests. Commission staff working document.

ISTAT, 2022. Classificazione dei Comuni in base alla densità turistica come indicato dalla Legge 17 luglio 2020, n. 77, art. 182. Nota metodologica.

LUCATELLI, S., STORTI, D. 2019. La strategia nazionale aree interne e lo sviluppo rurale: scelte operate e criticità incontrate in vista del post 2020. *Agrieuropa.* Anno **15** n°56.

UNWTO, 2020. Recommendations on Tourism and Rural Development – A Guide to Making Tourism an Effective Tool for Rural Development. Madrid, 27 September 2020.

SARIMA MODELS WITH MULTIPLE SEASONALITY

Luisa Bisaglia [1], Francesco Lisi[1]

[1] Department of Statistical Sciences, University of Padua, Italy, (e-mail: luisa.bisaglia@unipd.it, francesco.lisi@unipd.it)

ABSTRACT: SARIMA models and exponential smoothing methods are classical approaches to account seasonal dynamics. However, they tipically allow to model just one periodic component, while many empirical time series data show multiple seasonality, possibly interlacing toghether. To face this case, different decomposition models have been proposed in literature, while SARIMA models have been quite neglected. To fill the gap, in this work we suggest a suitable specification of the SARIMA model, called mSARIMA, able to account multiple seasonality. To study its performance, we compare it with two popular seasonal-trend decomposition approaches, namely the TBATS and MSTL models. A simulation exercise shows that mSARIMA models are more effective in describing the the different seasonal components.

KEYWORDS: Time series, Multiple seasonality, mSARIMA, seasonal-trend decomposition models.

1 Introduction

Typically, a SARIMA model allows to account just one periodic component. When multiple cycles arise, REG-SARIMA or SARIMAX models are often considered. In this case, only one seasonal component is treated as stochastic while the other ones are deterministically described using dummy variables, trigonometric functions or spline functions. Alternatively, a large body of literature focuses on time series decomposition techniques such as the Seasonal-Trend decomposition by regression (STR, Documentov & Hyndman, 2022), the Trigonometric Exponential Smoothing State Space model (TBATS, A.M. *et al.* , 2011) and the Multiple Seasonal Trend decomposition using Loess (MSTL, Bandara *et al.* , n.d.). The present work aims at showing that a suitable specification of SARIMA models allows to consider multiple seasonal components and effectively estimate them. We denote this class of models as Multiple Seasonality ARIMA models, briefly mSARIMA. We note, however, that these models are nothing but suitably constrained specifications of the general SARIMA models from which they inherit all properties.

2 The Multiple Seasonality ARIMA model

Let ε_t be a zero-mean white noise process with variance σ^2. We denote by mSARMA$(p,q)\times(P_1,Q_1)_{S_1}\times\ldots\times(P_m,Q_m)_{S_m}$ the stationary process Y_t

$$\phi(B)\prod_{i=1}^{m}\Phi(B^{S_i})Y_t=\theta(B)\prod_{i=1}^{m}\Theta(B^{S_i})\varepsilon_t, \tag{1}$$

where $\phi(B)$ and $\theta(B)$ are the usual AR and MA polynomials in B of degrees, respectively, p and q, $\Phi(B^{S_i})=(1-\Phi_{i,1}B^{S_i}-\ldots-\Phi_{i,P_i}B^{P_iS_i})$ is the i-th seasonal AR polynomial of degree P_i in B^{S_i} (i=1,..,m) while $\Theta(B^{S_i})=(1-\Theta_{i,1}B^{S_i}-\ldots-\Theta_{i,Q_i}B^{Q_iS_i})$ is the correspondig MA seasonal polynomial of degree Q_i. The polynomials $\phi(B)$ and $\theta(B)$ describe the non-periodic serial dependence of the time series, while the polynomials $\Phi(B^{S_i})$ and $\Theta(B^{S_i})$ model the periodic correlation for the m seasonal components of period S_i, $(i=1,...,m)$.
Just to give an example, the $2-SARIMA(1,0)\times(1,0)_4\times(1,0)_7$ is given by:

$$\begin{aligned}Y_t &= \phi_1Y_{t-1}+\Phi_{1,1}Y_{t-4}-\phi_1\Phi_{1,1}Y_{t-5}+\Phi_{2,1}Y_{t-7}-\phi_1\Phi_{2,1}Y_{t-8}-\\ &\quad \Phi_{1,1}\Phi_{2,1}Y_{t-11}+\phi_1\Phi_{1,1}\Phi_{2,1}Y_{t-12}+\varepsilon_t.\end{aligned}$$

It is clear that, although 7 different lags are involved, there are only 3 parameters to be estimated and that it can be also thought as a particular constrained (S)ARMA model. This implies that the stationary conditions for model (1) are those of a standard (S)ARMA, once the constraints are considered. The same holds also for the invertibility conditions. Model (1) can be straightforwardly generalized to the non-stationary case by including suitable unit root polynomials. We can define the non-stationary mSARIMA$(p,d,q)\times(P_1,D_1,Q_1)_{S_1}\times\ldots\times(P_m,D_m,Q_m)_{S_m}$ process, Y_t, by:

$$\phi(B)(1-B)^d\prod_{i=1}^{m}\Phi(B^{S_i})(1-B^{S_i})^{d_i}Y_t=\theta(B)\prod_{i=1}^{m}\Theta(B^{S_i})\varepsilon_t \tag{2}$$

where $(1-B^S)^d=(Y_t-Y_{t-S})^d$ is the $d-$seasonal difference of Y_t. As for standard ARIMA models, the process $X_t=(1-B)^d\prod_{i=1}^{m}(1-B^{S_i})^{d_i}$ is a stationary mSARMA. For building an mSARIMA model, the classical Box-Jenkins approach (Box & Jenkins, 1976) can be applied, with some simple and intuitive modifications needed to account the presence of more than one seasonal component. For the estimation step maximum likelihood methods can still be used taking care that the mSARMA model is a constrained one. This implies that the user has to write the specific likelihood to be maximized.

3 A comparison with other models

In this section we compare the mSARMA with two popular seasonal-trend decomposition models, namely the TBATS and MSTL models. We analyse, through a simple Monte Carlo exercise, their ability in whitening the residuals' autocorrelation function. To this end we simulated 500 independent realizations of length $n = 200$ and 500 from the three following mSARMA specifications. All models include two seasonal components: of periods 4 and 7, the first two, and of periods 4 and 12, the third one. In the last case, cycles overlap.

1. Model 1: $2-SARIMA(1,0,0)\times(1,0,0)_4\times(1,0,0)_7$, with $\phi_1 = 0.4$, $\Phi_{1,1} = 0.3$, $\Phi_{2,1} = 0.35$ and $\sigma^2 = 1$;
2. Model 2: $2-SARIMA(1,0,0)\times(1,0,0)_4\times(2,0,0)_7$, with $\phi_1 = 0.4$, $\Phi_{1,1} = 0.3$, $\Phi_{2,1} = 0.25$, $\Phi_{2,2} = 0.35$ and $\sigma^2 = 1$;
3. Model 3: $2-SARIMA(1,0,0)\times(1,0,0)_4\times(1,0,0)_12$, with $\phi_1 = 0.4$, $\Phi_{1,1} = 0.3$, $\Phi_{2,1} = 0.4$ and $\sigma^2 = 1$

For each series we estimated an mSARMA model, a MSTL model and a TBATS model and we analyzed the residuals time series to check if the multi-seasonal serial dependence has been completely accounted. To assess the residuals' appropriateness we propose a modification of the Pierce test (Pierce, 1978) able to account for multiple seasonality. When 2 periodic components are present, the hypothesis system to be verified is $H_0 : \rho_{S_1} = ... = \rho_{k\cdot S_1} = \rho_{S_2} = ... = \rho_{k\cdot S_2} = 0$ against $H_1 : \bar{H}_0$. When applied to the residuals of a model, it tests the model's adequacy in describing both seasonal components. The test statistics is:

$$mQ_{S_1,S_2}(k) = n\cdot(n+2)\left(\sum_{j=1}^{k}\frac{1}{n-j\cdot S_1}\rho^2_{j\cdot S_1} + \sum_{j=1}^{k}\frac{1}{n-j\cdot S_2}\rho^2_{j\cdot S_2}\right) \qquad (3)$$

where ρ_j is the correlation coefficient at lag j of the considered series. Under the null hypothesis of no seasonal autocorrelation, it follows a χ^2_{2k-par} distribution, where *par* is the number of estimated parameters.
In our exercise we computed $mQ_{4,7}(5)$ (for the first two models) and $mQ_{4,12}(5)$ (for the third model) on the residual series of our three models, i.e. mSARIMA, MSTL and TBATS, and we counted the percentage of times the null hypothesis is not rejected at a significance level of 5%. Results are given in Table **??**: it is clear that the mSARIMA model produce (sesaonally) uncorrelated residuals most of times, while the other two models do not, particularly when the sample

size increases. One could argue that it is not fair considering time series generated only by mSARMA models: we agree with this point. These results are very preliminary and other generating processes must be taken into account, in particular processes with deterministic components. However, we showed that, when multiple seasonality is generated by an ARMA process, MSTL and TBATS model are not appropriate to describe this dynamics.

Table 1. *Percentage of times, the Pierce test for multiple seasonality does not reject the hypothesis of seasonal uncorrelation in the residuals of the mSARMA, MSTL and TBATS models. The level of the test is* $\alpha = 5\%$.

		mSARIMA	MSTL	TBATS
Model 1	$n = 200$	93.2	2.3	42.3
	$n = 500$	92.7	0.0	17.3
Model 2	$n = 200$	92.5	0.3	6.3
	$n = 500$	91.6	0.0	0.0
Model 3	$n = 200$	93.9	0.3	1.4
	$n = 500$	94.8	0.0	0.0

References

A.M., DE LIVERA, R.J., HYNDMAN, & R.D., SNYDER. 2011. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association.*, **106**, 1513–1527.

BANDARA, K., HYNDMAN, R.J., & C., BERGMEIR. MSTL: A Seasonal-Trend Decomposition Algorithm for Time Series with Multiple Seasonal Patterns. *International J Operational Research.*

BOX, G.E.P., & JENKINS, G.M. 1976. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

DOCUMENTOV, A., & HYNDMAN, R.J. 2022. STR: Seasonal-Trend decomposition using Regression. *INFORMS Journal on Data Science.*, **1**, 50–62.

PIERCE, D.A. 1978. *Seasonal adjustment when both deterministic and stochastic seasonality are present. In: Seasonal Analysis of Time series (edited by Zellner)*. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census.

ADOPTION OF 4.0 TECHNOLOGIES AND RELATED OBSTACLES. APPLICATION OF A MULTIVARIATE NONPARAMETRIC TEST FOR CATEGORICAL VARIABLES

Stefano Bonnini[1] and Michela Borghesi[1]

[1] Department of Economics and Management, University of Ferrara,
(e-mail: stefano.bonnini@unife.it, michela.borghesi@unife.it)

ABSTRACT: The goal of this work is to investigate the relationship between the adoption of Industry 4.0 technologies and organizational, economic and financial obstacles. A combined permutation test for ordinal categorical variables was applied to an original dataset from a survey carried out in Italy.

KEYWORDS: nonparametric test, Anderson-Darling, categorical variables, industry 4.0.

1 Introduction

The case study of this work concerns the adoption of Industry 4.0 technologies by Italian small and medium-sized enterprises (SMEs). Initially, Industry 4.0 was basically considered the fourth revolution in the manufacturing industry. For this reason, in recent years, this topic has been much debated (Ghobakhloo, 2020). Besides the technical implementation, aspects such as lacking trust, unclear benefits for suppliers and different perceptions of Industry 4.0 hamper digital information sharing, especially by SMEs (Müller et al., 2020).

In the literature, there are several works dedicated to Industry 4.0 and related technologies, as well as to the relationship between implementation of these new technologies and policy incentives (Oztemel, 2020). Our interest concerns the degree of relevance of the obstacles to the adoption of 4.0 technologies in the two-year period 2018-2019 by Italian manufacturing SMEs. From a methodological point of view, we applied a multivariate permutation test for ordinal categorical responses (Pesarin and Salmaso, 2010).

This work represents a contribution in the empirical literature on the adoption of 4.0 technologies by SMEs and related obstacles. Section 2 focuses on the methodological approach. Section 3 is dedicated to the application and section 4 includes the final conclusions.

2 Methodological approach

Let us consider the multivariate two-sample problem for ordered categorical variables with a one-sided alternative hypothesis of stochastic dominance. Such a complex problem has not an easy parametric solution (Cohen & Sackrowitz, 1998) and the asymptotic distribution of the test statistic of such solution under the null hypothesis depends on unknown parameters.

Let A_h^v be the h-th category for the v-th variable with $h = 1, \dots, c_v$ and $v = 1, \dots, k$. The null hypothesis in terms of CDFs (cumulative distribution functions) is

$$H_0: F_1^v(A_h^v) = F_2^v(A_h^v)\ \forall h, v,$$

where $F_i^v(x)$ is the CDF of the i-th population, with $i = 1,2$. The alternative hypothesis of stochastic dominance can be formalized as

$$H_1: F_1^v(A_h^v) \leq F_2^v(A_h^v)\ \forall h, v \text{ and } \exists h, v: F_1^v(A_h^v) < F_2^v(A_h^v).$$

A suitable test statistic for the partial hypothesis concerning the v-th variable might be that of Anderson-Darling:

$$T_v = \sum_{h=1}^{c_v - 1} \left[\hat{F}_2^v(A_h^v) - \hat{F}_1^v(A_h^v)\right] \left[\hat{F}_\cdot^v(A_h^v)\left(1 - \hat{F}_\cdot^v(A_h^v)\right)\right]^{-1/2},$$

where $\hat{F}_i^v(x)$, for $i = 1,2$, is the empirical cumulative distribution function of sample i and $\hat{F}_\cdot^v(x)$ is the marginal empirical distribution function for the v-th variable (Bonnini et al., 2014).

H_0 must be rejected in favor of the alternative hypothesis for large values of the test statistics. A test statistic for the multivariate problem can be obtained by combining the p-values of the univariate tests according to the Fisher combining function $T_{mul} = -\sum_v \log \lambda_v$, where λ_v is the p-value of the v-th partial test. The p-value of such a test is computed as

$$\lambda_v = P[T_v \geq t_v | H_0] = [\#(T_v \geq t_v) + 0.5]/(B + 1),$$

where t_v is a generic value taken by T_v, $\#(T_v \geq t_v)$ is the number of times T_v is greater than or equal to t_v according to the permutation distribution and B is the number of permutations.

3 Application

The permutation test for categorical data presented in the previous section was applied to original data collected in a sample survey carried out in January 2022. The survey was conducted in the northern regions of Italy by the Department of Economics and

Management of the University of Ferrara. It was aimed at manufacturing enterprises of the North Italy. The total number of interviewed companies is 3924. For the selection of the companies, a stratified random sampling was applied.

The goal is to compare companies that have adopted at least one 4.0 technology with those that have not adopted any 4.0 technology, in terms of the degree of relevance of the obstacles to the adoption of such technologies in the two-year period 2018-2019. The hypothesis under test is that the relevance of obstacles is higher for the companies that didn't adopt 4.0 technologies. The significance level is $\alpha = 0.05$.

The technologies considered were: advanced manufacturing solutions (interconnected and programmable robots), additive manufacturing (3D printers connected to digital development software), augmented reality (to support production processes), simulation (between interconnected machines for process optimization), horizontal integration (integration of information along the production process stages), vertical integration (sharing of information along the value chain/supply chain with suppliers and customers), industrial internet (multidirectional communication between production processes and products), cloud computing (data management on open systems), cyber-security (during network operations on open systems) and big data/analytics (for the optimization of products and production processes).

On the other hand, the barriers to the adoption of the aforementioned technologies are:

- lack of internal economic resources,
- difficulty in obtaining credit,
- difficulty in accessing public funding (subsidies),
- high costs,
- lack of internal skills.

For each category of obstacles, the degree of relevance was expressed in a Likert scale from 1 to 4, where 1="not at all", 2="somewhat", 3="very" and 4="extremely".

The application of the permutation test presented in the previous section, leads to the overall p-value **0.006**. Hence, at the significance level $\alpha = 0.05$, we have empirical evidence in favor of the hypothesis that the relevance of obstacles is higher for the companies that didn't adopt 4.0 technologies.

Given the overall result, we can attribute the significance to one or more specific partial tests, after adjustment of the partial p-values to control the family-wise error (FWE) with the Bonferroni-Holm method (Westfall & Young, 1992).

Table 1: Table of adjusted p-values with Bonferroni-Holm method (significant in bold).

	Lack of resources	*Difficulty obtaining credit*	*Difficulty access subsidies*	*High costs*	*Lack of skills*
Partial p-values	**0.010**	**0.010**	**0.034**	0.961	**0.021**

According to Table 1, we can conclude that the most significant obstacles to the introduction to Industry 4.0 technologies are the lack of internal economic resources, the difficulty to obtain credit, the difficulty in accessing public funding and the lack of internal skills.

4 Conclusions

The nonparametric approach used in this work, based on the application of a combined permutation test, is a robust and flexible statistical solution for multivariate tests in the presence of categorical data. Its application to an original dataset concerning a survey about Italian enterprises leads to the conclusion that the main obstacles that determine the decision of adopting 4.0 technologies are the lack of internal economic resources, the difficulty to obtain credit, the difficulty in accessing public funding and furthermore the lack of internal skills.

Acknowledgments

Authors thank University of Ferrara that funded the project entitled "Public policies, 4.0 technologies and enterprise performance. Empirical analyses on a representative sample of manufacturing enterprises of northern Italy (Politiche pubbliche, tecnologie 4.0 e performance d'impresa. Analisi empiriche su un campione rappresentativo di imprese manifatturiere del Nord Italia)" for the period 2022-2024, with the Departmental Research Incentive Fund - FIRD 2022.

References

BONNINI, S., CORAIN, L., MAROZZI, M., & SALMASO, L. 2014. *Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R*. Chichester, UK: Wiley.

COHEN, A., & SACKROWITZ, H.B. 1998. Directional tests for one-sided alternatives in multivariate models. *Annals of Statistics*, **26**, 2321-2338.

GHOBAKHLOO, M. 2020. Industry 4.0, digitization, and opportunities for sustainability. *Journal of Cleaner Production*, **252**, 119869.

MÜLLER, J.M., VEILE, J.W.& VOIGT, K. 2020. Prerequisites and incentives for digital information sharing in Industry 4.0 - An international comparison across data types. *Computers & Industrial Engineering*, **148**, 106733.

OZTEMEL, E., & GURSEV, S. 2020. Literature review of Industry 4.0 and related technologies. *Journal of Intelligent Manufacturing*, **31**, 127-182.

PESARIN, F., & SALMASO, L. 2010. *Permutation tests for complex data. Theory, applications and software*. Chichester, UK: Wiley.

WESTFALL, P.H., & YOUNG, S.S. 1992. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York, NY, USA: Wiley-Interscience.

AN APPLICATION OF ASYMMETRIC MULTIDIMENSIONAL SCALING TO THE VQR 2015-2019 DATA

Giuseppe Bove[1]

[1] Dipartimento di Scienze della Formazione, Università degli Studi Roma Tre,
(e-mail: giuseppe.bove@uniroma3.it)

ABSTRACT: Proximity matrices are frequently asymmetric and analysed by using the additive decomposition into symmetric and skew-symmetric components. A preliminary graphical description of the two components can allows to detect interesting relationships in the data. An application to the matrix of flows of scientific products between GEV's in the VQR 2015-2019 is presented to emphasize the advantages of the graphical approach.

KEYWORDS: proximity data, asymmetric multidimensional scaling, research assessment.

1 Introduction

Asymmetric proximities between pairs of entities (e.g., import-export data, sociomatrices, brand switching, flows and migration data, etc.) are analysed in economics, sociology, marketing research and other behavioural sciences by using a variety of models and methods to detect meaningful relationships. When no substantive model is readily available and there is not a priori reason for preferring any particular model, we might look for graphical displays as preliminary investigation of the proximity data.

In this presentation, the approach analysing separately by diagrams the symmetric and the skew-symmetric components of the proximity matrix is applied to the matrix of flows of scientific products between groups of experts for evaluation (GEV), in the process of research quality assessment of the Italian research institutions regarding the years 2015-2019 conducted by the agency ANVUR (Italian Ministry of University and Research - MUR).

In the next sections, first the asymmetric multidimensional scaling (MDS) models for symmetry and skew-symmetry are presented, focalizing on the diagrams that are obtained and their interpretation. Follows the description of the proximity data published by ANVUR and the presentation of the diagrams obtained by the application of asymmetric MDS.

2 Models for symmetry and skew-symmetry

A ($n \times n$) proximity matrix $\mathbf{\Omega} = (\omega_{ij})$ between pairs of entities (*i,j)* (*i,j=1,2*, ….., *n*) can always be decomposed additively as $\mathbf{\Omega} = \mathbf{M} + \mathbf{N}$, with $\mathbf{M} = (m_{ij})$ a symmetric component and $\mathbf{N} = (n_{ij})$ a skew-symmetric component, defined respectively, $m_{ij} = (\omega_{ij} + \omega_{ji})/2$ and $n_{ij} = (\omega_{ij} - \omega_{ji})/2$. The elements n_{ij} are the deviations of the proximities from symmetry, because $n_{ij} = \omega_{ij} - m_{ij}$. Constantine and Gower (1978) remark that the symmetric component **M** and the skew-symmetric component **N** of any square matrix **Ω** are uncorrelated and the following orthogonal breakdown of the sum of squares (**SS**) holds, **SS(Ω)** = **SS(M)** + **SS(N)**, reflecting the uniqueness of the additive decomposition (for a review, see Bove, Okada and Vicari (2021)). Thus, the two components can be viewed independently, and studied by separate models. An advantage of the separate analysis is that one can deal with the representation problem by adopting different kinds of models (e.g., spatial vs nonspatial) and different kinds of transformations (e.g., metric vs nonmetric) for the two components.

The symmetric component **M** can be represented by a Euclidean distance model in *r* dimensions

$$f(m_{ij}) = \hat{d}_{ij} = d_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j) + e_{ij} = \sqrt{\sum_{s=1}^{r}(x_{is} - x_{js})^2} + e_{ij}, \quad (1)$$

where f is a monotone transformation, mapping the proximities ω_{ij} into a set of transformed values $\hat{d}_{ij} = f(m_{ij})$, x_{is} and x_{js} are the coordinates of row (column) *i* and row (column) *j* on dimension *s*, respectively, and the symmetric property holds ($d_{ij} = d_{ji}$). The estimation of the model provides a map where entries m_{ij} are approximated by distances in *r* dimensions: when the entries m_{ij} are similarities, the larger the values, the smaller the distances. The model can be estimated using standard statistical software for symmetric multidimensional scaling.

The skew-symmetric matrix **N** can be represented in a diagram by the two-step method proposed in Bove and Vicari (2023). At the first step, the sizes of the skew-symmetries $|n_{ij}|$ are collected in a symmetric matrix $\mathbf{T} = (t_{ij}) = (|n_{ij}|)$, a standard dissimilarity matrix that can be approximated in a *low-dimensional* (usually two-dimensional) Euclidean space by the distance model (1) where m_{ij} is replaced with t_{ij}. At the second step, the signs of the skew-symmetries $sign(n_{ij})$ can be represented in the configuration obtained at the first step, by the drift vector method proposed in Borg and Groenen (2005, par. 23.5) applied to matrix $\mathbf{K}^{sign} = (k_{ij}^{sign}) = (sign(n_{ij}))$. From each point *i* an arrow is drawn to each other point *j* so that its length and direction correspond to the values in row *i* of the skew-symmetric matrix $\mathbf{K}^{sign}$. For each pair (*i,j*), if n_{ij} is positive, the arrow goes from point *i* to point *j*, while, when n_{ij} is negative, the arrow points in the opposite direction. When the number of rows of the proximity matrix is large, the representation of all arrows may result into cluttered pictures, and it can be convenient to draw only the average vector (drift vector) of the

arrow bundle attached to each point. The length of the drift vector represents the homogeneity of the directions of the (n-1) vectors emitting from the point. The arrows representing the drift vectors go towards areas of the diagram containing points having more frequently negative skew-symmetries and they can exhibit systematic asymmetric trends.

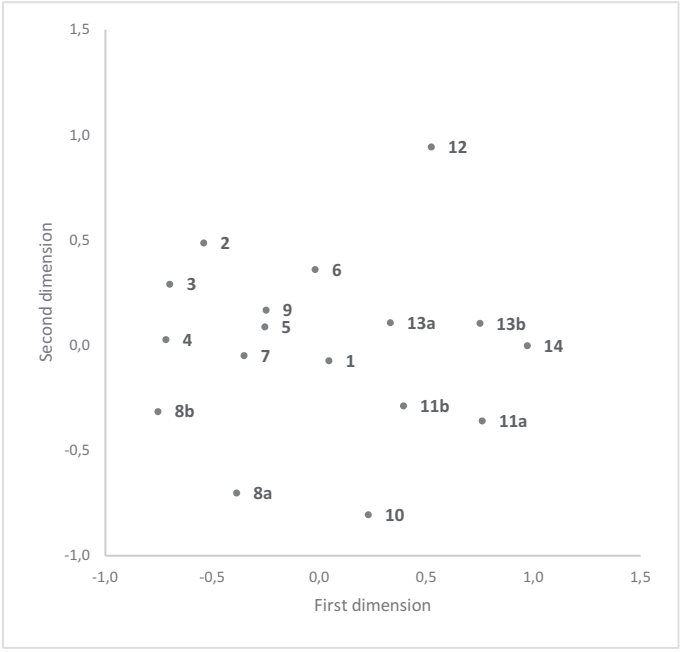


Fig. 1 - MDS of the symmetric component of the transformed 2015-2019 VQR data

3 Application to 2015-2019 VQR data

The methods for symmetry and skew-symmetry presented in section 2 are now applied to a proximity matrix published on-line by ANVUR in the VQR 2015-2019 Final Report - Statistics and summary results (Table 2.11). Entries of the matrix are the flows of scientific products between the seventeen GEV's (their labels and names are listed in Table 2.1 of the Final Report) and can be considered as measures of similarity between GEV's. Diagonal entries are strongly dominant (97.3% of the total flows) and represent the scientific products evaluated inside the GEV's. Nonetheless, off-diagonal entries seem interesting to detect affinity relationships between GEV's. A preliminary transformation was applied to remove the main effects, which reflects the influence of the row and column totals. A simple correction for such main effects is to equalize all self-similarities by dividing the entry in each cell (i,j) by the square

root of the product of the entries in diagonal cells (*i*,*i*) and (*j*,*j*) (self-similarities). This transformation does not affect the asymmetry, the ratios between off-diagonal entries in cells (*i*,*j*) and (*j*,*i*) for ($i \neq j$) remain the same.

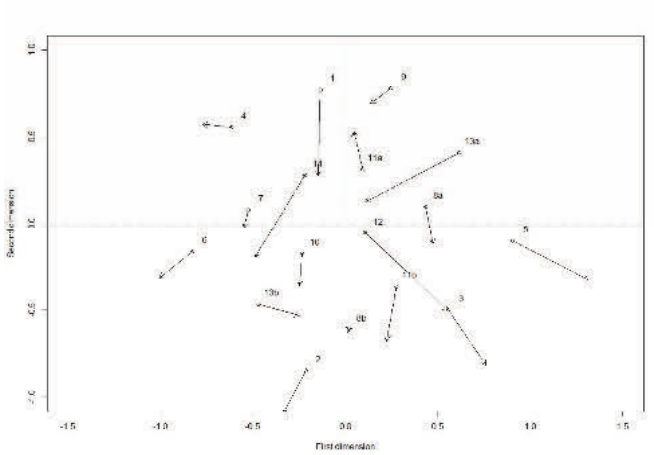


Figure 1 provides the representation of the GEV's average flows (symmetric component), small distances represent high average flows. GEV's positioned far from the origin (e.g., GEV 10 - Antiquity, Philologic.-Literary and Historical-Artistic sciences and GEV 12 – Legal sciences) have small or null flows with the other GEV's. GEV's close to the origin (e.g., GEV 1 - Mathematics and computer sciences and GEV 5 – Biological sciences) have high flows with several other GEV's. Figure 2 provides the representation of the skew-symmetric component, large distances indicate high imbalances (e.g., GEV 2 – Physical science and GEV 9 – Industrial and information engineering, in this case the direction of the arrows indicate that the skew-symmetry is positive from GEV 9 to GEV 2). Homogeneity of the directions characterize GEV's with long arrows (e.g., GEV 13a – Economics and statistics).
Fig. 2 – Drift vectors for skew-symmetry in the transformed 2015-2019 VQR data

References

BORG, I., GROENEN, P.J.F. 2005. *Modern Multidimensional Scaling. Theory and Applications*. (Second Edition). New York: Springer.

BOVE, G., OKADA, A. & VICARI, D. 2021. *Methods for the analysis of asymmetric proximity data*. Singapore: Springer Nature.

BOVE, G., VICARI, D. 2023. Graphical Analysis and Clustering of Asymmetric Proximities. In: Okada, A. et al. (eds.) *Facets of Behaviormetrics: The 50th Anniversary of the Behaviormetric Society*. Singapore: Springer Nature (to appear)

CONSTANTINE, A.G., GOWER, J.C. 1978. Graphical representation of asymmetric matrices. *Applied Statistics*, **3**, 297-304.

IMPROVING CLUSTERING IN TEMPORAL NETWORKS THROUGH AN EVOLUTIONARY ALGORITHM

Luca Brusa[1] and Fulvia Pennoni[1]

[1] Department of Statistics and Quantitative Methods, University of Milano-Biccoca
(e-mail: luca.brusa@unimib.it, fulvia.pennoni@unimib.it)

ABSTRACT: The dynamic stochastic blockmodel is commonly used to analyze longitudinal network data when multiple snapshots are observed over time. The variational expectation-maximization (VEM) algorithm is typically employed for maximum likelihood inference to allocate nodes to groups dynamically. To address the problem of multiple local maxima, which may arise in this context, we propose modifying the VEM according to an evolutionary algorithm to explore the whole parameter space. A simulation study on dynamic networks and an application illustrate the proposal comparing the performance with that of the VEM algorithm.

KEYWORDS: local maxima, longitudinal networks, node classification, stochastic blockmodel, variational expectation-maximization algorithm.

1 Introduction

The dynamic stochastic blockmodel (Matias & Miele, 2017) extends the stochastic blockmodel (SB, Nowicki & Snijders, 2001) for the analysis of longitudinal network data when multiple snapshots are observed over time. This model aims to identify homogeneous blocks of nodes and to analyze interactions between nodes and their evolution. At each time occasion, nodes are partitioned into a set of groups whose number is estimated; the probability of observing an edge between a couple of nodes depends on the assigned groups.

In the inferential context, the variational expectation-maximization (VEM, Jordan *et al.*, 1999) algorithm has been proposed for maximum likelihood estimation. However, a drawback of this method is that it can be trapped in one of the multiple local maxima. To account for this problem we propose a modified version of the VEM through an evolutionary algorithm (EA, Ashlock, 2004). We perform a Monte Carlo simulation study to evaluate the performance of the proposed evolutionary VEM (EVEM) algorithm in avoiding local maxima and improving the accuracy of the posterior classification. We also show an application estimating the dynamic SB with data related to face-to-face contacts between employees to investigate transmission of an infectious disease.

2 Notation and inference in dynamic stochastic blockmodel

Considering n nodes observed at T discrete times, let $\mathbf{Y}$ denote an adjacency array of dimensions $n \times n \times T$, where $\mathbf{Y}^{(t)}$ is the adjacency matrix at time t and $Y_{ij}^{(t)} = 1$ if there is an edge between nodes i and j (symmetric association) at time t and $Y_{ij}^{(t)} = 0$ otherwise ($i, j = 1, \ldots, n$, $i \neq j$). The dynamic SB assumes that block membership depends on a set of independent and identically distributed discrete latent variables $Z_i^{(t)}$ following a Markov chain with k support points. In this way, each node is partitioned into one of k latent blocks at every time occasion according to the initial and the transition probabilities denoted as α_u and π_{uv}, $u, v = 1, \ldots, k$, respectively. Under the local independence assumption and conditionally on the latent blocks to which nodes i and j belong at time t, the variables $Y_{ij}^{(t)}$ are assumed to be independent and Bernoulli distributed with connection probabilities denoted as β_{uv}.

For maximum likelihood inference of SB the VEM was proposed in Matias & Miele, 2017 to maximize a lower bound of the log-likelihood function denoted as $\mathcal{J}(\theta)$, where θ collects the model parameters. More recently, Bartolucci & Pandolfi, 2020, proposed an exact formulation of the VEM algorithm to improve clustering units across time occasions. They initialize the starting values for the model parameters through the k-means method since random initialization is usually ineffective in this context. However, this approach does not prevent the VEM algorithm from being trapped in the local maxima that frequently arise with complex data structures.

3 Proposed evolutionary VEM algorithm

The proposed EVEM algorithm is defined by the following features: (*i*) an initial "population" denoted as P_0 of N candidate solutions for the maximization problem at issue, here specified as possible arrays of cluster memberships; (*ii*) a mutation operator that introduces variations to the existing candidates and generates new solutions by randomly selecting an observation and providing an updated cluster membership; (*iii*) selection of the best solutions based on a quality measure that favors candidates with higher values of $\mathcal{J}(\theta)$.

In order to explore the whole parameter space the first candidate for population P_0 is obtained according to the k-means deterministic initialization; in particular, the adjacency matrices $\mathbf{Y}^{(t)}$ for $t = 1, \ldots, T$ are row-concatenated together, and the k-means algorithm is applied on the rows of the resulting $nT \times n$ matrix. Then, the remaining $N - 1$ candidates are obtained through

mutation. The procedure alternates the following steps until convergence:

1. $P_1 \leftarrow$ **Update**(P_0): perform a small number of iterations of the VEM algorithm on each individual of population P_0.
2. $P_2 \leftarrow$ **Mutate**(P_1): add variation in each individual of population P_1 to encourage a broader exploration of the parameter space.
3. $P_3 \leftarrow$ **Update**(P_2): perform a small number of iterations of the VEM algorithm on each individual of population P_2.
4. $P_4 \leftarrow$ **Select**$(P_1 \cup P_3)$: consider individuals of both populations P_1 and P_3, and retain the N showing the highest value of $\mathcal{J}(\theta)$ for the next generation.

Convergence is assessed considering the best solution of population P_4, analyzing the relative difference of $\mathcal{J}(\theta)$ at two consecutive steps and that between the corresponding parameter vectors.

4 Simulation study and application

In analogy with the design used in Bartolucci & Pandolfi, 2020, a Monte Carlo simulation study is conducted, varying the number of nodes ($n = 20, 50$), the number of latent blocks ($k = 2, 3$), the block persistence (high or low), and the connectivity parameters (intra-group greater or smaller than inter-group). For each of the 16 resulting scenarios, we randomly draw 50 networks and estimate the dynamic SB with both the VEM and the EVEM algorithms. The effectiveness of the proposed approach is evaluated in terms of the Adjusted Rand Index (ARI, Hubert & Arabie, 1985) between the true and the estimated classification at each time occasion.

Simulation results show that the EVEM algorithm outperforms the existing VEM algorithm in most scenarios, especially those with higher complexity. For example, considering a scenario characterized by 50 nodes, 3 latent blocks, low persistence of latent states, and higher intra-group than inter-groups connection probabilities, the ARI equals 0.688 using the VEM algorithm and 0.761 with the EVEM algorithm. In another scenario, with the same features but opposite connectivity parameter setting, ARI is 0.707 with VEM and 0.784 with the EVEM. In both cases, the improvements are statistically significant. When using the EVEM algorithm, we also observe a decrease of the mean squared error between the estimated and true model parameters, computed as an aggregated measure over all the model parameters.

Real data refer to face-to-face contacts between $n = 90$ employees in a building of the *Institut de veille sanitaire* (French Institute for Public Health

Surveillance) for ten working days ($T = 10$), from June 24 to July 3, 2013 (data are available at the website: `http://www.sociopatterns.org/datasets/contacts-in-a-workplace/`). The building hosts three scientific departments ("DISQ", "DMCT", and "DES"), logistics ("SFLE") and human resources ("SRH"). The adjacency array is built by setting each element $Y_{ij}^{(t)}$ equal to 1 if at least one face-to-face contact was registered between employees i and j at time t, and 0 otherwise.

A dynamic SB with 5 latent blocks is estimated using both VEM and EVEM algorithms. The resulting classification of employees helps understand how a certain infectious disease may spread across different departments of the same building. We observe that the value of $\mathcal{J}(\hat{\theta})$ at convergence increases from -2613 to -2600 when the EVEM algorithm is employed. This is reflected in a more accurate classification of the employees in each group of the network. The EVEM algorithm identifies a specific latent block for employees from the "DISQ" department, while the VEM algorithm allocates them with employees from the "DMCT" department. Additionally, the EVEM algorithm correctly assigns all employees from the "DSE" department to a single latent block, whereas the VEM algorithm splits them into two distinct blocks.

References

ASHLOCK, D. 2004. *Evolutionary Computation for Modeling and Optimization*. Springer, New York.

BARTOLUCCI, F., & PANDOLFI, S. 2020. An exact algorithm for time-dependent variational inference for the dynamic stochastic block model. *Pattern Recognit. Lett.*, **138**, 362–369.

HUBERT, L., & ARABIE, P. 1985. Comparing partitions. *J. Classif.*, **2**, 193–218.

JORDAN, M.I., GHAHRAMANI, Z., JAAKKOLA, T.S., & SAUL, L.K. 1999. An introduction to variational methods for graphical models. *Mach. Learn.*, **37**, 183–233.

MATIAS, C., & MIELE, V. 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Series B*, **79**, 1119–1141.

NOWICKI, K., & SNIJDERS, T.A.B. 2001. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.*, **96**, 1077–1087.

A SUPPORT VECTOR MACHINE APPROACH TO CREATE OBLIQUE DECISION TREES FOR REGRESSION

Andrea Carta [1]

[1] Department of Business and Economics, University of Cagliari, Cagliari, Italy, (e-mail: andrea.carta88@unica.it)

ABSTRACT: Decision trees are a popular statistical learning algorithm for classification and regression that recursively split the data based on the most informative characteristics. Unfortunately, they do not have a high predictive power with respect to other statistical learning methods. To enhances their performances, this paper proposes a support vector machine approach to create oblique decision trees for regression problems. In this novel model, the split at each node is made through a weighted support vector machine classifier with a linear Kernel that minimizes the deviance of the split. We test the model with respect to the usual CART on four public datasets with numerical predictors on three global metrics: Root Mean Squared Error, Mean Absolute Deviation, and R^2. The results of repeated cross-validation show that the novel model can overperform the usual Decision trees.

KEYWORDS: Trees, Oblique Split, SVM, Regression, Oblique Trees

1 Introduction

Decision trees (DTs) are a popular statistical learning algorithm for classification and regression. They can be easily viewed and interpreted by humans, making them valuable assets in data. A DT is a tree structure in which each internal node represents a decision based on a specific characteristic of the data, and where each leaf node represents a prediction or result. The algorithm works by recursively splitting the data based on the most informative characteristics until a stopping criterion is met. Unfortunately, DTs are prone to overfitting and do not have a high predictive power with respect to other statistical learning methods. To improve their performances oblique DTs were introduced (Breiman, 2017), and lately, they are gaining interest in the research community. Unlike traditional DTs, in which each node corresponds to a single variable split and the separation between the branches is orthogonal to the axes, oblique DTs allow the definition of separation hyper-planes that can be inclined with respect to the Cartesian axes. In other words, oblique

DTs use linear combinations of multiple variables to define decision boundaries. However, to find the linear combination of variables to construct the best-suited hyperplane is an NP-hard problem, in fact, to split a node with n observations using an axis-aligned CART, an exhaustive search would require no more than $n \cdot p$ evaluations. On the other hand, oblique CART would require a significantly larger number of evaluations, specifically $2^p \binom{n}{p}$. Nevertheless, oblique DTs have the advantage of generally building smaller trees with better accuracy compared with axis parallel trees (Wickramarachchi *et al.*, 2016). In contrast to the Breiman's approach, we introduce Support Vector Machine Regression Oblique Tree (SVM-ROT). In the Breiman method, the algorithm optimizes the coefficients of oblique splits based on a coordinate descent method. This is an iterative approach where each coefficient is optimized individually while keeping the others fixed. On the other hand, in SVM-ROT the split at each node is determined through a weighted support vector machine (SVM) classifier with a linear Kernel that minimizes the deviance of the split. SVM is a supervised statistical learning method introduced by Vapnik, 1999 to solve pattern classification and regression problems, moreover, it can be linear or nonlinear but is most commonly the former. Essentially, SVM identifies a reproducible hyperplane that maximizes the margin between the support vectors of both class labels. To improve the performance of the SVM classifiers, Yang *et al.*, 2007 suggests adding different weights to observations to different data points such that the weighted SVM algorithm estimates the best hyperplane according to the relative importance of the observation in the training data set. This short paper is organized as follows. Section 2 introduces the model in detail, in Section 3 the model is tested on 4 datasets and some concluding comments are reported.

2 Model

SVM-ROT at each node separates the observations given the results of a SVM classifier. Let us consider N observations characterized by a continuous response Y and p continuous features. First, Y is transformed K times into a dichotomous variable, each time using a different quantile as the threshold for its partitioning. Then, for each of these dichotomized variables, a weighted SVM classifier with linear kernel is applied, and the algorithm saves the deviance reduction resulting from the two partitions. The algorithm then chooses the split that has the highest reduction in deviance. The weighting of the SVM is very important because when the algorithm dichotomizes the target variable much information is lost. To overcome this problem the absolute values of

the scaled elements of the target variable Y are used as weights in the classifiers. This process assures that the hyperplane takes into account the values of the original Y. The result of this process will be a set of coefficients $\mathbf{w}$ of length p, and an intercept b, which describe the separating hyperplane. The hyperplane will be then expressed in a decision rule similar to that one of the usual DT, creating the pair of half-spaces: $R_1(w,b) = \{\mathbf{X} \mid \mathbf{w} \cdot \mathbf{x} + b \leq 0\}$ and $R_2(w,b) = \{\mathbf{X} \mid \mathbf{w} \cdot \mathbf{x} + b > 0\}$, where $\mathbf{X}$ is the matrix of the p predictors.

The result will be the division of the feature space into two subsets. This operation is then applied in a recursive binary partition manner until a certain criterion is met. These stopping criteria can be the number of elements in a leaf, the number of elements in a node, or the complexity parameter given by the ratio between the resulting deviance after the split and the deviance in the parent node.

3 Application to real datasets

SVM-ROT has been applied to several real datasets using the software `R` (R Core Team, 2022). The first is "Body Fat" dataset from Penrose *et al.*, 1985. In this dataset, the response variable is the percentage of body fat and the eleven predictors represent several physiologic measurements related to 252 men. The second dataset, called BCF, comes from Grisoni *et al.*, 2016, here the target variable is the Bioconcentration Factor in log units of 779 chemicals, while the independent variables are nine molecular numerical descriptors. The third data set is Auto MPG dataset from Dua & Graff, 2017 consisting of 398 observations, but in which only the seven numerical predictors have been used. Finally, the last dataset is from Ancell, 2021, it is made up of 413 instances and contains the 50 year ground snow load at a variety of measurement stations together with four numerical predictors. The performance of the SVM-ROT is compared to the one of a CART. Both models were tuned for the complexity parameter with 10-fold cross-validation, and the most parsimonious model with the one standard error rule was chosen. Then we performed 10 times repeated 10-fold cross-validation. The overall performance is computed by Root Mean Squared Error (RMSE), Mean Absolute Deviation (MAD), and R^2. For the first two metrics, lower values result in better predictive models. However, RMSE is more sensitive to high errors. R^2 is the proportion of variance explained by the model, this means that a value close to one indicates that the model explains most of the variance. Table 1 shows the results of the experiments.

In BCF and MPG SVM-ROT shows a better performance with respect to

	Body Fat		BCF		MPG		Snow	
	SVM-ROT	CART	SVM-ROT	CART	SVM-ROT	CART	SVM-ROT	CART
RMSE	**5.385**$_{(0.151)}$	5.396$_{(0.198)}$	**0.776**$_{(0.008)}$	0.795$_{(0.008)}$	**3.245**$_{(0.071)}$	3.367$_{(0.073)}$	1.506$_{(0.0521)}$	**1.445**$_{(0.058)}$
MAD	**4.422**$_{(0.109)}$	4.430$_{(0.170)}$	**0.597**$_{(0.008)}$	0.613$_{(0.005)}$	**2.404**$_{(0.061)}$	2.460$_{(0.068)}$	**0.898**$_{(0.027)}$	0.940$_{(0.027)}$
R^2	**0.604**$_{(0.022)}$	0.602$_{(0.031)}$	**0.674**$_{(0.008)}$	0.656$_{(0.005)}$	**0.833**$_{(0.008)}$	0.819$_{(0.008)}$	0.861$_{(0.007)}$	**0.871**$_{(0.011)}$

Table 1. *Results of SVM-ROT and CART for all four dataset. The means (standard errors) of the* 10*-times* 10*-fold cross-validation of the three metrics are reported. In bold the best model for each metric and dataset.*

CART for all three global metrics. Instead, in "Snow" the improvement is only for MAD, whilst for "Body Fat" the results are almost identical. Nevertheless, as at each node, the SVM-ROT splits the predictor space using all the covariates at once, so SVM-ROT is prone to overfit the data. In the future, it will be then interesting to use this novel model with an ensemble learning approach such as random forests or gradient boosting, or to apply a kind of feature selection at each split.

References

ANCELL, ETHAN. 2021. *autocart: Autocorrelation Regression Trees*. R package version 1.4.5.

BREIMAN, LEO. 2017. *Classification and regression trees*. Routledge.

DUA, DHEERU, & GRAFF, CASEY. 2017. *UCI Machine Learning Repository*.

GRISONI, FRANCESCA, CONSONNI, VIVIANA, VIGHI, MARCO, VILLA, SARA, & TODESCHINI, ROBERTO. 2016. Investigating the mechanisms of bioconcentration through QSAR classification trees. *Environment international*, **88**, 198–205.

PENROSE, KEITH W, NELSON, AG, & FISHER, AG. 1985. Generalized body composition prediction equation for men using simple measurement techniques. *Medicine & Science in Sports & Exercise*, **17**(2), 189.

R CORE TEAM. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

VAPNIK, VLADIMIR. 1999. *The nature of statistical learning theory*. Springer science & business media.

WICKRAMARACHCHI, D.C., ROBERTSON, B.L., REALE, M., PRICE, C.J., & BROWN, J. 2016. HHCART: An oblique decision tree. *Computational Statistics Data Analysis*, **96**, 12–23.

YANG, XULEI, SONG, QING, & WANG, YUE. 2007. A weighted support vector machine for data classification. *International Journal of Pattern Recognition and Artificial Intelligence*, **21**(05), 961–976.

COMPARING SOFT CLASSIFICATION METHODS FOR THE RARE TYPE MATCH PROBLEM

Giulia Cereda [1], Fabio Corradi[1] and Cecilia Viscardi[1]

[1] Department of Statistics, Computer Science and Applications, University of Firenze, (e-mail: giulia.cereda@unifi.it, fabio.corradi@unifi.it, cecilia.viscardi@unifi.it)

ABSTRACT: In this work, we compare five different methods proposed in forensic statistics to cope with the rare type match problem. This problem arises when the DNA profile of a suspect coincides with the profile from a crime sample, but it is not present in the available database collected from the population of reference. We compare the methods designed to evaluate the likelihood ratio in this framework by using a set of supervised cases and by considering each method as a classifier that provides the posterior probabilities of two alternative hypotheses, those of the prosecution and the defense, starting from a grid of prior probabilities. We compare them using the value of the posterior cross entropy and decompose it into two terms quantifying their calibration and refinement loss.

KEYWORDS: Forensic statistics, Soft Decisions, Empirical Cross Entropy

1 Introduction

The rare type match problem is the challenging situation faced by a forensic statistician who has to provide the value of a match between the characteristic ($\tilde{y}$) of a crime stain and that of a suspect when $\tilde{y}$ is not in a database of reference of size n. The information provided by evidence, y, is evaluated through a likelihood ratio that can lead to the posterior odds of the hypotheses formulated by the prosecution and the defense, $H \in \{h_p, h_d\}$, for a grid of prior probabilities. Several methods have been designed in the literature to cope with this problem when evidence consists of Y-STR profiles. We aim to compare these methods according to Bayesian decision theory, evaluating the expected cost of decisions expressed as posterior probabilities for the two hypotheses. Using strictly proper scoring rules as cost functions, the expected cost can be decomposed into two components corresponding to calibration and refinement, features of a classifier useful to guide the choice among alternative methods.

Y-STR are polymorphic loci on the Y-chromosome containing a repeated sequence of nucleotides. Individuals differ by the number of times the se-

quence appears at each locus. A Y-STR profile is a list of the numbers of repetitions at a finite number (typically 7 to 23) of loci. The Y-chromosome is only contributed by the father so that there is no recombination; the loci are dependent and cannot be modeled separately. For this reason, a profile must be considered as a whole, and, in case of a rare type match, no frequencies are available from the database to estimate the rarity of the $\tilde{y}$ profile.

2 Proposals for the LR evaluation in case of a rare type match

We want to compare methods that address the rare type match problem differently. We restrict ourselves to five methods assuming that the observed profiles are an i.i.d. sample and not assuming any genetic model; other possibilities exist, e.g. (Andersen *et al.*, 2013), but are not directly comparable.

A first group of methods copes with the rare type match problem by including the suspect profile in the reference data base:

- Augmented Count (AC), is a frequentist method for which:

$$\mathrm{LR}_{AC} = (n+1)/(n_{\tilde{y}}+1) = n+1,$$

  with $n_{\tilde{y}}$ equal to the frequency of $\tilde{y}$ in the database.
- The Bayesian AC, B-AC, (Cereda, 2017a) assumes that the frequency of $\tilde{y}$ in the database is distributed according to a $\mathrm{Bin}(n, \phi_{\tilde{y}})$, with $\phi_{\tilde{y}}$, the unknown probability of $\tilde{y}$ in the population, distributed according to a $\mathrm{Beta}(1,1)$ distribution. These assumptions yield to:

$$\mathrm{LR}_{B-AC} \quad = \quad (n+3)/(n_{\tilde{y}}+2) = (n+3)/2.$$

A second group looks at the list of profiles in the data, including the suspect profile, as partitioned into subsets containing the same Y-STR profile. Building upon different assumptions, the methods evaluate the LR by summarizing the data through π_{n+1}, the vector containing the cardinality of the subsets.

- The two-parameter Poisson Dirichlet method (2PD) (Cereda *et al.*, 2023) assumes an infinite number of Y-STR profiles in the population and that the vector of their ordered relative frequencies follows a 2PD distribution with parameters $\alpha \in (0,1)$ and $\theta > -\alpha$. Thus, the LR becomes:

$$\mathrm{LR}_{2PD} \quad = \quad \left[\int \frac{1-\alpha}{n+1+\theta} \, p(\alpha, \theta \mid \pi_{n+1}) d\alpha d\theta \right]^{-1}.$$

- The Generalized Good (GG) method (Cereda, 2017b) evaluates the LR:

$$\mathrm{LR}_{GG} = n n_1 / 2 n_2,$$

  with n_1 and n_2 the number of singletons and doublets in the database.

- The Brenner's kappa method (Bk) (Brenner, 2010) evaluates the LR as:

$$\mathrm{LR}_{Bk} = (n+1)^2/(n-n_1).$$

3 The posterior cross entropy and its decomposition

We use tools developed by Bayesian decision and information theory to evaluate the five reviewed proposals. Starting from an LR provided by a method $m \in \{\text{AC, B-AC, 2PD, GG, Bk}\}$, LR_m, the evaluation concerns the distribution $p_m(H \mid y)$ with $p_m(h_p \mid y) = \frac{LR_m O(H)}{1+LR_m O(H)}$, where $O(H) = \frac{p(h_p)}{p(h_d)}$ is the prior odds. We consider the log cost function, acting when H is known:

$$C[p_m(h|y)] = \begin{cases} -\log_2(p_m(h|y)) & \text{if } H = h \\ -\log_2(1-p_m(h|y)) & \text{if } H \neq h\,. \end{cases}$$

Since H is usually unknown, we must consider the expected cost corresponding to Shannon's Entropy of $H|y$.

In comparing methods, the mixing distribution of the costs, $p(\cdot|h)$, can be thought of as how Nature expresses the uncertainty on $Y|h$ and, consequently, via Bayes' theorem, on $H|y$. Moreover, we are interested in an average over all the possible evidence y which could arise from the population. This leads to the posterior cross entropy:

$$\mathcal{CE}_{p,p_m}(H \mid Y) = -\sum_{h\in\{h_p,h_d\}} p(h) \sum_{y\in\mathcal{Y}} p(y \mid h)\log(p_m(h|y)) = \mathcal{D}_{p,p_m}(H \mid Y) + \mathcal{E}_p(H \mid Y).$$

As a result, $\mathcal{CE}_{p,p_m}(H \mid Y)$ is the primary criterion of evaluation. The two other criteria are a) $\mathcal{D}_{p,p_m}(h \mid Y)$, the Kullback-Leibler divergence that quantifies the calibration loss, i.e., how the method puts forward posteriors on H in agree with Nature; b) $\mathcal{E}_p(H \mid Y)$, the posterior entropy that quantifies the refinement loss, i.e., the degree of sharpness provided in discriminating hypotheses. We denote the evidence generically by y, but different methods provide probability distributions based on different statistics. Unfortunately, we cannot directly compute the two terms in the decomposition since we have no access to $p(y|H)$, so we provide empirical estimates that require a strategy for building a database of supervised cases starting from a large sample from the population. The proposed solution is based on a Monte Carlo approach and relies on a Pool-Adjacent-Violators (PAV) algorithm that provides an approximate solution. Our results can be presented as the so-called $\mathcal{ECE}$-plot, showing each method's empirical posterior cross-entropy evaluated for different prior probabilities $p(h)$. An example is in Fig 3, where we can compare the

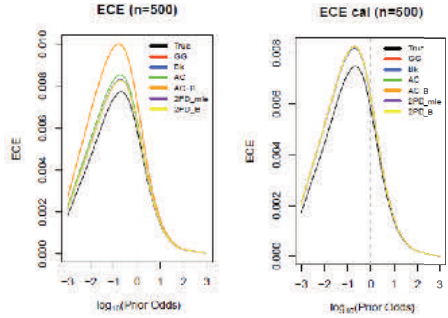


Figure 1. *$\mathcal{ECE}$ before (LHS) and after (RHS) the application the PAV algorithm.*

$\mathcal{ECE}$ of the five methods before and after applying the PAV algorithm. Fig 3 (left) shows that 2PD-B, exploiting π_{n+1}, achieves the smallest $\mathcal{ECE}$; while, the worst method is AC-B which uses only the size of the data and makes the lazy assumption of a flat prior distribution on the probability of the "rare" characteristic. Fig 3 (right) shows that, once recalibrated, all the methods have almost the same refinement so that the main differences attain calibration.

References

ANDERSEN, M. M., ERIKSEN, P. S., & MORLING, N. 2013. The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, **329**(7), 39–51.

BRENNER, C. H. 2010. Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics.*, **4**, 281–291.

BRÜMMER, N. 2010. *Measuring, refining and calibrating speaker and language information extracted from speech*. Ph.D. thesis, Stellenbosch University.

CEREDA, G. 2017a. Bayesian approach to LR in case of rare type match. *Statistica Neerlandica.*, **71**, 141–164.

CEREDA, G. 2017b. Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach. *Scandinavian Journal of Statistics.*, **44**, 230–248.

CEREDA, G., CORRADI, F., & VISCARDI, C. 2023. Learning the two parameters of the Poisson-Dirichlet distribution with a forensic application. *Scandinavian Journal of Statistics.*, **50**(1), 120 – 141.

BAYESIAN SHANNON ENTROPY ESTIMATION UNDER NORMALIZED INVERSE GAUSSIAN PRIORS VIA MONTE CARLO SAMPLING

Annalisa Cerquetti [1]

[1] (e-mail: annalisa.cerquetti@gmail.com,)

ABSTRACT: An analytical solution to Bayesian Shannon entropy estimation under general Gibbs-type priors has been devised in 2014 as a limiting case of Bayesian Tsallis entropy estimation. Here we propose a different approach and derive a Monte Carlo solution under normalized Inverse Gaussian prior relying on known results for its stick-breaking representation.

KEYWORDS: Shannon entropy, stick-breaking constructions, Bayesian nonparametrics, normalized Inverse Gaussian, Monte Carlo estimation.

1 Introduction

Interest in Shannon entropy $H_1(p) = -\sum_i p_i \log p_i$ as an index of diversity of a population of species arises in many different fields spanning from ecology, genetics, information theory, computer science, cryptography, neuroscience, linguistics and many others. It stands out among diversity indices for being "additive" and being, with its monotonic transformations, the only measure which weighs species (categories) in proportion to their population abundances p_i. The typical problem in estimating diversity indices from a finite set of experimental data is that relative abundances are *a priori* unknown and replacing them by sample relative frequencies, as in the maximum likelihood approach, produces negatively biased estimators, especially in biological communities where a large number of species has relatively small abundances and many of the rare species remain unobserved. A wide range of estimation methods have been proposed to overcome this drawback both in the Bayesian like in the frequentist approach. See Cerquetti (2014) and references therein for an account.

The normalized Inverse Gaussian prior, as introduced in Pitman (2003), is a Gibbs-type prior of parameter $\alpha = 1/2$ whose EPPF has the characteristic product form $p_{\alpha,V}(n_1,\ldots,n_k) = V_{n,k}\prod_{j=1}^{k}(1-\alpha)_{n_j-1}$, for $\alpha \in (-\infty,1)$ and Gibbs weights $V = (V_{n,k})$ satisfying the backward recursive relation $V_{n,k} =$

$(n-k\alpha)V_{n+1,k}+V_{n+1,k+1}$ where $V_{1,1}=1$ and $(x)_y=(x)(x+1)\cdots(x+y-1)$ is the usual notation for rising factorials. This class has been largely investigated during the last twenty years both from a theoretical like from an applied perspective mostly with respect to hierarchical mixtures modelling (see e.g. Lijoi et al. 2005, Favaro and Teh, 2013). Actually its implementation in diversity estimation, as an alternative to the Dirichlet prior and to the two-parameter Pitman-Yor prior, has received less attention.

2 Results

Moving from some preliminary results in Archer et al. (2014) the first general solution to Bayesian nonparametric estimation of Shannon entropy under Gibbs-type priors has been devised in Cerquetti (2014) as a limiting case from a result for *m*-generalized Tsallis entropies.

Proposition 1. (Cerquetti, 2014). *Let $P=(P_i)_{i\geq 1}$ be a random discrete distribution belonging to the (α,V) Gibbs-type family, then, for $\psi_0(\cdot)$ the digamma function, prior expected Shannon entropy is given by*

$$E_{\alpha,V}\left[H_1(P)\right]=-\lim_{m\to 1}\frac{\partial}{\partial m}V_{m,1}-\psi_0(1-\alpha). \tag{1}$$

Let $n=(n_1,\ldots,n_k)$ be the multiplicities of the first k species observed in a random sample of size n from P then expected posterior Shannon entropy is given by

$$\begin{aligned} E_{V,\alpha}\left[H_1(P)\mid n\right] &= -\frac{\sum_j(n_j-\alpha)}{V_{n,k}}\left[\lim_{m\to 1}\frac{\partial}{\partial m}V_{m+n,k}+\psi_0\left(n_j-\alpha+1\right)V_{n+1,k}\right] \\ &\quad -\frac{1}{V_{n,k}}\left[\lim_{m\to 1}\frac{\partial}{\partial m}V_{n+m,k+1}+\psi_0(1-\alpha)V_{n+1,k+1}\right]. \end{aligned} \tag{2}$$

Equations (1) and (2) imply the availability of the Gibbs weights $V=(V_{n,k})$ in a sufficiently tractable analytical form. But this doesn't always happen in the Gibbs-type class. The weights of the normalized Inverse Gaussian partion model, for example, are notoriously difficult to handle both analytically and computationally (see Lijoi et al. 2005, Arbel and Favaro, 2021) nevertheless, in such a case, Shannon entropy estimation can be faced by a different route. Cerquetti (2014) shows that if the distribution of the size-biased atoms of the specific Gibbs prior are known explicitly, like e.g. when a stick-breaking construction has been devised, moments of Shannon entropy can also be obtained mimicking the approach in Archer et al. (2013) for the two parameter

Poisson-Dirichlet priors. In the following results we rely on the stick-breaking construction of the normalized Inverse Gaussian prior as devised in Cerquetti (2022) moving from results in Pitman (2003). For T an inverse Gaussian r.v. of parameters $(1/v,1)$ and χ_1^2 and χ_2^2 two independent chi-square (1) r.v.s independent from T then

$$\tilde{P}_{1,nig} \stackrel{d}{=} \frac{\chi_1^2}{T^{-1}+\chi_1^2} \tag{3}$$

and

$$\tilde{P}_{2,nig} \stackrel{d}{=} \left(\frac{\chi_2^2}{T^{-1}+\chi_1^2+\chi_2^2}\right)\left(1-\frac{\chi_1^2}{T^{-1}+\chi_1^2}\right). \tag{4}$$

Given the first two size-biased picks from a normalized Inverse Gaussian prior $(v,1)$, prior expected first and second moment of Shannon entropy can be easily derived, for example, as

$$E_P[H_1(P)] = -E_{P_1}[\log(\tilde{P}_1)] \tag{5}$$

$$E[H_1(P)]^2 = E[\tilde{P}_1\left(\log\tilde{P}_1\right)^2] + E[\log\tilde{P}_1\log\tilde{P}_2(1-\tilde{P}_1)] - E[\left(\log\tilde{P}_1\right)^2]. \tag{6}$$

A preliminary investigation of the prior behaviour of Shannon entropy under NIG prior for different values of the parameter v can then be performed via Monte Carlo sampling.

As for the posterior expectation, i.e. the Bayesian estimator under quadratic loss function, we can state the following result.

Proposition 3. *Let $(n_1,\ldots,n_k)$ be the vector of the multiplicities of the k different species observed in a sample of size n from $P \sim NIG(v,1)$. Let $R_{n,k}$ be the posterior missing mass, $\tilde{P}_j|n_j$, for $j=1,\ldots,k$ the posterior relative abundances of the k observed species in order of appearance and $\tilde{Q}_1$ the first size-biased pick from the unseen species proportions, then*

$$E[H_1(P)|n_1,\ldots,n_k] = -E_{P|n}\left[\sum_{j=1}^{k}\tilde{P}_j|n_j\log(\tilde{P}_j|n_j)\right] - E_{R,Q_1}\left[R_{n,k}\log(R_{n,k}\tilde{Q}_1)\right]. \tag{7}$$

Equation (7) can be evaluated relying on Monte Carlo sampling. We just need to be able to simulate from the posterior missing mass, i.e. the proportion of the unseen species in the sample, from the first size-biased pick from the unseen species proportions and from the posterior relative abundances of

the species seen. It can be shown that the availability of the stick-breaking construction is enough for the task.

Remark 1. Results (5), (6) and (7) follow from standard arguments in the theory of size-biased permutations and their corresponding stick-breaking constructions. Here we omit explicit proofs for space limits but those will be provided - together with the details and the algorithm of the Monte Carlo sampling - in an extended version of this contribution in preparation (Cerquetti, 2023).

References

ARBEL, J., AND FAVARO, S. 2021. Approximating Predictive Probabilities of Gibbs-Type Priors. *Sankhya A*, **83**(1), 496–519.

ARCHER, E., PARK, I., AND PILLOW, J.W. 2014. Bayesian Entropy Estimation for Countable Discrete Distributions. *Journal of Machine Learning Research*, **15**(81), 2833–2868.

CERQUETTI, A. 2014. Bayesian nonparametric estimation of Tsallis diversity indices under Gnedin-Pitman priors. *arXiv:1404.3441v2 [math.ST]*.

CERQUETTI, A. 2022. On the first two size-biased picks from the normalized Inverse Gaussian prior. *Proceedings of the 36th IWSM, Trieste, Italy, 18–22 July 2022*.

CERQUETTI, A. 2023. Posterior analyisis of diversities and richness under normalized Inverse Gaussian priors via Monte Carlo sampling. *In preparation*.

FAVARO, S. AND TEH, YW. 2013. MCMC for Normalized Random Measure Mixture Models. *Statistical Science*, **28**, 335 – 359.

LIJOI, A. MENA, R. H., & PRÜNSTER, I. 2005. Hierarchical Mixture Modeling With Normalized Inverse-Gaussian Priors. *JASA*, **100**, 1278–1291.

PITMAN, J. 2003. Poisson-Kingman partitions. *IMS, Lecture Notes Monograph Series, Institute of Mathematical Statistics, Hayward, CA.*, **40**, 1–34.

GOODNESS–OF–FIT TEST FOR SINGLE FUNCTIONAL INDEX MODEL

Lax Chan [1] and Aldo Goia[1]

[1] Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Via Perrone 18, 28100, Novara, Italy(e-mail: lax.chan@uniupo.it, aldo.goia@uniupo.it)

ABSTRACT: Motivated by a problem that commonly arise in the food industry, a methodology based on the Single Functional Index Model (SFIM) is proposed and a test procedure to specify the link function between the real response and the functional covariate is described and applied.

KEYWORDS: Functional regression, Goodness–of–fit test, Single Index Model

1 Introduction

In the food industry, to obtain the composition of a given substance in terms of protein, fat, moisture, etc. is an important task. Since a full–scale chemical analysis is often costly and time consuming, it is preferred to estimate that composition by using spectrometric curves which can be obtained easily as the absorption of a reflected light for various wavelengths. In that situation a regression model with a scalar response (the percentage of the component) and a functional covariate (the spectrometric curve, or a transformation of it) can be profitably used. Consider for instance the prediction of the fat proportion by using the near-infrared spectra of 39 milk specimens obtained by SCiO device recorded between 740 and 1070 nm in Figure 1 and originally considered in Riu *et al.*, 2020. This dataset has been used Di Brisco *et al.*, 2023, where some functional parametric and nonparametric regression models have been applied and compared.

One can note that, if full nonparametric approaches are exploratory but suffer of dimensionality problems, parametric models are easily interpreted but not flexible. A useful alternative in this research field can come from the semi-parametric regression approaches that combine flexibility and interpretability. In particular the class of Single Functional Index Model (SFIM) defines a relationship between the functional predictor X and the real–valued random variable Y through an unknown real link function g that acts on a projection of the functional predictor along an unknown direction θ, subject to an identifiability condition: $Y = g(\langle X, \theta \rangle) + \mathcal{E}$, where $\langle X, \theta \rangle = \int X(t)\,\theta(t)\,dt$, $\|\theta\|^2 = 1$ and

$\theta(t) > 0$ for a fixed t. A methodology which combines a spline approximation of the functional coefficient θ and the one-dimensional Nadaraya-Watson approach to estimate the link function g are proposed in Ferraty *et al.*, 2013. The main advantage in using SFIM is the possibility to work in the one dimensional analogue of an infinite dimensional problem, through the projective strategy, and hence to visualise an estimate of g from the observed data and hence suggests the nature of the relationship of X and Y. This allows to postulate a target link function g_0 and test its compatibility with the observed data at a significance level.

The new test procedure in the SFIM context based on the conditional moment test approach has been defined and analyzed in Chan *et al.*, 2023. This work aims to summarize the main features of such a test and apply it to the spectrometric example. In particular, after illustrating the basic principle of the test in Section 2, the application to the real data is discussed in Section 3.

2 The test principle

Consider the SFIM and define $\mathcal{G}_0 = \{g_0^\beta : \mathbb{R} \to \mathbb{R}, \beta \in \mathbb{R}^{d+1}\}$, where g_0^β is a known function depending on the parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_d) \in \mathbb{R}^{d+1}$, $d \geq 1$ integer. Consider then the following hypothesis:

$$H_0 : g \in \mathcal{G}_0 \qquad \text{vs.} \qquad H_1 : g \in \mathcal{G}_1$$

where $\mathcal{G}_1$ is a set of real functions g_1^β such that $\mathcal{G}_1 \cap \mathcal{G}_0 = \emptyset$.

Define $\mathcal{E} = Y - g_0^\beta(\langle X, \theta \rangle)$ and $\mathbb{E}[\mathcal{E}|X] = g(\langle X, \theta \rangle) - g_0^\beta(\langle X, \theta \rangle)$. The quantity $Q = \mathbb{E}[\mathcal{E}\mathbb{E}[\mathcal{E}|X]\, w(X)]$, where $w(X) > 0$ is a weight function, is null under H_0 and strictly positive under H_1.

To implement the test procedure, an empirical version of Q has to be derived from a sample (X_i, Y_i), $i = 1, \ldots, n$ drawn from (X, Y). Assuming the projection random variable $\langle X, \theta \rangle$ admits a positive probability density function f_θ, then a possible choice for the weight function is $w = f_\theta$. By taking a Nadaraya–Watson type nonparametric kernel estimate of $\mathbb{E}[\mathcal{E}|X]$ at the point X_i and a cross–validated kernel estimate of f_θ, the empirical version of Q is:

$$Q_n\left(\widehat{\theta}\right) = \frac{1}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \widehat{\mathcal{E}}_i \widehat{\mathcal{E}}_j K_{ij}^{\widehat{\theta}},$$

where $\widehat{\theta}$ is an estimate of θ and $\widehat{\mathcal{E}}_i = Y - g_0^{\widehat{\beta}}\left(\left\langle X_i, \widehat{\theta} \right\rangle\right)$, where $\widehat{\beta}$ is an estimate

for β. The standardised test statistic is $T_n = n\sqrt{h}Q_n\left(\widehat{\theta}\right)/\nu_n\left(\widehat{\theta}\right)$ where

$$\nu_n^2\left(\widehat{\theta}\right) = \frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathcal{E}_i^2\mathcal{E}_j^2\left(K_{ij}^{\widehat{\theta}}\right)^2.$$

To compute the p-value and to derive the critical region of the test at the significance level α, the derivation of the asymptotic null distribution for T_n is required. Under appropriate assumptions, one can prove that $T_n \sim \mathcal{N}(0,1)$, as n diverges. Then one rejects the null hypothesis if $T_n \geq z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$-th quantile of the standard normal distribution. For further details, interested readers are invited to consult Chan *et al.*, 2023.

3 Application to spectrometric data

Consider the SFIM involving the original spectra as covariate and the quantitiy of fat as response. Some attempts with first and second derivatives of the spectrometric curves have been performed but with a deterioration in the quality of the prediction (and this is coherent with the models in Di Brisco *et al.*, 2023). In Figure 1 the estimates $\widehat{\theta}$ and $\widehat{g}$ of the direction θ and link function g are plotted. Observing the shape of the former, it seems that the relevant part of the spectrum in predicting the fat content is between about 950 and 1070 nm, whereas the latter suggests that a linear specification for the model seems not reasonable. For what concerns the prediction ability of that model, one used the RMSE, that is $\sum_i (y_i - \widehat{y}_i)^2 / \sum_i y_i^2$, and the MAPE, that is $\sum_i |y_i - \widehat{y}_i| / y_i$; the first index equals 0.015 and the second one 0.096.

At this stage it is possible to carry out the specification test; in particular the following polynomial and logistic null models are considered:

$$H_0^p : g_0(u) = \beta_0 + \sum_{j=1}^{p}\beta_j u \qquad H_0^{\log} : g_0(u) = e^{\beta_0+\beta_1 u}/(1+e^{\beta_0+\beta_1 u})$$

where $u = \left\langle x, \widehat{\theta}\right\rangle$ and $p = 1,2,3$ (corresponding to linear, quadratic and cubic link espectively). Since all the real parameters β_j are unknown, they are estimated by an OLS approach under the null hypothesis. The p-values calculated by using the asymptotic null distribution are: 0 for H_0^1, 0.035 for H_0^2, 0.207 for H_0^3 and 0 for $H_0^{\log}$. One can conclude that the linear, quadratic as well as logistic assumptions on the link function are not compatible with the empirical evidence, whereas a cubic link could be a good choice to model

the relationship. Therefore, a model to predict the content of fat Y in milk specimens starting from the spectrometric curve X can be specified as follows: $Y = 0.014 - 0.69 \cdot \langle X, \widehat{\theta} \rangle + 10.6 \cdot \langle X, \widehat{\theta} \rangle^2 - 33.2 \cdot \langle X, \widehat{\theta} \rangle^3 + \mathcal{E}$.

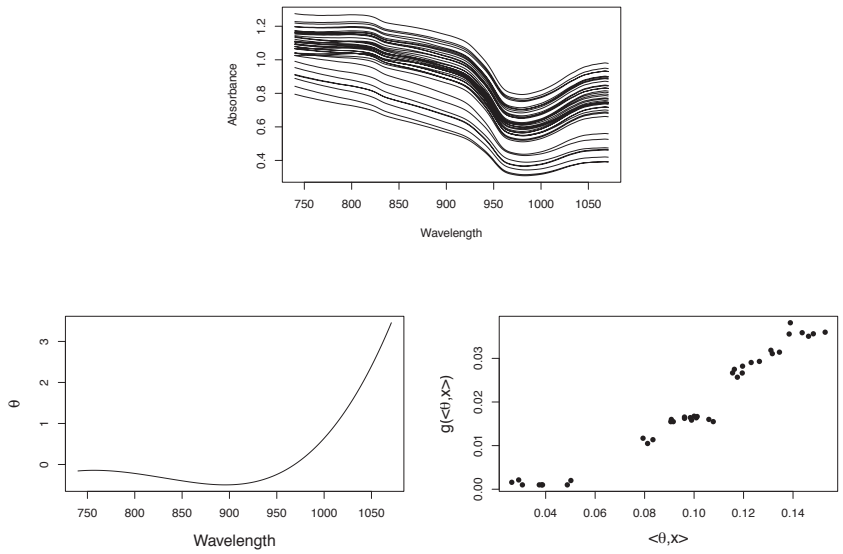


Figure 1. *Milk spectra recorded using SCiO device (top), Estimated direction* θ *(bottom left) and estimated link function (bottom right) for the SFIM.*

References

CHAN, L., DELSOL, L. & GOIA, A. 2023. A Link Function Specification Test in the Single Functional Index Model. *Advances in Data Analysis and Classification*. https://doi.org/10.1007/s11634-023-00545-7

DI BRISCO, A.M., BONGIORNO, E.G., GOIA, A. & MIGLIORATI, S. 2023. Bayesian flexible beta regression model with functional covariate. *Computational Statistics* **38**, 623-–645 .

FERRATY, F., GOIA, A., SALINELLI, E. & VIEU, P. 2013. Functional Projection Pursuit Regression, *Test* **22** , 293–320.

RIU, J., GORLA, G., CHAKIF, D., BOUQUÉ, R. & GIUSSANI, B. 2020. Rapid analysis of milk using low-cost pocket-size NIR spectrometers and multivariate analysis. *Foods*, **9** (8), 1090.

MULTILEVEL CROSS-CLASSIFIED LATENT CLASS MODELS

Silvia Columbu [1], Nicola Piras [1] and Jeroen K. Vermunt[2]

[1] Department of Mathematics and Computer Science, University of Cagliari, (e-mail: silvia.columbu@unica.it, nicola.piras97@unica.it)

[2] Department of Methodology and Statistic, Tilburg University, (e-mail: J.K.Vermunt@tilburguniversity.edu)

ABSTRACT: We propose an extension of latent class models to deal with multilevel cross-classified data structures, where each observation is considered simultaneously nested within two groups, such as for instance, children within both schools and neighborhoods. We show how such a situation can be dealt with by having a separate set of mixture components for each of the crossed classifications. Unfortunately, given the intractability of the derived loglikelihood, the EM algorithm can no longer be used in the estimation process. We therefore propose an approximate estimation of this model using a stochastic version of the EM algorithm similar to Gibbs sampling.

KEYWORDS: Latent class, cross-classified, Stochastic EM

1 Introduction

Latent class analysis (LCA) is a popular model-based approach for data clustering of units on the basis of observations arising from a set of categorical indicators. When the data have a multilevel hierarchical structure with units nested within higher level observations, such as children nested within schools, a possible extension (Laird, 1978) discussed in Vermunt, 2003 and Vermunt, 2008 takes two levels of clustering with separate latent variables for lower-level units and higher-level ones. Sometimes data have a cross-classified structure with units grouped within multiple higher level units, for example, children can be considered nested within both schools and neighborhoods. In this contribution we propose to extend Multilevel Latent Class analysis to handle cross-classification. Given the untractability of the derived likelihood the standard EM algorithm can not be applied in the estimation, and we propose to use a stochastic version of the EM algorithm that can handle the hierarchy of units but also their double cross-classification, similar to what done in Keribin *et al.*, 2015 for coclustering.

2 Model definition

Let Y_{ijkq} be the response on categorical indicator (or item) i ($i = 1,\ldots,I$) of individual or first level unit j ($j = 1,\ldots,n_{kq}$) belonging simultaneously to the group level units k ($k = 1,\ldots,K$) and q ($q = 1,\ldots,Q$). We denote with X_{jkq}, W_k and Z_q the discrete latent variables respectively for membership of level-1 units and for the two group level units. A particular latent class will be indicated with ℓ ($\ell = 1,\ldots,L$), for level-1 units, h ($h = 1,\ldots,H$) and r ($r = 1,\ldots,R$) for level-2 units. For ease of notation, we focus on binary indicators and denote with $\pi_{i|\ell}$ the probability distribution parameters of each item within the first level latent class. The data model consists of two parts, described through two separate equations, one for the level-2 cross-classified (or higher level) units and one for the level-1 (or lower level) units. Each of the two equations is a mixture of probabilities. The model for the higher part is described, in the complete data form, by

$$\begin{aligned} P(\mathbf{Y}_{kq}, W_k = h, Z_q = r) &= P(W_k = h, Z_q = r) P(\mathbf{Y}_{kq} | W_k = h, Z_q = r) \\ &= P(W_k = h, Z_q = r) \prod_{j=1}^{n_{kq}} P(\mathbf{Y}_{jkq} | W_k = h, Z_q = r) \\ &= P(W_k = h) P(Z_q = r) \prod_{j=1}^{n_{kq}} P(\mathbf{Y}_{jkq} | W_k = h, Z_q = r). \end{aligned}$$

We assumed independence of observations within a combination of groups given their belonging to the cross-classified latent classes, and also marginal independence of the two higher level latent classes W_k and Z_q.
The second part models the density of observations conditionally to their simultaneous belonging in higher level cross-classified latent classes, that is:

$$P(\mathbf{Y}_{jkq} | W_k = h, Z_q = r) = \sum_{\ell=1}^{L} P(X_{jkq} = \ell | W_k = h, Z_q = r) \prod_{i=1}^{I} P(Y_{ijkq} | X_{jkq} = \ell),$$

in which we have assumed the local independence of indicators within latent classes.

3 Parameters' Estimation

The estimation of model parameters $\boldsymbol{\theta} = \{\pi_{\ell|hr}, \pi_h, \pi_r, \pi_{i|\ell}\}$, requires the maximization of the observed likelihood of the model in the form

$$L(\boldsymbol{\theta}; \mathbf{y}) = \sum_{h_1=1}^{H} \sum_{h_2=1}^{H} \cdots \sum_{h_K=1}^{H} \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} \cdots \sum_{r_Q=1}^{R} \prod_{k=1}^{K} P(W_k = h_k) \prod_{q=1}^{Q} P(Z_q = r_q) \times$$

$$\prod_{j=1}^{n_{kq}} \left[\sum_{\ell=1}^{L} P(X_{jkq} = \ell | W_k = h_k, Z_q = r_q) \prod_{i=1}^{I} P(Y_{ijkq} | X_{jkq} = \ell) \right].$$

The presence of a double missing data structure at higher level, with W_k and Z_q unobserved, causes that the likelihood cannot factorize as a product of the mixing probabilities as for standard LC and multilevel LC models. The likelihood becomes easily untractable and standard EM algorithms cannot be directly applied for its maximization. We propose to consider a Stochastic version of the algorithm with the inclusion of a Gibbs sampling scheme between the E and the M step. The Stochastic step consists in the consecutive sampling from marginal posterior distributions of higher level and lower level latent classes, which reduces the computational burden.

E and S step

After initialization of $\pi_h = P(W_k = h)$, $\pi_r = P(Z_q = r)$, $\pi_{\ell|hr} = P(X_{jkq} = \ell | W_k = h, Z_q = r)$ and $\pi_{i|\ell}$ iterate the following sampling steps

1) Draw $\mathbf{w}^{(t)}$ from a Multinomial distribution with probabilities

$$P(W_k = h | \mathbf{y}_k, \mathbf{z}^{(t-1)}) = \frac{\pi_h P(\mathbf{Y}_k | \mathbf{z}^{(t-1)}, W_k = h)}{P(\mathbf{Y}_k | \mathbf{z}^{(t-1)})},$$

$$P(\mathbf{Y}_k | \mathbf{z}, W_k = h) = \prod_{q_k=1}^{Q_K} \prod_{r=1}^{R} \left[\prod_{j=1}^{n_{kq}} P(\mathbf{Y}_{jkq} | W_k = h, Z_q = r) \right]^{z_q^r};$$

2) Draw $\mathbf{z}^{(t)}$ from a Multinomial distribution with probabilities

$$P(Z_q = r | \mathbf{y}_q, \mathbf{w}^{(t)}) = \frac{\pi_r P(\mathbf{Y}_q | \mathbf{w}^{(t)}, Z_q = r)}{P(\mathbf{Y}_q | \mathbf{w}^{(t)})},$$

$$P(\mathbf{Y}_q | \mathbf{w}, Z_q = r) = \prod_{k_q=1}^{K_Q} \prod_{h=1}^{H} \left[\prod_{j=1}^{n_{kq}} P(\mathbf{Y}_{jkq} | W_k = h, Z_q = r) \right]^{w_k^h};$$

3) Draw $\mathbf{x}^{(t)}$ from a Multinomial distribution with probabilities

$$P(X_{jkq} = \ell | \mathbf{y}_{jkq}, \mathbf{w}^{(t)}, \mathbf{z}^{(t)}) = \frac{\left[\pi_{\ell|h,r} P(\mathbf{Y}_{jkq} | X_{jkq} = \ell) \right]^{w_{jk}^h z_{jq}^r}}{P(\mathbf{Y}_{jkq})},$$

where w_k^h, z_q^r, w_{jk}^h, z_{jq}^r and x_{jkq}^ℓ are all binary indicators of units' membership at different levels, in particular w_{jk}^h, z_{jq}^r are the expansion of higher level latent class indicators over the first level units j.

M step

$$\pi_h = \frac{\sum_{k=1}^{K} w_k^{h(t)}}{K}, \quad \pi_r = \frac{\sum_{q=1}^{Q} z_q^{r(t)}}{Q},$$

$$\pi_{\ell|hr} = \frac{\sum_{j=1}^{n} w_{jk}^{h(t)} z_{jq}^{r(t)} x_{jkq}^{\ell(t)}}{\sum_{j=1}^{n} w_{jk}^{h(t)} z_{jq}^{r(t)}}, \quad \pi_{i|\ell} = \frac{\sum_{j=1}^{n} x_{jkq}^{\ell(t)} y_{ijkq}}{\sum_{j=1}^{n} x_{jkq}^{\ell(t)}}.$$

Final estimates are calculated as the mean over the total number of iterations, burn-in period excluded.

Results from simulation studies with data generated under varying scenarios, prove that the estimators have satisfactory finite sample properties. In figure 1 is reported the error resulting from the estimation of $\pi_{\ell|h=1,r=1}$ over 50 binary simulated datasets with fixed number of classes L=4, H=R=2. Two scenarios of moderate increasing separation have been compared. It emerges that the average across replications is close to the true value, with an improvement with the increase of the number of groups. Similar results are observed for the other first-level and distribution parameters. Almost no error is observed for high-level latent class parameters. In the implementation of the SEM-Gibbs 150 iterations have been considered, including 50 burn-in. These are sufficient for convergence.

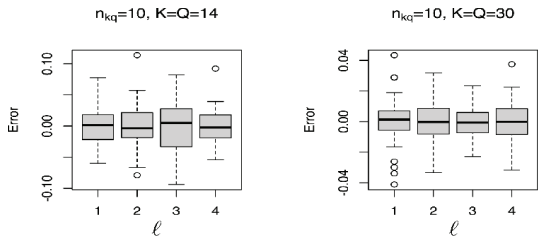


Figure 1. *Error on the estimation of* $\pi_{\ell|h=1,r=1}$.

References

KERIBIN, C., BRAULT, V., CELEUX, G., & GOVAERT, G. 2015. Estimation and selection for the latent block model on categorical data. *Statistics and Computing.*, **25**(6), 1201–1216.

LAIRD, N. 1978. Nonparametric Maximum-Likelihood Estimation of a Mixture Distribution. *Journal of the American Statistical Association.*, **73**, 805–811.

VERMUNT, J. K. 2003. Multilevel latent class models. *Sociological Methodology.*, **33**, 213–239.

VERMUNT, J. K. 2008. Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research.*, **17**, 33–51.

MULTIVARIATE REGRESSION TREE TO INVESTIGATE THE ITALIAN MORTALITY RATES

Giulia Contu [1], Luca Frigau[1], Marco Ortu [1] and Sara Pau [2]

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

[2] Department of Business and Economics, University of Sassari,(e-mail: spau@uniss.it)

ABSTRACT: Multivariate Regression Tree is a tree where univariate response variable has been substituted by a multivariate response variable. It has been proposed to investigate complex ecological data. We apply the Multivariate Regression Tree to investigate social and economical issues in order to comprehend if this method can be generalized and used in different research fields. We apply the Multivariate Regression Tree to identify the causes of death in Italian Counties in 2019. The first results evidence the capacity of Multivariate Regression Tree to define nodes characterized for specific causes of death and to classify together geographical areas with similar impact levels of the variables.

KEYWORDS: Multivariate regression tree; semi-supervised clustering; causes of deaths

1 Introduction

Tree-based methods define a wide set of methodologies finalized to partition the features' space in different areas to realize classification and regression analysis (De'ath & Fabricius, 2000). The aim is to obtain a subset more homogeneous compared to the initial set. A tree can be *univariate* or *multivariate*. The adjective "multivariate" is used to define two approaches. The first is related to the use of more than one attribute in the partition of the observations. The second is characterized by the introduction of the model of one outcome variable composed of more than one level.
In this paper, we focus on the second approach and on its possible use in the economical field. Specifically, we focus on the Multivariate regression tree (MRT) proposed by (De'Ath, 2002). It is *a natural extension of univariate regression trees, with the univariate response of the latter being replaced by a multivariate response* (De'Ath, 2002, p. 1106). The method has been proposed to investigate, describe and predict the relationship between the multi-species

data and the environmental characteristics. It is structured to analyze the community data without making assumptions about the form of relationship between the species and their environment. The nodes identified with MRT are characterized by the presence of a reduced number of species and a habitat with specific environmental characteristics. To our knowledge, this method has been used only to analyze complex ecological data. We attempt to use it to investigate a social, medical and economical issues. More in detail, we apply MRT to comprehend which aspects can impact on the number of deaths in a specific area. We focus on the data of Italian Counties in 2019.
Three sections, besides the introduction, complete this study. Firstly, the MRT methodology has been presented. Secondly, the results have been proposed and, finally, some concluding remarks are highlighted.

2 Methodology

MRT transforms the univariate tree into a multivariate including in the model a multivariate response and redefining the impurity of the node. De'Ath, 2002 has proposed two different measures of impurity. In the first case, MRT operates using an impurity measure called *sums of squared distances* (SSD) and minimizing the SSD of sites from the centroids of the nodes to which they belong. The sum of squares multivariate tree (SS-MRT) has been calculated through the following formula: $\sum_{ij}(x_{ij} - \bar{x_j})$, where x_{ij} is the species data for site i and species j and $\bar{x_j}$ is the mean. The measure can also be calculated considering the median value. In the second case, MRT is built using a dissimilarity matrix and considering the dissimilarities as a distance measure. The nodes are defined as minimizing the intersite sums of squared distances within the clusters. The impurity measure is defined as: $\sum_{i>kk} d_{ik}^2$, where d_{ik}^2 identifies the squared dissimilarities between sites i and k. The MRT built using the first impurity measure can be considered a form of the multivariate regression. Instead, the MRT built using distance measures can be considered a method of constrained clustering, because it allows obtaining clusters that are similar with respect to a measure of species dissimilarity. In both cases, it allows identifying nodes that are characterized for the presence of a reduced number of species and a habitat with specific environmental characteristics.
In this paper, we focus on SS-MRT. We define two different models: the response variable is defined by the number of deaths distinct for disease and gender, the covariates are related to the characteristics of the counties as the percentage of degrees, and the number of specialists, as explained in Table 1

Table 1. *Variables names and description. The AMR (Adjusted Mortality Rate) acronyms suffix denotes the target variable ($\mathcal{Y}$), all other variables are considered as predictors ($\mathcal{X}$).*

Variable name	**Extended label**
AMRm	Adjusted mortality rate from diseases (males)
AMRf	Adjusted mortality rate from system diseases (females)
Doctors rate	Rate of doctors enrolled in professional register
Graduates	Percentage of graduates over population
Eployment rate	Employment rate 15-64 M+F
Pop	Population
Aging index	(Pop65+)/(Pop0-14)*100
% specialists	Percentage of active doctors per indicated specialization in the health system per 10,000 inhabitants
VA	Value added per person (current prices)

3 Conclusion

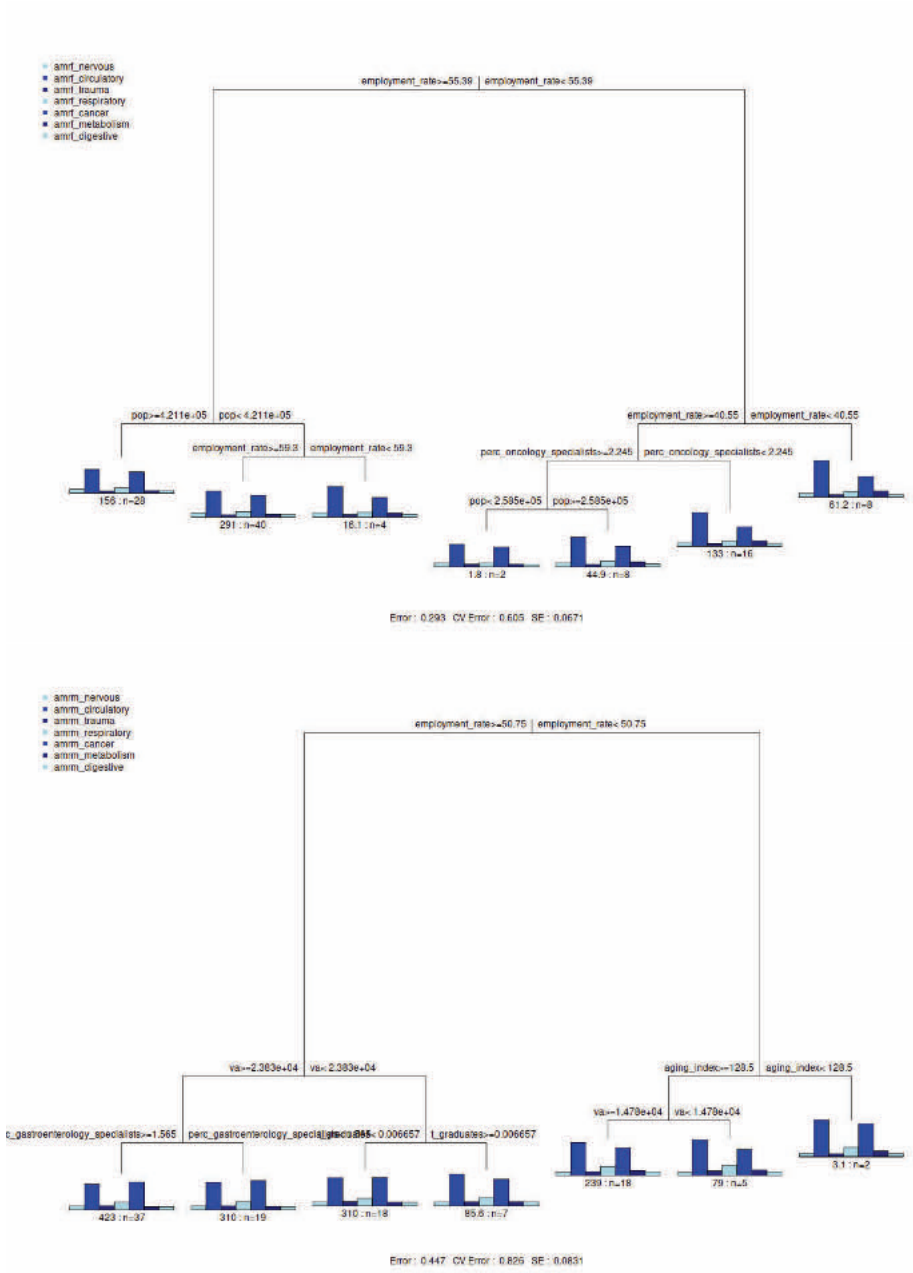
We use the MRT to investigate the elements that can impact the number of death in Italian counties. From a methodological point of view, our study highlights the importance of using advanced statistical methods to analyze the complex dataset and interpret the findings to obtain meaningful insights. From a managerial perspective, our results highlight which aspect can reduce the mortality rates and support the healthcare policy in allocation decisions.

References

DE'ATH, GLENN. 2002. Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*, **83**(4), 1105–1117.

DE'ATH, GLENN, & FABRICIUS, KATHARINA E. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**(11), 3178–3192.

Figure 1. *The causes of death of Italian women and men, 2019*



EMPIRICAL ANALYSIS OF THE QUADRATIC SCORING FOR SELECTING CLUSTERING SOLUTIONS

Luca Coraggio [1], Pietro Coretto[2]

[1] Department of Economics and Statistics, University of Naples Federico II, (e-mail: luca.coraggio@unina.it)

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

ABSTRACT: Selecting an optimal clustering solutions is a difficult problem, and there exist many data-driven validation strategies to perform this task. In this paper, we focus on a recent proposal, the BQH and BQS criteria, based on quadratic discriminant scores and bootstrap resampling. We provide more insight on these criteria, comparing them with a likelihood-based alternative and using different resampling schemes.

KEYWORDS: cluster validation, mixture models, model-based clustering, resampling methods

1 Quadratic scoring, likelihood-based scoring, and resampling

Selecting an optimal clustering solution is not an easy task (von Luxburg *et al.*, 2012). Recently, in Coraggio & Coretto, 2023, we proposed a novel validation index aimed at selecting clustering solutions in cases where clusters can be expected to have elliptic-symmetric shapes, or to be separable by quadratic boundaries.

Let $\mathbb{X}_n$ indicate sample data, and $\mathcal{G}^{(m)} = \left\{ G_k^{(m)},\ k = 1,\ldots,K_m \right\}$ be a clustering solution, obtained running clustering method $m \in \mathcal{M}$. We assume that $\mathcal{G}^{(m)}$ can be meaningfully described by K_m triplets $\boldsymbol{\theta}^{(m)} = \left\{ \boldsymbol{\theta}_k^{(m)},\ k = 1,\ldots,K_m \right\}$, each collecting unique elements of *(i)* π_k, the expected fraction of points belonging to the *k-th* group; *(ii)* $\boldsymbol{\mu}_k \in \mathbb{R}^p$, the *k-th* cluster's center; *(iii)* $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ a positive definite scatter matrix. For a point $\boldsymbol{x}$ and a triplet $\boldsymbol{\theta}_k$, we define the quadratic score (inspired to Quadratic Discriminant Analysis; e.g., see Hastie *et al.*, 2009) of point $\boldsymbol{x}$ for the *k-th* cluster as

$$\mathrm{qs}(x, \boldsymbol{\theta}_k) = \log(\pi_k) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma}_k)) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k); \qquad (1)$$

it can be seen as a measure of how well point $\boldsymbol{x}$ is accommodated into cluster k. The hard (QH) and (QS) smooth scores are based on (1), and are essentially

Algorithm 1 Bootstrap likelihood-based scoring

input: observed sample $\mathbb{X}_n$ (with ecdf $\mathbb{F}_n$), $\alpha \in (0,1)$; clustering method $m \in \mathcal{M}$; integers $B > 0$

output: bootstrap likelihood-based scoring for method m: $\widetilde{L}_n^{(m)}$.

(to ease notation, dependence on m is dropped and reintroduced in step 3)

for $b \in \{1,\ldots,B\}$ do

(step 1.1) $\mathbb{X}_n^{(b)} \leftarrow$ non-parametric bootstrap resample from $\mathbb{X}_n$ (sample of size n from $\mathbb{F}_n$)

(step 1.2) $\hat{\boldsymbol{\theta}}_n^{(b)} \leftarrow$ triplets of parameters from clustering solution m fitted on $\mathbb{X}_n^{(b)}$

(step 1.3) $S_n^{(b)} \leftarrow l(\hat{\boldsymbol{\theta}}_n^{(b)}; \mathbb{X}_n)$ (score solution on $\mathbb{X}_n$)

end for

(step 2) $\widetilde{W}_n \leftarrow \frac{1}{B}\sum_{b=1}^{B} S_n^{(b)} \qquad R_n^{(b)} \leftarrow \sqrt{n}\left(S_n^{(b)} - \widetilde{W}_n\right)$

(step 3) Compute $(\alpha/2)$-level and $(1-\alpha/2)$-level empirical quantiles:

$$\widetilde{L}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{R_n^{*(b)} \le t\right\} \ge \frac{\alpha}{2} \right\}; \qquad \widetilde{U}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left\{R_n^{*(b)} \le t\right\} \ge 1-\frac{\alpha}{2} \right\}$$

weighted averages of the quadratic score (see Coraggio & Coretto, 2023 for details). The quadratic score (1) is strongly connected to likelihood theory, and it is easy to show that it is proportional to the Gaussian density function. Thus, as a natural alternative to the scoring criteria we use the following likelihood function

$$l(\boldsymbol{\theta}^{(m)};\, \mathbb{X}_n) = \frac{1}{n}\sum_{\boldsymbol{x}\in\mathbb{X}_n} \log\left(\sum_{k=1}^{K^{(m)}} \pi_k^{(m)} \phi(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)})\right), \tag{2}$$

where $\phi(\boldsymbol{x}, \boldsymbol{\theta}_k^{(m)})$ is the density function of a multi-variate Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$.

Choosing the solution that maximizes (2) may give poor results: since the sample data $\mathbb{X}_n$ is used both to estimate $\boldsymbol{\theta}^{(m)}$ and for scoring, overly-complex solutions may be selected due to overoptimism in the evaluation process. Thus, we use the same resampling scheme used for the BQH and BQS scores, proposed in Coraggio & Coretto, 2023, that is to estimate clustering solutions on non-parametric bootstrap resamples (Efron, 1979) from $\mathbb{X}_n$, while using the full data to evaluate the score. The procedure is reviewed in Algorithm 1 for the likelihood-based scoring criterion.

2 Empirical analysis

The experimental analysis is a scaled-down version of that in Coraggio & Coretto, 2023, using the Pentagon5, T510D and Uniform simulated data sets.

Table 1: Selected solution by selection criteria (left-most column). Each subtable shows results from a data set: the first column shows the selected solution, and the second column reports its ARI, computed against true classes.

	(b) Pentagon5		(c) T510D		(d) Uniform	
Criterion	Selected m	ARI	Selected m	ARI	Selected m	ARI
QH	M, K=3, VVV	**0.86**	O, K=10, γ=10^4	0.51	O, K=10, γ=10^3	0
QS	M, K=3, VVV	**0.86**	O, K=10, γ=10^4	0.51	M, K=8, VVV	0
LK	O, K=10, γ=10^3	0.44	O, K=10, γ=10^4	0.51	O, K=10, γ=10^3	0
CVQH	M, K=3, EVE	**0.86**	O, K=6, γ=1	0.73	O, K=5, γ=10^2	0
CVQS	M, K=3, EVE	**0.86**	O, K=5, γ=1	0.97	M, K=1, EEI	**1**
CVLK	M, K=5, EVI	0.86	O, K=8, γ=1	0.60	M, K=7, VEE	0
BQH	M, K=3, EVE	**0.86**	O, K=8, γ=5	0.57	O, K=9, γ=10^4	0
BQS	M, K=3, EVE	**0.86**	O, K=5, γ=5	**0.98**	M, K=1, EEI	**1**
BLK	O, K=5, γ=1	0.85	O, K=8, γ=5	0.57	M, K=10, VVI	0

Since likelihood-based scoring is only justified for model-based clustering, $\mathcal{M}$ includes: *(i)* 140 Gaussian mixture models with covariance matrices restrictions (Banfield & Raftery, 1993), implemented with the Mclust (M) software (Scrucca *et al.*, 2016; setting $K = 1, \ldots, 10$, and 14 covariance models); *(ii)* 180 Gaussian mixture models with eigen-ratio contraints (ERC; Ingrassia, 2004), implemented with Otrimle (O) software (Coretto & Hennig, 2017, Coretto & Hennig, 2021; setting $K \in \{1, \ldots, 10\}$, ERC $\gamma \in \{1, 5, 10, 10^2, 10^3, 10^4\}$, and 3 initialization methods). The criteria compared to select optimal solutions are as follows. QH, QS, and LK: clustering solutions are estimated and scored using the full data, $\mathbb{X}_n$; CVQH, CVQS, CVLK: clustering solutions are estimated on a "train set" and scored on a non-overlapping "test set", using a 10-fold cross-validation scheme, as in Smyth, 2000. BQH, BQS, BLK: clustering solution are estimated and scored according to Algorithm 1, selecting the method m maximizing $\widetilde{L}_n^{(m)}$. For each criterion, the selected solutions are evaluated against the true class labels, reporting the achieved Adjusted Rand Index (ARI, Hubert & Arabie, 1985).

Results are presented in Table 1. The comparison gives a better understanding on the mechanism that lies behind the effectiveness of the BQH and BQS criteria. First, notice that all criteria where solutions are estimated and scored on the full data (QH, QS, LK) always select overly-complex solutions. The extra penalization of the smooth score on overlapping clusters is key to select better solutions in more complicated settings (T510D and Uniform). Finally, the bootstrap scheme improves on the cross-validation. Overall, both the quadratic scores, QH and QS, and the resampling scheme in Algorithm 1 seem

equally important to consistently achieve good results.

3 Conclusion

In this paper, we run an empirical comparison of the BQH and BQS procedures from Coraggio & Coretto, 2023 with a likelihood-based alternative, using different resampling schemes. Our experiments provide new insights on the criteria, showing that both the bootstrap resampling scheme and the quadratic scores contribute equally to the procedure: *(*i) the penalization for clusters' overlap from the quadratic scores allows achieving better results in cases where clusters are not well separated; *(ii)* the bootstrap resampling scheme allows to effectively take into account clustering methods' variability, better than cross-validation would (likely better suited for prediction settings).

References

BANFIELD, JEFFREY D., & RAFTERY, ADRIAN E. 1993. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, **49**(3), 803.

CORAGGIO, LUCA, & CORETTO, PIETRO. 2023. Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score. *Journal of Multivariate Analysis*, **196**(July), 105181.

CORETTO, PIETRO, & HENNIG, CHRISTIAN. 2017. Consistency, Breakdown Robustness, and Algorithms for Robust Improper Maximum Likelihood Clustering. *Journal of Machine Learning Research*, **18**(142), 1–39.

CORETTO, PIETRO, & HENNIG, CHRISTIAN. 2021. *otrimle: Robust Model-Based Clustering*. R package version 2.0.

EFRON, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**(1).

HASTIE, TREVOR, TIBSHIRANI, ROBERT, & FRIEDMAN, JEROME. 2009. *The Elements of Statistical Learning*. 2 edn. Springer Series in Statistics (SSS). Springer New York.

HUBERT, LAWRENCE, & ARABIE, PHIPPS. 1985. Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

INGRASSIA, SALVATORE. 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods & Applications*, **13**(2).

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), 289–317.

SMYTH, PADHRAIC. 2000. *Statistics and Computing*, **10**(1), 63–72.

VON LUXBURG, ULRIKE, WILLIAMSON, ROBERT C., & GUYON, ISABELLE. 2012. Clustering: Science or Art? Proceedings of Machine Learning Research, vol. 27. Bellevue, Washington, USA: PMLR.

CLASSIFICATION OF DAILY STREAMFLOW DATA: A STUDY ON REGIME CHANGES

Corduas Marcella [1], Domenico Piccolo[1]

[1] Department of Political Sciences, University of Naples Federico II
(e-mail: marcella.corduas@unina.it, domenico.piccolo@unina.it)

ABSTRACT: This contribution presents a classification strategy, based on widely available statistical tools, for detecting time series that have changed flow regime in recent years. The results from the analysis of 221 time series of unregulated streamflows in the United States is discussed.

KEYWORDS: time series, classification, flow regime, AR metric.

1 Introduction

The climate change often affects the variability and persistence of river discharges that may show an alterated balance between snow and rainfall and an intensification of extreme hydrological events. Such climate-induced hydrologic changes may have relevant consequences on the freshwater ecosystem (Dhungel *et al.* , 2016). The search of simple but effective tools for river regime classification is still a topic of interest in order to investigate variations in flow regimes and evaluate future climate impact (Yang & Olivera, 2023). In this article, we present a procedure for classifying streamflow time series according to their underlying dynamic structures. We illustrate our approach analyzing streamflow data from 221 unregulated catchments in the United States (Newman & al., 2015).

2 Methods

Streamflow time series are typically characterized by a marked seasonal pattern, due to the alternating of wet and dry periods, and a persistent or long term component. The seasonality often appears as a deterministic component in the spectrum. This makes the time series unsuitable for stochastic modelling, because the marked seasonal pattern obscures the other dynamic components. At this stage, we assume that the effect of data skewness, calendar effects, outliers and missing value have already been removed by preliminary analysis

and transformations and that the time series W_t has zero mean. Thus, W_t is described by the harmonic regression model:

$$W_t = \sum_{j=1}^{[s/2]} [\alpha_{wj} \sin(2\pi jt/s) + \beta_{wj} \cos(2\pi jt/s)] + Z_t \quad (1)$$

where s denotes the seasonal period, and Z_t follows a stationary Autoregressive model, *AR(p)*:

$$\varphi(B) Z_t = a_t, \quad (2)$$

where a_t is a Gaussian White Noise (WN) process with constant variance σ^2_{aw}. It is well known that any process with an absolutely continuous spectrum can be adequately approximated by an Autoregressive model, then (2) describes both short and long memory stationary components. The order p can be selected by BIC criterion, so that parsimonious models are preferred. Thus, the time series W_t is characterized by the coefficients estimated by GLS: $\widehat{\delta}_w = (\hat{\alpha}_{w1}, ..., \hat{\alpha}_{wk}, \hat{\beta}_{w1}, ..., \hat{\beta}_{wk})'$ and $\widehat{\varphi}_w = (\hat{\varphi}_{w1}, ..., \hat{\varphi}_{wp})'$.

Given two independent time series W_t and Y_t, the dissimilarity will be measured by comparing the seasonal and non-seasonal coefficients separately because, as already mentioned, the two components (seasonality and inertia) have a very different weight in determining the dynamics of the series.

Seasonal components are compared by evaluating the Mahalanobis distance: $M_{wy} = (\widehat{\delta}_w - \widehat{\delta}_y)'(\sigma^2_{aw}\Omega_w + \sigma^2_{ay}\Omega_y)^{-1}(\widehat{\delta}_w - \widehat{\delta}_y)$, where $\sigma^2_{a\bullet}\Omega_\bullet$ is the covariance matrix of $\widehat{\delta}_\bullet$. The dissimilarity between the residual components is measured by means of the *AR* metric (Piccolo, 1990; Corduas & Piccolo, 2008): $D_{wy} = \sqrt{\sum_{j=1}^{\infty} (\varphi_{wj} - \varphi_{yj})^2}$.

Then, the corresponding distance matrices $\mathcal{M}$ and $\mathcal{D}$ are objects of a clustering algorithm in order to identify groups of time series having similar seasonal pattern and different level of inertia. Here, we use the complete linkage method because it does not require the preliminary specification of the cluster number and produces compact clusters.

3 Results

The analysis has been conducted on 221 time series of mean daily discharge (feet^3/sec) of unregulated streamflows in the United States (available from the US Geological Survey at https://waterdata.usgs.gov/nwis/). Two non overlapping reference periods have been considered: from 1930.10.01 (or later, depending on data availability) to 1974.09.30 and from 2000.10.01 to 2021.09.30.

The complete link clustering of the Mahalanobis distance matrices, $\mathcal{M}$, evaluated in the two reference periods, leads to the identification of six clusters.

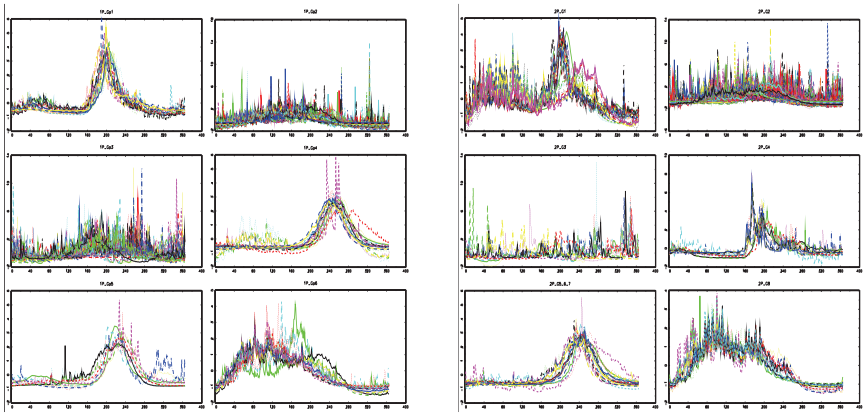


Figure 1: Average daily discharge of clustered time series (1st period-left panels; 2nd period-right panels)

In particular, the clusters describe: strong fall/spring regime (G1: mostly in the North Atlantic and Pacific NW coast); intermittent winter/spring regime (G2: mid Atlantic coast and central valleys); intermittent regime (G3: Gulf coast); weak winter regime (G4: upper Great lakes and Northern Great Planes); melt regimes (G5: mostly in the Rocky mountain and Northern Great planes); strong winter regime (G6: mostly in the NW coast). Fig.1 illustrates the average daily discharge of the series belonging to each cluster in the two reference periods. The fundamental features of the seasonal patterns are rather stable in the two periods, but a number of series (33%) have changed their class memberships. Changes are due to various factors: the anticipation of the seasonal peak due to early snow-melt, the increase of the winter rainfall, the increase of dry periods and flashy peaks. Moreover, the analysis of residual components by means of the AR metric identifies three clusters of series with increasing level of inertia (low, moderate, high). The parametric spectral densities of the cluster centroids help to define these level of inertia. However, the long term dynamics does not change remarkably in the two periods. This may be due to the fact that the residual components are heavily affected by specific physio-characteristics of the basins (for example, the slope).

At the end of the procedure, each time series is characterized by two labels specifying the seasonal regime and the level of inertia. These features can be summarized in a two-way table. In the period 2000-2021, there are 13 clusters

(Table 1) and most rivers show an intermittent regime with peaks in winter or spring and a low/medium level of inertia. The intermittent regime gather numerous series that have changed their class memberships in recent years.

Table 1: Final classification for the dataset observed the period 2000-2021

Seasonality	Inertia		
	high	medium	low
strong fall/spring	2	17	6
intermittent winter/spring	0	72	61
intermittent	0	6	6
weak winter	0	10	5
snow-melt	0	15	5
strong winter	0	15	1

4 Final remarks

The results that we have achieved using widely applicable statistical tools provide a useful basis for further discussion about the relationships of streamflow regimes with physiographic and climate indices, and for determining the future regime changes according to simulated scenarios from models driven by climate data.

References

CORDUAS, M., & PICCOLO, D. 2008. Time series clustering and classification by the autoregressive metric. *Computational Statistics & Data Analysis*, **52**, 1860–1872.

DHUNGEL, S., TARBOTON, D.G., JIN, J., & HAWKINS, C.P. 2016. Potential effects of climate change on ecologically relevant streamflow regimes. *River Research and Applications*, **32**, 1827–1840.

NEWMAN, A.J., & AL. 2015. Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: dataset characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, **19**, 209–223.

PICCOLO, D. 1990. A distance measure for classifying ARIMA models. *Journal of time series analysis*, **11**, 153–164.

YANG, M., & OLIVERA, F. 2023. Classification of watersheds in the conterminous United States using shape-based time-series clustering and Random Forests. *Journal of Hydrology*, **620**, 129409.

MODAL CLUSTERING FOR CATEGORICAL DATA

Noemi Corsini [1] and Giovanna Menardi[1]

[1] Department of Statistical Sciences, University of Padova,
(e-mail: noemi.corsini@phd.unipd.it, menardi@stat.unipd.it)

ABSTRACT: Despite the ill-posedness of the clustering task, in the continuous setting a broad consensus is overall acknowledged in defining the concept of cluster. Conversely, a general notion of cluster remains controversial in the presence of categorical data. We propose a novel notion of cluster hinging on the twofold concept of high frequency and association between variables. The former concept, in fact, complies with the cluster notion described by the modal formulation of the clustering problem, which we take advantage of to borrow some operational tools to propose an operational procedure.

KEYWORDS: association, contingency table, graph

1 Introduction

The importance of clustering in statistics has never been questioned over the years, thanks to the many fields in which it finds relevant applications. However, more than to its wide applicability, the proliferation of a voluminous amount of literature on this topic is perhaps due to the ill-posedness of the problem, which is inherent with its unsupervised nature. In fact, when numerical data are at hand, a general agreement is met across alternative notions of cluster, which collectively fall under the heading of groups of similar subjects. Even when more sophisticated density-based cluster formulations are considered, indeed, the underlying notion of cluster implies the observations to be somewhat close to each other.

Conversely, this does not apply to categorical data. While, in principle, a natural clustering gathers subjects within the observed cross-categories of the variables, such description turns out to lack parsimony when either the number of variable and/or the number of categories grows. On the other hand, the lack of a total order among categories makes somewhat controversial even the notion of distance, and increases the arbitrariness in the subsequent definition of cluster, which, in the literature about clustering categorical data, is usually left unspecified.

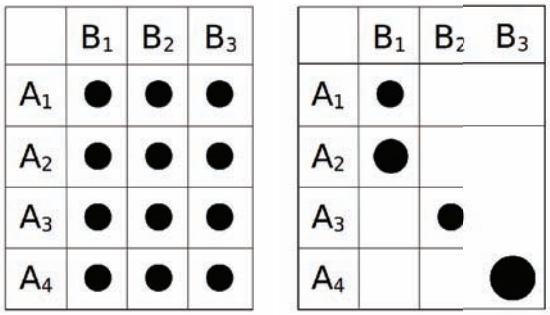


Figure 1. *Graphical representation of two contingency tables where the diameter of each circle is set as proportional to the frequency of the cell.*

In this work we attempt the ambitious aim of filling this gap by proposing a novel notion of cluster within the setting of categorical data, along with an operational procedure to identify clusters.

Consider the two toy examples in Figure 1, which describe two alternative frequency patterns of a number of subjects observed with reference to two variables. Even in the lack of a cluster definition, we feel highly shareable to acknowledge that the left panel, where variables are independent, identifies a configuration without clusters (or formed by 12 clusters, as the number of cross-categories); on the other hand, the right panel, characterized by a strong association pattern between the two variables, aggregates the subjects in three clusters. This intuition leads us to build a novel notion of cluster hinging on the twofold concept of high frequency and association between variables, i.e. groups arise as highly populated (aggregations of) cross-categories of variables leading a large contribution of mutual information. The former concept, in fact, complies with the cluster notion described by the *nonparametric* or *modal* formulation of the clustering problem (see Menardi, 2016, for a review), which we shall use to borrow some operational tools to identify groups. Note that a similar idea is implicitly acknowledged by one of the most widespread approaches to clustering categorical data, i.e. the $k-$modes (Huang, 1998).

2 Method

According to the nonparametric formulation of the clustering problem, groups are intended as the domains of attraction of the modes of the density underlying data. In the continuous setting, such regions are operationally identified

either as the set of points whose direction of the steepest gradient ascent path converges to the same mode, or as connected sets with density above a threshold.

In the categorical setting, defining both a density or its gradient is precluded and, at the same time, there is no obvious method to define the connectedness of a region. Nevertheless, we shall define a procedure that jointly extends both these ideas, if not formally, at least conceptually. To this aim, we build a directed weighted graph where each node represents a cross-category. The idea of steepest gradient ascent path is translated into a sequence of links between nodes driving in the direction of the node with the locally highest (estimated) probability. On the other hand, the connectedness of a region, intended as a set cross-categories, is evaluated by the weight of the links.

Consider again the example in Figure 1: two nodes identified by the cross-categories (A_r, B_c) and (A_s, B_c) shall be considered as highly connected not only because they share the same level for variable B but also when, given that $B = B_c$, both A_r and A_s become more likely, that is, when

$$\frac{P(A_r|Bc)}{P(A_r)} \quad \text{and} \quad \frac{P(A_s|B_c)}{P(A_s)} \tag{1}$$

are high. This results in providing the link with a weight set to the minimum between the probabilities (1), or, for example, to their mean, a choice selected hereafter to avoid ties. The direction of the link will point toward the maximum between the two probabilities. In fact, note that

$$\frac{P(A_r|Bc)}{P(A_r)} < \frac{P(A_s|B_c)}{P(A_s)} \Leftrightarrow \frac{P(A_r, Bc)}{P(A_r)P(B_c)} < \frac{P(A_s, B_c)}{P(A_s)P(B_c)},$$

that is, the path of each node moves toward the direction where the ratio between the joint probability of the cell and the expected probability under the hypothesis of independence is the maximum.

With this toolkit at hand, clusters can be formed as high density upper-level sets or, at the same time, as domains of attractions of the density modes, where the concept of density is here intended as a measure of how each cross-category occurs more frequently than it would do if the variables were independent. The outlined ideas easily extend to an arbitrary number of variables.

3 Application

Figure 2 outlines a synthetic illustrative example that cross-classifies 460 individuals according to their religion and geographic area of origin. The two

	Christianity	Islam	Eastern Religions
Europe	85	4	1
America	110	2	3
Africa	20	30	4
Asia-Pacific	1	120	80
Europa	42.3	30.5	17.2
America	54.0	39.0	22.0
Africa	25.4	18.3	10.3
Asia-Pacific	94.4	68.2	38.5

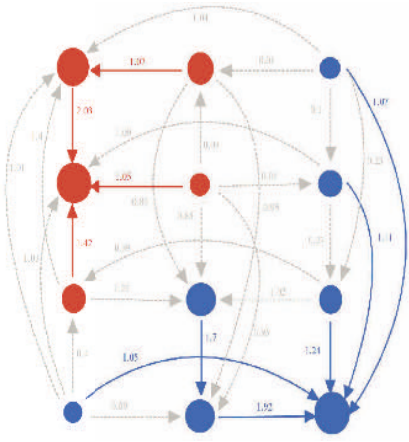


Figure 2. *Contingency table of religion and geographic area of origin for 460 observations (top), associated table expected under the hypothesis of independence (bottom), and graph built based on the proposed method, with highlighted the resulting clusters in different colors.*

variables exhibit a high dependency structure, with certain cross-categories presenting a higher frequency than expected under the assumption of independence. We built the graph according to the presented procedure, by estimating the involved joint probabilities by their empirical counterpart and the expected ones under the hypothesis of independence, by a suitable log-linear model. The graph, displayed on the right side of Figure 2 reports the direction addressed by the nodes, along with the intensity of the connections between them (shaded colors describe outgoing links whose weight is not the highest). While, for example, the cross-categories in the first column share a common level for the religion variable, they are not connected with the same strength, because being Christian increases the probability of coming from America and Europe, whereas the same does not apply to the Asia-Pacific region. By following the path of each node, two clusters are revealed, attracted by subjects of Asiatic origin of Eastern religions and by Christians from America.

References

HUANG, Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data min. Know. Disc.*, **2**, 283–304.

MENARDI, G. 2016. A review on modal clustering. *Int.Stat.Rev.*, **84**, 413–433.

THE USE OF PRINCIPAL COMPONENTS IN QUANTILE REGRESSION: A SIMULATION STUDY

Cristina Davino [1], Tormod Næs[2], Rosaria Romano[1] and Domenico Vistocco[3]

[1] Department of Economics and Statistics, University of Naples Federico II, (e-mail: c.davino@unina.it, rosaroma@unina.it)

[2] Nofima AS, Norway, (e-mail: tormod.næs@nofima.no)

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

ABSTRACT: Least squares regression is highly unreliable when a strong collinearity structure is present among the predictors. Among several proposals introduced in the literature, principal component regression is a straightforward method to overcome the problem, even if it introduces a slight bias in the parameter estimation. This paper presents a simulation study to evaluate the use of principal component regression in the context of quantile regression and, focusing on the variability of the estimates and the model's prediction ability.

KEYWORDS: mutlicollinearity, principal component regression, quantile regression.

1 Introduction

In classical multiple linear regression applications, multicollinearity occurs very often, i.e. whenever two or more predictors are strongly correlated with each other. Such an issue can affect least-squares (LS) regression coefficients, their standard deviation, and consequently the associated t-tests, fitted values, and predictions.

Although multicollinearity has been extensively covered in the linear regression literature (Weisberg, 2005, Martens & Næs, 1992), little attention has been devoted to its effects in the context of quantile regression (QR) (Koenker & Hallock, 2001, Davino *et al.*, 2013, Furno & Vistocco, 2018). Possible solutions to the problem have been proposed from the ridge regression viewpoint (Bager, 2018), or focusing on variable selection techniques (Zaikarina *et al.*, 2016), for instance. However, an alternative approach addresses the problem of multicollinearity from a different perspective: the entire set of variables is preserved but replaced by some synthetic variables defined as *principal components*. This alternative approach is known as regression on latent variables

(James *et al.*, 2013), the variants of which differ in how these latent variables are obtained. Among these, the best-known method is the principal component regression (PCR)(Massy, 1965), from which the technique of quantile on principal component regression (QPCR) (Davino *et al.*, (2022)) originated.

The contribution of this article is to investigate the multicollinearity issue in the QR by evaluating its effects and deepening the study of the QPCR method.

2 Methods

In formal notation, the multiple linear regression model can be expressed as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is the $(n \times 1)$ vector of the dependent variable, $\mathbf{X}$ is a $(n \times K)$ fixed matrix representing the independent variables, $\boldsymbol{\beta}$ is a $(K \times 1)$ vector of unknown regression coefficients, and $\mathbf{e}$ is a $(n \times 1)$ vector of errors assumed to be normally distributed, with $\mathbf{E}(\mathbf{e}) = \mathbf{0}$, and $\mathbf{E}(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I}_n$. In the following, without loss of generality, we assume that $\mathbf{X}$ and $\mathbf{y}$ are centered columnwise. The LS estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2}$$

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is equal to

$$cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{3}$$

and can be also formulated in terms of the singular value decomposition of the $\mathbf{X}'\mathbf{X}$ matrix as

$$cov(\hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{k=1}^{K} \mathbf{p}_k(1/\lambda_k)\mathbf{p}_k', \tag{4}$$

where $\mathbf{p}$ and λ are the eigenvectors and the eigenvalues of $\mathbf{X}'\mathbf{X}$, respectively (Næs & Mevik, 2001). Equation (4) highlights how, in presence of collinearity among the predictors, i.e. when some eigenvalues are very small, the variance of the regression coefficients increases.

The LS predictor $\hat{y}$ is unbiased, and the related Mean Squared Error (MSE), written using the eigenvector and eigenvalue decomposition of $\mathbf{X}'\mathbf{X}$, is

$$MSE(\hat{y}) = \sigma^2/N + \sigma^2 \sum_{k=1}^{K} t_k^2/\lambda_k + \sigma^2, \tag{5}$$

where $t_k = \boldsymbol{x}'\mathbf{p}_k$ is the score of $\boldsymbol{x}$ along eigenvector k. Equation (5) shows that the MSE depends not only on the magnitude of the eigenvalue but also on the t-score, i.e., on how much the new observations fall within the range of variability of the observed data along the different axes.

PCR finds some linear combinations of the original variables and use them as regressors to predict $\mathbf{y}$. Specifically, principal components analysis is applied to the matrix of predictors $\mathbf{X}$ to extract the A most dominating principal components. The PCR model structure is given by the following two equations

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}' + \mathbf{E}, \\ \mathbf{y} &= \mathbf{T}\mathbf{q} + \mathbf{f}, \end{aligned} \tag{6}$$

where $\mathbf{T}$ is called scores matrix and collects the A dimensions responsible for the systematic variation in $\mathbf{X}$, $\mathbf{P}$ and $\mathbf{q}$ are called loadings and describe how the variables in $\mathbf{T}$ are related to the original variables in $\mathbf{X}$ and $\mathbf{y}$, respectively. The PCR estimator is no longer unbiased since only the main dimensions are retained, while the less relevant ones are discarded. The MSE of the predictor $\hat{y}_{PCR}$ is

$$MSE(\hat{y}) = \sigma^2/N + \sigma^2 \sum_{k=1}^{A} t_k^2/\lambda_k + \left(- \sum_{k=A+1}^{K} (t_k/\sqrt{\lambda_k})\alpha_k \right)^2 + \sigma^2. \tag{7}$$

It has been empirically demonstrated (Næs & Mevik, 2001) that in situations of collinearity among the predictors, the PCR predictor performs better than the LS predictor in terms of MSE. Equation (7) suggests that a more considerable contribution of the variance along the eigenvectors with small eigenvalues ($a = A+1, \ldots, K$) for the LS predictor is replaced in the case of the PCR predictor by a more negligible bias contribution.

The extension of the PCR to the context of the QR is straightforward, as shown in Davino *et al.*, (2022). The model structure for the so-called QPCR is given by the following two equations:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}' + \mathbf{E} \\ Q_\theta &(\hat{\mathbf{y}}|\mathbf{T}) = \mathbf{T}\hat{\boldsymbol{\beta}}(\theta) \end{aligned} \tag{8}$$

where $Q_\theta(.|.)$ is the conditional quantile function for the θ–th conditional quantile with $0 < \theta < 1$. It is worth noting that QPCR can produce the same numerical and graphical outputs as PCR, for each selected θ.

3 Simulation study

The simulation study aims to investigate the QPCR properties assessing:

- the variability of the regression coefficients in terms of MSE, given that the PCR estimator is biased;
- the prediction ability of the model both in the case of new cases within the range of the sampled data (i.e. to interpolate) and in the case of new data outside such a range (i.e. to extrapolate).

References

BAGER, AS. 2018. Ridge parameter in quantile regression models: An application in biostatistics. *International Journal of Statistics and Applications*, **8**(2), 72–78.

DAVINO, C, FURNO, M, & VISTOCCO, D. 2013. *Quantile regression: theory and applications*. Vol. 988. John Wiley & Sons.

DAVINO, C, ROMANO, R, & VISTOCCO, D. (2022). Handling multicollinearity in quantile regression through the use of principal component regression. *METRON*, **80**(2), 153–174.

FURNO, M, & VISTOCCO, D. 2018. *Quantile regression: estimation and simulation, Volume 2*. Vol. 216. John Wiley & Sons.

JAMES, G, WITTEN, D, HASTIE, T, & TIBSHIRANI, R. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

KOENKER, R, & HALLOCK, KF. 2001. Quantile regression. *Journal of economic perspectives*, **15**(4), 143–156.

MARTENS, H, & NÆS, T. 1992. *Multivariate calibration*. John Wiley & Sons.

MASSY, WF. 1965. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, **60**(309), 234–256.

MCCULLAGH, P, & NELDER, JA. 1989. Binary data. *Pages 98–148 of: Generalized linear models*. Springer.

NÆS, T, & MEVIK, B-H. 2001. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, **15**(4), 413–426.

WEISBERG, S. 2005. *Applied linear regression*. Vol. 528. John Wiley & Sons.

ZAIKARINA, H, DJURAIDAH, A, & WIGENA, AH. 2016. Lasso and ridge quantile regression using cross validation to estimate extreme rainfall. *Global Journal of Pure and Applied Mathematics*, **12**(3), 3305–3314.

AN INTERDISCIPLINARY METHODOLOGY FOR SOCIO-ECONOMIC SEGREGATION ANALYSIS

Antonio De Falco [1], Antonio Irpino[2]

[1] Department of Economics and Statistics, University of Napoli Federico II, (e-mail: antonio.defalco3@unina.it)

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

ABSTRACT: This paper proposes an original methodology for the analysis of socio-economic residential segregation. The strategy involves employing areal interpolation methods to create population grids, applying a compositional data approach to quantify categorical distributions, and utilising principal component analysis to define an index of socio-economic class composition for each cell in the study area. By combining index values with spatial autocorrelation tools, it is possible to identify and map segregated areas. To test our method, we rely on the latest UK census data (2021) for the Liverpool metropolitan area, using social groups defined according to the National Statistics Socio-economic Classification.

KEYWORDS: residential segregation analysis, grid cells, compositional data analysis, PCA, spatial analysis

1 Introduction

Residential segregation refers to the spatial separation of social groups within urban areas based on factors such as socio-economic status or ethnicity. While not inherently negative, segregation can lead to the formation of urban areas with distinct social compositions and unequal distribution of resources and services. These factors shape the opportunity/constraint structure of individuals, perpetuating and transmitting social inequalities (Musterd, 2020). Over the years, different indices have been proposed to measure the phenomenon according to its dimensions. However, recent re-conceptualisations and more effective measures have introduced new analysis approaches. Within this framework, there is particular interest in developing indices that incorporate the spatial dimension as they are better able to capture population patterns and the variability of segregation across urban space. Segregation measures typically rely on categorical data provided by national statistical agencies and reported

for different spatial units, such as census tracts. As a result, segregation studies often use ecological or aggregated units for analysis. Summary statistics describing these spatial units often involve compositional data with a fixed-sum constraint. However, applying standard statistical methods designed for unconstrained data to compositional data, which are constrained to a simplex can introduce bias (Aitchison, 1986). Additionally, the use of aggregated units poses challenges in spatial analysis due to their arbitrary scale of aggregation and delineation of boundaries, which may not align with meaningful divisions relevant to the studied phenomenon. This issue is commonly known as the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984). Moreover, employing aggregated units with irregular and changing geometries further complicates the comparison of urban areas over time or synchronously. To overcome these challenges, the next section presents a novel methodology for analysing socio-economic segregation, addressing the measurement complexities, and ensuring comparability across urban areas.

2 Methodology

The methodological proposal is based on an interdisciplinary approach, incorporating statistical, sociological, and geographical knowledge. The first phase involves using areal interpolation methods, commonly employed in quantitative geography to improve the estimation of population distribution across a territory. Starting with census aggregated units, a dasymetric binary interpolation procedure (Langford, 2013) using satellite data on land use and land cover is applied to enhance the estimation process. This procedure defines a new set of regular hexagonal grids with higher spatial resolution. The use of grid cells allows for diachronic and/or synchronic comparative analyses between urban areas that report different administrative subdivisions. Utilising grid cells instead of standard units provides a flexible tool to effectively address the MAUP, as the spatial resolution of the cells can be easily modified according to the research objectives. After estimating population grid data, a compositional data analysis strategy, as defined by Aitchison (1986), is employed in the second phase. Population data categorical distributions are quantified by performing the clr-logratio transformation, enabling a subsequent correlation-based statistical analysis. Next, the strategy for measuring socio-economic segregation is implemented in the third phase. A weighted principal component analysis (PCA) (Greenacre, 2018) is performed on the clr-coordinates to synthesise the distributions of socio-economic classes into a single factor while reducing the influence of sparsely populated grid cells on the results. Subsequently, the so-

cioeconomic composition index is defined using the scores derived from the first component. For ease of interpretation, the scores are normalised to values ranging from 0 to 100. In the fourth phase, to detect the spatial structure of the index, a spatial autocorrelation analysis is conducted using the Moran index (Moran, 1948). Index values are used as input data, and a different definition of proximity is incorporated in the spatial weight matrix to define the spatial relationships between areal units. This criterion, based on temporal distances, utilises the median time taken by an individual to travel from one cell to another using four different modes of transportation (walking, biking, driving, and transit). This approach may offer a more realistic representation of the degree of connection between areal units and the potential spatial interaction between social groups compared to criteria based on adjacency and geographical distance. Furthermore, the Moran index can consider the degree of clustering of only one population group at a time, but this limitation is overcome by using an index that summarises the distributions of all socio-economic groups. To assess the intensity of the spatial structure of the socio-economic composition index, the global Moran index is first calculated using different temporal distances as thresholds. Then, the local Moran index (Anselin, 1995) is applied to the index by selecting the specification of the spatial weight matrix that maximises the autocorrelation value.

3 Results

The proposal was applied to the metropolitan area of Liverpool. Data were collected from various sources, including the UK Census data for the year 2021, the UK Corine land cover dataset for the year 2018, the Traveline National Dataset (TNDS), and the Open Street Map data. The methodology was implemented in R. Figure 1 displays the local Moran map of the socio-economic composition index, illustrating the spatial patterns of socio-economic groups defined according to the UK National Statistics Socio-economic Classification in the study area. Based on significant local Moran values and PCA loadings, cells were classified into different categories: HH (High-High) spatial clusters indicate higher socio-economic class segregation, LL (Low-Low) spatial clusters indicate lower socio-economic class segregation, HL (High-Low) spatial outliers represent high index values surrounded by low index values, and LH (Low-High) spatial outliers represent low index values surrounded by high index values. Cells with no significant local Moran values are considered non-segregated areas.

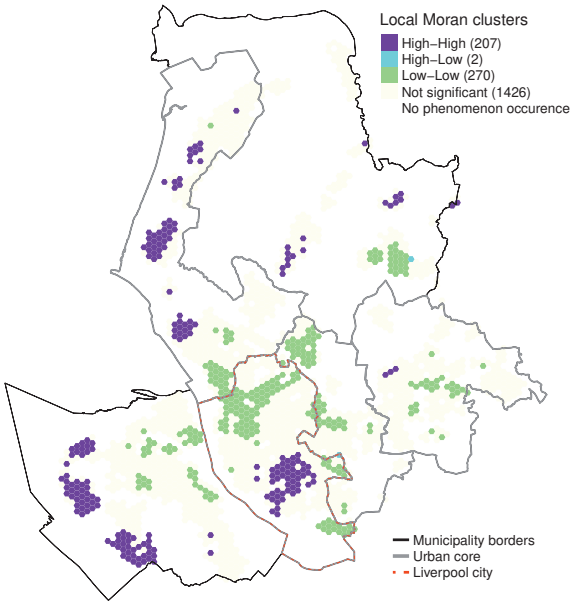


Figure 1. *Local Moran map of the socio-economic composition index. Liverpool, 2021*

References

AITCHISON, J. 1986. *The statistical analysis of compositional data*. London: Chapman and Hall.

ANSELIN, L. 1995. Local Indicators of Spatial Association – LISA. *Geographical Analysis*, **27**, 93–115.

GREENACRE, M. 2018. *Compositional Data Analysis in Practice*. New York: Chapman and Hall/ CRC.

LANGFORD, M. 2013. An evaluation of small area population estimation techniques using open access ancillary data. *Geographical Analysis*, **45**, 324–344.

MORAN, P. A. P. 1948. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society*, **10**, 243–251.

MUSTERD, S. 2020. *Handbook of Urban Segregation*. Cheltenham: Edward Elgar Publishing.

OPENSHAW, S. 1984. *The Modifiable Areal Unit Problem*. Norwich: GeoBooks.

VISUALIZING CLASSIFICATION RESULTS: GRAPHICAL TOOLS FOR DD-CLASSIFIERS

Houyem Demni [1] and Simona Balzano [1]

[1] Department of Economics and Law, University of Cassino and Southern Lazio , (e-mail: houyem.demni@unicas.it, s.balzano@unicas.it)

ABSTRACT: DD-classifiers have been widely used to perform classification tasks given that they are non-parametric and flexible and can also be applied in high-dimensional spaces when a suitable notion of depth is adopted. The aim of DD-classifiers is to assign new unlabeled observations to the labeled groups based on their depth values with respect to each group. Visualizing the cases being classified can be very interesting. It can reveal a clue about the data and the classification method as well, e.g. the causes for which some observations are misclassified or whether the classifier is appropriate to the data or not, which can be reflected by the posterior probability of the alternative class. For these reasons, rather than focusing on the mechanism of the DD-classification procedure itself, we investigate how the silhouette plot, the class map and the quasi-residual plot can be adopted to visualize the results of the DD-classifiers. Several real data examples are considered in order to illustrate the potential of these visualization tools. We also use the average silhouette width to compare the results of DD-classifiers exploiting different discriminant rules when associated with different depths for each data set.

KEYWORDS: Discriminant analysis, silhouette plot, class map, quasi residual plot.

DETECTING THE POSITIONS OF NONCONSESUS AMINO ACIDS IN HIV PATIENTS BY MARGINAL LIKELIHOOD THRESHOLDING

Claudia Di Caterina [1]

[1] Department of Economics, University of Verona, (e-mail: claudia.dicaterina@univr.it)

ABSTRACT: We show how marginal likelihood thresholding can be applied in the context of multiple hypothesis testing, proposing a rule to select the tuning parameter involved. For detecting the positions of nonconsensus amino acids in patients suffering from two different HIV variants, we use a logistic regression framework and see that our results are in line with those from standard and advanced procedures controlling the false discovery rate, i.e. the proportion of incorrectly rejected null hypotheses.

KEYWORDS: composite marginal likelihood, logistic regression, multiple testing.

1 Setup and Methods

Let Y be a $p \times 1$ random vector with probability mass or density function $f(y;\theta)$ indexed by the parameter $\theta = (\theta_1,\ldots,\theta_p)^\top$, which is sparse in the sense that a small number $p^* \ll p$ of its elements are different from zero. Suppose the full $f(y;\theta)$ is difficult to specify or compute, but we can identify the p conditional univariate marginal distributions of the single Y_js, $f_j(y|x;\theta_j)$ $(j = 1,\ldots,p)$ where x is a k-vector of covariates. Specifically, we assume a generalized linear model $\mu_j = E(Y_j) = g^{-1}(\alpha_j + \theta_j x)$ with link function $g(\cdot)$ and dispersion parameter $\phi > 0$.

Given independent observations $(Y^{(i)},x^{(i)})$ $(i = 1,\ldots,n)$, the composite marginal likelihood (CML) estimator $\tilde{\theta}$ (Varin *et al.*, 2011) maximizes

$$\ell(\theta;Y^{(1)},\ldots,Y^{(n)}) = \sum_{j=1}^{p} w_j \ell_j(\theta_j;Y^{(1)},\ldots,Y^{(n)}), \tag{1}$$

where $\ell_j(\theta_j;Y^{(1)},\ldots,Y^{(n)}) = \sum_{i=1}^{n} \log f_j(Y^{(i)}|x^{(i)};\theta_j)$ is the jth marginal log-likelihood and $w = (w_1,\ldots,w_p)^\top$ is the design vector of weights that determines which margins are included in (1). Finally, we assume that p grows with the sample size n, but at a slower rate.

1.1 Marginal likelihood thresholding

We review here the method presented in Di Caterina & Ferrari, 2022, for the current setting. Since the marginal log-likelihoods depend on separate parameters, we have $\tilde{\theta}_j = \left\{\theta_j : \sum_{i=1}^n u_j(\theta_j; Y^{(i)}) = 0\right\}$ $(j = 1, \ldots, p)$ where $u_j(\theta_j; y) = \partial \ell_j(\theta_j; y)/\partial \theta_j$ denotes the jth marginal score. Sparsity in the final estimator $\hat{\theta}$ is induced via the marginal likelihood thresholding (MLT)

$$\hat{\theta}_j = \begin{cases} \tilde{\theta}_j & \text{if} \quad \hat{w}_j \neq 0 \\ 0 & \text{if} \quad \hat{w}_j = 0 \end{cases} \qquad (j = 1, \ldots, p),$$

where $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_p)^\top$ is a sparse design vector, selected by minimizing for some $\lambda > 0$ the convex criterion that balances statistical efficiency and sparsity:

$$\hat{d}_\lambda(w) = \frac{1}{2} w^\top \hat{C} w - w^\top \mathrm{diag}(\hat{C}) + \frac{\lambda}{n} \sum_{j=1}^p \frac{|w_j|}{\tilde{\theta}_j^2}, \tag{2}$$

where $\hat{C}$ is the sample covariance matrix of the marginal scores and, if $g(\cdot)$ takes canonical form, has entries $\hat{C}_{jk} = \sum_{i=1}^n (Y_j^{(i)} - \tilde{\mu}_j^{(i)})(Y_k^{(i)} - \tilde{\mu}_k^{(i)})(x^{(i)})^2/(\phi^2 n)$ with $\tilde{\mu}_j^{(i)} = g^{-1}(\hat{\alpha}_j + \tilde{\theta}_j x^{(i)})$.

1.2 Selection of the tuning parameter

The tuning parameter λ is crucial in determining the proportion of nonzero elements in the final MLT estimator $\hat{\theta}$. From the Karush-Kuhn-Tucker (KKT) first-order conditions for the minimization of (2), we find that $\hat{\theta}_j$ is set to zero if the corresponding rescaled z-statistic is smaller than $\sqrt{\lambda}$. This condition is an acceptance region for the null hypothesis $\theta_j = 0$ and suggests that λ may be selected by some form of error control for multiple tests based on the family of hypotheses $\mathcal{H}_\lambda = \{H_0^j : \theta_j = 0 \text{ vs } H_a^j : \theta_j \neq 0, j \in \hat{\mathcal{A}}_\lambda\}$, where $\hat{\mathcal{A}}_\lambda = \{j : \hat{w}_j \neq 0\}$. Rejecting all the hypotheses in $\mathcal{H}_\lambda$ indicates that the selected parameters are probably useful and a larger model could be considered by decreasing λ.

By this rationale, using the asymptotic normality of the z-statistic for θ_j, Slutsky's Theorem and the KKT conditions, if the false discovery rate (FDR) is set equal to $\alpha \in (0, 1)$ we obtain the following selection rule for λ:

$$\hat{\lambda} = \inf \left\{ \lambda : \frac{\tilde{\theta}_j^2}{SE_j^2} > q_\alpha, \text{ for all } j \in \hat{\mathcal{A}}_\lambda \right\}, \tag{3}$$

where $SE_j = \phi\{\sum_{i=1}^n (Y_j^{(i)} - \tilde{\mu}_j) x^{(i)}\}^{-1}$ if $g(\cdot)$ is canonical and q_α is the upper α-quantile of the χ_1^2 distribution.

2 Analysis of HIV data

We analyze data from Gilbert, 2005, to investigate differences between two variants of HIV. The gag p24 amino acid sequence with $p = 118$ positions was obtained from $n = 146$ individuals, half infected with subtype C (group 1, $n_1 = 73$) and half infected with subtype B (group 2, $n_2 = 73$). For each jth position, the number of subjects with a nonconsesus amino acid was recorded in groups 1 and 2. Our aim is to detect the differentially polymorphic positions, where the probability of a nonconsensus amino acid differs in the two groups.

Both Gilbert, 2005, and Chen *et al.*, 2018, §5, assumed the counts per position were distributed as $Bin(\tau_{jg}, n_g)$ in the gth group $(g = 1,2)$, computed Fisher's exact statistics to test the null hypotheses $H_0^j : \tau_{j1} = \tau_{j2}$ for $j = 1,\dots 118$, and adjusted for multiple comparison. They discussed that the Benjamini-Hochberg (BH) method (Benjamini & Hochberg, 1995), which here finds 12 relevant positions controlling the FDR at level $\alpha = 5\%$, has less power and possibly yield unreliable results in discrete settings. Because the first 50 positions have Fisher's exact test statistics with p-values almost surely equal to 1, the BH procedure is expected to be extremely conservative here, meaning to have a FDR much lower than α.

Instead, we model the presence/absence of a nonconsensus amino acid in subject i on position j as $Y_j^{(i)} \sim Ber(\pi(i)_j)$ with $\pi(i)_j = \text{logit}^{-1}(\alpha_j + \theta_j x^{(i)})$, where $x(i)$ is a dummy variable encoding the ith subject's group $(i = 1,\dots,n)$. We can then apply the MLT method to such logistic regression scenario using $p = 118$ univariate marginal likelihoods: a nonzero estimate of the logit coefficient θ_j will indicate to reject the hypothesis $H_0^j : \theta_j = 0$ and so will identify the jth position as differentially polymorphic.

Since quasi-complete separation occurs when fitting the logistic regression in some positions, it is convenient to set the marginal $\tilde{\theta}_j$s equal to the equally consistent bias-reduced estimates (Firth, 1993). If we choose $\hat{\lambda}$ as described in (3) with $\alpha = 5\%$, we select $\hat{p}^* = 15$ nonzero parameters corresponding to 15 differentially polymorphic positions. This is in line with what found by Gilbert, 2005, via their modified BH procedure. Chen *et al.*, 2018, §5, noticed that the classical BH method applied after excluding the first 50 positions also leads to the same conclusion. In terms of positions selected by MLT, Table 1 shows that 13 out of 15 were identified also by at least another multiple testing procedure conducted on this data set in Gilbert, 2005, and Chen *et al.*, 2018, §5, controlling the FDR at level $\alpha = 5\%$. Note that, when the complete data are analyzed, neither of the FDR-controlling procedures considered selects any

Table 1. *Number of positions selected via MLT classified by alternative FDR-controlling method. A tick indicates the corresponding method also selects those positions at level* $\alpha = 5\%$*. The* * *marks methods run after excluding the first 50 positions.*

# Selected positions by MLT	BH	BH*	Modified BH (Gilbert, 2005)	adaptive BH* (Chen *et al.*, 2018)	adaptive BH-Heyse* (Chen *et al.*, 2018)
7	✓	✓		✓	✓
3			✓		
1		✓		✓	✓
1				✓	✓
1					✓
2					
Tot: 15					

of the first 50 positions. It would then appear sensible that results did not change once those were discarded. Yet this sort of robustness holds only for the modified BH procedure and our proposal.

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 B*, **57**, 289–300.

CHEN, X., DOERGE, R. W., & HEYSE, J. F. 2018. Multiple testing with discrete data: proportion of true null hypotheses and two adaptive FDR procedures. *Biometrical Journal*, **60**, 761–779.

DI CATERINA, C., & FERRARI, D. 2022. Sparse composite likelihood selection. *Pages 423–426 of:* TORELLI, N., BELLIO, R., & MUGGEO, V. (eds), *Proceedings of the 36th International Workshop on Statistical Modelling*, vol. 3.

FIRTH, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

GILBERT, P. B. 2005. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society C*, **54**, 143–158.

VARIN, C., REID, N., & FIRTH, D. 2011. An Overview of Composite Likelihood Methods. *Statist. Sinica*, **21**, 5–42.

ONE-INFLATED BAYESIAN MIXTURES FOR POPULATION SIZE ESTIMATION

Davide Di Cecco [1], Andrea Tancredi[1] and Tiziana Tuoto[2]

[1] Sapienza University of Rome, (e-mail: davide.dicecco@uniroma1.it, andrea.tancredi@uniroma1.it)

[2] ISTAT, (e-mail: tuoto@istat.it)

ABSTRACT:

The phenomenon of one-inflation frequently affects the estimates of population size when the available dare are represented by frequencies of counts. A particular behavioral effect preventing subsequent captures after the first one may be the reason for such an effect. We consider a Bayesian semi-parametric approach by fitting a truncated Dirichlet process mixture model as a base tool for modeling repeated count data and extend this class to include one–inflation. The proposed methodology is briefly illustrated via a real data application.

KEYWORDS: capture-recapture, Dirichlet process mixture, repeated count data

1 Introduction

Consider a closed population composed of N individuals. Suppose that N is unknown, n distinct units have been identified for a fixed amount of time and a given unit may be identified exactly once or observed twice, three times, or more. Under time–homogeneity, without individual covariates, the data can be simply summarized as counts of units captured j times, $j = 1, 2, ...,$ commonly called "repeated count data" The common parametric approach for estimating N is to define a counting distribution for the number of captures in the population. In the absence of any additional individual information, it is crucial to model the unobserved heterogeneity. A well–established approach to this end is represented by the use of mixtures of counting distributions, see, Böhning *et al.* (2005).

Mixtures of Poisson distributions are a standard choice both for repeated captures and species sampling problems but they present several issues related to the selection of the number of components and the instability of the N estimator. The choice of the number of mixture components has been usually addressed by the use of the nonparametric maximum likelihood estimation

(NPMLE) approach (Norris & Pollock (1996)) which maximizes the likelihood of an over-fitting finite mixture model. The frequentist properties of the NPMLE approach have been discussed in Wang & Lindsay (2005) where a penalized NPMLE estimator of N with better inferential performance is proposed. Another critical issue that has been recently addressed for the N estimation problem with repeated counts is that the collected data set frequently exhibit an elevated number of individuals captured exactly once. See, for example, Godwin & Böhning (2017). This excess of singletons is also termed as "one–inflation". Failing to identify and model in the analysis such a mechanism implies a (possibly severe) overestimation of the total population count.

Bayesian semi-parametric approaches underlying population size estimation have already been proposed. Guindani *et al.* (2014) handled the heterogeneity problem proposing a Dirichlet process mixture (DPM) of Poisson distributions for modeling gene expression sequence abundance and estimating the number of different unique sequences. The DPM approach, as the NPMLE, avoids to fix the number of components and, by averaging over mixtures of different order, has the advantage of properly accounting for the clustering process uncertainty in the final estimate of N. A DPM latent class model has been also proposed in the context of multiple systems estimation by Manrique-Vallier (2016) under capture heterogeneity and list dependence. In this paper, we present an application of the DPM approach handling the presence of one inflation with repeated count data.

2 One-inflated mixture distributions

Let Y_i, $i = 1,\ldots,N$, be the integer-valued random variable representing the number of times a given unit has been captured. We assume that

$$Y_i|\lambda_i \overset{ind}{\sim} \text{Poisson}(\lambda_i) \quad \lambda_i|G \overset{iid}{\sim} \Lambda \quad \Lambda \sim DP(\phi\Lambda_0) \tag{1}$$

where $\Lambda \sim DP(\phi\Lambda_0)$ denotes a distribution generated by a Dirichlet process with base measure $\phi\Lambda_0$, see Guindani *et al.* (2014). Note that we only observe the n individuals which are captured at least once. Let n_j denote the number of units captured j times, such that $\sum_{j>0} n_j = n$. We want to estimate the number of uncaptured units n_0, or, equivalently, $N = n + n_0$. Considering the *truncated* version of the DPM model (1) (see Ishwaran & James (2001)), Y_i is a finite mixture of Poisson distributions with mixing weights $\pi_1,\ldots,\pi_k$ following a finite stick-breaking prior, that is, $\pi_1 = V_1$ and

$$\pi_i = (1-V_1)(1-V_2)\cdots(1-V_{i-1})V_i \quad i = 2,\ldots k \tag{2}$$

where V_i for $i = 1, \ldots, k-1$ are independent Beta$(1, \phi)$ random variables and $V_k = 1$. Denote as $f(j|\lambda_i)$ the probability $\lambda_i^j e^{-\lambda_i}/j!$ of being captured j times in the i–th component defined by the parameter λ_i and denote as θ the set of all parameters. The truncated Poisson DPM model is defined as $P(Y = j) = f(j|\theta) = \sum_{i=1}^{k} \pi_i f(j|\lambda_i)$ for $j = 0, 1, \ldots$ with the mixing weights given by (2).

Under the hypothesis of one–inflation caused by a specific behavioral effect, an individual that, without that effect, would face multiple captures, under this effect will be captured just once. The hypothesis can be modeled as follows: let B be the latent indicator variable identifying the units having this behavior. Each individual has a marginal probability ω of belonging to this subpopulation. Denote as Y^* the latent number of captures of a given unit that we would observe in absence of the behavioral mechanism and let $f^*(j|\theta) = P(Y^* = j\,|\,\theta)$ be its probability distribution. By assuming $P(Y = j|B = 0) = f^*(j|\theta)$ for all j and $P(Y = j|B = 1) = f^*(0|\theta)$ for $j = 0$ and $1 - f^*(0|\theta)$ for $j = 1$ the resulting distribution for Y is the one–inflated model defined as:

$$P(Y = j\,|\,\theta, \omega) = \begin{cases} f^*(0|\theta) & \text{if } j = 0; \\ (1-\omega) f^*(1|\theta) + \omega(1 - f^*(0|\theta)) & \text{if } j = 1; \\ (1-\omega) f^*(j|\theta) & \text{if } j > 1. \end{cases} \tag{3}$$

The one-inflated Poisson DPM model is then obtained by assuming for the baseline distribution $f^*(j|\theta)$ in (3) a Poisson DPM model.

3 Application

In this Section, we briefly illustrate the proposed methodology. We consider a data set that contains counts of treatment episodes by heroin users in Bangkok, see Godwin (2017). Upon visiting a treatment center, heroin users may find the treatment less pleasant than expected, and decide never to return, thus giving rise to one-inflation. Figure 1 shows the data set, the posterior distributions for N under the truncated DPM and the one-inflated version, the posterior distributions for the number of observed clusters, and the one-inflation parameters. These posterior distributions have been obtained via MCMC methods and a prior on the DPM parameter ϕ penalizing the overestimation of the number of clusters. As expected, the one-inflated model produces lower estimates of the population count by assigning a greater number of captures to a portion of singletons. The estimation for N and ω under the one-inflated DPM are comparable to those obtained by Godwin (2017) confirming that the DPM and its one-inflated counterpart represent valid competitors in this setting.

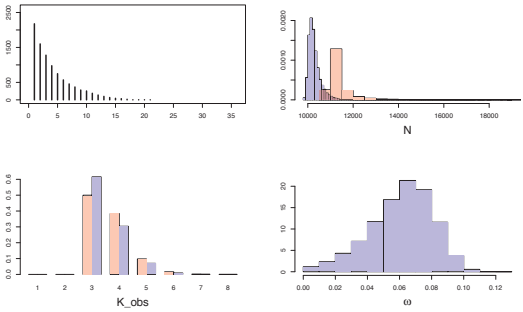


Figure 1. *Heroin users data set. Top left: count distribution. Top right: posterior distribution for N under the DPM (red) and one-inflated DPM (blue). Lower left: posterior distribution for the number of observed clusters. Lower right: posterior distribution for the DPM one-inflation parameter* ω

References

BÖHNING, D., DIETZ, E., KUHNERT, R., & SCHÖN, D. 2005. Mixture models for capture-recapture count data. *Statistical Methods and Applications*, **14**, 29–43.

GODWIN, R.T. 2017. One-inflation and unobserved heterogeneity in population size estimation. *Biometrical Journal*, **59**, 79–93.

GODWIN, R.T., & BÖHNING, D. 2017. Estimation of the population size by using the one-inflated positive Poisson model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **66**, 425–448.

GUINDANI, M., SEPÚLVEDA, N., PAULINO, C.D., & MÜLLER, P. 2014. A Bayesian semi-parametric approach for the Differential Analysis of Sequence Counts Data. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, **63**, 385.

ISHWARAN, H., & JAMES, L. 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Stat. Association*, **96**, 161–173.

MANRIQUE-VALLIER, D. 2016. Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*, **72**, 1246–1254.

NORRIS, J., & POLLOCK, K. 1996. Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics*, 639–649.

WANG, J. Z., & LINDSAY, B. G. 2005. A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association*, **100**, 942–959.

CLUSTER ANALYSIS AND CONDITIONAL COPULA: A JOINT APPROACH TO ANALYSE ENERGY DEMAND

F. Marta L. Di Lascio[1] and Roberta Pappadà[2]

[1] Faculty of Economics and Management, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy, (e-mail: marta.dilascio@unibz.it)

[2] Department of Economics, Business, Mathematics and Statistics "B. de Finetti", University of Trieste, Italy, (e-mail: rpappada@units.it)

ABSTRACT: In this work we investigate the thermal energy demand (TED) in urban areas through a copula-based approach. The proposed method enables the characterization of the probability law of TED under extreme weather conditions and for specific groups of buildings. In particular, we show how building characteristics, such as energy class and heating surface, may worsen or mitigate the impact of extreme scenarios.

KEYWORDS: Ali-Mikhail-Haq copula, cluster analysis, conditional copula, thermal energy demand

1 Introduction

The analysis of thermal consumption in urban areas is crucial to increase the sustainability and efficiency of energy systems (Menapace *et al.*, 2021) and reduce the impact of climate change. One major issue is the study of the complex dependence between thermal energy demand (TED) and weather conditions. Focusing on district heating (DH) – a heat distribution system representing a key technology to reduce waste of energy in urban areas – Di Lascio *et al.*, 2020 and Di Lascio *et al.*, 2021 analysed the temporal dynamics of TED and its relationships with meteorological variables. In particular, they assessed the effect of extreme values of solar radiation (SR, in W/m^2) and outdoor temperature (OT, in $^\circ C$) on TED by using a conditional copula-based approach. In addition, Di Lascio *et al.*, 202X recently proposed a copula-based dissimilarity measure to group buildings according to their TED, which turns out to be strongly influenced by building characteristics, such as energy class, age class, and heating surface.

In this paper, we refine the study of the impact that meteorological variables have on TED, by merging the proposals in Di Lascio *et al.*, 2021 and Di Las-

cio *et al.*, 202X. Sect. 2 presents the background and the proposed methodology, while Sect. 3 illustrates the application and discusses the results.

2 Methodology

Our proposal is to exploit conditional copula to describe the probability law of TED (X_3) given extreme scenarios (i.e. very low quantiles of SR (X_1) and OT (X_2)) (Di Lascio *et al.*, 2021), after performing a suitable cluster analysis to identify buildings with similar TED profile, as done in Di Lascio *et al.*, 202X.

The proposed methodology is grounded on (i) the following conditional distribution function (see Di Lascio *et al.*, 2021 for details)

$$P(X_3 > x_3 | X_1 < x_1, X_2 < x_2) = 1 - C(F_3(x_3)|F_1(x_1), F_2(x_2)) \tag{1}$$

where $C(F_3(x_3)|F_1(x_1), F_2(x_2)) = C(U_3|U_1, U_2)$ is the conditional copula defined using Bayes' rule (Trivedi & Zimmer, 2005), and (ii) the following copula-based spatially-weighted dissimilarity measure between TED time series at different sites (see Di Lascio *et al.*, 202X for details)

$$d_{jj'} = c_{jj'} \sqrt{2(1-\theta_{jj'})} \tag{2}$$

where $\theta_{jj'} \in [-1, 1[$ is the parameter of the Ali-Mikhail-Haq copula model (Ali *et al.*, 1978) of TED time series at sites j and j', and $c_{jj'} = \exp(g_{jj'} / \max(G))$, $\forall j \neq j'$, where $G = (g_{jj'})$ is matrix of geographical distances, is the spatial weight. The final clustering here is obtained via the hierarchical method with complete linkage rule.

In what follows, an application of the conditional-copula approach to the partition obtained via the clustering procedure based on Eq. (2) is illustrated.

3 Empirical analysis and discussion

We use hourly time series of TED of 41 residential users (i.e., one or more aggregated buildings) in Bozen-Bolzano during two intermediate weeks in January 2016. We first estimate the following dynamic panel regression model

$$\mathrm{TED}_{it} = \rho_1 \mathrm{TED}_{i(t-1)} + \rho_2 \mathrm{TED}_{i(t-24)} + \beta_1 \mathrm{SR}_{it} + \beta_2 \mathrm{OT}_{it} + \beta_3 \mathrm{OT}_{i(t-3)} + \mu_i + \varepsilon_{it}$$

where $i = 1, \ldots, 41$, $t = 1, \ldots, T = 366$, $\mu_i \sim \mathrm{iidN}(0, \sigma^2_\mu)$, $\varepsilon_{it} \sim \mathrm{iidN}(0, \sigma^2_\varepsilon)$, with μ_i and ε_{it} independent. Secondly, the hierarchical clustering method with complete linkage rule and dissimilarity in Eq. (2) is applied to the 41 TED

residual time series, yielding $K = 3$ clusters of users (selecting K through the average silhouette width). Fig. 1 shows the time invariant characteristics of DH users for the final clusters. Based on the obtained partition, we model the

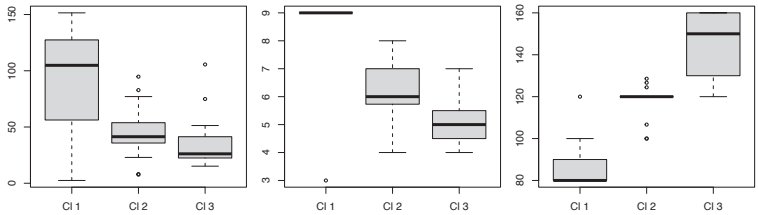


Figure 1. *Heating surface (left panel), age class (middle panel), energy class (right panel) for each cluster, from Cl 1 to Cl 3.*

temporal dynamics of TED aggregated by cluster through a suitable SARIMA model, identified via the AIC and validated by checking for residual autocorrelation ($SARIMA(2,0,0)(1,1,1)$ with a drift for SR, $SARIMA(0,1,2)(2,1,0)$ for OT, and $SARIMA(1,1,2)(0,1,1)$ for TED in the first and third cluster, $SARIMA(1,1,1)(1,1,2)$ for TED in the second cluster). Finally, we model the dependence relationship between each residual series and SR and OT via the conditional copula in Eq. (1), where a parametric copula model is chosen among the Elliptical, the Archimedean and the Joe family (Durante & Sempi, 2015). For all the three clusters, the resulting copula model is the Student-t, which is selected on the basis of the AIC and estimated via maximum likelihood. Fig. 2 shows the behaviour of TED for extreme values (low quantiles) of OT and SR for each identified cluster. While weather conditions have a clear

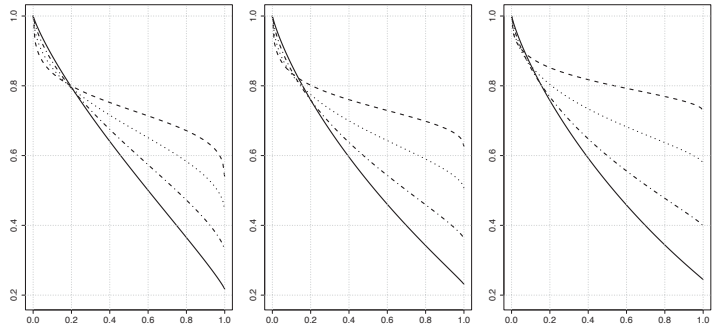


Figure 2. *Copula-based conditional probability function in Eq. (1) with* $(U_1, U_2) < 0.3$ *(solid line),* < 0.15 *(dash-dotted line),* < 0.05 *(dotted line),* < 0.01 *(dash line) for each identified cluster: residual TED quantile (x-axis).*

impact regardless of the cluster considered, it is also evident that TED behaves differently in the three clusters. For instance, the probability of exceeding the 80-th percentile of TED given that SR and OT are smaller than their 0.01-th quantile is 0.671, 0.698, and 0.769 for the first, second and third cluster, respectively. Moreover, the conditional probability increases more quickly as the weather tends to a more extreme scenario for the third cluster in comparison to the other two clusters. Indeed, the first cluster, which includes large, new and efficient buildings, is characterized by the lowest impact of extreme events of TED; the third cluster shows the strongest effect, including old and small buildings, with the lowest energy performance; the second seems to be characterized by an intermediate behaviour (the buildings are small, with medium energy class and heterogeneous age).

These findings can contribute to the study of the impact of meteorological conditions on the energy needs of buildings in the urban area, thus supporting the efficient management and production of thermal energy.

References

ALI, M., MIKHAIL, N.N., & HAQ, M.S. 1978. A class of bivariate distributions including the bivariate logistic. *J. Multivar. Anal.*, **8**(3), 405–412.

DI LASCIO, F.M.L., MENAPACE, A., & RIGHETTI, M. 2020. Joint and conditional dependence modelling of peak district heating demand and outdoor temperature: a copula-based approach. *Stat. Methods Appt.*, **29**, 373–395.

DI LASCIO, F.M.L., MENAPACE, A., & RIGHETTI, M. 2021. Analysing the relationship between district heating demand and weather conditions through conditional mixture copula. *Environ. Ecol. Stat.*, **28(1)**, 53–72.

DI LASCIO, F.M.L., MENAPACE, A., & PAPPADÀ, R. 202X. A spatial AMH copula-based dissimilarity measure to cluster variables in panel data. *BEMPS wp. Under review.*, 1–18. https://econpapers.repec.org/paper/bznwpaper/bemps89.htm.

DURANTE, F., & SEMPI, C. 2015. *Principles of Copula Theory*. Boca Raton: CRC Press.

MENAPACE, ANDREA, SANTOPIETRO, SIMONE, GARGANO, RUDY, & RIGHETTI, MAURIZIO. 2021. Stochastic Generation of District Heat Load. *Energies*, **14**(17).

TRIVEDI, P. K., & ZIMMER, D. M. 2005. Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, **1(1)**, 1–111.

HIERARCHICAL PERCENTILE CLUSTERING TO ANALYSE GREENHOUSE GAS EMISSIONS FROM AGRICOLTURE IN EUROPEAN UNION

F. Marta L. Di Lascio [2], Fabrizio Durante [1], Aurora Gatto [2]

[1] Department of Economic Sciences, University of Salento, Centro Ecotekne, 73100, Lecce, Italy, (e-mail: fabrizio.durante@unisalento.it)

[2] Faculty of Economics and Management, Free University of Bozen-Bolzano, Piazza Università, 1-39100, Bozen-Bolzano, Italy, (e-mail: marta.dilascio@unibz.it, aurora.gatto@unibz.it)

ABSTRACT: One of the key issues in the European Union's environmental policy concerns greenhouse gas emissions reduction, which is a challenge task to mitigate climate change. In this paper we investigate the emissions of different greenhouse gases from agriculture in the European Union countries through the innovative agglomerative hierarchical percentile clustering algorithm.

KEYWORDS: agriculture, greenhouse gas emission, hierarchical percentile clustering.

1 Introduction

Clustering methods are unsupervised techniques useful to identify structure underlying the data with the aim of gaining insights into their generating process. Durante *et al.*, 2021 proposed a method called *Agglomerative Hierachical Percentile Clustering* (AHPC) that allows us to deal with experimental errors and/or uncertainty affecting the observations. Hence, AHPC algorithm aims to cluster objects represented by repeated measurements on a set of variables.

Recently, European Union (EU) established to reduce greenhouse gas (GHG) emissions by at least 55% by 2030 compared to 1990 levels, and achieve climate neutrality by 2050. In particular, the emissions from agriculture account for about 11% of EU-27 emissions and in 2020 were about the same as in 2005. Hence, in the coming years substantial GHG emission reductions across all the sectors of the economy, including agriculture, are expected.

Our interest is to investigate the possible different behavior of EU countries in GHG emissions from agriculture through the innovative AHPC. To this aim, Sect. 2 presents the AHPC algorithm, while Sect. 3 contains the empirical analysis and a discussion of the findings. Finally, Sect. 4 concludes the paper.

2 The AHPC algorithm

The percentile clustering (PC) is a clustering method based on a dissimilarity matrix computed according to the percentile approach suggested in a seminal paper by Janowitz & Schweizer, 1989. Recently, Durante *et al.*, 2021 have developed the PC method in the hierarchical clustering framework and investigated its performance on both simulated and observed data.

Suppose to cluster d objects, i.e. statistical units, on which p different variables are observed. Also assume that each object i is associated with a set of n_i repeated observations, e.g. observations over time, and thus represented by a $(n_i \times p)$-dimensional matrix $\mathbf{X}_i$. The AHPC algorithm can be summarized as follows:

1. for each pair of objects i and j, with $i \neq j$ and $i, j = 1, \ldots, d$, i.e. for each pair of matrices $\mathbf{X}_i$ and $\mathbf{X}_j$:
   (a) compute the Euclidean distance $d_{k\ell}^{ij}$ between the k-th row of $\mathbf{X}_i$ and the ℓ-th row of $\mathbf{X}_j$ for every $k = 1, \ldots, n_i$, $\ell = 1, \ldots, n_j$;
   (b) compute the dissimilarity p_{ij} as the α-quantile of the $d_{k\ell}^{ij}$ distances computed at the step (a) by varying k and ℓ; α can obviously assume value in $[0, 1]$;
2. create the $(d \times d)$-dimensional dissimilarity matrix $\mathbf{P} = (p_{ij})$ with $p_{ii} = 0$ and $p_{ij} = p_{ji}$;
3. apply the classical agglomerative hierarchical clustering algorithm (Everitt *et al.*, 2011) choosing a linkage rule among the classical ones, e.g. single, average and complete linkage, and using $\mathbf{P}$ as dissimilarity matrix.

Roughly speaking, the AHPC is a classical hierarchical clustering algorithm based on a dissimilarity matrix computed through the α-percentile of the distribution functions of Euclidean distances between each pair of considered objects. Hence, the AHPC has interesting features. Firstly, the AHPC allows us to exploit prior knowledge of each object that can be observed for a different number of times, i.e. $n_i \neq n_j$, when $i \neq j$. Secondly, the possible presence of missing values does not prevent the application of the method. Thirdly, it should be noted that the dissimilarity matrix based on the α-percentile is actually an ordered weighted aggregation function (see, e.g., Yager, 2000). However, it is important to stress that considering percentiles as dissimilarities only takes into consideration the ranks of the objects (see, e.g., Cena & Gagolewski, 2020).

3 Case study

In order to illustrate the usefulness of the AHPC method, we analyze the GHGs emissions from agriculture. We consider the methane-CH4, carbon dioxide-CO2 and nitrous oxide-N2O per capita emissions (expressed in CO2 equivalent) for the EU countries (excluding Malta due to lack of available CO2 data) over the period 2012-2020 (source: European Environment Agency). We standardise each variable to ease the interpretation of findings. Next, we apply the AHPC algorithm using the complete linkage and the percentile level $\alpha = 0.75$ as suggested by the Monte Carlo simulation results in Durante *et al.*, 2021. Ignoring the two-cluster solution that is low informative due to the presence of a singleton cluster, i.e. Ireland, we consider the second highest value of the Average Silhouette Width (Rousseeuw, 1987) (0.346) that suggests a partition in five clusters.

To interpret the obtained results we consider the boxplots of each analysed GHG variable according to the obtained partition (see Fig. 1). We do not represent the fifth cluster because it is only formed by Ireland whose average emissions are extremely higher than those of the other EU countries for all the considered gases (average values of CH4, CO2, and N2O are 4.632, 3.790, and 3.354, respectively). It appears that countries belonging to the second cluster are the ones with lowest GHGs emissions per capita, while the first and the third cluster only show a lower-than-average emissions per capita for two of the three considered gases. Finally, the countries in the fourth cluster are the least virtuous, with emission values for each gas considered higher than the corresponding mean value for all EU countries. It is interesting notice that the first cluster is those with the highest number of outliers, meaning that there are some countries with extreme GHG emissions per capita.

4 Conclusions

We have analysed the GHG emissions per capita in the EU, using the innovative AHPC algorithm. We have found that the GHG emissions are different across countries and kind of gas. This supports the need of adopting a common policy to reach the EU goal of reducing GHG emissions.

Acknowledgments

This study was carried out within the Agritech National Research Center and received funding from the European Union Next-Generation EU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1032 17/06/2022, CN00000022). In particular,

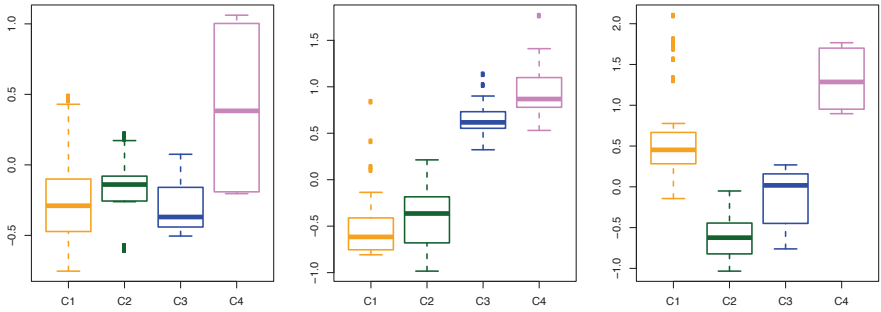


Figure 1. *Greenhouse gases emissions of CH4 (left panel), CO2 (middle panel), and N2O (right panel) by varying clusters from C1 to C4.*

our study represents an original paper related to the Spoke 4 "Multifunctional and resilient agriculture and forestry systems for the mitigation of climate change risks" and a baseline for the fulfilment of the milestones within the Task 4.2.3. "Big data analysis and decision support systems for the climate adaptation of agricultural and forestry systems". This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

FD has been supported by MIUR-PRIN 2017, Project "Stochastic Models for Complex Systems" (No. 2017JFFHSH). FD is also member of "ICSC - Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing".

References

CENA, ANNA, & GAGOLEWSKI, MAREK. 2020. Genie+OWA: Robustifying hierarchical clustering with OWA-based linkages. *Inf. Sci.*, **520**, 324–336.

DURANTE, FABRIZIO, GATTO, AURORA, & SAMINGER-PLATZ, SUSANNE. 2021. On Agglomerative Hierarchical Percentile Clustering. *Pages 616–623 of: Joint Proceedings of the 19th World Congress of IFSA, the 12th Conference of EUSFLAT, and the 11th International Summer School on AGOP*.

EVERITT, B. S., LANDAU, S., LEESE, M., & STAHL, D. 2011. *Cluster analysis*. Fifth edn. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

JANOWITZ, M. F., & SCHWEIZER, B. 1989. Ordinal and percentile clustering. *Math. Soc. Sci.*, **18**(2), 135–186.

ROUSSEEUW, P.J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. of Comp. and Appl. Math.*, **20**, 53–65.

YAGER, R.R. 2000. Intelligent control of the hierarchical agglomerative clustering process. *IEEE Trans. Syst. Man Cybern. Syst., Part B (Cybernetics)*, **30**(6), 835–845.

MAXIMUM LIKELIHOOD APPROACH TO PARAMETER SELECTION IN THE SPECTRAL CLUSTERING ALGORITHM

Cinzia Di Nuzzo[1], Salvatore Ingrassia[1]

[1] Department of Economics and Business, University of Catania, (e-mail: cinzia.dinuzzo@unict.it, salvatore.ingrassia@unict.it)

ABSTRACT: Automatic selection of the parameter in the spectral clustering algorithm through the mixture model approach has been considered. Specifically, a maximum likelihood approach using the Gaussian mixture model to select the proximity parameter in the self-tuning kernel function has been introduced.

KEYWORDS: Spectral clustering, parameters selection, gaussian mixture model.

1 Introduction

Spectral clustering methods are based on graph theory, where data are represented by the vertices of an undirected graph and the edges are weighted by the similarities between pairs of units, see von Luxburg, 2007, Shi & Malik, 2000, Ng *et al.* , 2001. Specifically, the spectral approach is based on the properties of the pairwise similarity matrix coming from a suitable kernel function. Then the clustering problem is reformulated as a graph partition problem.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathbb{R}^p$ be a set of units. In order to cluster $\mathbf{X}$ in K clusters, the first step of the spectral clustering algorithm concerns the definition of a symmetric and continuous function $\kappa : \mathbf{X} \times \mathbf{X} \to [0, \infty)$ called kernel function. Afterwards, a similarity matrix $\mathbf{W} = (w_{ij})$ can be assigned by setting $w_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$, for $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$. Specifically, here, we consider the following *self-tuning* kernel function (see Zelnik-Manor & Perona, 2004)

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{\varepsilon_i \varepsilon_j}\right), \qquad i, j = 1, \ldots, n, \tag{1}$$

with $\varepsilon_i = \|\mathbf{x}_i - \mathbf{x}_h\|$, where $\mathbf{x}_h$ is the h-th neighbour of point $\mathbf{x}_i$ (similarly for ε_j). Afterward, the normalized graph Laplacian is introduced as the $n \times n$ matrix $\mathbf{L}_{\text{sym}} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, where $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_n)$ is the *degree matrix*; d_i is the *degree* of the vertex $\mathbf{x}_i$ defined by $d_i = \sum_{j=1}^{n} w_{ij}$ and $\mathbf{I}$ denotes the $n \times n$ identity matrix. The spectral clustering algorithm works on the embedded space. Given K, let $\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K\}$ be the eigenvectors corresponding to the K

smallest eigenvalues of $\mathbf{L}_{\text{sym}}$. Then the normalized Laplacian embedding is defined as the map $\Phi_{\boldsymbol{\Gamma}} : \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \to \mathbb{R}^K$ given by $\Phi_{\boldsymbol{\Gamma}}(\mathbf{x}_i) = (\gamma_{1i}, \ldots, \gamma_{Ki})$, for $i = 1, \ldots, n$. Let $\mathbf{Y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')$ be the $n \times K$ matrix of the embedded data, where $\mathbf{y}_i = \Phi_{\boldsymbol{\Gamma}}(\mathbf{x}_i)$ for $i = 1, \ldots, n$. Finally, the embedded data $\mathbf{Y}$ are clustered according to some clustering procedure. Usually, this latter step is performed using the k-means algorithm, here, mixture models have been taken into account, since they are more robust approaches with respect to the choice of parameter of the spectral clustering algorithm, see Di Nuzzo & Ingrassia, 2022b for details.

As a matter of fact, in the spectral clustering algorithm, there are two free parameters to be tuned: the local scale parameter h in the kernel function (1) and the number of clusters K. Specifically, the kernel function plays an important role in the spectral clustering context because it affects the entire structure of the data. For this reason, the goal of many authors has been to find an automatic or heuristic way to select the kernel function with the corresponding scale parameter.

In this framework, given the number of clusters K, a proposal of an automatic method for parameter selection in the kernel function (1) via the Gaussian mixture model according to the maximum likelihood approach is introduced.

The rest is organized as follows: in Section 2 a maximum likelihood approach to select the parameter h in (1) is introduced; in order to confirm the validity of methodology, in Section 3 some numerical examples are shown.

2 Maximum likelihood approach to parameter selection

In this section, an automatic criterion to select the parameter h in the self-tuning kernel function (1) is introduced. Note that for the sake of simplicity, we introduce this approach by using the self-tuning kernel function (1), but it can be extended to other kernel functions proposed in the spectral clustering context, see e.g Zhang & Yu, 2011, John C.R., 2020, Park S., 2021.

The parameter h in (1) has a key role in pre-processing data because it affects the geometrical structure of the graph in terms of weight associated with any pairs of vertices in the graph. Specifically, in Di Nuzzo & Ingrassia, 2022a a graphical approach to select the parameters of the spectral clustering algorithm has been considered. The results in Di Nuzzo & Ingrassia, 2022a show that by analysing the graphic features of the embedded space and the number of the diagonal blocks of the similarity matrix $\mathbf{W}$, an optimal number of groups K can be easily selected. However, the choice of the parameter h isn't always easy to select. Therefore, without a criterion to address this problem,

different values of h can be considered optimal choices.

More precisely, as h varies, we have different configurations of the data in the embedded space, so we select h such that the embedded data are fitted by a Gaussian mixture model as much as possible. Therefore, we don't apply the Gaussian mixture model for fitting a given data set, but we look for the parameter h such that the corresponding data set is fitted by the Gaussian mixture model as much as possible.

For this purpose, we analyse the maximum log-likelihood parameter estimates deriving from the Gaussian mixture model using the EM algorithm and set h according to the maximum log-likelihood. In other words, we fit a Gaussian mixture model (with a fixed number K of components), according to the maximum likelihood approach, to different data sets corresponding to different $h \in \mathcal{H}$, where $\mathcal{H} \subseteq \{1,\dots,n-1\}$ is the collection of possible parameters h considered in the numerical experiments. Then we get a set of maximum likelihood values $l_1,\dots,l_{|\mathcal{H}|}$ for each data set, and select h^* leading to the overall maximum likelihood value, i.e. $h^* = \text{argmax}_h l_h$. Our proposal is summarized in Algorithm 1.

Algorithm 1 Parameter selection h in (1)

1. $\forall h \in \mathcal{H}$, compute the spectral clustering algorithm where the last step is executed with Gaussian mixture model.
2. $\forall h \in \mathcal{H}$, compute the log-likelihood value using EM algorithm obtaining the log-likelihood set $\mathcal{L} = \left\{ l_1,\dots,l_{|\mathcal{H}|} \right\}$.
3. Select h according to the maximum log-likelihood value, i.e. h^* corresponds to $l^* = \max \mathcal{L}$.

3 Numerical examples

Numerical examples according to the proposed approach (Algorithm 1) are here presented.

Table 1. Toy data.

h	Acc	ARI	Lik
1	1	1	**3961.853**
2	1	1	2658.617
10	1	1	2463.996
20	0.9866	0.9444	2424.739

Table 2. Flame data.

h	Acc	ARI	Lik
2	0.9875	0.9501	**344.7159**
5	0.9125	0.6789	238.3863
10	0.9042	0.6517	307.1519
48	0.8583	0.5116	244.413

Toy. Toy data (`http://cs.joensuu.fi/sipu/datasets/`) consists of $n = 373$ units, $p = 2$ variables and $K = 2$ clusters. In Table 1 we list, for

some parameters, the accuracy, ARI, and the log-likelihood values, the optimal choice according to Algorithm 1 corresponds to $h = 1$.

Flame. The Flame data (`http://cs.joensuu.fi/sipu/datasets/`) consists of $n = 240$ units, $p = 2$ variables and $K = 3$ clusters. In Table 2 we list ARI and log-likelihood values for some h parameters. Also in this case, the maximum log-likelihood corresponds to the maximum value for accuracy and this confirms our proposal.

Acknowledgement

Acknowledgement of financial support from PNRR MUR project PE0000013-FAIR.

References

Di Nuzzo, C., & Ingrassia, S. 2022a. A graphical approach for the selection of the number of clusters in the spectral clustering algorithm. *Pages 31–44 of:* Salvati, Nicola, Perna, Cira, Marchetti, Stefano, & Chambers, Raymond (eds), *Studies in theoretical and applied statistics*. Cham: Springer International Publishing.

Di Nuzzo, C., & Ingrassia, S. 2022b. A mixture model approach to spectral clustering and application to textual data. *Statistical methods & applications*, **31**(4), 1071–1097.

John C.R., Watson D., Barnes M.R. Pitzalis C. Lewis M.J. 2020. Spectrum: fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics.*, **36**(4).

Ng, A. Y., Jordan, M. I., & Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Page 849–856 of: Proceedings of the 14th international conference on neural information processing systems: Natural and synthetic*. NIPS'01. Cambridge, MA, USA: MIT Press.

Park S., Xu H., Zhao H. 2021. Integrating multidimensional data for clustering analysis with applications to cancer patient data. *Journal of the american statistical association*, **116**(533), 14–26.

Shi, J., & Malik, J. 2000. Normalized cuts and image segmentation. *Ieee transactions on pattern analysis and machine intelligence*, **22**(8), 888–905.

von Luxburg, U. 2007. A tutorial on spectral clustering.tutorial on spectral clustering. *Statistics and computing*, **17**.

Zelnik-Manor, L., & Perona, P. 2004. Self-tuning spectral clustering. *In:* Saul, L., Weiss, Y., & Bottou, L. (eds), *Advances in neural information processing systems*, vol. 17. MIT Press.

Zhang, X., Li J., & Yu, H. 2011. Local density adaptive similarity measurement for spectral clustering. *Pattern recognition letters*, **32**(2), 352 – 358.

FINITE MIXTURE MODELS: A SYSTEMATIC REVIEW

José G. Dias [1]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisboa, Portugal, (e-mail: jose.dias@iscte-iul.pt)

ABSTRACT: Finite mixture models have been used in many fields for different purposes and under different names. Other non-exact names are model-based clustering and latent class models. This presentation gives an overview of this area. In particular, it summarizes the research that has been published both theoretical papers and in applications. A systematic literature review using the PRISMA methodology was used. The text mining analysis then identifies topics in the literature. Results show an explosion of the research in the field since 2000. Social and health sciences are the most prominent application areas, mainly focused on the detection of unobserved heterogeneity.

KEYWORDS: finite mixture models, latent class models, discrete latent variables, model-based clustering, systematic literature review.

Finite mixture (FM) models and related latent variable models are over one hundred years old. The origin of the FM model is usually attributed to Newcomb and Pearson (i.e., Newcomb, 1886; Pearson, 1894). Stigler, 1986, however, at least traces its origin back to the analysis of conviction rates by Poisson in the second quarter of the nineteenth century. Since 2000, the use of these models has grown exponentially. In the past few decades, advances in computer technology, FM modeling has proven to be a powerful tool for the analysis of a wide range of empirical problems. For instance, in the social sciences, which have a long tradition of latent class (LC) models, following the seminal work by Lazarsfeld and refinements notably by Goodman and Clogg (see, e.g., Goodman, 1974 and Clogg, 1995), more sophisticated models are gaining popularity. McLachlan & Peel, 2000 provide a good overview of the field until 2000. The exponential growth in the use of these models over the past two decades clearly shows that they are directly related to the democratization of statistical computation using fast personal computers (PCs) and increasing availability of software for their estimation.

This work presents an overview of the field using a systematic literature review. In addition to searching for articles using keywords to retrieve papers, we also used papers citing well-known references in the field (e.g., Titterington

et al. , 1985; McLachlan & Peel, 2000; Scrucca *et al.* , 2016). The extraction and selection of papers from the Web of Science follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology. A total of 38,997 papers were included in the analysis. Topic analysis, a special case of text mining, is used to identify topic clusters in the corpus.

Results show the diverse use of FMs in the literature. Most publications use FMs to identify clusters. However, in other applications and contexts, topics cover density estimation, defining prior probabilities in Bayesian statistics, discrete latent variables, the golden standard problem, speech modeling, imagine analysis, longitudinal and trajectory analysis, or social class analysis. This research establishes a typology in the field of FM methodology and shows its wide range and flexible use in statistical modeling.

References

CLOGG, C. C. 1995. Latent class models. *Pages 311–359 of:* ARMINGER, G., CLOGG, C.C., & SOBEL, M.E. (eds), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. New York: Plenum.

GOODMAN, L. A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**(2), 215–231.

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

NEWCOMB, S. 1886. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, **8**(4), 343–366.

PEARSON, K. 1894. Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, **185**, 71–110.

SCRUCCA, L., FOP, M., MURPHY, T. B., & RAFTERY, A. E. 2016. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, **8**(1), 289–317.

STIGLER, S.M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA and London: Belkap Press of Harvard University Press.

TITTERINGTON, D.M., SMITH, A.F.M., & MAKOV, U.E. 1985. *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley & Sons.

MEASUREMENT INVARIANCE: A METHOD BASED ON LATENT MARKOV MODELS

Francesco Dotto[1], Roberto Di Mari[2], Alessio Farcomeni[3] and Antonio Punzo[2]

[1] Department of Economics, University of Roma Tre, (e-mail: francesco.dotto@uniroma3.it

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

[3] Department of Economics and Finance, University of Tor Vergata, (e-mail: alessio.farcomeni@uniroma2.it)

ABSTRACT: We define differential item functioning in the context of panel data. We then present a general approach to detect measurement non-invariance cases in this context. We use a model selection procedure based on the Bayesian information criterion (BIC). A real data application and a simulation study are presented to illustrate and motivate the methods.

KEYWORDS: latent Markov model, measurement invariance, panel data

1 Introduction

Much empirical work in general social sciences leverages on questionnaire data analysis to measure possibly unobserved (latent) traits. In these contexts crucial working assumptions are that items have the same discriminatory power, unidimensionality of the latent trait, and measurement equivalence (invariance) in the scale.We focus on the assessment of potential violations of measurement equivalence in longitudinal studies. Namely, when respondents are repeatedly measured over time, and the model for the latent trait is not strong enough to describe the dependence structure among the items and external variables. This phenomenon is also known as differential item functioning (DIF), and we study it in connection with latent Markov models (see, e.g., Bartolucci *et al.*, 2013). We specify distinct notions of DIF, combining and extending ideas from Kankaraš *et al.*, 2018 and Masyn, 2017. Effectively, we develop a toolkit based on a classical model selection tool - i.e., the Bayesian Information Criterion - to select the most appropriate DIF configuration. An extended presentation of the technical framework, and of both numerical and real-data results is available in Di Mari *et al.*, 2022.

2 Mathematical Formulation

Let Y_{ith}, $h = 1,\ldots,H$, be the h-th dichotomous indicator, measured for the i-th subject, $i = 1,\ldots,n$, at time t, $t = 1,\ldots,T$; observed alongside a vector of time-specific covariates $\mathbf{X}_{it}$. In addition, let U_{it} denote a discrete latent variable with support $\{1,\ldots,K\}$, which follows a possibly inhomogeneous first-order Markov chain. In case of measurement invariance (no DIF), we assume the data arise from the following model

$$\begin{cases} P(Y_{it1} = y_1,\ldots,Y_{itH} = y_H \mid U_{it} = k) = \prod_{h=1}^{H} \phi_{h|k}^{y_h}(1-\phi_{h|k})^{1-y_h}, \\ \log\left[\dfrac{P(U_{i1} = k \mid X_{i1})}{P(U_{i1} = 1 \mid \mathbf{X}_{i1})}\right] = \alpha_{1k} + \beta_{1k}\mathbf{X}_{i1}, \\ \log\left[\dfrac{P(U_{it} = k \mid U_{i,t-1} = j, \mathbf{X}_{it})}{P(U_{it} = j \mid U_{i,t-1} = j, \mathbf{X}_{it})}\right] = \alpha_{kj} + \beta_{kj}\mathbf{X}_{it}, \end{cases} \tag{1}$$

where the first equation denotes the measurement model which involves the item specific probabilities $\phi_{h|k}^{y_h}$. The two remaining equations define structural models for the initial and transition probabilities. The parameters α and β model the effect of the covariates on both the initial and transition probabilities. For the sake of simplicity we assume, as commonly done within this context, that such regression coefficients are time constant letting the covariate values be the driver of time heterogeneity. If DIF is allowed, the measurement model depends on $\mathbf{X}_{it}$ as

$$\text{logit}(\phi_{h|k}) = \gamma_{hk} + \eta_{htk}\mathbf{X}_{it}. \tag{2}$$

where γ is the intercept term and η_{htk} represents the direct effect of the covariate on the item specific probabilities. From equation (2) other DIF scenarios can be derived:

1. No DIF: The covariates only affect transition probabilities but they do not affect item specific probabilities.
2. Full DIF: The covariates affect both the transition probabilities and the item specific probabilities. The η_{htk} vector varies across items, time, and class.
3. Time-Constant DIF: The η_{htk} vector varies across items and class, but remains fixed across time.
4. State-Constant DIF: The η_{htk} vector varies across time and item, but remains fixed across latent states.
5. State- and Time- constant DIF: The η_{htk} vector is homogeneous across time and latent states.

3 Results

We analyse show syntetic simulation results and a real data analysis. These somewhat summarize the results reported in Di Mari *et al.*, 2022.

3.1 A simulation study

Table 1 reports the performance of the methodology in terms of rate of correct classification over 500 replicates for each setting. The fabricated data sets are based on $n = 500$, $T = 4$, $K = 3$, $H = 10$, with a single standard Gaussian covariate. It can be seen that the proper model is always selected with high probability.

3.2 General social survey: Measuring tolerance toward non-conformity

Data are taken from the American General Social Survey (GSS), a survey of the English-speaking, non-institutionalized adult population of the United States. The $H = 5$ binary items are formulated as follows: "Suppose ... wanted to make a speech in your community. Should he be allowed to speak?" and are referred to communists, atheists, militarists, homosexuals, and racists.

We include the covariate "Education", which we re-code into three categories. The best fit reveals a direct effect of Education on items, pointing out that to a higher education corresponds, on average, a higher probability to allow "Atheists" "Communists" "Homosexuals" "Militarists" and "Muslims" to speak in public.

The lowest BIC is attained at time- and state-constant DIF (DIF 4), i.e., for differing levels of education, individuals have varying probabilities of scoring "Yes" to the items, regardless of the underlying tolerance (latent) type, and record time (see Figure 1).

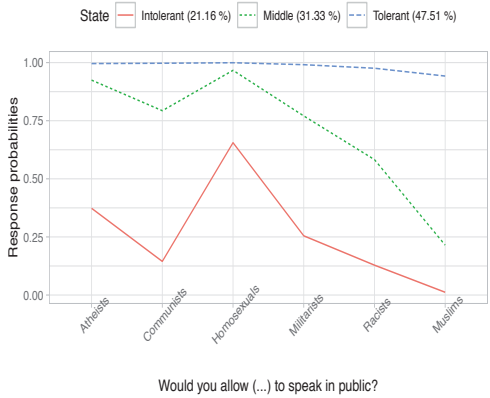
References

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2013. *Latent Markov models for longitudinal data*. Chapman and Hall / CRC Press, New York.

DI MARI, R., DOTTO, F., FARCOMENI, A., & PUNZO, A. 2022. Assessing measurement invariance for longitudinal data through latent Markov models. *Structural Equation Modeling: A Multidisciplinary Journal*, **29**, 381–393.

BIC	True Model				
	No DIF	Full DIF	Time Constant DIF	State Constant DIF	State Time Constant DIF
No DIF	1.00	0.00	0.01	0.00	0.00
Full DIF	0.00	1.00	0.00	0.00	0.00
Time constant DIF	0.00	0.00	0.99	0.00	0.00
State constant DIF	0.00	0.00	0.00	1.00	0.00
State Time constant DIF	0.00	0.00	0.00	0.00	1.00

Table 1. *Confusion matrix normalized by column to evaluate the BIC performance.*



	DIF 4
$\eta_{1..}$	0.76***
("Atheists")	(0.17)
$\eta_{2..}$	1.12***
("Communists")	(0.20)
$\eta_{3..}$	0.85***
("Homosexuals")	(0.15)
$\eta_{4..}$	0.39***
("Militarists")	(0.14)
$\eta_{5..}$	0.09
("Racists")	(0.14)
$\eta_{6..}$	1.16***
("Muslims")	(0.23)

Figure 1. *GSS data: Estimated response probabilities to answer "Yes" given state membership (on the left) and estimates of the direct effect η_h of the covariate "Education" on the six items available according to the time- state-constant DIF (on the right). Standard errors in parentheses are based on the observed Information matrix.*

KANKARAŠ, MILOŠ, MOORS, GUY, & VERMUNT, JEROEN K. 2018. Testing for measurement invariance with latent class analysis. *Pages 393–419 of: Cross-Cultural Analysis*. Routledge.

MASYN, K. E. 2017. Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, **24**, 180–197.

A Comparison between the Varying-Thresholds Model and Quantile Regression

Niccolò Ducci [1], Leonardo Grilli[2] and Marta Pittavino[2]

[1] Agenzia delle Entrate (e-mail: niccolo.ducci.it@gmail.com)

[2] Department of Statistics, Computer Science, Applications "Giuseppe Parenti", University of Florence (e-mail: leonardo.grilli@unifi.it, marta.pittavino@unifi.it)

Abstract: The varying-thresholds model is a new modelling approach capable of estimating the whole conditional distribution of a response variable in a regression setting. The varying-thresholds model can be used for continuous, ordinal and count responses. Conditional quantiles estimated through the varying-thresholds method are compared to those of quantile regression. The comparison is based on models' simulations to assess the performance of the two methodologies regarding the coverage and width of prediction intervals. The simulation study encompasses eight different settings with several functional forms and types of errors. In addition, a discrete variation of the continuous ranked probability score is proposed as a way to choose the best link function for the binary models used to estimate the varying-thresholds model. The comparison shows that the varying thresholds model performs better whenever the functional form of the true data generating model is non-linear.

Keywords: varying-thresholds model, quantile regression, robit, prediction intervals, continuous ranked probability score

1 The Varying-Thresholds Model

The varying-thresholds model is a novel methodology proposed by Tutz, 2021 that can estimate the whole conditional distribution of a response variable in a regression setting. Estimating the conditional distribution allows one to obtain values of interest such as the conditional expected value, standard error, or quantiles. The general form of the Varying-Thresholds Model can be written as follows:

$$P(Y > \theta \,|\, \mathbf{x}) = F(\,\eta(\theta, \mathbf{x})\,) \tag{1}$$

where Y is the response variable, $\mathbf{x}$ is a vector of covariates, F is a distribution function and $\eta(\theta, \mathbf{x})$ is a predictor function. The predictor function can take

Table 1. *All types of data generating models used in the comparison between quantile regression and the varying-thresholds model. Every model comprise a single covariate and a response variable.*

Model	*Functional Form*	Error Distribution	Covariate Distribution
Model 1	$\beta_0 + \beta_1 x$	$\varepsilon_N \sim N(0,1)$	$X \sim N(5,1)$
Model 2	$\beta_0 + \beta_1 x$	$\varepsilon_{\chi^2} \sim \chi^2(df = 3)$	$X \sim N(5,1)$
Model 3	$\beta_0 + \beta_1 x$	$\varepsilon_N \sim e^{(x-5)} \cdot N(0,1)$	$X \sim N(5,1)$
Model 4	$\beta_0 + \beta_1 x$	$\varepsilon_N \sim e^{(5-x)} \cdot N(0,1)$	$X \sim N(5,1)$
Model 5	$\beta_0 + \beta_1 x + \beta_2 x^2$	$\varepsilon_N \sim N(0,1)$	$X \sim U(-2,12)$
Model 6	$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$	$\varepsilon_N \sim N(0,1)$	$X \sim U(-3,8)$
Model 7	$\beta_0 + \beta_1 x$	$\varepsilon_{CN} \sim 0.9N(0,5) + 0.1N(50,5)$	$X \sim N(0,5)$
Model 8	$\beta_0 + \beta_1 x + \beta_2 x^2$	$\varepsilon_t \sim t(df = 3)$	$X \sim U(0,6)$

many forms: linear, non-linear or non-parametric. In this work, we consider a single covariate x and adopt a linear specification $\eta(\theta, \mathbf{x}) = \beta_0^\theta + \beta_1^\theta x$. The response variable Y can be ordinal or continuous. The varying-thresholds model is estimated using a series of binary regression models: for every threshold θ in a prespecified grid of values, the response variable Y is dichotomized to become binary, then a model is fitted to the data as described in equation 1. This method allows for the estimation of varying coefficients, indexed by θ, that are then used to compute the conditional distribution of the response variable [*].

2 Data Generating Models and Simulation Settings

The varying-thresholds model and quantile regression are compared using a variety of error assumptions and different functional forms. Quantile regression is fitted as $Q_{Y|x}(\theta) = \beta_0(\theta) + \beta_1(\theta)x$, likewise the varying-thresholds model is estimated using the predictor function $\eta(\theta, \mathbf{x}) = \beta_0^\theta + \beta_1^\theta x$. All the data generating models are reported in Table 1. Model's errors mimic the

[*]Note that, even if the predictor is linear, the binary response model is repeatedly fitted with different thresholds, thus the regression function is estimated in a data-driven way.

latent response approach , i.e. $Y^* = functional form + error$ and $Y = 1$ if and only if $Y^* > 0$, e.g., a model with normally distributed error corresponds to the probit model. The errors are always standardized to ensure comparability of the regression coefficients. Quantile regression and the varying-thresholds model are compared through the empirical coverage of their estimated prediction intervals computed at a $(1-\alpha) = 80\%$ level conditioned on a given value of $X = x$. This interval is computed by estimating the first and ninth conditional decile. The empirical coverage is calculated through a simulation. The simulation has 1000 iterations, each time a different sample of $n = 1000$ observations is drawn from the generating model. After each iteration the two methodologies compute the intervals; then, a new observation is sampled from the generating model; the proportion of times the new observation falls within the prediction interval is the empirical coverage level. Quantile regression is estimated with the R package `quantreg`, Koenker, 2022.

3 Simulation Results and Link Selection

Table 2 reports the results of the simulations for prediction intervals conditioned on the median value of X. The comparison shows that the varying-thresholds model performs better whenever the functional form of the true data generating model is non-linear. The lack of assumptions about the functional relationship makes the varying-thresholds model a very flexible approach, capable of detecting non-linear effects without specifying a non-linear effect in the predictor function $\eta(\theta, \mathbf{x})$. If the functional relationship between variables is known in advance and it is correctly specified quantile regression generally yields better results. The choice of the link function for the binary models used to estimate the varying-thresholds model is crucial; a discrete approximation of the continuous ranked probability score (CRPS), Jordan *et al.*, 2019; Gneiting & Raftery, 2007, is used to select the best link function. Both out-of-sample or in-sample approaches seems to be valid with this metric. In $Model8$ the robit link function with three degrees of freedom is selected through the CRPS and yields better results than other links.

4 Conclusions

The varying-thresholds model performs better, regarding prediction intervals, than quantile regression when there are non-linear effects and the relationship between variables is not correctly specified. Link function selection for the binary models' estimation method can be facilitated using the CRPS. Areas of

Table 2. *Empirical coverage and average width of prediction intervals at* 80% *level on* 1000 *simulations from Model* $1-8$ *at the median value of X. The varying-thresholds model is fitted with probit link function except for Model* 8 *where it is fitted with robit*[a] *link function with three degrees of freedom.*

Model	*Quantile Regression*		*Varying-Thresholds Model*	
	Coverage	*Avg. Width*	*Coverage*	*Avg. Width*
Model 1	0.783	2.562	0.783	2.567
Model 2	0.820	5.670	0.822	5.788
Model 3	0.926	3.759	0.930	4.124
Model 4	0.937	3.755	0.939	4.121
Model 5	0.704	4.651	0.865	3.306
Model 6	1.000	8.559	0.883	3.354
Model 7	0.801	33.221	0.844	37.694
Model 8	0.649	7.964	0.810	3.442

[a]The robit link function is related to the t-distribution, see Liu, 2004.

future research may include different types of response variables such as count and ordinal data.

References

GNEITING, T., & RAFTERY, A. E. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**(477), 359–378.

JORDAN, A., KRÜGER, F., & LERCH, S. 2019. Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software*, **90**(12), 1–37.

KOENKER, R.W. 2022. *quantreg: Quantile Regression*. R package version 5.94.

LIU, C. 2004. *Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression*. John Wiley & Sons, Ltd. Chap. 21, pages 227–238.

TUTZ, G. 2021. *Flexible Predictive Distributions from Varying-Thresholds Modelling*. arXiv:2103.13324.

EFFICIENT COMPUTATION OF PREDICTIVE PROBABILITIES IN PROBIT MODELS VIA EXPECTATION PROPAGATION

Augusto Fasano[1], Niccolò Anceschi[2], Beatrice Franzolini[3] and Giovanni Rebaudo[1,4]

[1] Collegio Carlo Alberto, Turin, IT (augusto.fasano@carloalberto.org)

[2] Duke University, Durham, USA (niccolo.anceschi@duke.edu)

[3] A*STAR, Singapore, SG (beatricef@sics.a-star.edu.sg)

[4] University of Turin, Turin, IT (giovanni.rebaudo@unito.it)

ABSTRACT: Binary regression models represent a popular model-based approach for binary classification. In the Bayesian framework, computational challenges in the form of the posterior distribution motivate still-ongoing fruitful research. Here, we focus on the computation of predictive probabilities in Bayesian probit models via expectation propagation (EP). Leveraging more general results in recent literature, we show that such predictive probabilities admit a closed-form expression. Improvements over state-of-the-art approaches are shown in a simulation study.

KEYWORDS: probit model, expectation propagation, Bayesian inference, extended multivariate skew-normal distribution

1 Introduction

Binary regression models represent a default model-based approach for binary classification. Although the theory in the frequentist setting is well established, flourishing research is still ongoing in the Bayesian framework, where such models are also used as benchmarks for posterior computations (Chopin & Ridgway, 2017). Here, we focus on the approximation of predictive probabilities via expectation propagation (EP) in the Bayesian probit model

$$y_i \mid \boldsymbol{\beta} \overset{ind}{\sim} \text{BERN}\left(\Phi\left(\mathbf{x}_i^\intercal \boldsymbol{\beta}\right)\right), i = 1, \ldots, n; \quad \boldsymbol{\beta} \sim \text{N}_p(\mathbf{0}, \nu^2 \mathbf{I}_p), \tag{1}$$

with $\boldsymbol{\beta} \in \mathbb{R}^p$ the unknown vector of parameters, $\mathbf{x}_i \in \mathbb{R}^p$ the covariate vector associated with observation i and $\mathbf{I}_p$ the identity matrix of dimension p. $\Phi(t)$ denotes the cumulative distribution function of a standard Gaussian random variable evaluated at t and $\phi_p(\mathbf{t}, \mathbf{S})$ will denote the density of a p-variate Gaussian random variable with mean $\mathbf{0}$ and covariance matrix $\mathbf{S}$, evaluated at $\mathbf{t}$.

We show that the EP approximate predictive probabilities admit a closed-form expression in terms of the output parameters returned by the EP routine. Such parameters can be obtained at per-iteration cost of $O(pn \cdot \min\{p,n\})$, as shown in Anceschi *et al.* (2023) for a broad class of models and derived in full detail for the probit model in Fasano *et al.* (2023).

2 Expectation Propagation (EP) review

Adapting more general results derived in Anceschi *et al.* (2023), Fasano *et al.* (2023) showed that, calling $\mathbf{y} = (y_1, \ldots, y_n)$, the EP approximation $q(\boldsymbol{\beta}) \propto \prod_{i=0}^{n} q_i(\boldsymbol{\beta})$ of the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y})$ for model (1) can be obtained by leveraging on extended skew-normal (SN) distributions (Azzalini & Capitanio, 2014). Except for $q_0(\boldsymbol{\beta})$, which is fixed equal to the prior $p(\boldsymbol{\beta})$, we take $q_i(\boldsymbol{\beta}) = \phi_p\left(\boldsymbol{\beta} - \mathbf{Q}_i^{-1}\mathbf{r}_i, \mathbf{Q}_i^{-1}\right)$, $i = 1, \ldots, n$, with the optimal $\mathbf{r}_i$'s and $\mathbf{Q}_i$'s to be obtained via the EP routine. Consequently, calling $\mathbf{r}_0 = \mathbf{0}$ and $\mathbf{Q}_0 = \nu^{-2}\mathbf{I}_p$, one gets $q(\boldsymbol{\beta}) = \phi_p(\boldsymbol{\beta} - \mathbf{Q}^{-1}\mathbf{r}, \mathbf{Q}^{-1})$, with $\mathbf{r} = \sum_{i=0}^{n} \mathbf{r}_i$, $\mathbf{Q} = \sum_{i=0}^{n} \mathbf{Q}_i$. At each EP cycle, the parameters $\mathbf{r}_i$ and $\mathbf{Q}_i$ of each site $i = 1, \ldots, n$ are updated by imposing that the first two moments of the global approximation $q(\boldsymbol{\beta})$ match the ones of the hybrid distribution

$$h_i(\boldsymbol{\beta}) \propto p(y_i \mid \boldsymbol{\beta}) \prod_{j \neq i} q_j(\boldsymbol{\beta}) = \Phi((2y_i - 1)\mathbf{x}_i^\intercal \boldsymbol{\beta}) \prod_{j \neq i} q_j(\boldsymbol{\beta}). \tag{2}$$

This is immediate after noticing that (2) coincides with the kernel of a multivariate extended skew-normal distribution $\text{SN}_p(\boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\alpha}_i, \tau_i)$, with

$$\boldsymbol{\xi}_i = \mathbf{Q}_{-i}^{-1}\mathbf{r}_{-i}, \;\; \boldsymbol{\Omega}_i = \mathbf{Q}_{-i}^{-1}, \;\; \boldsymbol{\alpha}_i = (2y_i - 1)\boldsymbol{\omega}_i \mathbf{x}_i, \;\; \tau_i = (2y_i - 1)(1 + \mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i)^{-1/2} \mathbf{x}_i^\intercal \boldsymbol{\xi}_i,$$

where $\mathbf{Q}_{-i} = \sum_{j \neq i} \mathbf{Q}_j$, $\mathbf{r}_{-i} = \sum_{j \neq i} \mathbf{r}_j$ and $\boldsymbol{\omega}_i = [\text{diag}\,(\boldsymbol{\Omega}_i)]^{1/2}$. Combining this with Woodbury's identity, Fasano *et al.* (2023) show that, for $i = 1 \ldots, n$, the updated quantities $\mathbf{Q}_i^{\text{NEW}}$ and $\mathbf{r}_i^{\text{NEW}}$ equal $k_i \mathbf{x}_i \mathbf{x}_i^\intercal$ and $m_i \mathbf{x}_i$, respectively, with $k_i = -\zeta_2(\tau_i) / \left(1 + \mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i + \zeta_2(\tau_i) \mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i\right)$ and $m_i = \zeta_1(\tau_i) s_i + k_i (\boldsymbol{\Omega}_i \mathbf{x}_i)^\intercal \mathbf{r}_{-i} + k_i \zeta_1(\tau_i) s_i \mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i$, having defined $\zeta_1(x) = \phi(x)/\Phi(x)$, $\zeta_2(x) = -\zeta_1(x)^2 - x\zeta_1(x)$ and $s_i = (2y_i - 1)(1 + \mathbf{x}_i^\intercal \boldsymbol{\Omega}_i \mathbf{x}_i)^{-1/2}$. These results, combined with the efficient computation of $\boldsymbol{\Omega}_i$ and update of the covariance matrix $\mathbf{Q}^{-1}$ of the Gaussian approximation $q(\boldsymbol{\beta})$, lead to an implementation of EP having a cost per iteration $O(p^2 n)$. When p is large, and especially when $p > n$, EP can be implemented at $O(pn^2)$ cost per iteration by storing and updating only the p-dimensional vectors $\mathbf{w}_i = \boldsymbol{\Omega}_i \mathbf{x}_i = \mathbf{Q}_{-i}^{-1} \mathbf{x}_i$ and $\mathbf{v}_i = \mathbf{Q}^{-1} \mathbf{x}_i$, $i = 1, \ldots, n$. Eventually, one can compute the full EP covariance matrix as

$$\mathbf{Q}^{-1} = \nu^2 \mathbf{I}_p - \nu^2 \mathbf{V}\mathbf{K}\mathbf{X}, \tag{3}$$

where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\intercal$ and $\mathbf{K} = \text{diag}(k_1, \ldots, k_n)$.

3 Closed-form EP predictive probabilities

One of the advantages of the Gaussian approximation provided by EP is that it results in a simple closed-form expression for the approximate predictive probability of observing $y_{\text{NEW}} = 1$ for a new statistical unit having covariate vector $\mathbf{x}_{\text{NEW}}$, namely $\Pr_{\text{EP}}[y_{\text{NEW}} = 1 \mid \mathbf{y}]$. Indeed, calling $\boldsymbol{\xi}_{\text{EP}} = \mathbf{Q}^{-1}\mathbf{r}$ and $\boldsymbol{\Omega}_{\text{EP}} = \mathbf{Q}^{-1}$ so that $q(\boldsymbol{\beta}) = \phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}_{\text{EP}}, \boldsymbol{\Omega}_{\text{EP}})$, it holds

$$\Pr_{\text{EP}}[y_{\text{NEW}} = 1 \mid \mathbf{y}] = \mathbb{E}_{q(\boldsymbol{\beta})}\left[\Phi\left(\mathbf{x}_{\text{NEW}}^{\intercal}\boldsymbol{\beta}\right)\right] = \Phi\left(\left(1+u\right)^{-1/2}\mathbf{x}_{\text{NEW}}^{\intercal}\boldsymbol{\xi}_{\text{EP}}\right), \qquad (4)$$

where $u = \mathbf{x}_{\text{NEW}}^{\intercal}\boldsymbol{\Omega}_{\text{EP}}\mathbf{x}_{\text{NEW}}$ and the last equality in (4) follows by Lemma 7.1 in Azzalini & Capitanio (2014). The only computationally relevant part in (4) is the computation of the quadratic form u. However, when $p < n$, $\boldsymbol{\Omega}_{\text{EP}}$ is directly returned by the algorithm, and u can be computed at cost $O(p^2)$. On the other hand, when $p > n$ (or in general when p is large), this direct computation can be avoided since, by (3), $u = \nu^2\left[\mathbf{x}_{\text{NEW}}^{\intercal}\mathbf{x}_{\text{NEW}} - \left(\mathbf{V}^{\intercal}\mathbf{x}_{\text{NEW}}\right)^{\intercal}\mathbf{K}\left(\mathbf{X}\mathbf{x}_{\text{NEW}}\right)\right]$, computable at cost $O(pn)$. Thus, Equation (4) provides an efficient closed-form approximation of the exact predictive probability $\Pr[y_{\text{NEW}} = 1 \mid \mathbf{y}]$, which can be computed at cost $O(p \cdot \min\{p, n\})$ from the EP parameters.

4 Simulation study

We show with a simulation study the advantages of combining the efficient EP implementation presented in Fasano *et al.* (2023) with the efficient computation of the predictive probabilities presented in Section 3. Fixing $n = 100$ and $\nu^2 = 25$, we compute the predictive probabilities for $\tilde{n} = 50$ test units in five different scenarios with synthetic data, for $p = 50, 100, 200, 400$ and 800. We compare the approximate predictive probabilities obtained with EP and with the partially-factorized variational approximation (PFM-VB) (Equation (9) in Fasano *et al.* (2022)) with the ones arising from a Monte Carlo approximation exploiting i.i.d. samples from the posterior (Durante, 2019). Figure 1 shows that EP can achieve superior accuracy for $p < 2n$, while in the other settings they provide comparable results. The EP running time ranges from 0.02 to 0.12 seconds, while for PFM-VB it ranges from 0.13 to 0.23. The slightly higher cost of PFM-VB is because, after convergence, the computation of predictive probabilities requires a sampling step that takes approximately 0.12 seconds. To conclude, the results presented in this work make the computation of EP approximate predictive probabilities feasible in settings where currently-available implementations are computationally impractical. Considering $p = 800$ for illustration, the function `EPprobit` from the `R` package `EPGLM`, requires 140

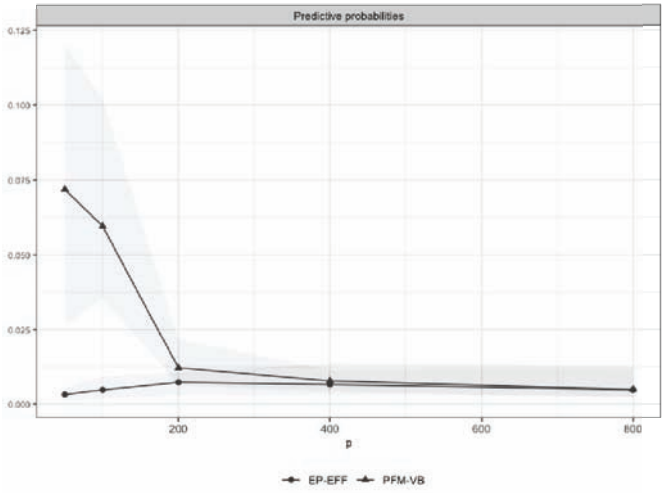


Figure 1. *For varying p, median absolute difference between the $\tilde{n} = 50$ predictive probabilities resulting from* 2000 *i.i.d. samples and the ones arising from* EP *and* PFM-VB *for probit regression with $n = 100$ and $\nu^2 = 25$. Grey areas denote the first and third quartiles.*

seconds, about 1000 times slower than the efficient implementation presented here. Code is available at https://github.com/augustofasano/EPprobit-SN.

References

ANCESCHI, N., FASANO, A., DURANTE, D., & ZANELLA, G. 2023. Bayesian conjugacy in probit, tobit, multinomial probit and extensions: a review and new results. *J. Am. Stat. Assoc.*, **118**, 1451–1469.

AZZALINI, A., & CAPITANIO, A. 2014. *The skew-normal and related families*. Cambridge Univ. Press.

CHOPIN, N., & RIDGWAY, J. 2017. Leave Pima Indians alone: binary regression as a benchmark for Bayesian computation. *Stat. Sci.*, **32**, 64–87.

DURANTE, D. 2019. Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, **106**, 765–779.

FASANO, A., DURANTE, D., & ZANELLA, G. 2022. Scalable and accurate variational Bayes for high-dimensional binary regression models. *Biometrika*, **109**, 901–919.

FASANO, A., ANCESCHI, N., FRANZOLINI, B., & REBAUDO, G. 2023. Efficient expectation propagation for posterior approximation in high-dimensional probit models. *Book of Short Papers - SIS 2023*, **in press**.

HOW WOMEN REACT TO THEIR PARTNERS' WORK INSTABILITY. THE ADDED-WORKER EFFECT

Donata Favaro[1], Anna Giraldo[2]

[1] Department of Economics and Management, University of Padova,
(e-mail: donata.favaro@unipd.it)

[2] Department of Statistical Sciences, University of Padova,
(e-mail: anna.giraldo@unipd.it)

ABSTRACT: In this paper we study the relationship between female and male labour supply within Italian households, focusing on the reaction of women to the transitions from employment to unemployment of their partner. In literature the phenomenon of women entering in the labour market as a consequence of an unemployment episode of their partner is called Added-Worker Effect. The analysis is carried out over a long period of time, 2004-2019 on data from the Italian Labour Force Survey (ILFS). To identify the Added-Worker Effect, we adopt a differences-in-differences methodology. By exploiting the richness of information contained in the ILFS data on unemployment status and unemployment risk, we were able to evaluate different "dimensions" of the Added-Worker Effect.

KEYWORDS: household labour supply, diff-in diffs, female labour force participation.

1 Introduction

The aim of this work is to study the relationship between female and male labour supply within Italian households. In particular, we evaluate whether an Added-Worker Effect (AWE) exists in Italy, focusing on the transition of women from labour market inactivity to unemployment when their male partners move from employment to unemployment, using a differences-in-differences methodology.

The literature on the topic goes back to the first contributions of Humphrey (1940) and Woytinsky (1940) and the empirical studies on the Added-Worker Effect (AWE) by Mincer (1962) and Heckman and Macurdy (1980). The AWE has been revived with the recent economic crisis. Gong (2011), focusing on Australia, found a significant AWE in terms of increased full-time employment and working hours. Starr (2014) found, for the USA, that employment rates of women whose husbands were non-employed rose significantly during the recession. As to the Italian case, Ghignoni and Verashchagina (2016) found that an AWE exists, even if only in cases of serious hardship. More recently, Baldini et al. (2018), studying Italy in the years 2004-2014 using EU-SILC data, found a strong and robust evidence that households hit by an employment shock do respond by increasing labour supply.

2 Data and methodology

Our study evaluates the AWE in Italy over a long period of time, 2004-2019, employing the ILFS, a rotating panel provided by Istat. The longitudinal data of the ILFS observe individuals across couples of years (t_0, t_1): in the quarter of entrance in the panel, in the subsequent one (first two quarters of observation), and in the 5th and 6th quarter. In each quarter, new individuals enter the survey, for a share of one fourth of the total sample. The available data are from 2004-05 to 2018-19. Unfortunately, until 2012 Istat only makes available the panel data related to the first quarter of each year. Thus, for reasons of balance and comparability between samples, we carried out our analysis on the first quarter only. This means that the database is thus made by 15 panels of individuals observed in the first quarter of year t_0 and in the first quarter of year t_1. Our analysis focuses on couples – married or cohabiting – with or without children, with partners not retired and not unable to work, in the age range 25-54. To our purpose, we focus on households with unchanged composition in the two occasions (t_0 and t_1).

To identify AWE, we employ a differences-in-differences methodology (DD). Our first definition of treated women includes those women whose partners became unemployed between t_0 and t_1. AWE occurs when the probability of changing employment status from inactive to unemployed is significantly different between treated and untreated women. Then, the equation we estimate to detect AWE is the following linear probability model (Angrist and Pischke, 2009):

$$ES_{it} = \beta_0 + \beta_1 D_t + \beta_2 T_i + \beta_3 T_i D_t + \beta_4 X_{it_0} + \varepsilon_{it} \quad (1)$$

where ES_{it} is the employment status of female i at time t, D_t is a dummy with value equal to 0 in t_0 and 1 in t_1, T_i is a dummy that captures whether the woman is treated or not. D_tT_i is our variable of interest: the parameter β_3 captures the effect of being treated, compared to not being treated, on the change of employment status of females. X_{it_0} includes several covariates (female age cohort, male nationality, male and female educational level, number of children, number of children under 15 years old, number of children employed, male type of job, male sector of activity and dummies that capture male unemployment risk) all evaluated at t_0.

Exploiting the richness of the ILFS, Equation 1 is estimated under different specifications of the treatment (T_i) and the outcome (ES_{it}). The different definitions of treatment we adopt allow us to consider situations that might reveal an increased risk of losing one's job or a reduction in the household available income, which may affect the decision of female partners to enter the labour market. As regards women's outcomes, we adopted different definitions of the dependent variable *ES* in order to capture different 'degrees' of exit from inactivity and changes in preferences for work involvement. These different treatments and outcomes are explained in the next section.

3 Results

In this section, we discuss the results of the estimates of Equation 1 for the different treatment and work transitions of the woman and her partner (see Table 1). The first treatment we have considered is the partner's transition from employed to unemployed (*T1*). Then, to capture the emergence of a risk in the partner's employment stability and a worsening in the economic condition of the family, we have considered men's transitions either from employment to CIG[1] or from working full hours in t_0 to reduced hours in t_1, and men losing jobs other than the main one between the two years (*T2*). To complete the analysis, we have considered male transitions from employment to non-employment—either unemployment or inactivity (*T3*).

Table 1: *The AWE with different treatments and outcomes*

Outcomes	Treatments		
	T1	*T2*	*T3*
ES1	0.149***	0.023*	0.060***
ES2	0.123***	0.020	0.051***
ES3	0.138***	0.020	0.093***
ES4	0.038***	0.027***	0.028***
ES5	0.014	0.051***	0.023

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

For women, we adopted different definitions of the dependent variable *ES* in order to capture different 'degrees' of exit from inactivity and changes in preferences for work involvement. *ES1* is the dependent variable definition that captures women's transitions from inactivity to unemployment. *ES2* evaluates transitions from inactivity to the labour force—either unemployed of employed. *ES3* captures the intention to work and the change of the status from 'not searching' to 'searching' for work. Thus, the sample of women is restricted to the only ones who are inactive and 'not searching' for work in period t_0. *ES4* assesses preferences for greater work involvement, i.e. more working hours; in this case, the sample consists of women who are already employed and who do not wish more hours of work in t_0. We also considered a further specification to assess whether the market allows women to work more. In this case, we selected only women with a part-time job in period t_0 and evaluated the significance of the transition from part-time to full-time work between the two years (*ES5*).

[1] Cassa Integrazione Guadagni (CIG) is an institution under Italian law consisting of an economic benefit, provided by the Italian Security System, for workers suspended from the obligation to perform work or working reduced hours. It is paid to workers when companies are in temporary difficulty.

The results show that an AWE exists within Italian families. The classical definition of AWE, measured by women's transitions from inactivity to unemployment as their partners move from employment to unemployment, compared to women whose partners do not move to unemployment (ES1-T1), is significant and positive. In fact, treated women are 15% more likely to enter the labour market than untreated women. The effect is still positive— even if the size is smaller (2.3%)— when the treated group includes only families whose men are 'at risk of unemployment' (ES1-T2). This means that women react to changes in the family economic situation. Interestingly, working women are willing to work more hours (ES4) when the partner loses his job or he is at risk of losing it; this effect is significant for all treatment types, ranging from 2.8 to 3.8%. However, the transition from part-time to full-time (ES5) is 5% higher for treated women only in the case of T2 treatment. Possible explanations for this result could be the existence of labour market rigidities in the transformation of part-time work into full-time work or constraints on the supply side of the labour market (due to care activities that limit women's working hours).

Acknowledgments: the authors acknowledge the financial support provided by the PRIN project «The Great Demographic Recession – GDR» financed by the Italian MIUR under the PRIN 2017 research, grant agreement n. 2017W5B55Y-003 (PI: Daniele Vignoli).

References

Angrist, J. & Pischke, S. 2009. Mostly harmless econometric, An empiricist's companion. Princeton: Princeton University Press.

Baldini, M., Torricelli, C. & Urzì Brancati, M.C. 2018. Family ties: Labor supply responses to cope with a household employment shock. *Review of Economics of the Household*. **16**, 809-832.

Ghignoni, E. & Verashchagina A. 2016. Added worker effect during the great depression: evidence from Italy. *International Journal of Manpower*. **37**, 1264-1285.

Gong, X. 2011. The added worker effect for married women in Australia. *The Economic Record*. **87**, 414-426.

Heckman, J.J. & Macurdy, T. 1980. A life cycle model of female labour supply. *The Review of Economics Studies*. **47**, 47-74.

Humphrey, D.D. 1940. Alleged "additional workers" in the measurement of unemployment. *Journal of Political Economy*. **48**, 412-419.

Mincer, J. 1962. Labor force participation of married women: A study of labor supply, in Universities-National Bureau Committee for Economic Research, *Aspects of Labor Economics*, Princeton: Princeton University Press

Woytinsky, W.S. 1940. Additional workers on the labor market in depressions: A reply to Mr. Humphrey. *Journal of Political Economy*. **48**, 735-739.

INFERENCE ON THE STATE DISTRIBUTION IN PERIODIC HIDDEN MARKOV MODELS

Carlina C. Feldmann[1], Sina Mews[1], Rouven Michels[1] and Roland Langrock[1]

[1] Department of Business Adminstration and Economics, Bielefeld University, (e-mail: carlina.feldmann@uni-bielefeld.de, sina.mews@uni-bielefeld.de, r.michels@uni-bielefeld.de, roland.langrock@uni-bielefeld.de)

ABSTRACT: We present an exact solution for the time-varying state distribution in hidden Markov models (HMMs) with periodic state-switching dynamics. In a case study using African elephant data, the approach is shown to be superior to commonly applied alternatives.

KEYWORDS: Markov chain, movement ecology, periodic stationarity.

1 Introduction

When inferring latent states and their dynamics from an observed time series, periodic effects such as diel variation or seasonality are often of primary interest. In applications such as movement ecology or climatology, hidden Markov models (HMMs) with cyclic components are commonly used to address periodic variation in the latent state process (see, e.g., Nagel *et al.*, 2021). Inference then often focuses on the periodically varying probabilities of occupying the different states. These can, in principle, be taken as the empirical distribution of states per time point, as obtained using decoding algorithms such as Viterbi (see, e.g., Schwarz *et al.*, 2021).

To avoid the noise associated with this approach, especially for shorter time series, it may however be desirable to instead evaluate the time-varying state distribution as implied under the fitted model. Here we show how to exploit the periodic stationarity of corresponding HMMs to arrive at an analytic solution for the time-varying state distribution. In a case study on elephant movement, we demonstrate the superiority of our approach over commonly applied alternatives.

2 Methods

We consider an HMM comprising a state-dependent process $\{X_t\}_{t=1,\ldots,T}$ (where X_t can be a vector) and a latent state process $\{S_t\}_{t=1,\ldots,T}$, with S_t selecting which of N possible component distributions generates X_t. The state process $\{S_t\}$ is assumed to be an N–state Markov chain, characterised by its initial state distribution and the time-varying transition probability matrix (t.p.m.)

$$\Gamma^{(t)} = (\gamma_{ij}^{(t)}), \text{ with } \gamma_{ij}^{(t)} = \Pr(S_t = j | S_{t-i} = i),$$

$t = 1, \ldots, T$. We consider a setting with periodically varying state-switching dynamics, such that

$$\Gamma^{(t)} = \Gamma^{(t+L)} \tag{1}$$

for all $t = 1, \ldots, T$, with L denoting the length of a cycle. For hourly data and $N = 2$, we could for example model time-of-day variation ($L = 24$) as

$$\text{logit}\left(\gamma_{ij}^{(t)}\right) = \beta_1^{(ij)} \sin\left(\frac{2\pi t}{24}\right) + \beta_2^{(ij)} \cos\left(\frac{2\pi t}{24}\right), \text{ for } i \neq j. \tag{2}$$

The interpretation of such transition probabilities as functions of time can be tedious, especially when $N > 2$. Therefore, it has become common practice to instead consider a simpler summary statistic, namely the (periodically varying) distribution of the states at time t,

$$\delta^{(t)} = \left(\Pr(S_t = 1), \ldots, \Pr(S_t = N)\right),$$

as a function of time $t = 1, \ldots, L$. The latter is usually approximated by the hypothetical stationary distribution $\rho^{(t)}$ of the Markov chain that would result if the t.p.m. was homogeneous with $\Gamma = \Gamma^{(t)}$, which is the solution to $\rho^{(t)} = \rho^{(t)}\Gamma$ subject to $\sum_{i=1}^{N} \rho_i^{(t)} = 1$ (Patterson *et al.*, 2009). This approximation of $\delta^{(t)}$ will in general be biased because it ignores the preceding process dynamics as implied by $\Gamma^{(t-1)}, \Gamma^{(t-2)}, \ldots$ and instead pretends that the process has been following the dynamics as implied by a constant $\Gamma^{(t)}$ for a considerable time.

However, for periodically inhomogeneous Markov chains as defined in (1), there is in fact no need for such an approximation. To see this, consider for fixed t the thinned Markov chain $S_t, S_{t+L}, S_{t+2L}, \ldots$, which is homogeneous with constant t.p.m.

$$\tilde{\Gamma}_t = \Gamma^{(t+1)} \cdot \ldots \cdot \Gamma^{(t+L)}.$$

Provided that this thinned Markov chain is irreducible, it has a unique stationary distribution $\delta^{(t)}$, which is the solution to

$$\delta^{(t)} = \delta^{(t)}\tilde{\Gamma}_t$$

(see also Ge *et al.*, 2006 and Kargapolova & Ogorodnikov, 2012). Provided that the Markov chain starts in its stationary distribution, $\delta^{(t)}$ is the state distribution at time t we are interested in (and otherwise it will be at least approximately correct as the thinned Markov chain will converge to its stationary distribution).

3 Case study: elephant movement

We consider a complete movement track of an African elephant with hourly GPS data between October 2008 and June 2009. Based on consecutive locations, we calculate the Euclidean step lengths as well as the turning angles and model them in a 3-state HMM with gamma and von Mises distributions, respectively. To investigate diel variation in the state-switching dynamics we model the transition probabilities as trigonometric functions of the time of day (see Equation 2). The fitted model features an "encamped" state with short step lengths and frequent reversals in direction (state 1), an "exploratory" state with higher persistence in direction and medium step lengths (state 2), and a "travelling" state with highly directed and fast movement (state 3).

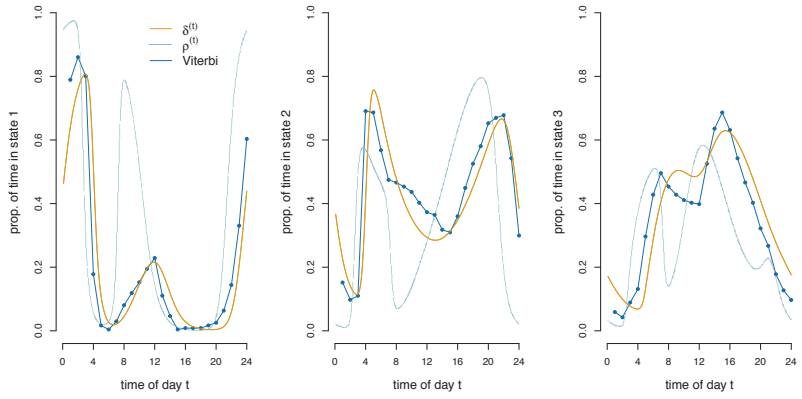


Figure 1. *Proportion of time spent in each state according to the model-implied periodic stationary distribution, the approximated stationary distribution, and the Viterbi state decoding.*

Based on the fitted HMM, we derive the proportions of time spent in each state using the model-implied periodic stationary distribution $\delta^{(t)}$, the approximated stationary distribution $\rho^{(t)}$, and the Viterbi-decoded states. The corresponding results are compared in Figure 1. The hypothetical stationary distribution $\rho^{(t)}$ differs greatly from the exact solution $\delta^{(t)}$ and is therefore a poor approximation in this example. Concerning the proportion of time spent in each state obtained using the Viterbi algorithm, the results are similar to the analytically derived periodic stationary distribution $\delta^{(t)}$. The advantage of the latter, however, is that it is less affected by noise in the data and instead offers a smooth function of time, even for short time series.

References

GE, HAO, JIANG, DA-QUAN, & QIAN, MIN. 2006. A Simple Discrete Model of Brownian Motors: Time-periodic Markov Chains. *Journal of Statistical Physics*, **123**(4), 831–859.

KARGAPOLOVA, N. A., & OGORODNIKOV, V. A. 2012. Inhomogeneous Markov chains with periodic matrices of transition probabilities and their application to simulation of meteorological processes. *Russian Journal of Numerical Analysis and Mathematical Modelling*, **27**(3), 213–228.

NAGEL, REBECCA, MEWS, SINA, ADAM, TIMO, STAINFIELD, CLAIRE, FOX-CLARKE, CAMERON, TOSCANI, CAMILLE, LANGROCK, ROLAND, FORCADA, JAUME, & HOFFMAN, JOSEPH I. 2021. Movement patterns and activity levels are shaped by the neonatal environment in Antarctic fur seal pups. *Scientific Reports*, **11**(1), 14323.

PATTERSON, TOBY A., BASSON, MARINELLE, BRAVINGTON, MARK V., & GUNN, JOHN S. 2009. Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *Journal of Animal Ecology*, **78**(6), 1113–1123.

SCHWARZ, JONAS F. L., MEWS, SINA, DERANGO, EUGENE J., LANGROCK, ROLAND, PIEDRAHITA, PAOLO, PÁEZ-ROSAS, DIEGO, & KRÜGER, OLIVER. 2021. Individuality counts: A new comprehensive approach to foraging strategies of a tropical marine predator. *Oecologia*, **195**(2), 313–325.

ZUCCHINI, W., MACDONALD, I.L., & LANGROCK, R. 2016. *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton, FL: Chapman and Hall/CRC Press.

TESTING CLUSTERS OF LOCATIONS IN SPATIAL DYNAMIC PANEL DATA MODELS

Feo G. [1], Giordano F.[1], Niglio M.[1], Milito S.[1] and Parrella M.L.[(*)1]

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

ABSTRACT: The *SDPD* (Spatial Dynamic Panel Data) models have been proposed in the socio-econometric literature to analyze spatio-temporal data. In this paper we consider a particular version of such models, where the set of spatial units is assumed to be partitioned into clusters and the parameters of the model are assumed to be constant within clusters and not constant across clusters. We propose a mutiple testing procedure that helps to choose the best model for a dataset by testing a given partition of clusters assumed under the null hypothesis.

KEYWORDS: spatial dynamic panel data models, model selection, spatial clustering.

1 Introduction

Let us consider a multivariate stationary process $\{\mathbf{y}_t, t = 1,2,\ldots\}$ of dimension p, where the vector $\mathbf{y}_t$ collects the observations at time t from p different locations (=*spatial units*). In this framework, the dependence between the p time series is usually due to spatial correlation.

The following model, in equation (1), belongs to the so called *SDPD* class of models, proposed in the socio-econometric literature (see Lee & Yu, 2010, Dou *et al.*, 2016 and references therein)

$$\begin{aligned}\mathbf{y}_t &= D(\lambda_0)\mathbf{W}\mathbf{y}_t + D(\lambda_1)\mathbf{y}_{t-1} + D(\lambda_2)\mathbf{W}\mathbf{y}_{t-1} + D(\beta_1)\mathbf{x}_t^{(1)} + \ldots \\ &\quad \ldots + D(\beta_k)\mathbf{x}_t^{(k)} + \mathbf{c} + \varepsilon_t.\end{aligned} \tag{1}$$

A typical feature of these models is the presence of the *spatial matrix*, denoted by $\mathbf{W}$, a known weight matrix with zero main diagonal, reflecting the physical distances between spatial units. It is used to deal with spatial correlation.

The parameters of the model are collected in the diagonal matrices $D(\lambda_j)$ and $D(\beta_l)$, with $j = 0,1,2$ and $l = 1,\ldots,k$, where the vectors $\lambda_j = (\lambda_{j1},\ldots,\lambda_{jp})'$ and $\beta_l = (\beta_{l1},\ldots,\beta_{lp})'$ assure that each location has its own parameter (*i.e.*, the model is *spatially heterogeneous*). Model (1) is characterized by the sum of

several components: *a)* a *spatial component*, $D(\lambda_0)\mathbf{W}\mathbf{y}_t$, for spatial correlation; *b)* a *dynamic component*, $D(\lambda_1)\mathbf{y}_{t-1}$, for serial correlation; *c)* a *spatial–dynamic component*, $D(\lambda_2)\mathbf{W}\mathbf{y}_{t-1}$, for the interactions between spatial and serial correlation; *d)* the component $D(\beta_l)\mathbf{x}_t^{(l)}$, for the effects of some covariates on the time series data $\mathbf{y}_t$ (the vector $\mathbf{x}_t^{(l)}$ collects the data observed at time t on the p locations and for a given covariate l, with $l = 1,\ldots,k$). Finally, $\mathbf{c}$ contains the fixed effects while $\varepsilon_t \sim i.i.d.$ with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \Sigma_\varepsilon$.

It is important to note that the number of parameters in model (1) is equal to $(4+k)p$ and may explode, since the number of locations p is allowed to increase to infinity asymptotically with the time series length. Many variants of *SDPD* models can be formulated starting from model (1) and considering some restrictions on the parameters. First of all, not all the components *a)-d)* are always active in the model. For example, in the well-known *SAR* model, only the parameters of the *spatial component a)* are active, while other parameters are zero. Moreover, sometimes the vectors λ_j and β_l may have constant parameters (*spatial homogeneity*, as Lee & Yu, 2010 and references therein), other times they are not constant (*spatial heterogeneity*, as in Dou *et al.*, 2016).

In this paper we consider a hybrid *SDPD* model, a cross between homogeneous and heterogeneous spatial models. By imagining that the spatial units can be subdivided into clusters, we assume that the model has parameters that are homogeneous within clusters and heterogeneous between clusters. This model has not yet been considered in the spatial econometric literature, as far as we know, and will be referred to as the *clusterized SDPD* model. It can be estimated by adapting the estimation procedure proposed in Dou *et al.*, 2016. But in order to estimate this model consistently and efficiently, one has to know the clustering structure (how many clusters there are and which locations are included in each cluster). The aim here is to propose a testing procedure which allows to test if a given partition of clusters assumed under H_0 can be accepted, so that one can use this information to estimate the *clusterized SDPD* model. The proposed testing procedure is briefly described in the following section.

2 The multiple testing procedure in a nutshell

Giordano *et al.*, 2023 propose a strategy to test a specific version of *SDPD* model for a given spatio-temporal dataset. The idea underlying their method is based on comparing two setups: *A)* the general version of the spatial model, shown in equation (1) and assumed under the alternative hypothesis (unrestricted model); *B)* a nested model, assumed under the null (restricted model).

Here we extend the procedure in Giordano *et al.*, 2023 to the case of a *clusterized SDPD* model. Denote with S the number of clusters assumed under H_0 and let $\{G_s,\, s = 1,\ldots,S\}$, be a partition of $\{1,\ldots,p\}$ with p_s the number of units in the s-th cluster, G_s. So, it is $\sum_{s=1}^{S} p_s = p$. The testing procedure is based on the following test-statistics

$$\widehat{\delta}_{jis} = \widehat{\theta}_{ji}^{(u)} - \widehat{\theta}_{js}^{(r)} \qquad j = 1,\ldots,3+k; i \in G_s; s = 1,\ldots,S; \tag{2}$$

where $\widehat{\theta}_{ji}^{(u)}$ is the *unrestricted estimator* of the j-th parameter in the vector $\theta_i = (\lambda_{0i}, \lambda_{1i}, \lambda_{2i}, \beta_{1i}, \ldots, \beta_{ki})'$, while $\widehat{\theta}_{js}^{(r)}$ is the restricted estimator, derived under the null hypothesis as

$$\widehat{\theta}_{js}^{(r)} = \frac{1}{p_s} \sum_{i \in G_s} \widehat{\theta}_{ji}^{(u)}, \tag{3}$$

that is the average of the unrestricted estimated values for the spatial units in the s-th cluster. These estimators are described in details in Giordano *et al.*, 2023. When the true *SDPD* model is the one assumed under H_0 (i.e., the assumed clustering partition is correct), the two estimators $\widehat{\theta}_{ji}^{(u)}$ and $\widehat{\theta}_{js}^{(r)}$ are expected to produce similar results (in mean) and the statistics $\widehat{\delta}_{ji}$ are expected to be centered around zero. A graphical evidence is given in Figure 1, where we simulated 200 replications of a *clusterized SDPD* model with $p = 10$ locations (on the x-axis) and $S = 4$ clusters. The clusters are shown by colours, but note that we assume only 3 clusters under the null hypothesis (more specifically, H_0 is true for the first two clusters while the last two clusters are erroneously assumed to be one). The boxplots summarize the unrestricted $\widehat{\theta}_{ji}^{(u)}$ (on the left) and restricted $\widehat{\theta}_{js}^{(r)}$ (in the center) estimations of the parameters. On the right, the values of the test-statistics $\widehat{\delta}_{ji}$, for each location. As evident from the figure, the test-statistics correctly deviate from the null hypothesis for the last two clusters. Note that the procedure is organized as a mutiple test (one test for each location), where a Bonferroni-type correction is used to calibrate the global size (details are reported in Giordano *et al.*, 2023).

In the simulation study we have further considered different values of dimension $p = (10, 50, 100)$ and sample size $T = (100, 500, 1000)$. Other settings are fixed as in Giordano *et al.*, 2023. The results are consistent in terms of False Positive Rate (*i.e.*, the average proportion of locations for which we wrongly reject H_0; note that it is not equivalent to the global size) and False Negative Rate (the average proportion of locations for which we wrongly accept H_0), as reported in the following table for the parameter λ_{1i}.

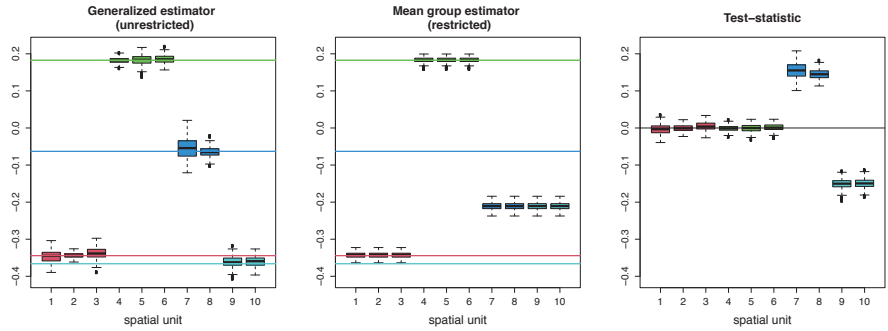


Figure 1. *For a* clusterized SDPD *model with 10 locations (x-axis), the boxplots summarize the unrestricted (left) and restricted (center) estimations of the parameter* λ_{1i}*. On the right, the values of the test-statistics for each location. There are 4 clusters (=colours) in the true model, but we assume only 3 clusters under the null hypothesis (so,* H_0 *is true for the first two clusters while it is false for the last two).*

	False Positive Rate			False Negative Rate		
$T =$	100	500	1000	100	500	1000
$p = 10$	0	0	0	0.53	0.27	0.15
50	0	0	0	0.09	0.06	0.06
100	0	0	0	0.12	0.06	0.04

There are many real cases where one can apply our testing procedure. For example, one may consider spatial data observed in a country and may want to test if the *SDPD* model is homogeneous within counties and heterogeneous between counties. In such a case, the clusters are the counties and the units in each cluster are perfectly identified under H_0. Our procedure allows to test if the assumed *clusterized SDPD* model is a good model for the dataset at hand.

References

DOU, B., PARRELLA, M.L., & YAO, Q. 2016. Generalized Yule-Walker Estimation for Spatio-Temporal Models with Unknown Diagonal Coefficients. *Journal of Econometrics*, **194**, 369–382.

GIORDANO, F., NIGLIO, M., & PARRELLA, M.L. 2023. Model structure identification in spatial dynamic panel data models. *Submitted*.

LEE, L.F., & YU, J. 2010. Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics*, **154**, 165–185.

BAYESIAN FORECASTING OF MULTIVARIATE LONGITUDINAL ZERO-INFLATED COUNTS: AN APPLICATION TO CIVIL CONFLICT

Beatrice Franzolini[1], Laura Bondi[2], Augusto Fasano[3], and Giovanni Rebaudo[4]

[1] Singapore Institute for Clinical Sciences (SICS), Agency for Science, Technology and Research (A*STAR), Singapore (e-mail: franzolini@pm.me)

[2] Medical Research Council (MRC) Biostatistics Unit (BSU), Cambridge University, United Kingdom (e-mail: laura.bondi@mrc-bsu.cam.ac.uk)

[3] Collegio Carlo Alberto, Italy (e-mail: augusto.fasano@carloalberto.org)

[4] Department of Economics, Social Studies, Applied Mathematics and Statistics (ESOMAS), University of Turin, Italy (e-mail: giovanni.rebaudo@unito.it)

ABSTRACT: Forecasting multiple dependent zero-inflated count processes is a problem encountered in many statistical applications. Standard parametric approaches typically rely on independence assumptions that fail to capture dependence structures. Here a Bayesian nonparametric approach is proposed to overcome this problem and showcased on a real dataset of civil conflicts in Asia. The forecasting model is obtained by generalizing the clustering methods proposed in Franzolini *et al.* (2023).

KEYWORDS: clustering, enriched Dirichlet, excess of zeros, mixtures of finite mixtures, rare events

1 Introduction

In statistical applications involving count data, it is common to encounter datasets showcasing a large number of zeros. Analyzing zero-inflated data requires statistical models that extend beyond standard count distributions, such as Binomial, Poisson, or Negative Binomial. Adding to the likelihood function a parameter specifically controlling the probability of observing a zero count is a popular strategy (Mullahy, 1986; Lambert, 1992), but this approach still relies on strong parametric assumptions regarding positive counts and is difficult to extend to multivariate count data: it requires a large number of parameters to avoid simplistic independence assumptions between multiple processes. Furthermore, when predicting future outcomes, the likelihood function is complicated by covariate values or autoregressive components, adding

to the complexity of the multivariate distribution of many zero-inflated processes. One flexible, yet parsimonious, solution has been recently proposed by Franzolini *et al.* (2023). They model joint probabilities of zero-inflation using a Bayesian enriched mixture of finite mixtures, obtained by combining the works of Wade *et al.* (2011) and Argiento & De Iorio (2022). The strength of the method relies on the fact that, within each mixture component, different processes are modeled with an independent kernel and the dependence across multiple count processes is captured by the underlying clustering structure. Thanks to the prior on the number of components of the mixture, the model automatically adjusts its complexity (measured by the number of parameters to be estimated) based on the data, ultimately requiring fewer parameters than traditional multivariate approaches when the data suggest so. Lastly, the method provides an additional interesting inferential outcome, i.e., a two-level clustering of subjects, based on the patterns of zero/non-zero counts (outer clustering) and values of positive counts (inner clustering). In Franzolini *et al.* (2023), the inferential goal is to detect groups of subjects with different count patterns and the data are cross-sectional. In this work, we extend their approach including in the model subject/time-specific covariates, autoregressive components, and random effects, aiming at predicting multiple longitudinal zero-inflated outcomes. We name the resulting model zero-inflated enriched mixture (ZIEM) regression.

2 ZIEM regression

The ZIEM regression is presented for a bivariate count process $(X_{i,t}, Y_{i,t})$, with multivariate predictors $Z_{i,t}$, where i and t denote subjects and time, respectively. The zero/non-zero components of the responses, i.e., $\tilde{X}_{i,t} = \mathbb{1}(X_{i,t} > 0)$ and $\tilde{Y}_{i,t} = \mathbb{1}(Y_{i,t} > 0)$, are modeled through a finite mixture model with bivariate kernel where mixture components are defined by the parameters of a logit regression with an autoregressive component, i.e.,

$$
\begin{aligned}
(\tilde{X}_{i,t}, \tilde{Y}_{i,t}) \mid p_{i,t}, q_{i,t} &\overset{ind}{\sim} \text{Bern}(p_{i,t}) \times \text{Bern}(q_{i,t}) \\
\text{logit}(p_{i,t}) &= \alpha_i^{(x)} + \beta_i^{(x)} X_{i,t-1} + (\eta^{(x)})^T Z_{i,t} \qquad (1) \\
\text{logit}(q_{i,t}) &= \alpha_i^{(y)} + \beta_i^{(y)} Y_{i,t-1} + (\eta^{(y)})^T Z_{i,t} \\
(\alpha_i^{(x)}, \alpha_i^{(y)}, \beta_i^{(x)}, \beta_i^{(y)}) \mid M_0, w, \theta &\overset{iid}{\sim} \sum_{m=1}^{M_0} w_m \delta_{\theta_m} \qquad (\eta^{(x)}, \eta^{(y)}) \sim \mathcal{N}(0, I) \\
\theta_m \mid M_0 \overset{iid}{\sim} \mathcal{N}(0, I) \quad w \mid M_0 &\sim \text{Dirichlet}_{M_0}(\gamma_0, \ldots, \gamma_0) \quad M_0 \sim \text{Poi}_0(\lambda_0)
\end{aligned}
$$

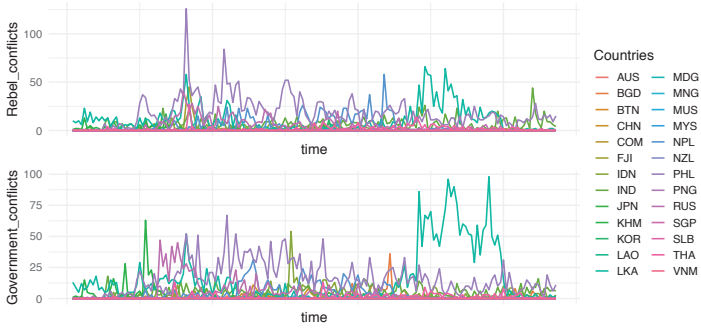


Figure 1. *Civil conflict data: data are part of a Defense Advanced Research Project Agency (DARPA) funded project which has created a dataset of over 2 million machine-coded daily events occurring between actors within the Asia-Pacific region.*

where $\theta = (\theta_1, \ldots, \theta_{M_0})$, with $\theta_m \in \mathbb{R}^4$, and Poi_0 denotes a shifted Poisson distribution on $\{1, 2, \ldots, \}$. The model (1) induces an outer clustering structure of the subjects denoted by i. Then, within each outer cluster m and independently across outer clusters, the positive component of the responses is modeled through a finite mixture model with bivariate Poisson kernel, i.e.,

$$\begin{aligned}(X_{i,t}, Y_{i,t}) \mid (X_{i,t} > 0), (Y_{i,t} > 0), \mu_i, \nu_i &\overset{ind}{\sim} \text{Poi}_0(\mu_i) \times \text{Poi}_0(\nu_i) \\ (\mu_i, \nu_i) \mid M_m, q_m, \xi_m &\overset{iid}{\sim} \sum_{m=1}^{M_m} q_{m,s} \delta_{\xi_{m,s}} \qquad (2) \\ q_m \mid M_m \sim \text{Dirichlet}_{M_m}(\gamma, \ldots, \gamma), \;\; \xi_{m,s} \mid M_m &\overset{iid}{\sim} Q_0, \;\; M_m \sim \text{Poi}_0(\lambda)\end{aligned}$$

where $\xi_m = (\xi_{m,1}, \ldots, \xi_{m,M_m})$, with $\xi_{m,s} \in (\mathbb{R}^+)^2$ and Q_0 is a bivariate Log-normal distribution with independent components. The extension to processes with dimension $d > 2$ is straightforward.

3 An application to civil conflict

We test the out-of-sample predictive performance of our model on a monthly bi-variate dataset concerning domestic civil conflicts from 1997 to 2010 in $n = 26$ countries in Asia. The observed responses are plotted in Figure 1. For a detailed description of the dataset, we refer to Bagozzi (2015). Monthly data from 1997 to 2009 are used to train our model (ZIEM regression), a zero-inflated Poisson (ZIP) regression, and a zero-inflated Negative Binomial re-

gression (ZINB) regression. ZIP and ZINB regressions are estimated with the `R` package `pscl` (Zeileis *et al.*, 2008). Data from the year 2010 are used to evaluate the prediction performance. All three models include an autoregressive component and three covariates (i.e., log-GDP per capita, GDP growth, log-population), which are used to predict the occurrence of a non-zero count. Table 1 summarizes the predictive performance of the three models, based on which we conclude that ZIEM regression outperforms the competitors.

Table 1. *Out-of-sample predictive performance: root mean squared error (RMSE), normalized root mean squared error (NRMSE), maximum squared error (*$\max\{\hat{e}^2\}$*), and squared error (e^2) distributions' quantiles. Bold values denote the best performance.*

				x s.t. $\hat{\text{pr}}(e^2 > x) = p$		
Model	RMSE	NRMSE	$\max\{\hat{e}^2\}$	p=0.15	p=0.25	p=0.50
ZIEM reg.	**6.72**	**0.1527**	**26.07**	**7.44**	**4.77**	**1.18**
ZIP reg.	7.52	0.1708	35.74	8.22	8.00	1.72
ZINB reg.	8.07	0.1834	36.90	8.11	7.10	5.26

References

ARGIENTO, R., & DE IORIO, M. 2022. Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *The Annals of Statistics*, **50**, 2641–2663.

BAGOZZI, B. E. 2015. Forecasting civil conflict with zero-inflated count models. *Civil Wars*, **17**, 1–24.

FRANZOLINI, B., CREMASCHI, A., VAN DEN BOOM, W., & DE IORIO, M. 2023. Bayesian clustering of multiple zero-inflated outcomes. *Philosophical Transactions of the Royal Society A*, **381**, 20220145.

LAMBERT, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

MULLAHY, J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

WADE, S., MONGELLUZZO, S., & PETRONE, S. 2011. An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis*, **6**, 359–385.

ZEILEIS, A., KLEIBER, C., & JACKMAN, S. 2008. Regression models for count data in R. *Journal of Statistical Software*, **27**, 1–25.

EFFICIENT DISENTANGLING γ-RAY SOURCES FROM DIFFUSE BACKGROUND IN THE SKY MAP

Francesco Freni [1] and Giovanna Menardi[1]

[1] Department of Statistical Science, University of Padova,
(e-mail: francesco.freni6@gmail.com, menardi@stat.unipd.it)

ABSTRACT: Searching for as yet undetected γ−ray sources is a major target of the Fermi LAT Collaboration. We address the problem by clustering the directions of the high-energy photon emissions detected by the telescope onboard the Fermi spacecraft. Putative sources are identified as the excess mass of disconnected high density regions on a sphere mesh, which allows for their joint discrimination from the diffuse γ−ray background spreading over the entire area. Density is estimated nonparametrically via binned directional kernel methods. The identification is accomplished by breaking the problem into independent subregions of the sphere separated by empty bins, thus leading to a remarkable gain in efficiency.

KEYWORDS: astrostatistics, directional data, modal clustering

1 Introduction

The Large Area Telescope (LAT) is an imaging γ−ray detector onboard the Fermi spacecraft, designed to perform an all-sky survey and gain a deeper comprehension of the processes responsible for generating and boosting γ−ray particles discharged by celestial bodies. Discovering and locating such sources is one of main purposes of the survey, and a declared target of the Fermi LAT collaboration. A main challenge, however, is the to separate the signal of the putative emitting sources from the diffuse γ−ray background which spreads over the entire area observed by the telescope. Furthermore, it is required to handle a remarkable computational burden, due to the huge amount of data recorded by the LAT.

Since γ−ray sources shall be intended as peaks of energy arising from a diffuse background, the underlying intuition complies with the *nonparametric*, or *modal* formulation of a clustering problem, which is here efficiently adapted to the considered framework. Modal clustering relies on the assumption that a probability density underlies the data, and clusters are defined as the domains of attraction of the density modes. With respect to most clustering methods,

relying on heuristic ideas of similarity between objects, the modal formulation is built on a probabilistic framework, which allows, for instance, a natural application of inferential tools. Additionally, the number of clusters is inherent to the data density and determined itself within the estimation process.

In this work we discuss a nonparametric method specifically conceived for high-energy $\gamma-$ray sources detection and discrimination from the background noise (Section 2). Its application is illustrated on a set of data drawn from one of the catalogues released by the Fermi LAT Collaboration (Section 3).

2 Nonparametric clustering for $\gamma-$ray sources detection

Nonparametric, or *modal* clustering hinges on the assumption that the data $(x_1,\dots,x_n)'$ are sampled from a probability density function f. The modes of f represent the archetypes of the clusters, in turn described by the surrounding regions. An indirect route to identify clusters, without attempting the explicit task of mode detection, is through disconnected (upper) density level sets of the sample space. Specifically, any section of f, at a level λ, singles out the set

$$L(\lambda) = \{x \in \mathbb{R}^d : f(x) \geq \lambda\}, \quad 0 \leq \lambda \leq \max f$$

which may be connected or disconnected. In the latter case, it consists of a number of connected components, associated with a cluster at the level λ.

While there may not exist a single λ catching all the modal regions, any connected component of $L(\lambda)$ includes at least one mode of the density. On the other hand, for each mode there exists some λ for which one of the connected components of the associated $L(\lambda)$ includes that mode at most and identifies the *excess mass* of that mode (Müller & Sawitzki, 1991). Hence, all the modal regions may be detected as the connected components of $L(\lambda)$ by varying λ (Figure 1). Points belonging to the surrounding regions are usually allocated to the clusters subsequently. See Menardi, 2016 for a detailed review.

Within the framework of $\gamma-$ray source detection, the data typically consist of an event list which gives the direction in the sky of each detected photon along with additional information. If the distance to the emitting source is not relevant, the data points are placed on the celestial sphere with Earth at its center and unit radius, as shown in the left panel of Figure 2. Directions are expressed in polar coordinates, that is, co-latitude (θ) and longitude (ϕ) in geographical terms, which can easily be back-transformed to Cartesian coordinates $x = (\cos\theta, \sin\theta\cos\phi, \sin\theta\sin\phi)$ on the unit sphere.

Due to the huge mole of available data, streamlining is firstly pursued via data discretization: rather than considering single photon emissions, the sphere

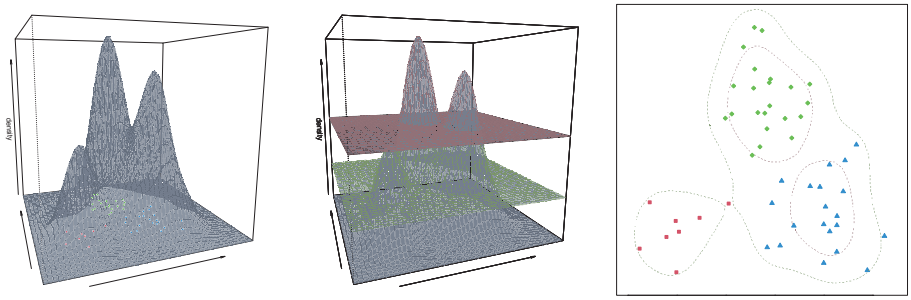


Figure 1. *A set of data from a trimodal density function (left), and the two sections of the density (center) for which the connected components of the associated $L(\lambda)$ identify the excess mass of the three modes (right).*

is partitioned into a thick triangle mesh, by recursively subdividing an icosahedron. Each of the B bins of the mesh is then associated with the count n_b of its inner photons. Density of photon emissions is then estimated nonparametrically, via binned directional kernel methods:

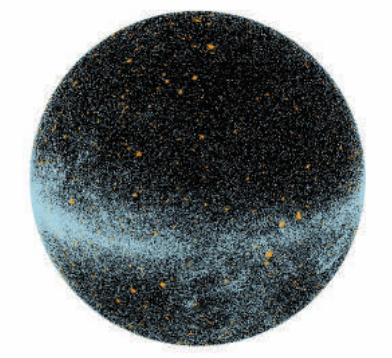
$$\hat{f}(x) = \frac{1}{n} \sum_{b=1}^{B} n_b K_h(x - m_b)$$

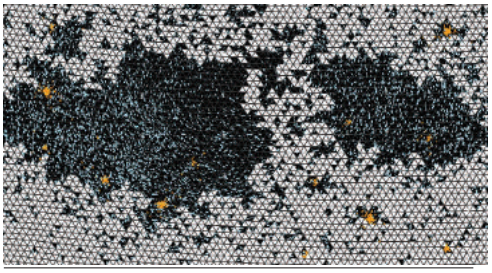
where $K_H(\cdot)$ is von Mises-Fisher kernel with concentration parameter $1/h^2$, n is the sample size, and m_b is the centroid of the b^{th} bin. This already produces by itself a computational gain of efficiency. Clustering is then built by identifying, for varying λ, the connected components of the upper level sets of the binned estimate, as union of edge-connected bins of the mesh. Once again, the mesh structure allows to accomplish the task efficiently on the whole sphere, by breaking the problem into independent subregions separated by empty bins.

The specific λ identifying the excess mass of each modal region allows for defining a source as the set of photons lying within the associated upper density level set. Outskirt photons are labeled as background.

3 Empirical analysis

We applied the proposed procedure to a set of data drawn from the 3FHL catalog of the Fermi LAT collaboration and spread on the whole sky map, along with the diffuse background. The sky distribution of the data, illustrated in Figure 2, is quite heterogeneous, with most of photon emission (around 84.4%) originated from a diffuse background noise, which mostly concentrates





Number of detected sources	
TPR	0.9013
FPR	0.2170

Figure 2. *Left: source data from the 3FHL catalog of the Fermi LAT collaboration (yellow), and the diffuse background (light blue). Right: a cut of the sphere highlighting the mesh built on the sphere and the opportunity of working on small separated regions separated by empty bins. The table reports the results of the proposed method.*

at the Galactic plane; in the same area, overlapping sources of various sizes arise while in the extragalactic sky sources are rather separated. The data set include 469784 photons, among which 73318 are emitted by 1529 sources, whose size range from 4 to 3572 photons.

Since the data are drawn from a catalogue of already detected sources, we may evaluate the performance of the procedure with respect to the knowledge of the pertaining source of each photon emission. As a summarizing measure of the quality of the association, we compute the True Positive Rate (TPR) and the False Positive Rate (FPR). The former index is defined as the proportion of true sources correctly detected, while the latter one corresponds to the proportion of estimated components formed in fact by background photons.

Results, summarised in Figure 2, show an overall good performance with respect to both the detection of sources, and the discrimination between photons emitted by sources and background. Future research will focus on providing the detected sources with a significance measure, as well as reducing the spread of the detected sources, since not reported results show a non negligible quote of misclassified background photons, lying nearby the sources.

References

MENARDI, G. 2016. A review on modal clustering. *Int.Stat.Rev.*, **84**, 413–433.

MÜLLER, D.W., & SAWITZKI, G. 1991. Excess mass estimates and tests for multimodality. *J. Am. Stat. Ass.*, **86**, 738–746.

A METHOD TO VALIDATE CLUSTERING PARTITIONS

Luca Frigau [1], Giulia Contu[1], Marco Ortu[1] and Andrea Carta[1]

[1] Department of Economics and Business Sciences, University of Cagliari, (e-mail: {frigau, giulia.contu, marco.ortu, andrea.carta88}@unica.it)

ABSTRACT: To evaluate the performance of clustering algorithms is challenging because typically the true classes are unknown. In this paper we propose a new cluster validity method that combines internal and relative criteria and employs Machine Learning algorithms to produce a relative validity ranking of partitions obtained from different clustering algorithms. Compared to other methods, the proposed approach considers the features' structure explicitly, can handle high-dimensional data, and can be applied to various clustering algorithms. The method has been tested on a simulated benchmark dataset, demonstrating its ability to rank correctly 11 classical clustering algorithms.

KEYWORDS: cluster validity, machine learning, simulation.

1 Introduction

Statistical learning methods can be categorized as supervised or unsupervised, according to on the availability of an associated response variable. In supervised methods the goodness of the estimated model is computed by comparing the prediction with the response variable, whereas in unsupervised methods the evaluation of their performance is very challenging because typically the true classes are unknown (Hastie *et al.*, 2009). Cluster analysis is an unsupervised method that deals with grouping a collection of objects into homogeneous clusters without having any information about the class of any object (Hennig *et al.*, 2015). There are several clustering algorithms available, none of which can be considered universally "best" in all circumstances. Therefore, it is common practice to compare the performance of several algorithms. The evaluation of a clustering algorithm's results is called cluster validity, which can be investigated through three main approaches: external, internal, and relative criteria. External criteria compare the obtained partition to externally known results, while internal criteria use only inherent quantities and features of the dataset, such as the proximity matrix. Relative criteria compare a set of defined partitions based on a pre-specified criterion. This paper proposes a

cluster validity method that combines internal and relative criteria, inspired by the validation of gray-level thresholding image segmentation algorithms. The proposed method employs Machine Learning algorithms to produce a relative validity ranking of partitions obtained from different clustering algorithms according to a predefined validity criterion. The goodness of the method's fit is evaluated through tests on a simulated clustering benchmark dataset.

The paper is structured as follows: Section 2 describes the proposed cluster validity approach's methodology. In Section 3, a simulation study is performed. Finally, Section 4 contains some concluding remarks and a discussion of future work.

2 Validation method

The aim of a clustering algorithm is to split observations into subsets based on a reasonable pattern in the data. The assigned classes express information about the pattern identified by the algorithm in the data, allowing to measure how much the identified pattern corresponds to the features' structure. As the true pattern is unknown, the quality of the identified pattern cannot be assessed absolutely, but it can be assessed relatively.

To evaluate the coherence between the assigned classes and the features' structure, Machine Learning algorithms (ML) are employed, using the classes as the response variable and the features as independent variables. The performance of ML, indicated as ρ, serves as a relative proxy for the reliability of the output of the clustering algorithm. The performance of ML can be measured by several indexes, such as accuracy, specificity, sensitivity, etc., according to which aspect the analyst wants to focus on. Particularly, ρ does not indicate how well the partition corresponds to the pattern in the data, but it indicates the clustering algorithm output's quality compared to that of other algorithms. Therefore, by ordering the different ρ obtained in each partition, it is possible to rank the clustering algorithms according to their capability to cluster the objects on the basis of the pattern in the data.

Compared to other cluster validity approaches (Arbelaitz *et al.*, 2013), the proposed method has some advantages. For example, external criteria approaches require externally known results, which may not always be available or applicable to the problem at hand. Internal criteria approaches only use quantities and features inherent to the data set and may not provide an accurate assessment of the clustering output's quality. Relative criteria approaches compare different partitions based on a pre-specified criterion, but they do not consider the features' structure explicitly. The proposed method, on the other

hand, uses Machine Learning algorithms to evaluate the coherence between the assigned classes and the features' structure and produces a relative validity ranking that takes this coherence into account. Moreover, the proposed method has the potentiality to be further developed to handle high-dimensional data and to be applied to various clustering algorithms, making it a versatile and robust method for cluster validity assessment.

3 Simulation study

To test the effectiveness of our method in ranking clustering partitions based on their ability to accurately reflect the data pattern, we conducted a simulation study. Our assumption was that the greater the noise in the data, the poorer the partition obtained by clustering algorithms. Therefore, we expected our method to rank the partitions based on the level of noise in the data.

For each clustering algorithm, we selected the best partition identified by the indexes included in the `R` function of `clusterCrit::intCriteria` (Desgraupes, 2018) within the range of 10-25 clusters. We then varied the level of noise in the data from 0% to 100% by randomly changing the classes of the partition. For instance, a noise level of 0% meant that no noise was added, and the classes of the partition remained the same. A noise level of 50% indicated that the classes of half of the observations were randomly assigned, while the classes of the other half were kept the same. In this simulation, we considered Support Vector Machine (Steinwart & Christmann, 2008) as Machine Learning algorithm, and 11 classical clustering algorithms.

Figure 1 shows that with the lower level of noise in the data, the method obtains higher values of ρ. So considering a partition is better when ρ is high, the method ranked correctly the partitions from the best (obtained in the data with no noise) to the worst (obtained in the data with the highest level of noise), for each of the 11 clustering algorithms. In that way, it is possible to use the method to rank different partitions without knowing the "true" one.

4 Conclusions

Validation of clustering algorithm output is of high interest due to the lack of a response variable to supervise the analysis. We have illustrated how the use of a Machine Learning algorithms-based method could allow for the ranking of clustering algorithms based on the proximity of their partitions to the unknown "true" partition. Using a simulated dataset, we showed that the method can rank the clustering algorithms among 11 different scenarios characterized by

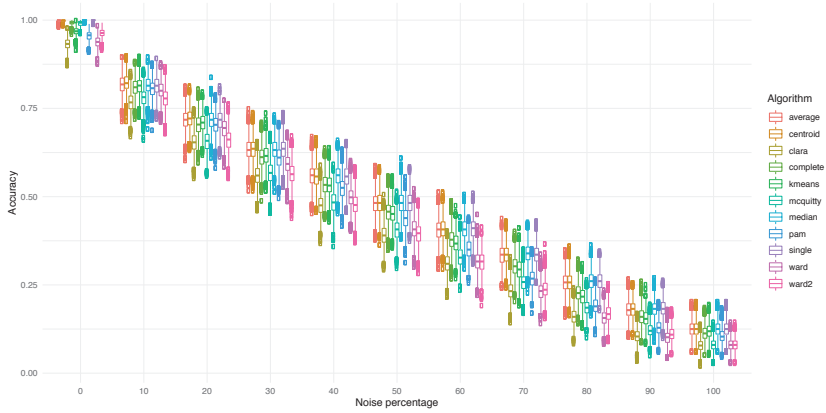


Figure 1. *Trend of performance of the validation method according to the level of noise added in the data.*

different noise levels. We believe that the proposed validation approach can enable the comparison of widely used clustering algorithms and help auditors choose the appropriate method for each situation.

As a potential extension, we are exploring the feasibility of applying the algorithm to big data scenario. In fact, many classical cluster validation indexes that already exist are characterized by high computational cost. Thus, it can be prohibitive to use them in big data scenarios.

References

ARBELAITZ, OLATZ, GURRUTXAGA, IBAI, MUGUERZA, JAVIER, PÉREZ, JESÚS M, & PERONA, IÑIGO. 2013. An extensive comparative study of cluster validity indices. *Pattern recognition*, **46**(1), 243–256.

DESGRAUPES, BERNARD. 2018. *clusterCrit: Clustering Indices*. R package version 1.2.8.

HASTIE, TREVOR, TIBSHIRANI, ROBERT, FRIEDMAN, JEROME H, & FRIEDMAN, JEROME H. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

HENNIG, CHRISTIAN, MEILA, MARINA, MURTAGH, FIONN, & ROCCI, ROBERTO. 2015. *Handbook of cluster analysis*. CRC press.

STEINWART, INGO, & CHRISTMANN, ANDREAS. 2008. *Support vector machines*. Springer Science & Business Media.

ANALYSIS OF THE NEED FOR WORKING TIMBER STARTING FROM ISTAT ANNUAL INDUSTRIAL PRODUCTION DATA

Flora Fullone[1], Gianmarco Farina[1], Enza Compagnone[1], Mirella Morrone[1], Gioacchino de Candia[1]

[1] Istat, Istituto Nazionale di Statistica,
(e-mail: enza.compagnone@istat.it, gianmarco.farina@istat.it, ffullone@istat.it, mimorron@istat.it, gioacchino.decandia@istat.it)

ABSTRACT: The consumption of timber represents a significant impact on the withdrawal of biomass and, more generally, raw materials in Italian regions. Timber is the essential raw material for the production of wood products, paper, and cardboard, as well as being a renewable fuel used in all Italian regions. This study reconstructs the timber supply chain used in industrial production for the creation of wooden semi-finished products, which will be used in subsequent productions, using annual data from Istat's ProdCom[1] survey. The aim is to estimate the timber consumption of the Italian industry and indirectly the withdrawals in the forest.

KEYWORDS: wood, industrial production, forest

1 Proposed methodology

The work aims to evaluate the amount of timber used in the Italian timber supply chain, starting from the production data of the "Annual Survey of Industrial Production (ProdCom)" by Istat. The survey is carried out on all productive local units (about 62 thousand establishments) of companies with at least 20 employees and on a representative sample of companies with 3 to 19 employees classified in the Nace Rev.2 divisions from 07 to 33 - excluding division 09 and group 19.2 - and division 38. The analysis focuses on the production of wood semi-finished products, which use both wood logs taken in Italian forests and those imported from abroad as the raw material input of the process, referring to the year 2020 ProdCom code. These activities are linked to Nace division 16 "Wood and wood products; except furniture; manufacture of articles of straw and plaiting materials" represented in detail in Table 1. The timber supply chain involves a first phase of activity in the forest, Nace division

[1] Annual Survey of Industrial Production

02 "Forestry and logging", which concerns tree felling operations, debarking, primary processing, and transportation to the timber sawmills. The quantities taken in this phase of the chain are not currently detected by a direct statistical survey. The indications received from a selected panel of companies operating in Nace division "02 - Forestry and logging" suggest that the quantity of logs selected for primary processing in sawmills represents a percentage ranging from 40% to 50% compared to the volume of felled trees. The remaining part generally undergoes a chipping process directly in the forest and is largely delivered to bioenergy generators.

Table 1: Economic activities due to the first phase of the wood supply chain.

Nace 02	**Forestry and use of forest areas**
02.10	Forestry and other forestry activities
02.20	Use of forest areas
Nace 16	**Wood industry**
16.10	Sawing and planing of wood
16.21	Manufacture of veneer sheets and wood-based panels
16.23	Manufacture of other wooden structural and joinery products for construction
16.24	Manufacture of wooden packaging
16.29	Manufacture of other wood products; manufacture of cork, straw, and weaving materials articles

For each of these activity classes, the different wood processing stages were revised and analysed, and the semi-finished wood products and primary products obtained directly from log processing were identified based on the 8-digit ProdCom classification codes, 6-digit CPA classification codes, and 4-digit Nace classification codes (Table 2). The ProdCom survey annually publishes, in reference to the list of ProdCom codes in force during the examined period, the data on industrial production achieved during the year in terms of value and quantity/volume.

Table 2. *Primary and semi-finished products, obtained directly from the processing of logs; ProdCom, Cpa, Nace coding.*

NACE	**CPA**	**PRODCOM**	**Descri**ption	**U. mis**
16.10	16.10.11	16.10.11.34	Spruce wood, fir wood sawn or chipped lengthwise, sliced or peeled, of a thickness > 6 mm (en)	m3
16.10	16.10.11	16.10.11.36	Pine wood sawn or chipped lengthwise, sliced or peeled, of a thickness > 6 mm	m3
16.10	16.10.11	16.10.11.38	Coniferous wood sawn or chipped lengthwise, sliced or peeled, of a thickness of > 6 mm	m3
16.10	16.10.12	16.10.12.50	Wood, sawn or chipped lengthwise, sliced or peeled, of a thickness > 6 mm	m3
16.10	16.10.12	16.10.12.71	Tropical wood, sawn or chipped lengthwise, sliced or peeled, end-jointed or planed/sanded, of a thickness > 6 mm	m3
16.10	16.10.21	16.10.21.10	Coniferous wood continuously shaped	Kg
16.10	16.10.23	16.10.23.00	Wood, incl. strips and friezes for parquet flooring, not assembled, continuously shaped "tongued, grooved, rebated, chamfered,	Kg

16.10	16.10.24	16.10.24.00	Wood wool; wood flour	Kg
16.10	16.10.25	16.10.25.03	Coniferous wood in chips or particles	Kg
16.10	16.10.25	16.10.25.05	Non-coniferous wood in chips or particles	Kg
16.10	16.10.39	16.10.39.00	Other wood in the rough, including split poles and pickets	m3
16.21	16.21.11	16.21.11.00	Plywood, veneered panels and similar laminated wood, of bamboo	m3
16.21	16.21.16	16.21.16.00	Other plywood, veneered panels and similar laminated wood, of coniferous wood	m3
16.21	16.21.17	16.21.17.11	Plywood consisting solely of sheets of wood, each ply not exceeding 6 mm thickness, with at least one outer ply of tropical wood	m3
16.21	16.21.18	16.21.18.00	Other plywood, veneered panels and similar laminated wood, of other wood	m3
16.21	16.21.22	16.21.22.10	Veneer sheets and sheets for plywood and other wood	m3
16.21	16.21.23	16.21.23.00	Veneer sheets and sheets for plywood and other wood	m3
16.21	16.21.24	16.21.24.00	Veneer sheets and sheets for plywood and other wood	m3
16.23	16.23.12	16.23.12.00	Shuttering for concrete constructional work, shingles and shakes, of wood	Kg
16.23	16.23.19	16.23.19.00	Builders' joinery and carpentry of wood	Kg
16.24	16.24.12	16.24.12.00	asks, barrels, vats, tubs, and coopers products and parts thereof of wood	Kg
16.24	16.24.13	16.24.13.20	Cases, boxes, crates, drums and similar packings of wood	Kg
16.29	16.29.14	16.29.14.95	Pellets	t

The published quantities are directly due to the timber used, through a simple mathematical model that takes into account some factors such as the moisture content of the raw material compared to the semi-finished products, the coefficients of use of the logs, and other aspects related to wood processing phases. The quantity of timber available for industrial production is estimated based on the relationship (1).
The timber used as raw material, estimated through the reconstruction of the supply chain, is partly imported from abroad, with reference to the commodity category of roughly squared logs or blocks. Consequently, the timber for industrial use taken in Italy can be estimated (2).

$$Tav = \sum_i^n p_i * \alpha_i * \frac{100}{\eta_i} \quad (1)$$

$$T_{Italy} = Tav - T_{imp} + T_{exp} \quad (2)$$

T_{av}: Quantity of timber available for industrial production used for semi-finished and primary products, expressed in tons at standard moisture content of 12-15%.
p_i: Quantity annually published by ProdCom for each product identified in Table 2 of primary codes, based on the unit of measurement reported in the table.
α_i: Conversion coefficient of the unit of measurement of the ith ProdCom code into tons
η_i: Yield of logs in sawmills. The expert panel declares a performance of around 90%.
$T_{imp/exp}$: Raw imported/exported timber
T_{Italy}: Raw timber taken in Italy, and considered with a standard moisture content of 12-15%;

Starting from the timber for industrial use of Italian origin (Tav), the volumes of trees cut in Italy are estimated (3). Moreover, it should be considered that both the estimation of the quantity of timber available for industry (Tav) and the estimation of imported and exported timber (Timp/exp) have been made considering a standard moisture content of wood equal to 12% - 15% of moisture[2].

$$Tf_{Italy} = \left(\frac{100+u_f}{100+u_s} * T_{Italy}\right) * \frac{100}{\eta_f} \quad (3)$$

Tf_{Italy} : Quantity in tons of trees cut, considering an average moisture content of wood in a range of 60% - 150%.

u_f : Moisture content of the timber at the time of extraction, in a range of 60% - 150%

u_s : Standard moisture content of the timber 12% - 15%, considering standard conditions of 20°C temperature and 65% relative humidity.

T_{Italy}: Raw timber taken in Italy, and considered with a standard moisture content of 12-15%;

η_f : Percentage ranging from 40% to 50% that represents the logs selected for primary processing in sawmills compared to the volume of felled trees, as indicated by the experts' panel.

In conclusion, the developed metodology provides annually an estimate of the quantity of timber used in the Italian timber supply chain, starting from the production data of the "Annual Survey of Industrial Production (ProdCom)" by Istat. The estimation is made by analysing the different phases of wood processing and identifying the wood semi-finished products and primary products obtained directly from the processing of logs. The estimation takes into account the quantity of timber available for industrial production, the conversion coefficient of the unit of measurement of the ProdCom codes into tons, and the yield of logs in sawmills. Moreover, the estimation considers the raw imported and exported timber and the moisture content of the timber at the time of extraction and under standard conditions. The result of the analysis is an estimate of trees cut in Italy, considering an average moisture content of wood in a range of 60% - 150%.

References

ZANUTTINI, R. 2014. Il Legno Massiccio, *Materiale per un'edilizia sostenibile*.

SEMERARO, N. 2021. Rapporto Rilegno 2021, *Progetti, innovazioni e prospettive*.

SCRUCCA F., RINALDI C., MORARA E, AGNANI A. 2021. Rapporto Arcadia, Studio di filiera del cippato forestale.

EUROSTAT, 2020. ProdCom List.

[2] $u = (m\_f - m\_o)/m\_o * 100$

u: percentage of wood moisture content; m_f weight of wood referred to the moisture percentage u; m_o weight of the wood when dried.

STRATIFIED SAMPLING ON DATA NUGGETS: A STRATEGY FOR DATA REDUCTION

Ravi Kumar Gangadharan, Vanessa Petrarca, Maria Chiara Pagliarella, Giovanni Porzio

Department of Economics and Law, University of Cassino and Southern Lazio,
(e-mail: mc.pagliarella@unicas.it, porzio@unicas.it)

ABSTRACT: The increased volume and velocity of data production has been causing a growing cost in storing and analysing data. Thus, due to this continuously increasing phenomenon, the urgency of data reduction technique arises.
Data reduction aims at decreasing storage and computational costs for data analysis. In order to tackle with this very large and complex issue, many techniques have been developed and employed (such as clustering, principal points, support points, prototypes, etc.). Among the many, this work focuses on a recently introduced specific type of data reduction method which has been called Data Nuggets.
Data Nuggets reduces huge datasets and compresses the observations into few points, by saving essential information on the data structure. In parallel with standard classic procedures, Data Nuggets splits a dataset in several subsets (called Nuggets) which are defined by three main components: a Center, a Weight (representing the number of observations within each subset), and a Scale (representing the average Nugget within variance).
Particularly, our work aims at investigating to what extent Data Nuggets can be used as a tool to obtain stratified samples from large datasets so that some computational cost can be gained. A comparison in terms of efficiency with respect to statistical techniques applied to a simple random sample drawn from the same large dataset will be provided.

KEYWORDS: Large datasets, computational effort, data partition.

References

BEAVERS, T., CABRERA, J., & LUBOMIRSKI, M. 2020. datanugget: Create, Refine, and Cluster Data Nuggets. *R package version 1.0.0,* https://CRAN.R-project.org/package=datanugget.

CHERASIA K.E., CABRERA J., FERNHOLZ L.T., & FERNHOLZ R. 2023. Data Nuggets in Supervised Learning. In M. Yi, K. Nordhausen (eds.), *Robust and Multivariate Statistical Methods*, Springer, https://doi.org/10.1007/978-3-031-22687-8_20.

IS THE SUBJECTIVE FINANCIAL WELL-BEING OF POLISH FAMILIES CHANGING WITH TIME? AN EMPIRICAL STUDY BASED ON CONSTRAINED LATENT MARKOV MODELS

Ewa Genge [1]

[1] Department of Economic and Financial Analysis, University of Economics in Katowice, e-mail: ewa.genge@ue.katowice.pl

ABSTRACT: Poland is one of the EU countries with the lowest level of perceived financial position, according to most recent Eurostat data. To investigate the problem of such a low level of subjective well-being and to show the changing behaviour of Polish families, we apply the dynamic latent variable models in which families can change the latent class over time. We compare the models with different numbers of latent states, different types of constraints and we study the transitions between latent structures at different points in time. We present the tendency of self-reporting income position in each wave of the survey with a special focus on the results for the respondents behaviour in waves preceding and following economic crisis. The study is based on the national longitudinal project (Social Diagnosis) using software of R.

KEYWORDS: constrained latent Markov model, material well-being, transition matrix

1 Introduction

Poland is a country of Central and Eastern Europe which have just been through a structural and economic transition, characterised by relatively good growth performance along with rather small (compared to Ukraine, Lithuania, Latvia and Russia) increase in income inequality. However, Poles tend to be very unhappy with their financial situation. The population of Poland is described by the lower than EU-28 average rating for subjective assessment of the material condition, ranked at the 22 position, in accordance with the latest Eurostat data (European-Commission, 2021).

To evaluate the financial assessment of Polish households we base our study on Social Diagonsis (Social-Diagnosis, 2015) panel research with all, eight waves being taken in the following years 2000, 2003, 2005, 2007, 2009,

2011, 2013, 2015. We rely our study on the sample of individual responses represented by the heads of each household. The substantive research question, addressed by the presented analysis concerns the evolution of the subjective assessment of the financial satisfaction in Poland. We present the tendency of self-reporting income position in each wave of the survey with a special focus on the results for the respondents behaviour in waves preceding and following economic crisis.

To show the changing behaviour of Polish families, we apply the dynamic latent variable models in which respondents are allowed to switch from one to another latent class over time. We adopt the latent Markov (LM) models (Bartolucci *et al.*, 2013), extended to include the survey weights (see also Pennoni & Genge, 2020). We compare the models with different numbers of latent structures, different types of constraints and we study the transitions between latent structures at different points in time.

2 Latent Markov Models

We conceive the income perception of families as a non-observable, latent feature, evaluated through the questionnaire items. Then, latent Markov (LM) models enable to conceive self-reporting income position as a time-varying latent trait denoted as $\mathbf{S} = (S^{(1)}, \ldots, S^{(T)})$, which is assumed as a hidden stochastic process of first-order having a discrete distribution with latent states.

In our analysis we observe a categorical response variable $X^{(t)}$, for each time occasion t, with $t = 1, \ldots, T$ ($T = 8$ waves in our case). The response variable $X^{(t)}$ is designed to monitor financial satisfaction of Polish families and has l_j categories ($l_j = 5$ in our study), labeled from 0 to $l_j - 1$. We denote by $\mathbf{X}$ the vector with elements $X^{(1)}, \ldots X^{(T)}$, which usually, is referred to repeated measurements on the same respondents at different points in time.

The probability mass function of $\mathbf{S}$ may be expressed as

$$p(\mathbf{S} = \mathbf{s}) = \pi_{s^{(1)}} \prod_{t=2}^{T} \pi^{(t)}_{s^{(t)}|s^{(t-1)}}, \tag{1}$$

where $\mathbf{s}$ denotes a realization of $\mathbf{S}$, with elements $s^{(1)}, \ldots, s^{(T)}$;
$\pi_{s^{(1)}} = p(S^{(1)} = s)$ is the initial and $\pi^{(t)}_{s^{(t)}|s^{(t-1)}} = p(S^{(t)} = s|S^{(t-1)} = \bar{s})$ is the transition probability of the model.

In the results, we compare different variants of the LM model (that is, with different types of constraints posed on transition matrix), such as separate, heterogeneous transition matrices for each year, partial time-homogeneous ma-

trices based on two different transitions (one until occasion T^* and the other for transitions after this occasion) as well as homogeneous with one common transition matrix for all years (see Bartolucci *et al.*, 2013, p. 86-96, for details).

3 Results

At the first stage of our analysis we select the number of latent states and then we try to simplify the LM model by adopting certain constraints on its parameters. We observe that the lowest *BIC* value (equal to 6, 221.312) is reached for the *LM-part-hetero* with $T^* = 6$ and three number of latent states ($s = 3$). We note also that, the *BIC* value both for *LM-hetero* and *LM-part-hetero* with $T^* = 6$ is lower than the value of this criterion achieved for traditional LC model (see Genge, 2021, Table 7, p.13). On the basis of the estimated conditional probabilities we classify the Polish households to three latent states: S_1 – households with the lowest income perception, S_2 – households generally satisfied and S_3 – households with the highest self-reported financial status.

Interestingly, on the basis of the estimated transition probabilities we can see the difference between the evolution of the income assessment from the first to the fifth wave and from the sixth to the last wave. Notably, the first transition matrix (concerning years before crisis) corresponds to a considerably higher level of persistence in the third, the most positive latent state than the second transition matrix. We can observe also that in the years following the economic crisis the respondents are more prone to remain in the unsatisfied and rather satisfied groups of Poles (S_1 and S_2) and to switch from the highest to the state characterised by satisfaction of intermediate level. These results might confirm just slightly deterioration of Polish moods reflecting the economic crisis. This feature of Polish nation was already noted by Helliwell *et al.* (2014) or Chzhen (2016). They found that subjective well-being decreased in the EU countries that were heavily affected by the crisis (Greece, Ireland, Italy, Portugal, and Spain). We note that they considered only the early impact of the economic crisis between 2008 and 2011. However, these results might suggest that the crisis affected Polish families in various forms, not only related to cutting their budgets. In a further stage of our study (Genge, 2023) we compare the results with the homogenous version of the LM model, extended to include also time-varying covariates allowing for better characteristics of changing family behaviours. This approach help us to identify the types of families (characterised by different socio-economic features) who are in need of greater social protection.

Ewa Genge acknowledges the research grant (SONATA 12, UMO-2016/23/D/HS4/00989, "Latent variable models in the identification of homogenous structures in socio-economic longitudinal data") of the National Science Centre, Poland.

References

BARTOLUCCI, F, FARCOMENI, A, & PENNONI, F. 2013. *Latent Markov Models for Longitudinal Data*. New York: Chapman and Hall/CRC.

CHZHEN, Y. 2016. Perceptions of the economic crisis in Europe: Do adults in households with children feel a greater impact? *Social Indicators Research.*, **127**, 341–360.

EUROPEAN-COMMISSION. 2021. *Eurostat – Average rating of satisfaction by domain, sex, age and educational attainment level*. `https://ec.europa.eu/eurostat/databrowser/view/ilc_pw01/default/table?lang=en`.

GENGE, E. 2021. LC and LC-IRT models in the identification of Polish households with similar perception of financial position. *Sustainability.*, **13**, 1–22.

GENGE, E. 2023. An evaluation of self-reported material well-being using latent Markov models with covariates. *Longitudinal and Life Course Studies.*, 1–28.

HELLIWELL, J. F., HUANG, H., & WANG, S. 2014. Social capital and well-being in times of crisis. *Journal of Happiness Studies.*, **15**, 145–162.

PENNONI, F., & GENGE, E. 2020. Analysing the course of public trust via hidden Markov models: a focus on the Polish society. *Statistical Methods and Applications.*, **29**, 399–425.

SOCIAL-DIAGNOSIS. 2015. *Objective and subjective quality of live in Poland. Czapinski J., Panek T. (eds.)*. `http://www.diagnoza.com/index-en.html`.

VISUALIZATION OF PROXIMITY AND ROLE-BASED EMBEDDINGS IN A REGIONAL LABOUR FLOW NETWORK

Sara Geremia[1], Fabio Morea[2] and Domenico De Stefano[1]

[1] University of Trieste (e-mail: sara.geremia@phd.units.it)

[2] Area Science Park (e-mail: fabio.morea@areascienepark.it)

[1] University of Trieste (e-mail: ddestefano@units.it)

ABSTRACT: This study uses graph representation learning techniques to analyze a regional labor flow network. The methods employed, VGAE and Role2Vec, reveal community structures and centrality of universities and research institutions in the network. The study demonstrates the potential of such techniques for analyzing complex networks and uncovering hidden structures.

KEYWORDS: graph representation learning; labour flow data; VGAE; Role2Vec

1 Introduction

The mobility of workers creates a network of connections that reflects the interconnectivity between employers. Such network data can reveal insights into the structure of the relationships between employers, which can be used to identify communities of employers that share geographic location, industry, and workforce characteristics (Park *et al.*, 2019). Examining regular patterns in the network is a key step in understanding the role and position of employers in the labour market. The role of large public sector organizations, such as universities, in the economic context under study can be determined by their centrality and relationship with industries employing a high number of experienced professionals (Smallbone *et al.*, 2015).

This work aims to investigate the structure of a labour flow network in Friuli Venezia Giulia (FVG) (Morea & De Stefano, 2022). The labour market flows are collected from Regional data from the Compulsory Communication on Employment (RCCE).

Understanding the structure of a labour flow network involves dealing with graph data containing rich relational information. Traditional machine learning algorithms require hand-engineered feature representation which is labor-intensive and relies on domain-specific knowledge. Representation Learning

(RL) provides an alternative approach to automatically learn to represent graph data using low-dimensional vectors (Hamilton, 2020). The learned embeddings can be used with data visualization techniques to generate representations of graphs useful for discovering communities, hub nodes, and other hidden structures.

The graph RL task can be performed to assess the potential of universities and research institutions as drivers of economic development and innovation in FVG. The interest is in investigating whether exploring the network by looking at both relational proximity and regular patterns yields valuable insights.

2 Methods

In this work two methods for graph RL are employed: Variational Graph Auto-Encoder (VGAE) (Kipf & Welling, 2016), and Role2Vec (Ahmed *et al.*, 2018). The methods differ in their approaches to preserving the structure of the graph, indeed they are based on two different definitions of node structural similarity known as structural and regular equivalence, respectively. Two nodes are structurally equivalent if they are relationally close, while they are regular equivalent if they have similar roles or occupy similar positions in the network.

VGAE is designed to preserve the structural equivalence between nodes, which means that structurally similar nodes should be mapped to similar embeddings. This is achieved by using a graph convolutional neural network as the encoder in the model, that applies convolutional filters to the graph to aggregate information from neighbouring nodes. VGAE share the encoder-decoder structure with standard autoencoders and it is built to learn the generative distribution of data. The decoder is a simple inner product between the latent representations of nodes, that enforces the reconstruction of the original graph from the learned representations. Role2Vec is used to incorporate global regular equivalence information, which means that nodes that share regular patterns in the graph should be mapped to similar embeddings. The latent representations are learned using a feature-based random walk approach, where walks find similar nodes identified by structural properties and higher-order graph features (e.g. triangles, 4-cycles, etc.). Both VGAE and Role2Vec have been shown to achieve state-of-the-art performance on various graph RL tasks, but the choice of method depends on the nature of the dataset and the task at hand. The performance evaluation in this work is conducted without true labels for a supervision task, thus it is based on visualization. The results of the models are evaluated by exploring the network latent representations re-embedded with Uniform Manifold Approximation and Projection (UMAP).

UMAP is a dimensionality reduction technique that takes local structure into account, to increase the data representation quality in terms of clusterability. For data visualization the number of dimensions is set to two.

The RCCE data include science and engineering and information and communication technology occupations over a 8 years period, from 2014 to 2022. To investigate the transfer of experienced professionals (P), weights (W) are assigned to transitions from employer A to B under the assumption that the experience gained by P while working for A is transferred to B (Morea & De Stefano, 2022). The sum of W of adjacent nodes defines the *strength*, which is an attribute included for the interpretation of the results.

3 Results and Final Remarks

The RCCE network comprised 1084 nodes and 1641 edges. Figure 1 compares the two-dimensional UMAP visualizations of node embeddings learned with the methods. The distances between nodes in the embedding space reflect structural (left panel) and regular (right panel) equivalence in the original graph. The size of the nodes indicates the strength, thus employers employing a high number of experienced professionals are shown with bigger nodes. The colour of the nodes highlights universities and research institutions in the network. The left plot captures a community structure in which two components clearly detach. It also seizes proximity between coloured nodes and big nodes. The right plot shows the universities very close together in space, along with the research institutes, illustrating the similarity of their roles in the network. The UMAP visualizations reveal community structures and the centrality of universities and research institutions. In particular, the study finds that universities are closer to each other in the embedding space when considering regular equivalence, indicating that although they share qualified employers with different organizations, structurally they play the same role.

Overall, the study highlights the potential of graph RL techniques to analyze complex networks and uncover hidden structures and patterns, which provide new outlooks on economic development and innovation. It also suggests that examining different embedding algorithms tailored to specific tasks would be valuable in addressing different research inquiries.

In encoding large graphs both methods enable the utilization of node attributes, which can provide crucial information regarding a node's community membership and role. In future applications, exploring the impact of incorporating node attributes could yield valuable insights.

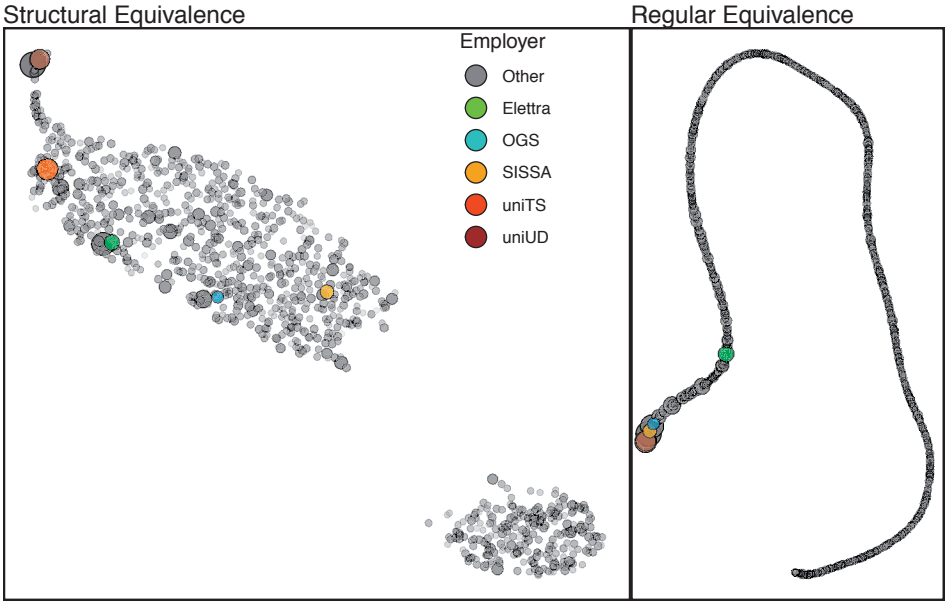


Figure 1. *Two- dimensional UMAP visualization of node embeddings generated from the RCCE network using VGAE (left panel) and Role2Vec (right panel).*

References

AHMED, N. K., ROSSI, R., LEE, J. B., WILLKE, T. L., ZHOU, R., KONG, X., & ELDARDIRY, H. 2018. Learning Role-based Graph Embeddings.

HAMILTON, W. L. 2020. Graph Representation Learning. **14**(3).

KIPF, T. N., & WELLING, M. 2016. Variational Graph Auto-Encoders.

MOREA, F., & DE STEFANO, D. 2022. Innovation patterns within a regional economy through consensus community detection on labour market network.

PARK, J., WOOD, I. B., JING, E., NEMATZADEH, A., GHOSH, S., CONOVER, M. D., & AHN, Y. Y. 2019. Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. *Nature Communications*, **10**(1).

SMALLBONE, D., KITCHING, J., BLACKBURN, R., & UKCES. 2015. Anchor institutions and small firms in the UK: A review of the literature on anchor institutions and their role in developing management and leadership skills in small firms.

METHOD FOR THE QUALITY CONTROL AND OPERATORS TRAINING IN MAINTENANCE ACTIVITIES

Massimiliano Giacalone[1], Vincenzo Dottorini[2], Giuseppe Oddo[3], Vito Santarcangelo[3], Angelo Romano[3]

[1] Department of Economics; University of Campania Luigi Vanvitelli

(e-mail: massimiliano.giacalone@unicampania.it)

[2] Ineltec srl (e-mail: dottorini@ineltec.it)

[3] iInformatica srl (e-mail: vito@iinformatica.it)

ABSTRACT: Maintenance activities are very important in the aim to prevent malfunctions and ensure the reliability, safety and performance of a productive process. In this work, a method for the evaluation of maintenance tasks is presented in the aim to give an objective evaluation of the maintenance technicians' skills and therefore for identifying critical areas in which intervene with appropriate training. For this purpose, a smart helmet for the training of the operators and for the control of the maintenance tasks was modified with cameras and sensors. The data collected were analysed with fuzzy logic approach and a score of the operators' skills was assigned in order to increase the quality of maintenance activities.

KEYWORDS: Fuzzy logic, Maintenance management, Quality control

1 Introduction

Maintenance activities are very important in the aim to prevent malfunctions and ensure the reliability, safety and performance of the equipment and the overall quality of a productive process [1]. Best practices of maintenance management include the development of a maintenance strategies through preventive intervention plans, which involve activities such as routine inspections and scheduled maintenance tasks [1]. If appropriate maintenance scheduling is the basis of good management practice in order to minimize the stop of the production, keep accurate records is fundamental in the aim to identify patterns of failure, predict maintenance needs and create statistical historical data in order to forecast interventions with accurate grade of precision [1]. At the same time, IoT technologies, such as sensors together with machine learning algorithms, can detect potential equipment failures and improve predictive maintenance activities [2]. However, best maintenance

practices should also include operators training for improving the efficiency and time of the maintenance operations [3]. In fact, providing training and appropriate know-how to the maintenance personnel can increase efficiency and effectiveness in maintenance tasks [3]. Higher knowledge in maintenance tasks could also promote a culture of safety ensuring awareness of the operators in the risks involved in the intervention tasks and thus implementing the necessary precautions to avoid injuries or accidents [3]. In this work, we present a case study of Ineltec srl, a firm that is specialised in projects and maintenance operations of electric plants together with iInformatica srl, for the research and development activities. The goal of this project was in particular to design a method for the evaluation of maintenance intervention through an objective evaluation of the maintenance personnel skills, and therefore for identifying critical areas in which intervene with appropriate training. The aim of the method was therefore to create a powerful tool based on a first control of the theoretic maintenance know-how of the operators, followed by a further analysis of their practical skills through sensors and camera mounted on a safety helmet. The method gives a score by analysing all the collected data with a fuzzy logic approach in order to give an objective evaluation of the overall operators' skills and improve constantly the maintenance quality intervention. For this purpose, a helmet for the training and control of maintenance activities of the operators was fabricated with the use of a 3D printed plastic shell filled by different cameras and sensors. Finally, an internal survey questionnaire was conducted in order to measure the operator feedback about the implemented method and their awareness about the importance of new technologies in their maintenance activities.

2. Method and results

In this work, a powerful method for evaluating maintenance activities was developed in the aim to improve constantly the quality of the intervention and training the operators in a safer manner. The designed method is based on a first evaluation of the theoretical know-how of the operators followed by a second control of the skills ability in solving the maintenance tasks. The evaluation of the theoretical know-how consists in an exercise in which the operators should choose the right sequence of actions in order to complete a maintenance activity. For the representation of a maintenance intervention we took inspiration from finite-state machine method in the aim to divide a global maintenance intervention in a sequence of events and actions. Taking into consideration the complexity of a maintenance intervention we developed a new method based on bubbles of events which describe not only a binary condition (such as close/open) but also an informative-semantic information of each events. In this way, by dividing and plotting maintenance activities through bubbles diagram we are able to represent complex and not linear activities such as maintenance intervention. Therefore, in order to evaluate theoretical know-how, the operators should choose the right sequence of actions by selecting the right bubbles in the right order (Panel in figure 1).

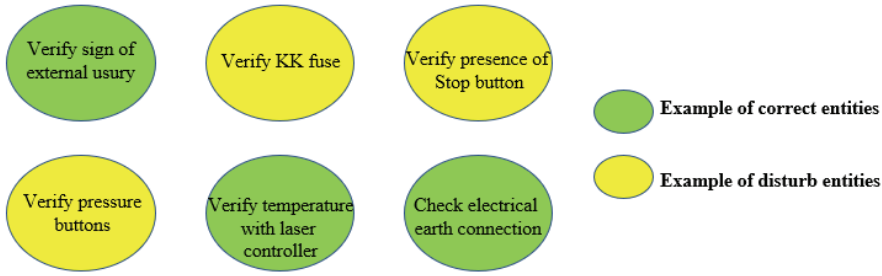


Figure 1. Bubbles maintenance test method

The implementation of also incorrect actions "disturb entities" was done in the aim to increase the difficulties of the test. The score is then attributed by considering the correct sequence of the selected actions (bubbles), decreased by the number of incorrect ones (disturb entities). The second phase of the evaluation consists in measuring operators' practical ability in solving the maintenance activities. With this purpose we fabricated a device for the remote real-time control of the operator intervention that consisted in a safety helmet modified with sensors and cameras. In particular, the helmet included a camera, a thermo camera and an endoscope for framing inaccessible places. Moreover, a mini screen and a mini sound box were mounted in order to communicate with remote operators for receiving instructions during the training activities. Through an opportune choice of each component, and choosing light 3D printed polymers for fabricating the outer shell of each sensors, the overall weight of the helmet was increased of only 320 grams in order to maintain a good wearability and operators' comfort during the intervention. In this way, by exploiting the cameras and sensors, a technician from remote can control the activities of the operators and give a feedback also on the practical abilities in solving the maintenance tasks and in case guiding the operators in critical situations. Thus, the final score considered both the know-how test and practical activity intervention taking also into consideration the boundary conditions in which the operator made the intervention through a fuzzy logic approach. The boundary conditions included: cleaning conditions (clean / intermediate / dirty), ventilation conditions (low / medium / high), Light conditions (low / medium / high), Spaciousness conditions (low/narrow/high), Noisy conditions (low/medium/high), customer stress conditions (low/medium/high) in the aim to increase the objectivity of the final score. An internal survey revealed that the totality of the operators was satisfied with the training activities conducted with the smart helmet and that they felt safer knowing that they could receive assistance remotely if needed. Considering the quality of the intervention, 75 % of the operators think that the helmet improved a lot their working activities and only the 25% think that the helmet had a small impact on their working activities. Finally, 100 % of the interviewers consider

important the adoption of new technologies in their work and think that new technologies will change drastically their work in the next years.

Conclusion

In this project, a method for analysing the quality of maintenance activities was designed in order to improve the training of the personnel involved and give a feedback of the quality of the maintenance activities. The method consisted in a first evaluation of the theoretical know-how of the operators followed by a second control of the skills ability in solving the maintenance tasks. A fuzzy logic analysis was implemented in the aim to consider also the boundary conditions in which the operators conducted the maintenance activities and thus assigning a more objective score of the performed task. For the real-time control of the activities, a safety helmet was modified with sensors and cameras in the aim to control the practical skills abilities of the operators. An internal survey questionnaire demonstrates an increase of operational speed, safety and quality of the maintenance intervention through the employed method, and a general awareness of the operators about the importance of new technologies adopted in their work.

References

[1] NAYARAN, V. 2012. Business performance and maintenance: How are safety, quality, reliability, productivity and maintenance related? *Journal of Quality in Maintenance Engineering*, **18**, 183-195.

[2] NUNES, P., SANTOS, J., ROCHA, E. 2023. Challenges in predictive maintenance – A review. *CIRP Journal of Manufacturing Science and Technology,* **40**, 53-67.

[3] SHERWIN, D. 2000. A review of overall models for maintenance management. *Journal of Quality in Maintenance Engineering*, **6**, 138-164.

BUILDING IMPROVED GENDER EQUALITY COMPOSITE INDICATORS BY OBJECT-ORIENTED BAYESIAN NETWORKS

Lorenzo Giammei [1], Flaminia Musella[2], Fulvia Mecatti[1] and Paola Vicard[3]

[1] Department of Sociology and Social Research, University of Milan-Bicocca, (e-mail: lorenzo.giammei@unimb.it, fulvia.mecatti@unimib.it)

[2] Department of Life Sciences, Health and Health Professions, Link Campus Unversity, (e-mail: f.musella@unilink.it)

[3] Department of Economics, Roma Tre University, (e-mail: paola.vicard@uniroma3.it)

ABSTRACT: This work proposes a novel methodology for constructing gender equality indicators using an Object-Oriented Bayesian Network (OOBN). The methodology is illustrated by focusing on the composite indicator known as Gender Equality Index, annually released by the European Institute of Gender Equality (EIGE). By using province-level ISTAT data, the index is re-constructed in a modern AI environment, able to enhance its information capacity and, at the same time, to preserve its original architecture. The modularity of the OOBN ensures a computational logic that is consistent with composite indicators, while also providing additional information about the relational structure of variables.

KEYWORDS: Object-Oriented Bayesian Networks, gender equality, composite indicator, regional indicator, sustainable development goals.

1 Introduction

Gender based inequalities represent a threat to socio-economic well-being on an individual level as well as for the society as a whole. Gender equality is one of the objectives pursued by the Sustainable Development Goals (SDG) of the UN Agenda 2030, as stated in the ambitious Goal 5: to achieve gender equality and empower all women and girls. In order to reach Goal 5, gender equality measurement plays a key role. The most widely used approach for measuring national gender equality is through a composite indicator. Composite indicators are useful in monitoring complex multidimensional phenomena, including gender equality, by providing a single value information, e.g. ranking of countries in their progresses toward SDG 5. However, composite indicators do not

show the process of how a country has reached its own level of national gender equality nor they allow for monitoring or predicting the effect of policies and interventions. In this work we aim at empowering the role of gender inequality data analysis, by proposing a new method to measure the gender gap that can be used alongside composite indicators to obtain a richer set of information. In particular we employ an OOBN that follows the idea behind the computation of the European Union's Gender Equality Index (EU-GEI) but takes into account the multivariate dependence structure among all the variables generating a certain level of gender equality.

2 Object Oriented Bayesian Networks

A Bayesian network (BN) (Cowell, 1998; Pearl, 1998) is a probabilistic statistical model representing the joint distribution of a set of variables by means of a directed acyclic graph (DAG). In a DAG, nodes represent variables and edges denote the influence of one variable to another one. Bayesian networks possess a relevant and crucial property, named *modularity*, by which a possibly complex multivariate relation structure can be decomposed into smaller modules encoding conditional independencies. In a sense, BN can serve as basic building blocks for an extended tool called Object-Oriented Bayesian Networks (Koller & Pfeffer, 2013). An OOBN is a multi-instances network made of objects and special nodes. Objects are also called instance nodes representing simpler networks; the flow of information between networks is allowed by two kinds of interface nodes: input nodes, that import information from the OOBN into the instance; output nodes, that broadcast information from the instance to the OOBN. Since an instance node is a BN encapsulated in the OOBN, these models take advantage of the statistical properties of Bayesian networks. The inference process of OOBNs is efficient due to conditional independence between standard nodes in the instance and the OOBN, given the interface nodes. The architecture of OOBNs is particularly useful for managing large and complex domains, particularly when hierarchical structures are present. In the literature, there are many applications of OOBNs; the most relevant contribution for our purpose arises from the managerial framework (Musella & Vicard, 2015) where different quality aspects have been combined to provide a synthetic global quality indicator.

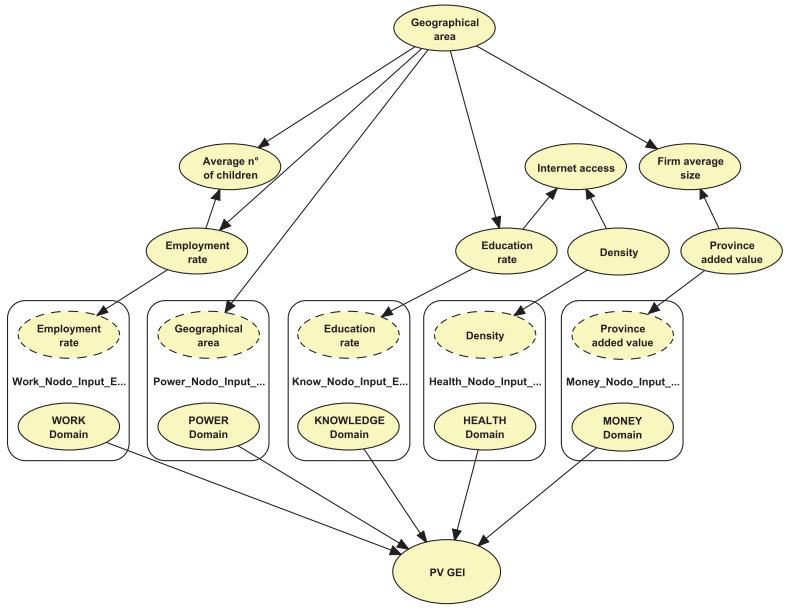


Figure 1. *Object-Oriented Bayesian Network model for the Italian province-level GEI*

3 Application to Real Data

In this work, we employ an OOBN to develop a gender equality indicator for Italian provinces based on the architecture of the EU-GEI. The EU-GEI, annualy released by the EIGE, is based on 6 domains: work, money, knowledge, time, power and health. We use province-level data from ISTAT to obtain a set of variables that is consistent with the one employed to compute the EU-GEI. Due to limited data availability, it is not possible to perfectly replicate the national GEI at the province level. To overcome the scarcity of gender-sensitive data at a fine granularity, proxy variables have been included if available. As a result all the EU-GEI domains are represented in the province-level GEI (PV-GEI), except for the Time domain. In addition, some extra socio-economic variables, such as *province added value* and *firm average size* are included to investigate their relationship with GEI ingredients. The resulting OOBN, obtained employing the statistical software Hugin, is depicted in Figure 1. Each box (rounded rectangle) in the OOBN represents an instance, which is a simpler network representing a specific PV-GEI domain that is linked to the whole OOBN. In each instance, the input node (represented as a node with a dashed

outline in the figure) is selected from among the extra variables, while the output node is the summary value of the domain. According to the EU-GEI methodology, this value is the geometric mean of the sub-domain measure, which is in turn the arithmetic mean of the ingredient variables. PV-GEI ingredients are not visible in the OOBN at this level of representation because they are part of the sub-networks given by the domain instances. The different domains are then aggregated to compute the PV-GEI as a weighted geometric mean of domain measures. This architecture allows information to flow from extra socio-economic variables to PV-GEI ingredients, which in turn generates a certain level of the domain nodes and of the PV-GEI. The resulting model is consistent with the EU-GEI architecture and, at the same time, constitutes a powerful tool to simulate scenarios of how the PV-GEI changes when ingredient or socio-economic variables take different values.

4 Discussion and future research

The proposed methodology enriches composite indicators and provides a new perspective on the analysis of the gender gap. By employing an OOBN, we not only obtain a measure of gender equality that is consistent with that of composite indicators but also gain insight into the multivariate relationships between ingredients and other variables of interest. In addition, a refined granularity can be reached depending on the available data. These findings can support policy decision-making by shedding additional light on the complex net of factors that affect the gender (in)equalities. Finally, the estimated OOBN provides a simulation engine to predict the effect of policies and intervention aiming at reducing gender-based inequalities in the Country.

References

COWELL, R. 1998. Introduction to inference for Bayesian networks. *Learning in graphical models*, 9–26.

KOLLER, D, & PFEFFER, A. 2013. Object-oriented Bayesian networks. *arXiv preprint arXiv:1302.1554*.

MUSELLA, F, & VICARD, P. 2015. Object-oriented Bayesian networks for complex quality management problems. *Quality & Quantity*, **49**, 115–133.

PEARL, J. 1998. *Probabilistic reasoning in intelligence systems: networks of plausible inference*. Los Altos, CA: Morgan Kaufmann.

A COMPARATIVE STUDY OF FINANCIAL LITERACY USING DATA FROM PISA SURVEY

Sabrina Giordano [1], Roberta Varriale[2] and Mariangela Zenga[3]

[1] Dipartimento di Economia, Statistica e Finanza "Giovanni Anania"- DESF, Università della Calabria, (e-mail: sabrina.giordano@unical.it)

[2] Dipartimento di Scienze Statistiche, Università Sapienza, (e-mail: roberta.varriale@uniroma1.it)

[3] Dipartimento di Statistica e Metodi Quantitativi, Università Milano Bicocca (e-mail: mariangela.zenga@unimib.it)

ABSTRACT: Financial literacy has become a crucial goal for countries' policymakers in recent decades. Since 2012, the PISA survey has been enriched by the assessment of financial literacy of adolescent students in various countries. Through the use of hierarchical models, we can account for variations in financial literacy levels between countries by examining the characteristics of students and their families, as well as analyzing the structure of formal education and policies within each country. This analysis is based on the 2018 year.

KEYWORDS: Financial Literacy, PISA survey, hierarchical model.

ON MODEL-BASED CLUSTERING FOR EQUITABLE AND SUSTAINABLE WELL-BEING AT LOCAL LEVEL: HOW MANY ITALIES?

Natalia Golini[1], Francesca Martella[2] and Antonello Maruotti[3]

[1] Department of Economics and Statistics "Cognetti de Martiis", University of Turin, (e-mail: natalia.golini@unito.it

[2] Department of Statistical Sciences, Sapienza University of Rome, (e-mail: francesca.martella@uniroma1.it)

[3] Department of Economic, Political Sciences and Modern Languages, Libera Università Maria Ss. Assunta, (e-mail: a.maruotti@lumsa.it)

ABSTRACT: The choice of an appropriate number of clusters is a key issue in model-based clustering framework. The most popular approaches are based on the information criteria. However, often the latter may likely overestimate the number of clusters even though a good density estimation is possible. Here, we provide a dynamic model-based clustering approach to identify homogeneous Italian NUTS3 areas based on their equitable and sustainable well-being indicators from 2004 to 2019. In particular, the proposed model allows NUTS3 areas to move between clusters over time and a local dimensional reduction within each cluster. The empirical results show a high heterogeneity among the NUTS3 areas, leading to a high number of clusters. Possible strategies for merging similar NUTS3 clusters are investigated.

KEYWORDS: dimensionality reduction, dynamic clustering, hidden Markov model, longitudinal data

1 Introduction

In Italy, the National Institute of Statistics (Istat) has developed a multidimensional approach to measure "equitable and sustainable well-being" (BES), having the aim to integrate the traditional economic indicators with the quality of life of people, environment, inequality and sustainability measures. These indicators, updated annually since 2004, are declined into 12 relevant domains. Recently, Istat has designed a system of equitable and sustainable well-being indicators at NUTS3 level*, i.e. at the 107 Italian provinces, to deepen the

*NUTS: Nomenclature of Territorial Units for Statistics; NUTS 3: small regions for specific diagnoses (https://ec.europa.eu/eurostat/web/nuts/background).

knowledge of the well-being distribution across Italy to assess inequalities across areas. Local indicators are consistent with the national BES measures.

This paper addresses the complex, often non-linear, correlation between the indicators, the heterogeneity characterizing the Italian NUTS3 areas and changes and shifts in society over time under a unified framework. We identify homogeneous NUTS3 areas which behave in a lifestyle-similar fashion while keeping track of changes over time. We consider a clustering approach because more structured than a suitable standard approach in socio-economic analyses. To accommodate the multivariate longitudinal structure of the data, we propose a parsimonious hidden Markov model (HMM) that allows NUTS3 areas to transit between clusters, i.e. different well-being levels, over time. In this respect, a first-order finite-state Markov chain has been used to consider the temporal dependence. Moreover, a factor model framework is considered to capture correlation among indicators. And finally, we allow such correlations to vary across clusters and times to make the model flexible enough to capture the longitudinal structure of the data. The model parameters have been estimated through an Alternating Expected Conditional Maximization (AECM; Meng & van Dyk, 1997) algorithm.

2 Data and methods

2.1 Data description

The motivating dataset is composed of 102 NUTS3 areas and 18 well-being selected indicators, declined in 7 domains, to monitor their dynamics during the period 2004 – 2019. This choice was made to consider the largest number of Italian NUTS3 areas without missing data during the observational period. Accordingly, four domains (Economic well-being, Social relationships, Landscape and cultural heritage and Innovation, research and creativity) and five NUTS3 areas (Barletta-Andria-Trani, Enna, Fermo, Monza e della Brianza and Sud Sardegna) are excluded from our analysis. The data are freely available at the Istat website[†]. A descriptive analysis confirms that socioeconomic divergence between the North and South of Italy continues, with the North more productive, rich and with a good health system, and the South/Islands, where the economy is mainly based on tourism, with higher unemployment

[†]`https://www.istat.it/en/well-being-and-sustainability/the-measurement-of-well-being/bes-at-local-level`.
This database contains data and metadata for the period 2004 – 2020.

rates. Moreover, each BES indicator is related to others differently: the correlation structure is rather heterogeneous, and patterns of nonlinear correlation are present.

2.2 Parsimonious hidden Markov models for longitudinal data

We consider an HMM for multivariate longitudinal data allowing the density of the observed process to follow a factorial model. In detail, the model is defined by an observed process $\{\mathbf{Y}_{it}, i = 1,\dots,n; t = 1,\dots,T\}$ and a hidden-dependent process $\{S_{it}, i = 1,\dots,n; t = 1,\dots,T\}$ defined on the cluster space $\{1,\dots,K\}$ such that $\Pr(S_{it} \mid S_{i1},\dots,S_{it-1}) = \Pr(S_{it} \mid S_{it-1})$. Regarding the observed process $\mathbf{Y}_{it} = \{Y_{it1},\dots, Y_{itP}\}$, Y_{itp} represents the p-th response variable given by the i-th units at time t $(i = 1,\dots,n; p = 1,\dots,P; t = 1,\dots,T)$ such that $f(\mathbf{Y}_{it}|\mathbf{Y}_{i1},\dots,\mathbf{Y}_{iT},S_{i1},\dots,S_{iT}) = f(\mathbf{Y}_{it}|S_{it})$. Moreover, we defined the initial probabilities $\pi_k = Pr(S_{i1} = k)$ $(i = 1,...,n; k = 1,...,K)$ and the transition probability matrix $\Pi = \{\pi_{k|j}\}$, where $\pi_{k|j} = Pr(S_{it} = k \mid S_{it-1} = j)$ $(i = 1,...,n; t = 1,...,T,; j,k = 1,...,K)$. In line with the idea proposed by Maruotti *et al.*, 2017, we assume that conditionally to the k-th cluster, the random vector $\mathbf{Y}_{it}$ is described by:

$$\mathbf{Y}_{it} = \mu_k + \Lambda_k \mathbf{f}_{itk} + \mathbf{e}_{itk}, \tag{1}$$

where $\mathbf{f}_{itk}$ is a q-dimensional vector of cluster-specific factors drawn from $\mathrm{N}_P(\mathbf{0}, \mathbf{I}_q)$, and $\mathbf{e}_{itk}$ is a p-dimensional vector of cluster-specific error terms drawn from $\mathrm{N}_P(\mathbf{0}, \Psi_k)$, where $\Psi_k = \mathrm{diag}(\psi_{k1},\dots,\psi_{kP})$, which is assumed to be independent of $\mathbf{f}_{itk}$. In other words, a unit i in cluster k follows a multivariate Gaussian density with cluster-dependent mean vector μ_k and covariance matrix $\Lambda_k\Lambda_k' + \Psi_k$. Notice that, by constraining whether $\Lambda_k = \Lambda$, $\Psi_k = \Psi$ and $\Psi_k = \psi_k \mathbf{I}_p$, a family of 8 different models can be derived. To fit the proposed models, we use the AECM algorithm and recursions widely used in the HMM literature. The simulation studies results have shown a very good model performance in terms of the accuracy of the parameter estimates, degree of agreement between two partitions, and the ability to detect the correct number of clusters.

3 Empirical results

We computed Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) for each of the eight fitted models and combination (K,q). All the information criteria select the

unconstrained model (volumes, shapes, and orientations of the clusters are variable among clusters) as the best solution and, in particular, BIC and ICL recommend the solution with $K = 15$ and $q = 4$ as the most reasonable one balancing fit and parsimony. The main results can be briefly summarized as follows: Italian NUTS3 areas are heterogeneous; the inferred clustering structure identifies homogeneous, well-separated *spatial* aggregations of NUTS3 areas, leading to *four Italies*; persistence is the norm, transitions across clusters are rare but still present; cluster-specific correlations among indicators are effectively observed; each cluster is strongly characterized by only a subset of well-being indicators.

4 Assessing separation between NUTS3 clusters

The estimated results show that Italy still has a significant way to go in achieving well-being convergence. The heterogeneity across NUTS3 areas is still relevant, leading to a high number of clusters. However, the fact that some clusters differ only by the values of a few indicators suggests that there may be opportunities to merge similar clusters to get a more accurate overall picture of well-being in Italy and keep the specificities for further policymakers interventions. On the basis of the most widely used approaches in the field (see Hennig, 2010; Baudry *et al.*, 2010; Melnykov, 2016 among others), this opportunity is investigated.

References

BAUDRY, JEAN-PATRICK, RAFTERY, ADRIAN E, CELEUX, GILLES, LO, KENNETH, & GOTTARDO, RAPHAEL. 2010. Combining mixture components for clustering. *Journal of computational and graphical statistics*, **19**(2), 332–353.

HENNIG, CHRISTIAN. 2010. Methods for merging Gaussian mixture components. *Advances in data analysis and classification*, **4**, 3–34.

MARUOTTI, A., BULLA, J., LAGONA, F., PICONE, M., & MARTELLA, F. 2017. Dynamic Mixture of factor analyzers to characterize multivariate air pollutant exposures. *Ann. Appl. Stat.*, **11**(3), 1617–1648.

MELNYKOV, VOLODYMYR. 2016. Merging mixture components for clustering through pairwise overlap. *Journal of Computational and Graphical Statistics*, **25**(1), 66–90.

MENG, X. L., & VAN DYK, D. A. 1997. The EM Algorithm? An Old Folksong Sung to a Fast New Tune. *J. R. Statist. Soc. B*, **59**(3), 511–567.

MODEL-BASED CLUSTERING FOR TORUS DATA

Luca Greco [1], Antonio Lucadamo [2] and Claudio Agostinelli [3],

[1] University G. Fortunato, Benevento, Italy, (e-mail: l.greco@unifortunato.eu)

[2] Department DEMM, University of Sannio, Italy (e-mail: antonio.lucadamo@unisannio.it)

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

ABSTRACT: Torus data are multivariate circular observations that arise as measurements on a periodic scale and are often recorded as angles. In this paper, we focus on parsimonious model based clustering for torus data by building on the `mclust` methodology. Therefore, covariance constraints are imposed on the completely general heterogeneous clustering model allowing a flexible and general framework to clustering torus data.

KEYWORDS: torus data, model-based clustering, wrapped distribution

1 Introduction

Torus data are multivariate circular observations. Many applications involve torus data in several fields: protein bioinformatics, wind directions, animal movements, people orientation, human motor resonance, robotics, astronomy, meteorology, geology, medicine, oceanography. Actually, multivariate circular data can be thought of as points on a p-torus $\mathbb{T}^p$, $p > 1$, whose surface is obtained by revolving the unit circle in a p-dimensional manifold. The multivariate wrapped normal (WN) distribution is a very attractive model for torus data (Mardia & Jupp, 2000). In Greco *et al.*, 2022, the WN distribution has been proved to be very useful in modeling mixtures of torus data and providing an effective tool for model based clustering and classification, but only under a completely general heterogeneous clustering model. In this paper, by paralleling a widely used methodology for *linear* data on $\mathbb{R}^p$, we focus on parsimonious model based clustering for torus data by building on the `mclust` methodology (Scrucca *et al.*, 2016).

2 Parsimonious model based clustering

Let us consider a sample of size n of torus data $y = (y_1, y_2, \ldots, y_n)$, from the finite mixture model with density function

$$f^{\circ}(y;\tau) = \sum_{g=1}^{G} \delta_g m^{\circ}(y;\theta_g), \tag{1}$$

where we set $\tau = (\delta_1, \ldots, \delta_G, \theta_1, \ldots, \theta_G)$, G denotes the number of groups, δ_g are membership probabilities, $\delta_g > 0$, $\sum_{g=1}^{G} \delta_g = 1$, $\theta_g = (\mu_g, \Sigma_g)$ are component specific location and scatter and $m^{\circ}(y;\theta_g) = \sum_{j \in \mathbb{Z}^p} m(y + 2\pi j;\theta)$ is the wrapped density function, where j is the vector of wrapping coefficients and $m(\cdot)$ the corresponding unwrapped density. Let $m^{\circ}(y;\theta_g)$ be the density of a WN distribution (being $m(\cdot)$ the normal density). Building on `mclust`, we enforce constraints on the scatter matrices Σ_g using the parsimonious models of Celeux & Govaert, 1995 that can be obtained by means of the eigenvalue decomposition of the covariance matrices of the form $\Sigma_g = \lambda_g D_g A_g D_g^{\top}$, where $\lambda_g = [\det(\Sigma_g)]^{1/d}, d = 1, 2, \ldots, p$, is a measure of the volume of the g^{th} cluster, A_g is a diagonal matrix with the eigenvalues of Σ_g, with $\det(A_g) = 1$, specifying the shape and D_g is an orthogonal matrix whose columns are given by the eigenvectors of Σ_g which determines the orientation.

In order to make estimation of wrapped models feasible, the infinite sum over $\mathbb{Z}^p$ is replaced by a sum over the Cartesian product $\mathcal{C}_J = \otimes \mathcal{J}^p$, $\mathcal{J} = (-J., -J+1, \ldots, 0, \ldots, J-1, J)$, for some J providing a good approximation. Then, maximum likelihood estimation of the model in (1) follows from the maximization of the mixture log-likelihood function.

$$\ell(\tau) = \sum_{i=1}^{n} \log f^{\circ}(y_i;\tau) = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \delta_g \sum_{j \in \mathcal{C}_J} m(y + 2\pi j;\theta_g) \right]. \tag{2}$$

The operations of mixing and wrapping commute, and (2) can be rewritten as

$$\ell(\tau) = \sum_{i=1}^{n} \log \left[\sum_{j \in \mathcal{C}_J} \sum_{g=1}^{G} \delta_g m(y + 2\pi j;\theta_g) \right] = \sum_{i=1}^{n} \log f(y_i + 2\pi j;\tau)$$

where $f(y + 2\pi j;\tau)$ is a mixture density for linear data.

Observe that the wrapping coefficients j are unknown. Then, they can be considered as latent variables and the observed torus data y as being incomplete. In the following, maximum likelihood estimation relies on a data augmentation

approach and is performed according to a suitable Classification Expectation Maximization algorithm. The point is that there are two sources of incompleteness in (2): one given by the wrapping coefficient vectors, the other from group memberships. The proposed algorithm iterates between an outer Classification Expectation (CE) step, in which the circular data are unwrapped to fitted linear data $\hat{x} = y + 2\pi\hat{j}$ (see Nodehi *et al.*, 2021), and an inner run of a classical EM algorithm for (linear) finite mixtures using the fitted linear data. Actually, the algorithm maximizes the (approximated) classification log-likelihood function based on the complete torus data (y, j):

$$\ell_c(\tau) = \sum_{i=1}^{n} \sum_{j \in C_J} v_{ij} \log \left[\sum_{g=1}^{G} \delta_g m(y_i + 2\pi j; \theta_g) \right] , \tag{3}$$

where $v_{ij} = 1$ or $v_{ij} = 0$ according to wheter y_i has $j \in C_J$ as wrapping coefficients vector.

Formal approaches to infer the number of clusters and select the best model among the available parsimonious alternatives can be based on the value of the penalized complete log-likelihood function (3) at convergence or, alternatively, of the incomplete data log-likelihood function (2). Classical model selection criteria are given by the Bayesian Information Criterion (BIC) or the integrated complete-data likelihood criterion (ICL).

3 A numerical example

Let us consider a synthetic data example to illustrate the proposed methodology. The sample size is $n = 500$. Data have been generated according to a bivariate WN mixture model with two components and unbalanced memberships probabilities, imposing an EII covariance structure. Starting values are driven from cluster-wise constrained maximum likelihood estimation under the assumed model from an initial partition obtained using the angular separation distance and the Ward agglomerative method. The BIC selects the right model, in this example. Cluster assignments are plotted in Figure 1. Tolerance ellipses are also given, based on the normal model. Note that the data have been represented on a flat torus, that is the same data structure repeats itself on the Euclidean space to account for the wraparound nature of the data. i.e. data are represented for different js. The procedure has been repeated 500 times. The model EII has been correctly selected in 95.6% of the simulations. The average Adjusted Rand Index (ARI) between the inferred partitions and the true component memberships is 0.963.

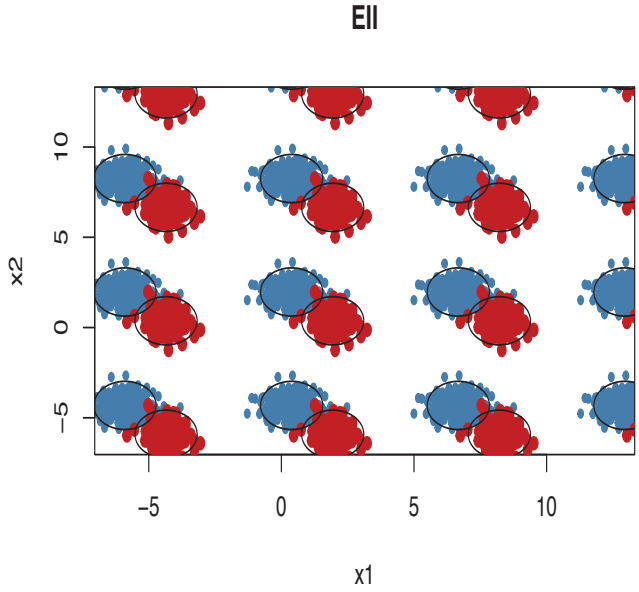


Figure 1. *Cluster assignments and tolerance ellipses under the EII model.*

References

CELEUX, G., & GOVAERT, G. 1995. Gaussian parsimonious clustering models. *Pattern recognition*, **28**(5), 781–793.

GRECO, L., NOVI INVERARDI, P., & AGOSTINELLI, C. 2022. Finite mixtures of multivariate Wrapped Normal distributions for model based clustering of p-torus data. *Journal of Computational and Graphical Statistics*.

MARDIA, K. V., & JUPP, P. E. 2000. *Directional statistics*. Wiley Online Library.

NODEHI, A., GOLALIZADEH, M., MAADOOLIAT, M., & AGOSTINELLI, C. 2021. Estimation of parameters in multivariate wrapped models for data on ap-torus. *Computational Statistics*, **36**, 193–215.

SCRUCCA, L., FOP, M., MURPHY, T. B., & RAFTERY, A. E. 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, **8**(1), 289.

AUTOSYNTH INDEX: A SYNTHETIC INDICATOR FOR SOCIO-ECONOMIC DEVELOPMENT BASED ON AUTOENCODERS

Giulio Grossi [1], Emilia Rocco[1]

[1] Department of Statistics, Informatics and Computer Science, University of Florence
(e-mail: giulio.grossi@unifi.it, emilia.rocco@unifi.it)

ABSTRACT: In this work, we propose a novel use for neural networks to build socioeconomic indicators, encoding a possible large information set, within single or multiple synthetic indexes, we call this proposal AutoSynth. In particular, we encode such information using an autoencoder, a neural network method to represent in a lower dimensionality space a matrix of features. We apply such a method to the evaluation of socio-economic developments of suburban areas in Florence, and we test the performance of our model against some golden standard methods using a stress test.

KEYWORDS: synthetic indicators, composite indicators, autoencoders, neural network, unsupervised learning

1 Introduction

Composite indicators are statistical measures that combine a set of elementary (or individual) indicators into a single measure of a complex phenomenon, such as the Human Development Index(HDI) or the Environmental performance index (EPI). See Commission *et al.*, 2008 for an account on the construction of synthetic indicators. Recently, Greco *et al.*, 2019 presents a review of the existent literature, focusing on the main goal of indicators construction and on the open challenges. The primary goal of a synthetic indicator should be the transmission of the information contained into each elementary indicator, with the lowest possible loss of such data. Moreover, such indicators rely on making transparent ranking that allows for spatial and temporal comparison between units and therefore are particularly suited to keep track of improvements in complex phenomena. With wider and more detailed sources of information, larger datasets are employed and feature extraction techniques are needed for accounting the amount of information that is considered. Golden standard approaches employ weighted averages or geometric averages to extract a single index from a matrix. An example is the Adjusted Mazziotta-Pareto index (AMPI) Mazziotta & Pareto, 2018, a novel synthetic indicator for measuring well-being. These methods are very transparent, yet it is not

completely clear what should be the weights accounted for, and often strong theoretical knowledge is required. This task could become very difficult in presence of large datasets, where describing the relationships between variables could be cumbersome. Unsupervised learning approaches for constructing composite indicators have been deployed during the last 30 years, such as Principal Component Analysis and Factor Analysis, see Greco *et al.*, 2019 and Commission *et al.*, 2008 for some review and comments. In this work, we propose a novel unsupervised framework for developing synthetic indicators. Exploiting modern methods for data analysis, we perform a data compression within a single index, with the minimum loss of information compared to previous approaches. We employ autoencoders based on neural networks that are able to grasp the relevant information in the dataset, even in presence of large datasets and without a backing theory. We apply this estimator to the evaluation of wellbeing in the suburban areas of Florence, and compare results from our methods with ones coming from previous approaches.

2 Methodology

Let $\mathbf{X}$ be a $N \times K$ normalized matrix of covariates, describing socioeconomic phenomena , observed for N units and K covariates. An autoencoder (Hinton & Zemel, 1993) is a type of neural network that consists of an encoder and a decoder, where the encoder maps the input data to a lower-dimensional latent space and the decoder maps the latent representation back to the original data space. The encoder can be seen as a probabilistic mapping function that generates a probability distribution over the latent variables, given the input data, Kramer, 1991 call it as nonlinear PCA, which is a quite familiar method into syntetic indexes literature. Therefore, let be ϕ an encoder function that maps $\mathbf{X}_{N\times K} \rightarrow \mathbf{R}_{N\times 1}$, and similarly let ψ be a decoder function mapping $\mathbf{R}_{N\times 1} \rightarrow \mathbf{X}_{N\times K}$. Thus the autoencoder is trained to minimize

$$\underset{\phi,\psi}{\operatorname{argmin}} ||\mathbf{X} - \psi(\phi(\mathbf{X}))||^2$$

In our application, as we wish to summarize covariates into a single vector, we are interested in calculating the code $\mathbf{R} = \phi(\mathbf{X})$. See figure 1 for a graphical representation. To assess AutoSynth performances we study stress values. Let $\theta = \sqrt{\frac{\sum_{i<j}(d_{ij}-\delta_{ij})^2}{\sum_{i<j} d_{ij}^2}}$ be a stress measure of the discrepancy between the distances in the original high-dimensional space ($d_{i,j}$) and the distances in the lower-dimensional space ($\delta_{i,j}$). Thus, The lower the value of θ, the higher the ability of the low-dimensional variables in representing the original data.

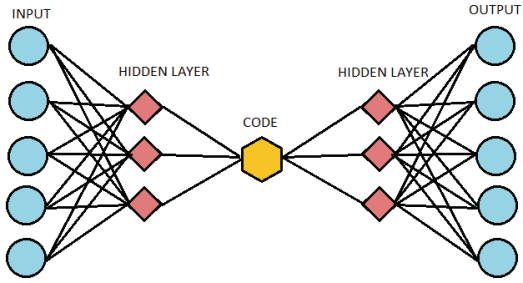


Figure 1: Basic scheme of autoencoders. In this application, inputs will be elementary indicators of socio-economic development, while the code will be the synthetic indicator

Table 1: Fragility dimensions of Florence - year 2021

Demographic	**Economic**	**Social**
% of elders in the population	% of inhabitants in poverty	% of minors in single-parent families
Natural balance	% of families in poverty	% of elders living alone
5-yr variation of inhabitants	% of rented residents	% of foreigners minors
	Median family income	% of graduates
		Permanent residents

3 Measuring Florentine fragilities

We applied our proposed method to the evaluation of fragilities into Florentine suburbs. Fragility can be represented into a composite indicator of three main dimensions: demographic fragility, economic fragility and social fragility. Moreover, we can identify some elementary indicators, previously used in this literature, to represent each of these dimensions. Table 1 shows the indicators used in the analysis, referred to 2021. In total, we collect information over the 74 suburbs that make up Florence. Using the elementary indicators in table 1, we first normalize the variables, as in Mazziotta & Pareto, 2018, and later we apply on the same dataset, AMPI, PCA and AutoSynth transformations, rescaling the compressed variables to the same ”goalposts”, as in Mazziotta & Pareto, 2018. Figure 2 and table 2 reports the fragility maps and the stress value for the three methods considered. From these results, we notice that our model has very noticeable performances in representing the input covariates, and thus is able to reproduce better the original dimensions into a single feature space.

4 Conclusion

Concluding, In this work, we propose to use Autoencoders to construct a synthetic indicator for socio-economic development and apply it to the evaluation

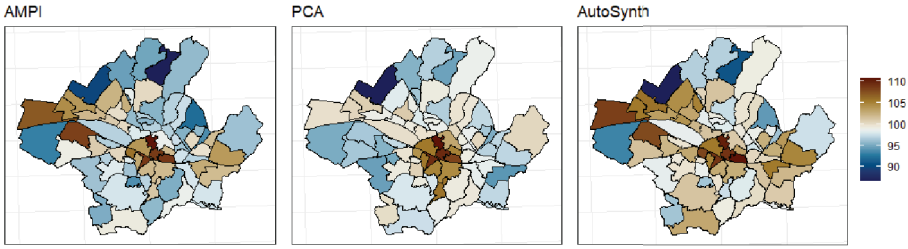


Figure 2: AMPI, PCA and AutoSynth Fragility Index for Florence

Table 2: Stress absolute values for each method considered and as fraction of the AMPI stress test

	AMPI	**PCA**	**AutoSynth**
θ	0.03497	0.00657	0.00447
	1	0.188	0.128

of fragility in the Florence suburbs. Results obtained from the stress values suggest an improved ability in dimension reduction, nevertheless, the maps comparison shows similar results with respect to the AMPI. Considering the wide flexibility of autoencoders, their application to the construction of synthetic indicators could become a promising area of study.

References

COMMISSION, JOINT RESEARCH CENTRE-EUROPEAN, *et al.* 2008. *Handbook on constructing composite indicators: methodology and user guide*. OECD publishing.

GRECO, SALVATORE, ISHIZAKA, ALESSIO, TASIOU, MENELAOS, & TORRISI, GIANPIERO. 2019. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social indicators research*, **141**, 61–94.

HINTON, GEOFFREY E, & ZEMEL, RICHARD. 1993. Autoencoders, minimum description length and Helmholtz free energy. *Advances in neural information processing systems*, **6**.

KRAMER, MARK A. 1991. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, **37**(2), 233–243.

MAZZIOTTA, MATTEO, & PARETO, ADRIANO. 2018. Measuring well-being over time: The adjusted Mazziotta–Pareto index versus other non-compensatory indices. *Social Indicators Research*, **136**, 967–976.

A STATISTICAL TEST TO ASSESS THE NON-NORMALITY OF THE LATENT VARIABLE DISTRIBUTION

Lucia Guastadisegni[1], Irini Moustaki[2], Silvia Cagnone[1] and Vassilis Vasdekis[3]

[1] Department of Statistical Sciences "Paolo Fortunati", University of Bologna, (e-mail: lucia.guastadisegni2@unibo.it, silvia.cagnone@unibo.it)

[2] Department of Statistics, London School of Economics and Political Science, (e-mail: i.moustaki@lse.ac.uk)

[3] Department of Statistics, Athens University of Economics and Business, (e-mail: vasdekis@aueb.gr)

ABSTRACT: This paper presents the generalized Hausman test to detect non-normality of the latent variable distribution in unidimensional Item Response Theory (IRT) models for binary data. The test is based on the estimators resulting from the two-parameter IRT model, that assumes normality of the latent variable, and the semi-nonparametric IRT model, that assumes a more flexible latent variable distribution. The performance of the test is evaluated through a simulation study, including the cases where the latent variable is generated from a skew-normal and mixture of normals. The results highlight the good performance of the test when the latent variable is generated from a mixture of normals and from a skew-normal only with many items and large sample sizes.

KEYWORDS: generalized Hausman test, SNP-IRT model, binary data

1 Introduction

In unidimensional IRT models for binary data, the latent variable is typically assumed standard normally distributed. However, assuming normality in the model when the true latent variable distribution has a different shape than the normal one can result in large biases in parameter estimates (Ma & Genton, 2010). IRT models that assume different form of the latent variable have been proposed (for example Irincheeva *et al.*, 2012) but detecting latent variable non-normality through a statistical test remains an open issue. In this paper, we consider the generalized Hausman (GH) test (White, 1982) to detect non-normality of the latent variable distribution in unidimensional IRT models for

binary data. The test is based on the maximum pairwise likelihood (PL) estimator (Lindsay, 1988) of the classical unidimensional IRT model for binary data, based on the normality assumption of the latent variable, and the quasi-maximum likelihood (ML) estimator of the unidimensional seminonparametric (SNP)-IRT model for binary data, that assumes a more flexible latent variable distribution (Irincheeva *et al.*, 2012). Some preliminary results on the performance of the GH test have been presented in Guastadisegni *et al.* (forthcoming). In details, the GH test has shown a good performance in terms of Type I error rates with many items and large sample size. The power of this test has only been evaluated when the latent variable is generated from a mixture of normals. In this paper, we evaluate the performance of the GH test also when the latent variable is generated from a skew-normal distribution.

2 The IRT models for binary data

Let $y_1,...,y_p$ denote a set of observed binary variables/items, n the number of individuals and z the latent variable with density function $h(z)$. The response probability for the i-th individual to the j-th item is modelled using a logistic model (measurement model)

$$P(y_{ij}=1|z_i)=\pi_{ij}(z_i)=\frac{\exp(\alpha_{0j}+\alpha_{1j}z_i)}{1+\exp(\alpha_{0j}+\alpha_{1j}z_i)}, \tag{1}$$

where α_{0j} is the item intercept and α_{1j} the item slope. For the classical IRT model, $h(z)=\phi(z)$, where $\phi(z)$ is the density of a standard normal. For the SNP-IRT model, the latent variable has the following SNP parametrization (Irincheeva *et al.*, 2012)

$$h(z_i)=P_L^2(z_i)\phi(z_i) \qquad P_L(z_i)=\sum_{0\le l\le L} a_i z_i^l. \tag{2}$$

$a_0,...,a_L$ are the real coefficients of the polynomial $P_L(z_i)$ and L is the polynomial degree. SNP_1 denotes the model for $L=1$, where $P_L(z)=a_0+a_1z$, $a_0=\sin\varphi_1$, $a_1=\cos\varphi_1$, $-\pi/2<\varphi_1\le\pi/2$. SNP_0 denotes the model for $L=0$, where the distribution of the latent variable reduces to the normal one. To implement the GH_T test, we consider the SNP_0 and the SNP_1 model.

3 The generalized Hausman test

Consider the maximum PL estimator $\tilde{\eta}_{SNP_0}$ of the SNP_0 model, that includes the item intercepts and slopes of dimension $2p\times 1$, where p is the number of

items. Under normality of the latent variable distribution, the maximum PL estimator $\tilde{\eta}_{SNP_0}$ converges in probability to the true parameter value η_0. Consider also the quasi-ML estimator $\hat{\theta}'_{SNP_1} = (\hat{\eta}'_{SNP_1}, \hat{\varphi}_1)$ of the SNP_1 model, of dimension $(2p+1) \times 1$. Under normal, multi-modal and asymmetric distributions of the latent variables and if the regularity conditions A2-A6 of White (1982) are satisfied, the quasi-ML estimator $\hat{\theta}'_{SNP_1} = (\hat{\eta}'_{SNP_1}, \hat{\varphi}_1)$ converges to $\theta'_{0*} = (\eta'_0, \varphi_1*)$, where φ_1* is the value of φ_1 that minimizes the Kullback-Leibler information criterion. The GH test is defined as

$$GH = (\hat{\eta}_{SNP_1} - \tilde{\eta}_{SNP_0})' \hat{S}(\tilde{\eta}_{SNP_0}, \hat{\theta}_{SNP_1})^{-1} (\hat{\eta}_{SNP_1} - \tilde{\eta}_{SNP_0}). \quad (3)$$

Details on the computation of the matrix $\hat{S}(\tilde{\eta}_{SNP_0}, \hat{\theta}_{SNP_1})$ can be found in Guastadisegni *at al.* (forthcoming). Under normality of the latent variable distribution, the GH test is asymptotically distributed as a χ^2_{2p}, where $2p$ are the degrees of freedom. To avoid the inversion of the matrix $\hat{S}(\tilde{\eta}_{SNP_0}, \hat{\theta}_{SNP_1})$ that is numerically unstable, we consider the following statistic

$$GH_T = (\hat{\eta}_{SNP_1} - \tilde{\eta}_{SNP_0})' (\hat{\eta}_{SNP_1} - \tilde{\eta}_{SNP_0}). \quad (4)$$

Under normality of the latent variable distribution, $GH_T \sim a\chi^2_b$, where $a = \frac{\sum_{l=1}^{d} \lambda_l^2}{\sum_{l=1}^{d} \lambda_l}$ and $b = \frac{(\sum_{l=1}^{d} \lambda_l)^2}{\sum_{l=1}^{d} \lambda_l^2}$, d is rank of $\hat{S}(\tilde{\eta}_{SNP_0}, \hat{\theta}_{SNP_1})$ and $\lambda_1, ..., \lambda_d$ are its non-zero eigenvalues.

4 Simulation study and results

The optimization of the SNP_1 model is achieved in R with direct maximization via the function "nlminb", that uses analytically computed gradient and Hessian matrix, while the SNP_0 model via the function "optim". We consider the following simulation conditions: number of items ($p = 4, 10, 20$), sample size ($n = 500, 1000$), 500 replications for each condition and $\alpha = 0.05$. Data are generated from a 2-PL model with the following latent variable distributions:

A $z \sim N(0,1)$
B $z \sim 0.7N(-1.5, 0.6) + 0.3N(1.5, 0.5)$, where z has an overall mean equal to -0.6 and variance equal to 2.217.
C $z \sim SN(\mu = 0, \sigma = 2.5, \lambda = 10)$, where z has mean 1.98 and variance 2.31.

Table 1 presents Type I error rates and power of the GH_T test for scenarios A, B and C. Overall, under scenario A, the GH_T test has good performance in terms

Table 1. *Type I error rates and power of the GH_T test for scenarios A, B, and C, $p = 4, 10, 20$, $n = 500, 1000$.*

		Type I error	Power	
p	n	**A**	**B**	**C**
4	500	0.016	0.796	0.03
	1000	0.086	0.92	0.234
10	500	0.018	1	0.388
	1000	0.044	1	0.59
20	500	0.056	0.986	0.744
	1000	0.06	1	0.918

of Type I error rates when the sample size is large and in general with many items. Under scenario B, the power of the GH_T test is high for most conditions. However, under scenario C, 4 and 10 items, the GH_T test has low power to detect non-normality of the latent variable distribution. It reaches a high power only with 20 items and large sample sizes. The low power of the test under scenario C can be due to the following reasons. First, the SNP_1 model does not approximate very well the skew-normal distributions (Irincheeva *et al.*, 2012). Second, the skew-normal distribution used in the simulations has a very high mean and this has a negative impact on the estimation of parameters.

References

GUASTADISEGNI, L., MOUSTAKI, I., VASDEKIS, V., & CAGNONE, S. Forthcoming. Detecting latent variable non-normality through the generalized Hausman test. *In: Quantitative Psychology: The 87th Annual Meeting of the Psychometric Society, Bologna, 2022.* Springer.

IRINCHEEVA, I., CANTONI, E., & GENTON, M. G. 2012. Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics*, **39**, 663–680.

LINDSAY, B. G. 1988. Composite likelihood methods. *Contemporary mathematics*, **80**, 221–239.

MA, Y., & GENTON, M. G. 2010. Explicit estimating equations for semiparametric generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 475–495.

WHITE, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.

QUANTIFYING VARIABLE IMPORTANCE IN CLUSTER ANALYSIS

Christian Hennig [1] and Keefe Murphy[2]

[1] Department of Statistical Science "Paolo Fortunati", University of Bologna, Italy (e-mail: christian.hennig@unibo.it)

[2] Hamilton Institute, Maynooth University, Ireland (e-mail: keefe.murphy@mu.ie)

ABSTRACT: We propose to measure the importance of variables when running a cluster analysis by measuring the similarity of a clustering using all variables with a clustering applying the same method leaving out one variable. If the resulting clustering is very similar, the left out variable does not have much impact. An alternative is to replace the variable by randomly permuted values. Beyond variable selection (on which we will not focus), variable importance measurement is useful for interpreting and understanding a clustering. Also we will use variable importance measurement to discuss whether clustering methods appropriately balance the impact of different variables in mixed type variables clustering

KEYWORDS: variable importance, adjusted rand index, permutation, mixed type variables clustering.

1 Introduction

The quantification of variable importance in cluster analysis is of interest in order to interpret and understand the impact of the variables on a clustering, and potentially also for variable selection.

Consider a data set of n observations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, where $x_{ij} \in \mathcal{X}_j$. $j = 1, \ldots, p$, where $\mathcal{X}_j$ is the sample space for variable j, with potentially different $\mathcal{X}_j$ for different j. Let $X_j = (x_{1j}, \ldots, x_{nj})$ denote variable j. As clusterings, we consider partitions $\mathcal{C} = (c_1, \ldots, c_n)$ of a data set with n observations, where $c_i \in \{1, \ldots, k\}$ indicates the cluster to which $\mathbf{x}_i$ belongs, with $k = \max \mathcal{C}$ the number of clusters. k is not necessarily known or fixed. Let C denote a general clustering (partitioning) method so that $C(\mathbf{X}) \in \{1, \ldots, k\}^n$.

2 Variable importance by leaving a variable out

In a case study regarding socioeconomic stratification based on mixed type (i.e., continuous, ordinal, categorical) variables, in order to assess the importance of the various variables for clustering, Hennig & Liao, 2013 re-ran their clusterings with each variable left out, and they computed the adjusted Rand index (ARI; Hubert & Arabie, 1985) between the clustering with one variable left out and the clustering based on the full data. The ARI takes the value of 1 if clusterings are identical and a value around 0 (that can in principle be negative) if clusterings behave like unrelated random draws of cluster labels; the closer to 1 the ARI, the more similar the clusterings.

Formally: Let $\mathbf{X}^{-j} = (\mathbf{x}_1^{-j}, \ldots, \mathbf{x}_n^{-j})$, where $\mathbf{x}_i^{-j} = \mathbf{x}_i$ with x_{ij} left out. Let $I_{C,\mathbf{X}}(j) = \text{ARI}(C(\mathbf{X}), C(\mathbf{X}^{-j}))$ be the inverse variable importance of j. The interpretation regarding variable importance is that if $I_{C,\mathbf{X}}(j)$ is large, i.e., close to 1, the variable importance is low (therefore "inverse"), because it means that leaving out variable j reproduces pretty much the same clustering. A low value of $I_{C,\mathbf{X}}(j)$ means that leaving out variable j changes the clustering a lot, i.e., X_j has a large impact.

This principle of measuring variable importance can be applied to general clustering methods, and in fact clusterings generated by different methods on the same data can be compared regarding the importance they give to the different variables. This can be particularly interesting when clustering mixed type variables data, as it is a known issue with methods for mixed type variables that they may balance the different variable types, particularly continuous and categorical variables, against each other systematically in different ways, arguably giving too much (or too little) influence to categorical variables in certain situations (Foss *et al.*, 2019).

It is important to note here that variable importance, measured in this way, applies to the empirical result of a clustering method. It can be informative not only about the "true" importance of the variables regarding any supposedly "true" clustering, but also about the way the different clustering methods treat the variables. The downside of this is that these two interpretations may be confounded with each other. This does not seem to be a problem with the proposed method in particular, but rather a general issue with defining and measuring variable importance in clustering. The user therefore needs to be careful when using variable importance measurements for variable selection. More generally, variable selection in clustering is a hard problem, because in general the clustering problem is not well defined, and various clusterings can be legitimate, for potentially different clustering aims, on the same data

set (Hennig, 2015). This means that there is no unique true set of relevant variables, rather the user's choice of involved variables determines the way the resulting clustering can be interpreted.

3 Variable importance by permutation

Breiman, 2001 proposed a scheme for measuring variable importance in random forests. The idea there was to replace a variable by a permutation of its values. This constitutes an alternative approach for measuring variable importance in clustering. For a permutation π on $\{1,\ldots,n\}$, let $\mathbf{X}^{j\pi} = (\mathbf{x}_1^{j\pi},\ldots,\mathbf{x}_n^{j\pi})$, where $\mathbf{x}_i^{j\pi} = \mathbf{x}_i$ except that X_j is replaced by $X_{j\pi} = (x_{\pi(1)j},\ldots,x_{\pi(n)j})$. As this depends on the specific permutation, it is advisable to run m random permutations (say $m = 100$) $\pi_1,\ldots,\pi_m$, and then average ARI-values over the permutations, i.e., define $I^*_{C,\mathbf{X}}(j) = \frac{1}{m}\sum_{h=1}^{m}\mathrm{ARI}(C(\mathbf{X}),C(\mathbf{X}^{j\pi_h}))$.

Both of these approaches (leave a variable out, "I", and permute its values, "I^*") have advantages and disadvantages. Advantages of I are:

- The approach is deterministic, fully reproducible, and computationally simpler.
- It is easy to think of a data set that has a variable left out as "realistic", whereas permuting values of variable X_j may lead to combinations with values of other variables that are unrealistic, due to potential dependence between variables. It may therefore be seen as irrelevant, in a real situation, what would be the effect of a permutation of the values.

Advantages of I^* are:

- Running the clustering method on $\mathbf{X}^{j\pi}$ is the same as running it on $\mathbf{X}$ in the sense that the variables are the same, whereas for I, C has to be run on a data set that has a variable fewer.
- We ran many simulations in which data were generated from Gaussian mixture models, with some variables intentionally generated as noise uninformative for clustering. The results show that I^* is clearly better at distinguishing informative from uninformative variables, i.e., I^*-values will be larger for the uninformative than for the informative variables with clearly larger probability than I-values, consistently over a fairly large number of simulation setups.

This indicates that I^* is preferable for variable selection and interpretation in terms of meaningful vs. noise variables, although it may not be preferable for

investigating the way different methods balance different variables. The most plausible explanation for the empirically superior performance of I^* is that for an informative variable it is worse to be permuted than to be left out, as permuting will replace good information with bad misinformation that can potentially (if a variable is clearly clustered on its own) actively indicate a wrong clustering. Therefore permutation makes more of a difference for variables with strong clustering information than leaving out the variable.

In the presentation we will show simulation results and examples and will discuss them in some detail.

References

BREIMAN, L. 2001. Random forests. *Machine Learning*, **45**, 5–32.

FOSS, ALEXANDER H., MARKATOU, MARIANTHI, & RAY, BONNIE. 2019. Distance Metrics and Clustering Methods for Mixed-type Data. *International Statistical Review*, **87**, 80–109.

HENNIG, C., & LIAO, T. F. 2013. Comparing latent class and dissimilarity based clustering for mixed type variables with application to social stratification. *Journal of the Royal Statistical Society, Series C*, **62**, 309–369.

HENNIG, CHRISTIAN. 2015. Clustering strategy and method selection. *Pages 703–730 of:* HENNIG, CHRISTIAN, MEILA, MARIAN, MURTAGH, FIONN, & ROCCI, ROBERTO (eds), *Handbook of Cluster Analysis*. CRC Press.

HUBERT, LAWRENCE, & ARABIE, PHIPPS. 1985. Comparing partitions. *Journal of Classification*, **2**, 193–218.

REAL-TIME DISCRIMINANT ANALYSIS IN THE PRESENCE OF LABEL AND MEASUREMENT NOISE

Mia Hubert[1], Iwein Vranckx[1], Jakob Raymaekers[2], Bart De Ketelaere[3] and Peter Rousseeuw[1]

[1] Section of Statistics and Data Science, Department of Mathematics, KU Leuven (e-mail: mia.hubert@kuleuven.be, peter@rousseeuw.net)

[2] Department of Quantitative Economics, Maastricht University, (e-mail: j.raymaekers@maastrichtuniversity.nl)

[3] Division of Mechatronics, Biostatistics and Sensors, (e-mail: bart.deketelaere@kuleuven.be)

ABSTRACT: Quadratic discriminant analysis (QDA) is a widely used classification technique. Based on a training dataset, each class in the data is characterized by an estimate of its center and shape, which can then be used to assign unseen observations to one of the classes. The traditional QDA rule relies on the empirical mean and covariance matrix. Unfortunately, these estimators are sensitive to label and measurement noise which often impairs the model's predictive ability. Robust estimators of location and scatter are resistant to this type of contamination. However, they have a prohibitive computational cost for large scale industrial experiments. We present a novel QDA method based on a real-time robust algorithm. We additionally integrate an anomaly detection step to classify the most atypical observations into a separate class of outliers. Finally, we introduce the classmap, a graphical display to identify label and measurement noise in the training data.

KEYWORDS: minimum covariance determinant, mislabeling, outliers, robust classification

A PROPOSAL TO EVALUATE THE SOLUTION OF FUZZY CLUSTERING ALGORITHMS

Carmela Iorio [1], Giuseppe Pandolfo[1] and Antonio D'Ambrosio[1]

[1] Department of Economics and Statistics, University of Naples Federico II, (e-mail: carmela.iorio@unina.it, giuseppe.pandolfo@unina.it, antdambr@unina.it)

ABSTRACT: When the aim is to evaluate the solution of a fuzzy clustering algorithm, the computation of the adjusted version of the Rand index requires converting the soft partitions to hard partitions. Furthermore, in comparing two fuzzy partitions from two different clustering methods, an external validation index should satisfy two desirable properties: *(i)* reflexivity, and *(ii)* a proper interpretation of correction for agreement due to chance. In this paper, we show an extension of the commonly used adjusted Rand index to fuzzy partitions based on normalized degree of concordance.

KEYWORDS: Cluster analysis, Cluster validity, External criteria, Adjusted Concordance Index.

1 Introduction

Cluster analysis is a data mining technique that groups units (or objects) into a finite set of clusters (or groups) based on a distance or a similarity. The purpose of clustering is to partition the objects into distinct groups so that observations within each cluster are similar to each other, while observations in different clusters are different from each other. Many clustering algorithms have been introduced In literature, and many of the methods do not produce a partition, but e.g.hierarchies, or posterior probabilities (e.g. model-based clustering). Furthermore, since groups can be formally seen as subsets of the entire data set, one possible classification of clustering methods can be done according to whether the subsets are crisp (hard) or fuzzy (soft). Hard clustering methods are based on classical set theory and restrict each object in the data set to belong to exactly one cluster. Soft clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership. In contrast to hard clustering, each object has a membership value in each cluster: the larger the value of the membership value for a given object with respect to a cluster, the larger the probability of that object being assigned

to that cluster. An extensive overview of cluster analysis can be found in Kaufman & Rousseeuw, 2005, Everitt *et al.*, 2011, Duran & Odell, 2013, Hennig & Meila, 2015. However, clustering is an unsupervised learning problem since the aim is to identify a structure in an unlabeled data set. As a consequence, an important issue in cluster analysis is the evaluation of clustering results. The procedure for evaluating the goodness of the results of a clustering algorithm is known as cluster validation. Generally, there are three approaches to assessing cluster validity involving internal, external, and relative criteria. Internal validation criteria use the information involving the data set used in the clustering process (e.g. Silhouette index). External validation criteria evaluate clustering results by comparing them to an externally known result. Relative validation criteria evaluate the clustering structure by comparing it to other clustering schemes, i.e. by varying different parameter values for the same algorithm. Several external validation criteria have been proposed in the literature to evaluate hard or soft clustering algorithms. Among them the most popular indexes are Rand Index proposed by Rand, 1971 and its corrected versions for fuzzy partitions (see e.g. Campello, 2007, Frigui *et al.*, 2007, Brouwer, 2009, Anderson *et al.*, 2010, Hüllermeier *et al.*, 2012). In this work, attention is put on external validation criteria to evaluate the goodness of fuzzy partitions.

2 The key idea

We think that, in comparing the partitions coming from two, different clustering methods, a good index to be used should satisfy at least two desirable properties: *(i)* reflexivity and *(ii)* a proper interpretation of correction for agreement due to chance. The problem with evaluating the solution of a fuzzy clustering algorithm with the original formulation of the Rand index (RI) is that it requires converting the soft partitions into hard partitions, thus losing information. As Meilă, 2007 and Morey & Agresti, 1984 pointed out, there are other known problems with RI. It approaches its upper limit as the number of groups increases; it is extremely sensitive to the number and size of groups considered in each partition as well as to the overall number of observations considered; the expected value of RI for two random partitions does not take a constant value. To overcome these drawbacks, Hubert & Arabie, 1985 has proposed an adjusted version of RI (ARI) assuming the generalized hypergeometric distribution as the randomness model. Besides the ARI, even the fuzzy generalizations of the RI proposed by Campello, 2007, Frigui *et al.*, 2007, Brouwer, 2009, and Anderson *et al.*, 2010 fail to satisfy reflexivity property and therefore cannot be considered a metric.

Since we are interested in comparing fuzzy partitions and ARI is still the most popular measure used for clustering comparison, we show an extension of ARI to fuzzy partitions. The proposed index, named Adjusted Concordance Index (ACI), is based on the fuzzy variant of the ARI proposed by Hüllermeier *et al.*, 2012. These authors based their proposal on the fuzzy equivalence relation and this allows us to rewrite every partition as a similarity matrix based on the normalized city block. Thus, the ACI is given by:

$$ACI = \frac{NDC - \overline{NDC}}{1 - \overline{NDC}},$$

where the normalized degree of concordance (NDC) is a direct generalization of the RI and $\overline{NDC}$, is the mean value of the NDC over all the permutations. Since, $NDC(\mathbf{P},\mathbf{Q}) = 1 - d(\mathbf{P},\mathbf{Q})$, where $\mathbf{P}$ and $\mathbf{Q}$ are two fuzzy partitions, the NDC is the only extension of the RI to the fuzzy partition which fulfills the reflexivity property that always guarantees that its maximum value is equal to one.
For further details and comments on ACI, the interested reader may refer to D'Ambrosio *et al.*, 2021.

3 Conclusion

To evaluate the fuzzy clustering results, the external validation criteria proposed in the literature fail two desiderata: reflexivity, and a proper expectation. To compare fuzzy clustering algorithms, the adjusted Rand index (ARI), is commonly used to measure agreement between partitions. Following similar reasoning to Hubert & Arabie, 1985, we have provided the adjusted version of the normalized degree of concordance (NDC) index defined by Hüllermeier *et al.*, 2012. We named it the adjusted concordance index (ACI). It normalizes the difference between NDC itself and the point estimate of its expected value. Since NDC is the only fuzzy extension of the Rand index that possess the reflexivity property, thus the resulting ACI is itself a reflexive index. In this regard, our proposal works with any raw fuzzy index, provided that the two above-mentioned desiderata are satisfied.

References

ANDERSON, D.T., BEZDEK, J.C., POPESCU, M., & KELLER, J.M. 2010. Comparing fuzzy, probabilistic, and possibilistic partitions. *Fuzzy Systems, IEEE Transactions on*, **18**(5), 906–918.

BROUWER, R.K. 2009. Extending the Rand, adjusted Rand and Jaccard indices to fuzzy partitions. *Journal of Intelligent Information Systems*, **32**(3), 213–235.

CAMPELLO, R. JGB. 2007. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, **28**(7), 833–841.

D'AMBROSIO, A., AMODIO, S., IORIO, C., PANDOLFO, G., & SICILIANO, R. 2021. Adjusted concordance index: an extensionl of the adjusted rand index to fuzzy partitions. *Journal of Classification*, **38**, 112–128.

DURAN, B.S., & ODELL, P.L. 2013. *Cluster analysis: a survey*. 2 edn. Heidelberg, Germany: Springer Science & Business Media.

EVERITT, B.S., LANDAU, S., LEESE, M., & STAHL, D. 2011. *Cluster analysis*. 5 edn. Chichester, UK: Wiley.

FOWLKES, E.B., & MALLOWS, C.L. 1983. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**(383), 553–569.

FRIGUI, H., HWANG, C., & RHEE, F.C.H. 2007. Clustering and aggregation of relational data with applications to image database categorization. *Pattern Recognition*, **40**(11), 3053–3068.

HENNIG, C., & MEILA, M. 2015. Cluster analysis: an overview. *Chap. 1, pages 1–20 of:* HENNIG, CHRISTIAN, MEILA, MARINA, MURTAGH, FIONN, & ROCCI, ROBERTO (eds), *Handbook of cluster analysis*. Boca Raton, FL: CRC Press.

HUBERT, L., & ARABIE, P. 1985. Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

HÜLLERMEIER, E., RIFQI, M., HENZGEN, S., & SENGE, R. 2012. Comparing fuzzy partitions: A generalization of the Rand index and related measures. *Fuzzy Systems, IEEE Transactions on*, **20**(3), 546–556.

KAUFMAN, L., & ROUSSEEUW, P.J. 2005. *Finding groups in data: an introduction to cluster analysis*. 2 edn. Hoboken, NJ: John Wiley & Sons.

MEILĂ, M. 2007. Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, **98**(5), 873–895.

MOREY, L.C., & AGRESTI, A. 1984. The measurement of classification agreement: an adjustment to the Rand statistic for chance agreement. *Educational and Psychological Measurement*, **44**(1), 33–37.

RAND, W.M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

A Fused-type Elastic Net Gaussian Graphical Model for Paired Data

Aazm Kheyri [1], Andriette Bekker [1] and Mohammad Arashi [12]

[1] Department of Statistics, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, South Africa, (e-mail: azam.kheyri@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: In many scientific and practical domains, it is common to have multiple groups of observations that share the same set of variables but are not independent and identically distributed. Traditional approaches to learning graphical models from such data usually assume that the groups correspond to different populations, which may not hold in many cases. To overcome this limitation, we propose a fused-type graphical elastic net for joint learning of graphical models from two dependent groups. The proposed method incorporates a fused-type penalty function that captures the shared and distinct network structures between the two groups while enforcing symmetrical constraints. We use elastic net regularization to balance the sparsity and stability of the estimated network.

Keywords: Elastic net penalty; Gaussian graphical model; Joint graphical model; Network

1 Literature Review

The previous research has highlighted the significance of graphical models in statistics and machine learning, offering a powerful tool to model complex relationships among variables based on some local relation between them. Undirected graphical models, specifically the Gaussian graphical model, have attracted attention for their ability to represent conditional dependencies between variables. This graphical model associates a multivariate Gaussian random vector with a graph, where the Markov property captures dependencies through the precision matrix. The relevance of Gaussian graphical models in diverse applications has been emphasized in the study conducted by Maathuis *et al.* (2019). In high-dimensional scenarios, estimating the precision matrix of a Gaussian graphical model is a fundamental and challenging task. Traditional statistical methods, e.g. likelihood estimation, are often impractical,

leading to the emergence of effective approaches based on penalized likelihood estimation. The graphical lasso, independently studied by Yuan & Lin (2007), Banerjee *et al.* (2008), and Friedman *et al.* (2008), incorporates an $L1$ penalty term into the log-likelihood function, promoting sparsity in precision matrix estimation. In situations where obtaining accurate representations of high-dimensional precision matrices is crucial, ridge regularization has been employed even without significant sparsity. Ridge regularization adds a Frobenius penalty term to the log-likelihood function, as explored recently by van Wieringen (2019). Recent studies, including Kovács *et al.* (2021) and Bernardini *et al.* (2022), have introduced the elastic net graphical model, which combines the graphical lasso and ridge penalties. The elastic net penalty strikes a balance between sparsity and precision matrix estimation, offering a versatile framework for graphical modelling and enhancing the accuracy of estimates.

In addition to the traditional Gaussian graphical model, colored graphical models introduced by Højsgaard & Lauritzen (2008) impose symmetry conditions on the covariance matrix using permutations. While prior research has mainly focused on independent groups with disconnected networks, the paired data analysis considers scenarios where two groups share variables and exhibit dependence. This analysis allows for the existence of symmetries across the groups. Building on prior work, our study extends the model proposed by Ranciati *et al.* (2020) by incorporating graphical elastic net models and symmetries across two dependent groups with a fused-type penalty function.

In the following section, we probe into the methodology of our proposed symmetric graphical elastic net method after providing a concise introduction to the Gaussian graphical model, graphical lasso, and symmetric graphical lasso.

2 Methodology

Consider the Gaussian graphical model $(Z_{\mathcal{V}}, \mathcal{G})$, where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote an undirected graph and $Z \sim \mathcal{N}_p(\mu, \mathbf{\Sigma})$ follows a p-variate normal distribution. The maximum likelihood estimation (ML) is a commonly used approach to estimate the precision matrix ($\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$). The ML method is formulated as

$$\hat{\mathbf{\Omega}} = \arg\max_{\mathbf{\Omega} \succ 0} \{\log\det(\mathbf{\Omega}) - tr(\boldsymbol{S}\Omega)\}, \tag{1}$$

where $\boldsymbol{S}$ is the sample covariance matrix based on the observed independent and identically distributed random with sample size n of $Z_{\mathcal{V}}$. Since the solution to the optimization problem (1) does not typically exhibit sparsity and may

not have a solution in high-dimensional cases to address these limitations, the graphical lasso method was introduced as follows:

$$\hat{\mathbf{\Omega}}_{glasso} = \arg\max_{\mathbf{\Omega}\succ 0} \{\log\det(\mathbf{\Omega}) - tr(\boldsymbol{S}\mathbf{\Omega}) - \rho||\mathbf{\Omega}||_1\}, \tag{2}$$

where $\rho > 0$ is a tuning parameter, and the L_1 norm, $||\cdot||_1$ is the sum of the absolute values of the elements.

Ranciati *et al.* (2020) considered the estimation of the precision matrix for paired data and order to promote sparsity in the graph structure and encourage similarity between the two dependent groups of data. The penalty encourages equality between the concentration values of relevant subgraphs. They proposed the symmetric graphical lasso estimator, denoted as sgl, which is obtained by solving the following optimization problem:

$$\hat{\mathbf{\Omega}}_{sgl} = \arg\min_{\mathbf{\Omega}} \{-\log\det(\mathbf{\Omega}) + \mathrm{tr}(S\mathbf{\Omega}) + \lambda_1\|\mathbf{\Omega}\|_1 + \lambda_2\|\mathbf{\Omega}_{11} - \mathbf{\Omega}_{22}\|_1\} \tag{3}$$

where λ_1 and λ_2 are non-negative regularization parameters controlling the amount of sparsity. The precision matrix $\mathbf{\Omega}$ is partitioned into four matrices, with $\mathbf{\Omega}_{11}$ and $\mathbf{\Omega}_{22}$ representing the diagonal submatrices. The penalty term L_1 encourages sparsity in the estimated precision matrix $\mathbf{\Omega}$, and the fused penalty term encourages the elements of $\mathbf{\Omega}_{11}$ to be identical to the corresponding elements of $\mathbf{\Omega}_{22}$.

This paper proposes modifying the optimization problem (3) by replacing the lasso penalty with an elastic net penalty. The elastic net penalty balances sparsity and accuracy in estimating the precision matrix. We then develop an alternating directions method of multipliers (ADMM) algorithm to solve the newly proposed optimization problem. The corresponding penalty term is as follows:

$$\alpha\lambda_1||\mathbf{\Omega}||_1 + \frac{(1-\alpha)\lambda_1}{2}||\mathbf{\Omega}||_F^2 + \lambda_2||\mathbf{\Omega}_{11} - \mathbf{\Omega}_{22}||_1, \tag{4}$$

where $\alpha \in [0,1]$, λ_1 and λ_2 are non-negative tuning parameters that control the sparsity and regularization strength and the Frobenius norm, $||\cdot||_F$ is the square root of the squared values of the elements.

3 Conclusion

In conclusion, the fused-type graphical elastic net method offers a novel approach for jointly learning graphical models from dependent groups. The proposed method balances sparsity and stability in estimated networks by incorporating a fused penalty function and elastic net regularization.

acknowledgments

This work was based upon research supported partly by the National Research Foundation (NRF) of South Africa, Ref.: RA211204653274 grant No. 151035. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF. Mohammad Arashi's work is based on the research supported in part by the Iran National Science Foundation (INSF) grant No. 4015320

References

BANERJEE, ONUREENA, EL GHAOUI, LAURENT, & D'ASPREMONT, ALEXANDRE. 2008. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, **9**, 485–516.

BERNARDINI, DAVIDE, PATERLINI, SANDRA, & TAUFER, EMANUELE. 2022. New estimation approaches for graphical models with elastic net penalty. *Econometrics and Statistics*.

FRIEDMAN, JEROME, HASTIE, TREVOR, & TIBSHIRANI, ROBERT. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

HØJSGAARD, SØREN, & LAURITZEN, STEFFEN L. 2008. Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(5), 1005–1027.

KOVÁCS, SOLT, RUCKSTUHL, TOBIAS, OBRIST, HELENA, & BÜHLMANN, PETER. 2021. Graphical Elastic Net and Target Matrices: Fast Algorithms and Software for Sparse Precision Matrix Estimation. *arXiv preprint arXiv:2101.02148*.

MAATHUIS, MARLOES H, DRTON, MATHIAS, LAURITZEN, STEFFEN, & WAINWRIGHT, MARTIN. 2019. Handbook of Graphical Models.

RANCIATI, SAVERIO, ROVERATO, ALBERTO, & LUATI, ALESSANDRA. 2020. Fused graphical lasso for brain networks with symmetries. *arXiv preprint arXiv:2005.11785*.

VAN WIERINGEN, WESSEL N. 2019. The generalized ridge estimator of the inverse covariance matrix. *Journal of Computational and Graphical Statistics*, **28**(4), 932–942.

YUAN, MING, & LIN, YI. 2007. Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**(1), 19–35.

COMPLETE RECORDS OVER INDEPENDENT FGM SEQUENCES

Amir Khorrami Chokami [1]

[1] ESOMAS Department, University of Torino and Collegio Carlo Alberto (e-mail: amir.khorramichokami@unito.it)

ABSTRACT: Records are defined as variables greater than all the preceding ones in a sequence. The stochastic behavior of subsequent records over sequences of independent and identically distributed random variables is well known. However, the extension to the multivariate framework is an extremely difficult task. In this work, we study the case of bivariate records over sequences of random vectors (rv) where the dependence among their components is described by the Farlie-Gumbel-Morgenstern (FGM) family of distributions.

KEYWORDS: complete records, standard max stable distribution, FGM copula.

1 Introduction

The study of multivariate maxima of rv is a challenging topic (see Resnick, 1987, Leonetti & Khorrami Chokami, 2022 among others), but records furnish a new way to tackle the problem as they give information on how often and to which extent maxima change. The theory on records is well developed in the case of independent and identically distributed (iid) sequences of random variables (see Galambos, 1987 and Falk *et al.*, 2018a). However, as soon as we relax the iid assumption to better reflect real-world data, the problem becomes immediately too difficult (we cite Falk *et al.*, 2020 for a study on univariate stationary Gaussian sequences). Multivariate records are an appealing topic of research. In $\mathbb{R}^d$, various definitions of records are possible. Here, we consider the so-called complete record (see Falk *et al.*, 2018b) and consider operations on vectors to be made componentwise. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be iid rv: $\boldsymbol{X}_n$ is a *complete record* (CR) if $\boldsymbol{X}_n > \boldsymbol{M}_{n-1} = \max(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{n-1})$ and the appearance of a CR at time n is indicated as $R_n := \mathbb{1}(\boldsymbol{X}_n > \boldsymbol{M}_{n-1})$.

This paper investigates the difficult problem of describing the appearance of CRs and their distribution, under the following hypothesis: we know that a vector is a CR, but we do not know which one. It is still an open problem to find such results in the case of iid sequences of rv with a general copula. Here,

we consider a sequence $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots \in \mathbb{R}^2$ of *standard max-stable* rv (*i.e.* with Negative-Exp(1) margins and such that $M_n \stackrel{d}{=} \boldsymbol{\eta}_1/n$) and FGM copula:

$$F(x,y) = \mathrm{e}^{(x+y)}\left(1+\lambda\left(1-\mathrm{e}^{x}\right)\left(1-\mathrm{e}^{y}\right)\right), \quad x,y \le 0 \text{ and } |\lambda| \le 1. \tag{1}$$

The usefulness of this copula lies in its manageable structure and intuitive interpretation of the parameter λ to describe dependence, which make the FGM distribution widely used in capital-allocation applications and in problems involving order statistics and their concomitants. A complete description of this copula is in Hashorva & Hüsler, 1999.

2 Complete Records

Theorem 1. *Let $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots$ be a sequence of bivariate standard max-stable rv with FGM copula. Then, the probability of appearance of a CR is*

$$\mathsf{P}(R_n = 1) = \frac{1}{n^2}\left(1+\lambda\left(\frac{n-1}{n+1}\right)^2\left(1+\frac{(n+1)^2+\lambda(n-2)^2}{(2n-1)^2}\right)\right) \tag{2}$$

for $n \in \mathbb{N}$ and the distribution of a rv given that it is a CR is

$$\mathsf{P}(\boldsymbol{\eta} \le \boldsymbol{z} \mid R_n = 1) = \frac{e^{n(z_1+z_2)}}{\mathsf{P}(R_n=1)}\left(\frac{1}{n^2}+\lambda\prod_{i=1}^{2}\left(\frac{2}{n+1}e^{z_i}-\frac{1}{n}\right)+\lambda\prod_{i=1}^{2}\left(\frac{1}{n}-\frac{e^{(n-1)z_i}}{2n-1}\right)+\right.$$
$$\left.+\lambda^2\prod_{i=1}^{2}\left(\frac{1}{n+1}e^{z_i}-\frac{1}{n}-\frac{e^{nz_i}}{n}+\frac{e^{(n-1)z_i}}{2n-1}\right)\right), \qquad \boldsymbol{z} \le 0. \tag{3}$$

Proof. Denote with $\eta^{(i)}$ the i-th component of $\boldsymbol{\eta}$. We firstly compute

$$\mathsf{P}(R_n=1) = \mathsf{P}(\boldsymbol{\eta}_n > \boldsymbol{M}_{n-1}) = \mathsf{P}\left(\boldsymbol{\eta}_n > \frac{\boldsymbol{\eta}_1}{n-1}\right) = \mathsf{P}(\boldsymbol{\eta}_1 < (n-1)\boldsymbol{\eta}_n)$$
$$= \int_{(-\infty,0]^2} \mathsf{P}\left(\eta_1^{(1)} < (n-1)x, \eta_2^{(1)} < (n-1)y \mid \boldsymbol{\eta}_n = (x,y)\right) f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$
$$= \int_{(-\infty,0]^2} \mathsf{P}\left(\eta_1^{(1)} < (n-1)x, \eta_2^{(1)} < (n-1)y\right) f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

The last equality follows by the independence assumption. Define

$$I(z_1,z_2) = \int_{-\infty}^{z_2}\int_{-\infty}^{z_1} \mathsf{P}\left(\eta_1^{(1)} < (n-1)x, \eta_2^{(1)} < (n-1)y\right) f(x,y)\,\mathrm{d}x\,\mathrm{d}y =$$
$$= \int_{-\infty}^{z_2}\int_{-\infty}^{z_1} \mathrm{e}^{n(x+y)}\left(1+\lambda\left(1-\mathrm{e}^{(n-1)x}\right)\left(1-\mathrm{e}^{(n-1)y}\right)\right)\left(1+\lambda\left(2\mathrm{e}^{x}-1\right)\left(2\mathrm{e}^{y}-1\right)\right)\mathrm{d}x\,\mathrm{d}y,$$

and note that $I(0,0) = \mathsf{P}(R_n = 1)$. After computations, we obtain

$$I(z_1,z_2) = \mathrm{e}^{n(z_1+z_2)}\left(\frac{1}{n^2} + \lambda\prod_{i=1}^{2}\left(\frac{2}{n+1}\mathrm{e}^{z_i} - \frac{1}{n}\right) + \lambda\prod_{i=1}^{2}\left(\frac{1}{n} - \frac{\mathrm{e}^{(n-1)z_i}}{2n-1}\right) + \right.$$
$$\left. + \lambda^2\prod_{i=1}^{2}\left(\frac{1}{n+1}\mathrm{e}^{z_i} - \frac{1}{n} - \frac{\mathrm{e}^{nz_i}}{n} + \frac{\mathrm{e}^{(n-1)z_i}}{2n-1}\right)\right)$$

and we have

$$I(0,0) = \frac{1}{n^2}\left(1 + \lambda\left(\frac{n-1}{n+1}\right)^2\left(1 + \frac{(n+1)^2 + \lambda(n-2)^2}{(2n-1)^2}\right)\right) = \mathsf{P}(R_n = 1)\,.$$

Equation (3) follows by noticing that, for $\boldsymbol{z} \leq \mathbf{0}$,

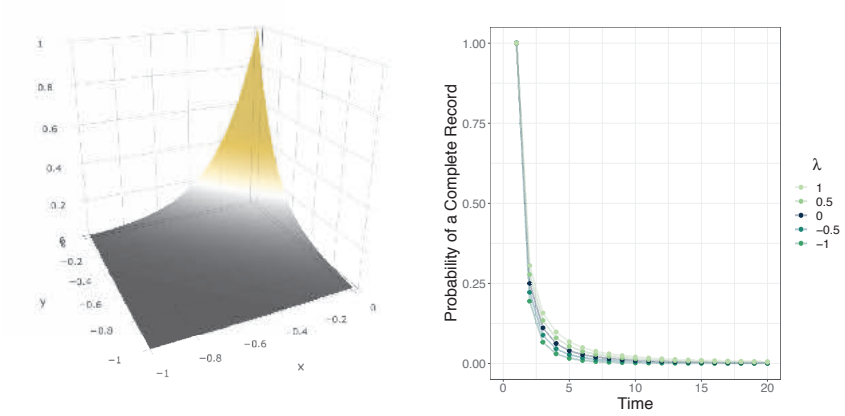
$$\mathsf{P}(\boldsymbol{\eta}_n \leq \boldsymbol{z} \mid R_n = 1) = \frac{\mathsf{P}(\boldsymbol{\eta}_n \leq \boldsymbol{z}, R_n = 1)}{\mathsf{P}(R_n = 1)} = \frac{\mathsf{P}(\boldsymbol{\eta}_n \leq \boldsymbol{z}, \eta_1 < (n-1)\eta_n)}{\mathsf{P}(R_n = 1)}$$
$$= \frac{1}{\mathsf{P}(R_n = 1)}\int_{(-\infty,\boldsymbol{z}]} \mathsf{P}\left(\eta_1^{(1)} < (n-1)x, \eta_2^{(1)} < (n-1)y \mid \boldsymbol{\eta}_n = (x,y)\right) f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

The thesis follows by noticing that $\mathsf{P}(\boldsymbol{\eta}_n \leq \boldsymbol{z} \mid R_n = 1) = I(0,0)^{-1} I(z_1,z_2)$. □

Figure 1a represents an example of the cdf of a CR at time 4 given by Equation (3), when the parameter λ is set to 0.8, while Figure 1b shows the decay of the appearance of a CR as n increases (from Equation (2)), for various choices of λ. Note that $\lambda = 0$ indicates independence of the components of $\boldsymbol{\eta}$.

Remark 1. *Let $N = \sum_{n=2}^{\infty} R_n$ be the number of records after the first vector (which is the first record by definition). From Equation* (2)*, it holds that $\mathsf{E}[N] = \sum_{n=2}^{\infty}\mathsf{P}(R_n = 1) < \infty$, which implies by the first Borel-Cantelli lemma that $\mathsf{P}(R_n = 1 i.o) = 0$, that is a finite number of records. This is coherent with Theorem 5.3 in Goldie & Resnick, 1989.*

To conclude, this work tackles the problem of studying CRs over iid bivariate sequences of standard max-stable rv with FGM copula, under the hypothesis of not knowing the position of the CRs in their sequence. This approach is proven to furnish handy results and links with the Extreme Value Theory (see Falk *et al.*, 2018b and Falk *et al.*, 2020). We highlight that Equation (2) is independent of the chosen marginal distribution function of the considered rv (say F_X), provided that it is continuous, as $\eta = \log(F_X(X))$ in distribution. However, the distribution of a CR does depend on the marginal distribution. The extension to CR on sequences with a general copula and unfixed continuous margins is an ongoing project of the author.



(a) *Bivariate cdf of a CR.* (b) *Decay of appearance of a CR.*

Figure 1: Panel (a) shows the bivariate cdf of $\boldsymbol{\eta}_4 \mid R_4 = 1$, with $\lambda = 0.8$. Panel (b) shows the decay of Equation (2), for different values of λ.

References

FALK, M., KHORRAMI CHOKAMI, A., & PADOAN, S. A. 2018a. On multivariate records from random vectors with independent components. *Journal of Applied Probability*, **55**(1), 43–53.

FALK, M., KHORRAMI CHOKAMI, A., & PADOAN, S. A. 2018b. Some results on joint record events. *Statistics and Probability Letters*, **135**, 11 – 19.

FALK, M., KHORRAMI CHOKAMI, A., & PADOAN, S. A. 2020. Records for time-dependent stationary Gaussian sequences. *Journal of Applied Probability*, **57**(03), 78–96.

GALAMBOS, J. 1987. *The Asymptotic Theory of Extreme Order Statistics*. 2 edn. Malabar: Krieger.

GOLDIE, CHARLES M., & RESNICK, SIDNEY I. 1989. Records in a partially ordered set. *Ann. Probab.*, **17**(2), 678–699.

HASHORVA, E, & HÜSLER, J. 1999. Extreme Values in FGM Random Sequences. *Journal of Multivariate Analysis*, **68**(2), 212–225.

LEONETTI, P., & KHORRAMI CHOKAMI, A. 2022. The maximum domain of attraction of multivariate extreme value distributions is small. *Electronic Communications in Probability*, **27**, 1 – 8.

RESNICK, SIDNEY I. 1987. *Extreme Values, Regular Variation, and Point Processes*. Applied Probability, vol. 4. New York: Springer.

NEW TOUR METHODS FOR VISUALIZING HIGH-DIMENSIONAL DATA

Ursula Laa [1] and Dianne Cook[2]

[1] University of Natural Resources and Life Sciences, Vienna, Institute of Statistics, (e-mail: ursula.laa@boku.ac.at)

[2] Department of Econometrics and Business Statistics, Monash University, (e-mail: dicook@monash.edu)

ABSTRACT: Tour methods visualize high-dimensional spaces as animated sequences of low-dimensional projections. Viewing the projected data allows us to uncover and understand shapes and patterns in such high-dimensional spaces. Typically we create this animation by first selecting a target plane, and then we interpolate to gradually move to the selected target. While several methods for target selection are available in the R package tourr, it currently only implements geodesic interpolation. Here we present recent developments in tour methods. We first describe a manual user-guided control for target selection, that also includes the interactive selection of sections in the context of a slice tour. We then present a new interpolation method for frame-to-frame transitions instead of plane-to-plane, important for projection pursuit applications.

KEYWORDS: data visualization, grand tour, dynamic graphics, projection pursuit

1 Introduction

The grand tour (Asimov, 1985) visualizes multivariate data distributions as animated sequences of interpolated low-dimensional projected views, and by following such an animation we can build intuition about a data distribution in a high-dimensional space. We may discover patterns of interest, for example, clustering, or we can detect outlying points. For a summary of the current state-of-the-art on tour methods and applications see Lee *et al.*, 2022.

Constructing the animation typically iterates between two steps: first, we select a target plane onto which the data should be projected (*target selection*), and then we compute the interpolated path between the current viewing plane and the selected target (*interpolation algorithm*).

Examples of target selection include a random selection (the *grand tour*, which provides an overview of the distribution across the full space), and a selection that optimizes a projection pursuit index (the *guided tour* (Cook *et al.*, 1995), which moves towards more "interesting" views of the data). The latter is important when patterns of interest may be hidden and only visible in a small

part of a much larger data space. These as well as additional approaches are implemented in the *tourr* R package (Wickham *et al.*, 2011).

The interpolation algorithm should then find a path from one target plane to the next, such that we can view the projected data as a smooth animation while gradually changing the viewing angle. Importantly each intermediate step also needs to be defined by an orthonormal projection matrix such that the data is not distorted in the visualization. The preferred approach is typically a geodesic interpolation (Buja *et al.*, 2005) which finds the shortest path between two planes, independent of the orientation of the target frame. This has the advantage that any within-plane rotation is avoided during the interpolation, but can limit applications of the guided tour when the considered projection pursuit index is not invariant to rotation within the plane (Laa & Cook, 2020).

2 Manual tour in Mathematica

With a manual tour, the user can alter the contribution of a selected variable to the target projection. This is, in particular, useful to interpret patterns found in one projection, for example through projection pursuit, to understand the sensitivity of the pattern to the input variables.

The previous approach to the manual tour (Cook & Buja, 1997; Spyrison & Cook, 2020) was overly complicated since it requires the construction of a manipulation space as an intermediate step. Here we present a simpler approach described in Laa *et al.*, 2023. The new method uses the interactive graphics interface available in Mathematica, for illustration, we show a screenshot in Fig. 1. We can manually change the projection by dragging one of the variables in an axes display of the projection matrix. The main part of the display shows the projected data (or slices of it, defined according to Laa *et al.*, 2020), tracking changes to the projection matrix while ensuring its orthonormality.

In the presentation, we will show an example of how the new approach can be used for a detailed inspection of a fitted classification model, by comparing the decision boundaries generated by two different models.

3 Alternative interpolation methods

We need an alternative to geodesic interpolation for a guided tour that is optimizing a projection pursuit index that is not rotation invariant. This situation is illustrated in Fig. 2: the data is simulated to have a functional dependence between two of the variables (V5 and V6). We define an index that captures functional dependence in the projection: we compute the residuals

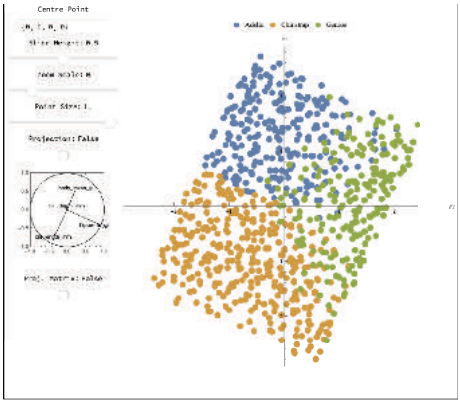


Figure 1. *Screenshot of the interactive Mathematica manual tour interface.*

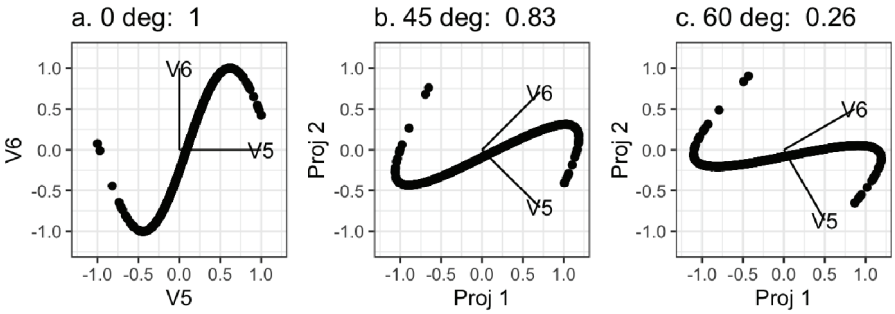


Figure 2. *Simple spline index computed on within-plane rotations of the same projection, resulting in very different index values.*

of a splines model using values along the projected x- and y-axis as independent and dependent variables, and normalize by the variance along the y-direction (Grimm, 2016). The index value changes dramatically when rotating within the plane: on the left we see the index taking its maximum value of 1, while on the right the index value has dropped to 0.26.

To offer an alternative interpolation method in those settings we have implemented a frame-to-frame interpolation based on Givens rotations, as suggested in Buja *et al.*, 2005. The algorithm is available through the R package *woylier*. The presentation will give a brief overview of the algorithm and its implementation. We then show how it can be used to improve the results of a guided tour for the example of exchange rate data.

References

ASIMOV, DANIEL. 1985. The Grand Tour: A Tool for Viewing Multidimensional Data. *SIAM Journal of Scientific and Statistical Computing*, **6**(1), 128–143.

BUJA, ANDREAS, COOK, DIANNE, ASIMOV, DANIEL, & HURLEY, CATHERINE. 2005. Computational Methods for High-Dimensional Rotations in Data Visualization. *Chap. 14, pages 391–413 of:* RAO, C R, WEGMAN, E J, & SOLKA, J L (eds), *Data Mining and Data Visualization*. Handbook of Statistics, vol. 24. Elsevier.

COOK, DIANNE, & BUJA, ANDREAS. 1997. Manual Controls for High-Dimensional Data Projections. *Journal of Computational and Graphical Statistics*, **6**(4), 464–480.

COOK, DIANNE, BUJA, ANDREAS, CABRERA, JAVIER, & HURLEY, CATHERINE. 1995. Grand Tour and Projection Pursuit. *Journal of Computational and Graphical Statistics*, **4**(3), 155–172.

GRIMM, KATRIN. 2016. *Kennzahlenbasierte Grafikauswahl*. doctoral thesis, Universität Augsburg.

LAA, URSULA, & COOK, DIANNE. 2020. Using tours to visually investigate properties of new projection pursuit indexes with application to problems in physics. *Comput Stat 35*, 1171–1205.

LAA, URSULA, COOK, DIANNE, & VALENCIA, GERMAN. 2020. A Slice Tour for Finding Hollowness in High-Dimensional Data. *Journal of Computational and Graphical Statistics*, **29**(3), 681–687.

LAA, URSULA, AUMANN, ALEX, COOK, DIANNE, & VALENCIA, GERMAN. 2023. New and simplified manual controls for projection and slice tours, with application to exploring classification boundaries in high dimensions. *Journal of Computational and Graphical Statistics*, to appear.

LEE, STUART, COOK, DIANNE, DA SILVA, NATALIA, LAA, URSULA, SPYRISON, NICHOLAS, WANG, EARO, & ZHANG, H. SHERRY. 2022. The state-of-the-art on tours for dynamic visualization of high-dimensional data. *WIREs Computational Statistics*, **14**(4), e1573.

SPYRISON, NICHOLAS, & COOK, DIANNE. 2020. spinifex: an R Package for Creating a Manual Tour of Low-dimensional Projections of Multivariate Data. *The R Journal*, **12**(1), 243.

WICKHAM, HADLEY, COOK, DIANNE, HOFMANN, HEIKE, & BUJA, ANDREAS. 2011. tourr: An R Package for Exploring Multivariate Data with Projections. *Journal of Statistical Software*, **40**(2), 1–18.

BAYESIAN AGGREGATION OF CROWD JUDGMENTS FOR QUANTITATIVE FACT CHECKING

M. Lambardi di San Miniato [1] , M. Battauz[1] , R. Bellio[1] and P. Vidoni[1]

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

ABSTRACT: Political fact-checking can be carried out by crowd workers, provided they are supervised by experts. We propose a Bayesian latent variable ordinal probit model for truthfulness rating data, to estimate workers' reliability, weigh in their contributions, and surrogate expert judgments. This is a notable example of aggregation function of an implicit type. This method may be used to dynamically assign workers to new tasks, as illustrated with an analysis of PolitiFact data.

KEYWORDS: Bayesian statistics, judgment aggregation, ordered probit.

1 Introduction

Fact-checking is about assessing the truthfulness of public statements to combat misinformation and improve debates. However, expert fact-checkers are few, while crowd workers are readily available but potentially biased. There is a stream of scientific research about how to surrogate expert judgements by means of workers, after some suitable calibration; see for example Roitero *et al.*, 2021. Latent traits of statements and workers are at stake, like truthfulness and political orientation. Methods from Item Response Theory (see for example Bartholomew *et al.*, 2011) can be adapted to this end. Here we adopt a Bayesian approach, which is suitable for the task.

We propose an ordinal probit model for quantitative fact-checking. An aggregation function is involved, which mimics expert judgments via the wisdom of crowds (Roitero *et al.*, 2021). The truthfulness of statements, even when encoded as an ordinal variable, is often treated as numeric. This allows to summarize ratings across workers by means of a simple average, but by using a generative model there is room for improvement. We argue that, as far as the aggregation function needs not be explicit, the Bayesian inferential approach always provides one, namely, the posterior distribution of expert judgments conditional to workers'. A different proposal, with some similarities, exists in the literature (Nguyen *et al.*, 2018).

2 Model

Let $i = 1,\dots,n$ and $j = 1,\dots,m$ be two indices to identify statements and workers, respectively. The truthfulness of statement i is rated as $Z_i = 1,\dots,k$ by a single expert and as $W_{ij} = 1,\dots,k$ by worker $j \in C_i$, with C_i a subset of workers that evaluate statement i. The aim is to predict Z_i through $(W_{ij})_{j\in C_i}$.

As typical in ordinal regression models, we think of Z_i as the observed discretization of a latent numeric variable Z_i^*, which is defined as

$$Z_i^* = \sigma_\xi \xi_i + \varepsilon_i\,, \quad \xi_i, \varepsilon_i \sim \mathcal{N}(0,1)\,.$$

Here, ε_i is a noise term, ξ_i is the truthfulness of the i-th statement and $\sigma_\xi > 0$ is a signal strength parameter. Analogously, we think of W_{ij} as the observed discretization of a latent numeric variable W_{ij}^*, which is defined as

$$W_{ij}^* = \alpha_j + \beta_j \xi_i + \eta_{ij}\,, \quad \alpha_j \sim \mathcal{N}(0,\sigma_\alpha^2)\,, \quad \beta_j \sim \mathcal{N}(0,\sigma_\beta^2)\,, \quad \eta_{ij} \sim \mathcal{N}(0,1)\,.$$

Here, η_{ij} is a noise term, while α_j and β_j are worker-specific parameters that affect their judging behavior. The worker-specific parameters α_j and β_j account for correlation within workers. All the terms $\xi_i, \varepsilon_i, \alpha_j, \beta_j, \eta_{ij}$ are assumed independent. Lastly, we define two sets of thresholds $(\gamma_h)_{h=0}^k$ and $(\delta_l)_{l=0}^k$ constrained as $\gamma_0 = \delta_0 = -\infty$, $\gamma_h < \gamma_{h+1}$, $\delta_l < \delta_{l+1}$, $\gamma_k = \delta_k = +\infty$, such that

$$\gamma_{h-1} < W_{ij}^* \le \gamma_h \iff W_{ij} = h\,, \quad \delta_{l-1} < Z_i^* \le \delta_l \iff Z_i = l\,.$$

Probit models are implied for Z_i and W_{ij}. As an original proposal, parameters α_j and β_j allow to represent the alignment with the experts. The model specification is then completed by assigning weakly informative priors to scale parameters and uniform priors on thresholds.

3 Example

We analyse a publicly available dataset (Roitero *et al.*, 2020), which includes expert ratings obtained from PolitiFact. Data relate to $m = 100$ workers and $n = 62$ public statements on COVID-19. The truthfulness ratings Z_i and W_{ij} have $k = 6$ levels, labeled as: "pants-on-fire", "false", "mostly-false", "half-true", "mostly-true" or "true". There were eight statements per worker and ten workers per statement, but two *gold* statements were rated by all the workers for control purposes. Gold statements have either $Z_i = 1$ or $Z_i = k$, while all

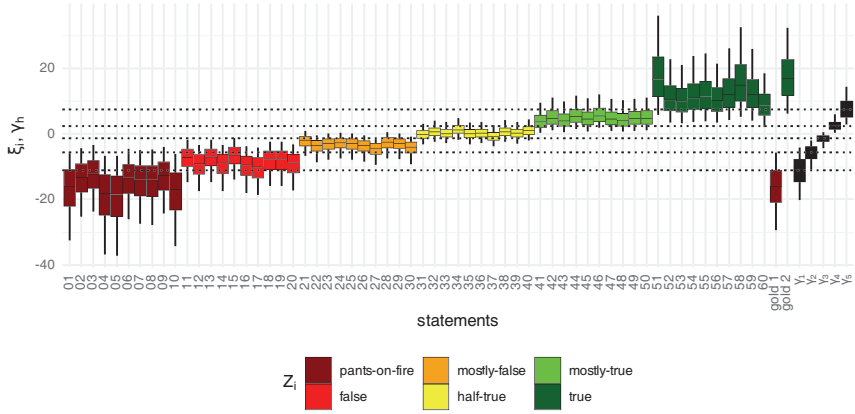


Figure 1. *Posterior percentiles (5%, 25%, 50%, 75% and 95% level) of truthfulness* ξ_i *and thresholds* γ_h.

the other statements administered to each worker cover different Z values. We estimate the model via the `R` interface to the `Stan` probabilistic programming language (Stan Development Team, 2023).

Figure 1 shows the posterior distribution of ξ_i, along with the thresholds γ_h. Were it only for the model on Z_i, the boxplots should all be similar, but the model on W_{ij} complements that information, so that there can be gradients of ξ within levels of Z.

Figure 2 summarizes the inferential results for α and β, which are affected by political orientation. Liberals tend to be more aligned with the truth (large β_j) and tend to give lower ratings (small α_j). Instead, conservatives seem more gullible (large α_j) and less aligned with the truth (small β_j). There are even two workers with negative β_j, who are detrimental on fact checking.

4 Conclusion

Our analyses support that Bayesian generative models may lead to important advances for crowd-based fact checking. Future research will focus on the usage of the model for prediction of Z given W, and on the extension to more complex settings.

Acknowledgments This work was supported by the Departmental Strategic Plan (PSD) of the University of Udine, Interdepartmental Project on Artificial

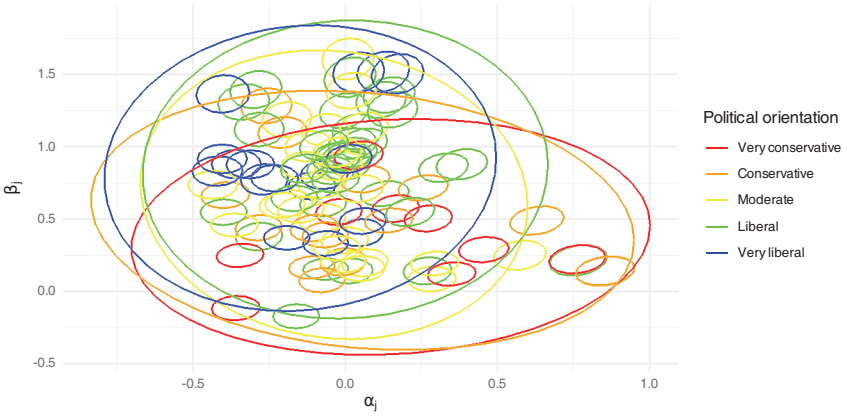


Figure 2. *Posterior normal ellipses of α_j and β_j (5%, smaller) grouped by political orientation of workers (90%, larger).*

Intelligence (2020-2025).

References

BARTHOLOMEW, D., KNOTT, M., & MOUSTAKI, I. 2011. *Latent Variable Models and Factor Analysis*. 3rd edn. Wiley.

NGUYEN, A., KHAROSEKAR, A., LEASE, M., & WALLACE, B. 2018. An interpretable joint graphical model for fact-checking from crowds. *In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, Part II*, vol. 32. New Orleans, USA: AAAI Press.

ROITERO, K., SOPRANO, S., PORTELLI, B., SPINA, D., DELLA MEA, V., SERRA, G., MIZZARO, S., & DEMARTINI, G. 2020. The COVID-19 infodemic: Can the crowd judge recent misinformation objectively? *In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. Virtual event, Ireland: ACM.

ROITERO, K., SOPRANO, M., PORTELLI, B., DE LUISE, M., SPINA, D., DELLA MEA, V., SERRA, G., MIZZARO, S., & DEMARTINI, G. 2021. Can the crowd judge truthfulness? A longitudinal study on recent misinformation about COVID-19. *Personal and Ubiquitous Computing*, **27**, 1–31.

STAN DEVELOPMENT TEAM. 2023. *RStan: The R interface to Stan*. R package version 2.21.8.

SUPERVISED CLASSIFICATION OF CURVES BY FUNCTIONAL DATA ANALYSIS: AN APPLICATION TO NEUROMARKETING DATA

Salvatore Latora [1] and Luigi Augugliaro[1]

[1] Department of Economics, Business and Statistics, University of Palermo, (e-mail: salvatore.latora@unipa.it, luigi.augugliaro@unipa.it)

ABSTRACT: In this paper we contribute to the functional data analysis literature by presenting a scalar-on-function penalized regression model with a multinomial response variable which takes into account possible information given by the phase variability. We also providing a practical application on neuromarketing data.

KEYWORDS: functional data, high-dimensional data, machine learning, sparse inference, supervised learning classification

1 Introduction

In recent decades, functional data analysis has played an increasingly important role in various scientific field, such as medicine, biology, engineering, and, above all, in the field of statistical research (see Ramsay & Silverman, 1997, Hsing & Eubank, 2015, Koner & Staicu, 2023 for some reference review). In this paper, we consider an application to neuromarketing data. *Neuromarketing* (Fisher *et al.*, 2010) is the application of neuroscientific methods to understand and analyse human behaviour in relation to markets and business needs. On the basis of different neurometrics, obtained by EEG recordings, taken on a sample of subjects while watching positive, negative, and neutral valence videos, to measure the α-*asymmetry* of the brain (a condition indicating the subject's attention to what he or she is observing, see Mazza & Pagano, 2017), the proposed methodology in this article aims to classify the valence of the video observed. The remaining part od this paper is organized as follows: in Section 2 we explain our proposal; in Section 3, the results obtained by analyzing the data introduced above are illustrated; Finally, conclusions are provided in Section 4.

2 Proposed model

Notation and definitions. By functional data, we mean a realization of a stochastic process. The functional data, i.e. the predictor, is modelled as: $f_{it_k} = f_i(t_k) + \varepsilon_{it_k}$, with $f_i \in \mathcal{F}$, where t_{ik} is the k-th time point detected on the i-th subject, with domain $[0,1]$, ε_{it_k} is an error term normally distributed, and f_{it_k} is an element of $\mathbb{L}^2_{[0,1]}$, where $\mathbb{L}^2_{[0,1]}$ denotes the space of square-integrable functions endowed with the standard inner product $\langle g_1 \, , \, g_2 \rangle = \int_0^1 g_1(t)\, g_2(t)\, dt$ and associated norm $\|g\| = \langle g \, , \, g \rangle^{\frac{1}{2}}$. Let us denote by Y_i, for $i = 1, \ldots, n$, a random variable distributed according to a Multinomial distribution, such that $Y_i \in \{-1, 0, 1\}$. Finally, by γ we denote a diffeomorphism, (*warping function*), belonging to the set $\Gamma = \{\gamma : [0,1] \to [0,1] \mid \gamma(0) = 0, \gamma(1) = 1\}$.

The propose model. The multinomial scalar-on-function regression model, belonging to the class of *FGLM* (James, 2002), takes the following form

$$\log\left\{\frac{Pr(Y_i = g \mid f_{it_k})}{Pr(Y_i = 0 \mid f_{it_k})}\right\} = \eta_{ig} = \beta_{0g} + \langle f_i \, , \, \beta_g \rangle, \tag{1}$$

where β_{0g} is the intercept of the g-th group and $\beta_g \in \mathbb{L}^2(t)$ is the regression coefficient function. Usually, for classification purposes, the phase variability of functional data is not taken into account, making it unitary during the pre-processing step through time warping (Ramsay & Silverman, 1997). However, as some authors show, (e.g., see Tucker *et al.*, 2013) phase variability may contain useful information for classification purposes. In this setting, time is expressed as $t_{ik} = \gamma_i(t_k)$, where $\gamma_i \in \Gamma$ is the warping function. Hence, the functional predictor to be used in (1) is expressed in a new re-parametrization of time as $f_{it_k} = f_i(\gamma_i(t_k)) = \tilde{f}_i(t_k)$, where $\tilde{f}_i(t_k) \in \mathbb{L}^2_{[0,1]}$ which only contains information on amplitude variability. Therefore, to use both phase and amplitude variability for our prediction problem, model (1) becomes

$$\log\left\{\frac{Pr(Y_i = g \mid f_{it_k})}{Pr(Y_i = 0 \mid f_{it_k})}\right\} = \beta_{0g} + \langle \tilde{f}_i \, , \, \beta_g \rangle + \langle \gamma_i \, , \, \theta_g \rangle, \tag{2}$$

where $\langle \gamma_i \, , \, \theta_g \rangle$ is the term contain information on the phase variability. Assuming that, both $\tilde{f}_i$ and γ_i are zero mean functions, and using by Karhunen–Loève expansion (Hsing & Eubank, 2015), i.e., $\tilde{f}_i(t) = \sum_{j=1}^{+\infty} X_{ij} \phi_j^f(t)$, and $\gamma_i(t) = \sum_{l=1}^{+\infty} Z_{il} \phi_l^\gamma(t)$. Model (2) can be expressed as follow:

$$\eta_{ig} = \beta_{0g} + \sum_{j=1}^{p} X_{ij} \langle \phi_j^f \, , \, \beta_g \rangle + \sum_{l=1}^{q} Z_{il} \langle \phi_l^\gamma \, , \, \theta_g \rangle, \tag{3}$$

where X_{ij} and Z_{il} are the *scores*, obtained by *FPCA*. In our application we use the *PACE* method (Yao *et al.*, 2005). The model becomes a classic multinomial regression model on scores, in which there are high dimensionality problems due to the choice of the number of basis by which to approximate both $\tilde{f}_i$ and γ_i. To overcome the problems from the high dimensional setting, we propose to minimize the penalised log-likelihood function $l_\lambda(\mathbf{b}) = l(\mathbf{b}) + n\lambda P(\mathbf{b})$, where $\mathbf{b}$ denote a vector of parameters for both amplitude and phase variability terms, whereas λ is the tuning parameter and $P(\mathbf{b})$ is the *Elastic-Net* penalty function (Zou & Hastie, 2005), i.e.: $P(\mathbf{b}) = \alpha\|\mathbf{b}\|_1 + \frac{(1-\alpha)}{2}\|\mathbf{b}\|_2^2$.

3 Application to Neuromarketing Data

The sample consists of $n = 60$ subjects who participated to a study, in which each subject was shown a video having positive, neutral, or negative valence. Through EEG signals, two indices, BIS and BAS (Davidson *et al.*, 1990), were obtained capable of capturing whether the subject showed attention when viewing the video. In the preprocessing step, all the curves were aligned. Subsequently, four separate FPCAs for each indicator and related warping functions were made to obtain the scores.

Table 1. *Hyper parameter values and model performance metrics on test set.*

α	λ	Accuracy	Precision[a]	Recall[a]
0.9797	0.0045	0.933	0.944	0.933

[a] Macro average was used

Table 1 shows the selected hyper-parameter: the selected α parameter allowed for a very selective model, which leads to a Lasso-type penalty function, however, the selected λ value is close to zero. Again Table 1 shows how the model achieves almost perfect classification ability on the test set, and thus excellent generalization ability.

4 Conclusions

The proposed approach allows the extraction and selection of relevant signals for classification, also taking into account the possible information of phase variability through a specific term in the linear predictor. The results show in

Section 3 highlight that the proposed model has achieved an excellent degree of generalization.

References

DAVIDSON, R. J., EKMAN, P., SARON, C. D., SENULIS, J. A., & FRIESEN, W. V. 1990. Approach-Withdrawal and Cerebral Asymmetry: Emotional Expression and Brain Physiology. I. *Journal of Personality and Social Psychology*, **58**(2), 330–341.

FISHER, CARL ERIK, CHIN, LISA, & KLITZMAN, ROBERT. 2010. Defining Neuromarketing: Practices and Professional Challenges. *Harvard Review of Psychiatry*, **18**(4), 230–237.

HSING, TAILEN, & EUBANK, RANDALL. 2015. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. First edn. Wiley Series in Probability and Statistics. Wiley.

JAMES, GARETH M. 2002. Generalized Linear Models with Functional Predictors. *Journal of the Royal Statistical Society: Series B*, **64**(3), 411–432.

KONER, SALIL, & STAICU, ANA-MARIA. 2023. Second-Generation Functional Data. *Annual Review of Statistics and Its Application*, **10**(1), 547–572.

MAZZA, VERONICA, & PAGANO, SILVIA. 2017. *Electroencephalographic Asymmetries in Human Cognition*. New York, NY: Springer New York. Pages 407–439.

RAMSAY, J. O., & SILVERMAN, B. W. 1997. *Functional Data Analysis*. Springer Series in Statistics. New York, NY: Springer New York.

TUCKER, J. DEREK, WU, WEI, & SRIVASTAVA, ANUJ. 2013. Generative Models for Functional Data Using Phase and Amplitude Separation. *Computational Statistics & Data Analysis*, **61**(May), 50–66.

YAO, FANG, MÜLLER, HANS-GEORG, & WANG, JANE-LING. 2005. Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, **100**(470), 577–590.

ZOU, HUI, & HASTIE, TREVOR. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B*, **67**(2), 301–320.

CAPTURING CORRELATED CLUSTERS USING MIXTURES OF LATENT CLASS MODELS

Gertraud Malsiner-Walli[1], Bettina Grün[1] and Sylvia Frühwirth-Schnatter[1]

[1] Institute of Statistics and Mathematics, Vienny University of Economics and Business
(gertraud.malsiner-walli@wu.ac.at, bettina.gruen@wu.ac.at, sylvia.fruehwirth-schnatter@wu.ac.at,)

ABSTRACT: Latent class models rely on the conditional independence assumption, i.e., it is assumed that the categorical variables are independent given the cluster memberships. Within the Bayesian framework, we propose a suitable specification of priors for the latent class model to identify the clusters in multivariate categorical data where the independence assumption is not fulfilled. Each cluster distribution is approximated by a latent class model, leading overall to a two-layer mixture of latent class models. By carefully specifying the priors on the model parameters, the Bayesian approach allows to identify the clusters and fit their cluster distributions using MCMC sampling. We provide suitable estimation and inference methods for the mixture of latent class models and illustrate the performance of this approach on a real data set containing patients suffering frome one of three types of low back pain.

KEYWORDS: Bayesian inference, model-based clustering, prior on the number of components, telescoping sampler

References

FOP, MICHAEL, SMART, KEITH M, & MURPHY, THOMAS BRENDAN. 2017. Variable selection for latent class analysis with application to low back pain diagnosis. *The Annals of Applied Statistics*, **11**(4), 2080–2110.

FRÜHWIRTH-SCHNATTER, SYLVIA, MALSINER-WALLI, GERTRAUD, & GRÜN, BETTINA. 2021. Generalized Mixtures of Finite Mixtures and Telescoping Sampling. *Bayesian Analysis*, **16**(4), 1279–1307.

MALSINER-WALLI, GERTRAUD, FRÜHWIRTH-SCHNATTER, SYLVIA, & GRÜN, BETTINA. 2017. Identifying mixtures of mixtures using Bayesian estimation. *Journal of Computational and Graphical Statistics*, **26**(2), 285–295.

A THREE-WAY "INDIRECT" REDUNDANCY ANALYSIS

Laura Marcis [1], Maria Chiara Pagliarella[1] and Renato Salvatore [1]

[1] Department of Economics and Law, University of Cassino and Southern Lazio, (e-mail: laura.marcis@unicas.it, mc.pagliarella@unicas.it, rsalvatore@unicas.it)

ABSTRACT: This work introduces a composite Three-Way application of the High Order Singular Value Decomposition. Two of the three component data matrices are processed by a standard Redundancy Analysis. The remaining "external" data matrix is related to the others in a heterogeneous system of relations, that can be well suited to tensor analysis. The external data are set to be linked with the first matrix, while with the second matrix the relations are explained only through multivariate linear regression. An application introduces the method, based on the official data from the Italian Equitable and Sustainable Well-being indicators.

KEYWORDS: Tucker decomposition, high order singular value decomposition, redundancy analysis.

1 Introduction and background

Tensor decomposition (Kolda & Bader, 2009) has the main objective of reducing complex information detected by higher dimensional arrays of data. From a pure statistical perspective, there are two important exploitations of the tensor analysis: the Candecomp/Parafac decomposition and the Tucker decomposition. They play the role of the extension to tensor objects of the principal component analysis (PCa), recognized as an explorative way to approach multidimensional information (Kroonenberg, 2008). In the literature, the most popular tensor decompositions are "Canonical Decomposition" and the "High Order SVD" (HOSVD, De Lathauwer *et al.*, 2000). The HOSVD decomposes an N-mode tensor, as a multidimensional array, in a core reduced-order tensor, multiplied by component matrices alongside each of the N modes. Three-way PCa was the first extension of the PCa to a three-way data set, giving the first useful employment of tensor analysis to explorative statistical analysis. In standard PCa, the components that come from the SVD that summarize individuals are uniquely related to the components that summarize variables. In a three-way PCa the components that summarize entities in each of the modes are related with the remaining two. Redundancy Analysis (RDA, Legendre

and Legendre, 2012) was originally introduced in order to capture the effect onto a reduced space $\widehat{\mathbf{Y}}_X = \mathbf{X}\widehat{\mathbf{B}}$ of the linear dependence by a set of criterion variables $\mathbf{Y}$ from a set of predictors $\mathbf{X}$, where $\widehat{\mathbf{B}}$ is the matrix of the ordinary least squares multivariate regression estimates. RDA provides a constrained analysis of the whole linear relations between the two sets of variables, and an unconstrained analysis given by the set of multivariate regression residuals. It can be considered as an extension of multivariate regression because models the effects of the explanatory variables on a response matrix. Partial RDA (pRDA) explores the effects of the predictors in $\mathbf{X}$ on the $\mathbf{Y}$ variables, given the covariates of some additional exploratory variables in a matrix $\mathbf{Z}$. It is a standard RDA performed taking into account the $\mathbf{X}$ variables as predictors on $\mathbf{Y} - \widehat{\mathbf{Y}}_Z$, with the "effect" by $\mathbf{Z}$ removed. Nevertheless, the relations between the variables $\mathbf{Y}$ and $\mathbf{Z}$ may be quite several. While remaining the same the role of the predictors $\mathbf{X}$ on $\mathbf{Y}$, a third set of variables $\mathbf{Z}$ may be related and depend on $\mathbf{Y}$, by an existing but not well defined dependence. Thus, applying multivariate regression may result hardly appropriate. Variables in $\mathbf{Z}$ in some cases can not be modeled on $\mathbf{Y}$ as predictors in a multivariate regression, while $\mathbf{X}$ predict $\mathbf{Y}$ and, indirectly through $\mathbf{Y}$, the variables in $\mathbf{Z}$. Residuals $\mathbf{Y} - \widehat{\mathbf{Y}}_X$ may take in account the role of $\mathbf{X}$ in the "indirect" explanation of $\mathbf{Z}$. This is somewhat different from pRDA, because $\mathbf{Y}$ is not regressed on $\mathbf{Z}$, as the external set of covariates from which we remove the effect on $\mathbf{Y}$, and also $\mathbf{Z}$ is not related with $\mathbf{Y}$ through linear regression. Given a 3rd-order tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, the Tucker decomposition through the HOSVD decomposes the tensor $\mathcal{X}$ into a core tensor $\mathcal{G}$ and factor matrices along each mode, as follows:

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$$

with the correspondent elementwise expression $x_{ijk} = \sum_{r=1}^{R} \sum_{s=1}^{S} \sum_{t=1}^{T} g_{rst} a_{ir} b_{js} c_{kt}$, with $i = 1, ..., I, j = 1, ..., J, k = 1, ..., K$. The factor matrices are columnwise orthonormal, $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_R], \mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_S], \mathbf{C} = [\mathbf{c}_1, ..., \mathbf{c}_T]$, with $r = 1, ..., R, s = 1, ..., S, t = 1, ..., T$. The matricized forms, one per mode, of the 3-way tensor $\mathcal{X}$ are:

$$\begin{aligned}
\mathbf{X}_{(1)} &\approx \mathbf{A}(\mathbf{C} \odot \mathbf{B})' = \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C} \otimes \mathbf{B})', \\
\mathbf{X}_{(2)} &\approx \mathbf{B}(\mathbf{C} \odot \mathbf{A})' = \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C} \otimes \mathbf{A})', \\
\mathbf{X}_{(3)} &\approx \mathbf{C}(\mathbf{B} \odot \mathbf{A})' = \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B} \otimes \mathbf{A})',
\end{aligned}$$

with the symbols $\odot$ and $\otimes$ that are the Khatri-Rao and Kronecker products, respectively. If $r_R(\mathcal{X})$ is the rank of the tensor $\mathcal{X}$ alongside one of the modes,

Table 1. *Description of the variables used for the application*

Variables	Description
S8	Age-standardised mortality rate for dementia and nervous system diseases
IF3	People having completed tertiary education (30-34 years old)
L12	Share of employed persons who feel satisfied with their work
REL4	Social participation
POL5	Trust in other institutions like the police and the fire brigade
SIC1	Homicide rate
BS3	Positive judgement for future perspectives
PATR9	Presence of Historic Parks/Gardens and other Urban Parks recognised of significant public interest
AMB9	Satisfaction for the environment - air, water, noise
INN1	Percentage of R&D expenditure on GDP
Q2	Children who benefited of early childhood services
BE1	Per capita adjusted disposable income
LBE1	Logarithm of Per capita adjusted disposable income

the HOSVD may uses Alternating Least Squares, in order to find:

$$\min_{\mathcal{G},\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \mathcal{X} - \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} \right\|.$$

Making the substitutions $\mathbf{A} = \mathbf{Y}$, $\mathbf{B} = \mathbf{Y} - \widehat{\mathbf{Y}}_X$, $\mathbf{C} = \mathbf{Z}$, with $I = J = K = n$, $R = S = r(\mathbf{Y}) = r(\mathbf{Y} - \widehat{\mathbf{Y}}_X)$, and $T = r(\mathbf{Z})$, we achieve the desired result, by finding a Three-Way version of the "indirect" RDA, with the proper data matrices. Like in the standard RDA, the data in $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{Z}$ have to be preprocessed by centering and standardazing their column vectors. This is requested before the application of the RDA of $\mathbf{Y}$ on $\mathbf{X}$.

2 Application study

The Equitable and Sustainable Well-being indicators (BES) are designed to define the economic policies which largely act on some fundamental aspects of the quality of life. Table 2 reports the description of these indicators. We use the latter as the predictor variable in the RDA that gives the constrained analysis in the subspace of $\widehat{\mathbf{Y}}_X$. Table 2 reports the correlation matrix between the column vectors of $\mathbf{Y}$, $\mathbf{Y}^*$, and $\mathbf{Z}$. Correlations in bold are significant. It is interesting to remark that in some cases the variables in $\mathbf{Z}$ are correlated with the columns of $\mathbf{Y}$, while they are generally poorly related with the RDA residuals vectors (given by the unconstrained RDA). In particular, the evidence is that even if $\mathbf{Z}$ may be regressed on $\mathbf{Y}$, for some variables the regression on $\mathbf{X}$ results inappropriate. One of the important cases is shown by the variable AMB9. This variable (Satisfaction for the environment - air, water, noise) is permanently correlated with the variable BS3 (Positive judgement for future

Table 2. *Correlations - Matrices* $\mathbf{Y}$*,* $\mathbf{Y}^{\star}$*, and* $\mathbf{Z}$

Variable	$Y1_{BS3}$	$Y2_{INN1}$	$Y3_{IF3}$	$Y4_{Q2}$	$Y5_{L12}$	$Y6_{S8}$
$Z1_{AMB9}$	**0,4029**	−0,0239	**0,4570**	**0,6852**	**0,8090**	**0,6926**
$Z2_{POL5}$	0,1906	**0,3629**	**0,2594**	**0,6395**	**0,6330**	**0,5973**
$Z3_{PATR9}$	0,1800	**0,3759**	0,0426	0,0353	0,0146	**0,2420**
$Z4_{REL4}$	**0,5133**	**0,2601**	**0,4413**	**0,7026**	**0,8380**	**0,6507**
$Z5_{SIC1}$	**−0,2215**	−0,1150	**−0,4665**	**−0,5397**	**−0,5925**	**−0,6343**
Variable	$Y1^{\star}_{BS3}$	$Y2^{\star}_{INN1}$	$Y3^{\star}_{IF3}$	$Y4^{\star}_{Q2}$	$Y5^{\star}_{L12}$	$Y6^{\star}_{S8}$
$Z1_{AMB9}$	**0,4605**	−0,1075	**0,2848**	0,1294	0,0423	−0,0119
$Z2_{POL5}$	0,0042	−0,1972	−0,0523	0,0662	−0,0624	−0,0755
$Z3_{PATR9}$	−0,1311	0,2081	**−0,2749**	**0,2794**	0,0053	0,1774
$Z4_{REL4}$	**0,3595**	−0,0025	−0,0056	0,0993	−0,1227	−0,1229
$Z5_{SIC1}$	−0,2029	−0,0184	**−0,3021**	−0,1787	−0,0291	−0,0234

perspectives), whatever is $\mathbf{y}$ or $\mathbf{y}^* = \mathbf{y} - \widehat{\mathbf{y}}_X$ (with $corr(y, y^*) = 0.7293$). We have a moderate correlation between the variable BS3 and the correspondent RDA residuals, and a moderate explanation of this variable is given by the BE1 (Per capita adjusted disposable income). Then, a tentative conclusion is that the "Satisfaction for the environment" (a $\mathbf{Z}$ variable) does not depend on the "Disposable income" (the RDA predictor $\mathbf{X}$). An opposite case occurs when we try to assess the same AMB9 variable, versus L12 (Share of employed persons who feel satisfied with their work). Even we have that $corr(y, y^*) = -0.2395$, AMB9 has the greatest correlation with the observed L12 (y), which reduces to be not significant in terms of L12 RDA residuals (y^*). Thus, even the "Share of employed persons who feel satisfied with their work" depends on the "Disposable income", and the "Satisfaction for the environment" can be explained by the relation with "People that feel satisfied with their work", the "Satisfaction for the environment" depends on the "Disposable income" through its relation with the "People that feel satisfied with their work".

References

DE LATHAUWER, LIEVEN, DE MOOR, BART, & VANDEWALLE, JOOS. 2000. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, **21**(4), 1253–1278.

KOLDA, TAMARA G, & BADER, BRETT W. 2009. Tensor decompositions and applications. *SIAM review*, **51**(3), 455–500.

KROONENBERG, PIETER M. 2008. *Applied multiway data analysis*. John Wiley & Sons.

LEGENDRE, PIERRE, & LEGENDRE, LOUIS. 2012. *Numerical ecology*. Elsevier.

MULTI-LEVEL STOCHASTIC BLOCKMODELS FOR MULTIPLEX NETWORKS

Maria Francesca Marino [1], Matteo Sani [1] and Monia Lupparelli [1]

[1] Department of Statistics, Computer Science, Applications "G.Parenti", University of Florence, (e-mail: mariafrancesca.marino@unifi.it, monia.lupparelli@unifi.it, matteo.sani@stud.unifi.it)

ABSTRACT: Multiplex arises when the network for the same set of nodes is repetitively observed on different layers that can represent, for instance, different statistical units or different criteria to connect the nodes. A multi-level Stochastic Blockmodel for multiplexes is introduced to provide a joint clustering of layers and nodes. This is achieved by considering two different sets of discrete latent variables. A former set allows us identifying groups of layers sharing similar connectivity patterns. A letter set of discrete latent variables, nested within the former, allows us identifying groups of nodes sharing similar relational features. A variational Expectation-Maximization algorithm is derived for estimation purposes.

KEYWORDS: network data, model-based clustering, finite mixtures, EM algorithm, variational inference.

1 Introduction

Uncover patterns underlying relations between nodes of a network is a complex task, especially when the network is repeatedly observed on a number of statistical units, or when different criteria to connect the nodes are available. For instance, connections between brain regions may be observed on a number of individuals, or imports/exports between countries may entail different types of products. In such cases, data provide a multilevel structure and multiplexes can be effectively used to describe, analyze, and model interactions between nodes (Barbillon *et al.*, 2017).

Stochastic blockmodels (SBMs - Daudin *et al.*, 2008) represent a valuable approach for identifying clusters of nodes sharing common relational features. These are identified by including in the model specification a set of node-specific, discrete, latent variables inducing nodes' partitioning. When multiplexes are available, one can decide to apply a SBM to each layer of the data structure, thus obtaining a separate clustering of nodes for each layer. As an alternative, the multivariate nature of dyadic relations may be properly taken

into consideration and nodes' clustering may be defined by fully exploiting the richness of the data at hand (Barbillon *et al.*, 2017).

We introduce a specification of the SBM for multiplexes that allows us to obtain a clustering of both layers and nodes. In detail, we introduce a multilevel SBM where layer-specific, discrete, latent variables allow us to cluster layers (i.e., the statistical units) sharing similar connectivity patterns. Within each of such clusters, nodes characterized by similar relational features are clustered by means of a further set of node-specific, discrete, latent variables. As typical of SBMs, Maximum Likelihood (ML) parameter estimates cannot be computed due to the intractability of the likelihood function. This makes infeasible the use of an Expectation-Maximization (EM - Dempster *et al.*, 1977) algorithm, as the posterior distribution of the random variables to compute at the E-step of the algorithm still requires the derivation of the likelihood function. To overcome the issue, we employ an extended variational EM algorithm, where the true, intractable, posterior distributions are substituted by their approximate versions, having a tractable form; see e.g., Blei *et al.*, 2017 for a thorough treatment of the topic.

2 Model definition

Let $\mathcal{G} = \{\mathcal{G}^k\}_{k\in(1,\ldots,K)}$ denote a multiplex characterized by K layers. Each graph $\mathcal{G}^k = (\mathcal{N}, \mathcal{E}^k) \in \mathcal{G}$ is defined by the same node set $\mathcal{N} = \{1,\ldots,n\}$ and the layer-specific edge set $\mathcal{E}^k$, with $k = 1,\ldots,K$. Equivalently, the multiplex $\mathcal{G}$ may be defined in terms of the *adjacency array* $\mathcal{Y} = \{Y^k\}_{k\in(1,\ldots,K)}$, with Y^k being the adjacency matrix associated to the k-th layer. Its generic element is

$$Y_{ij}^k = \begin{cases} 1 & \text{if the pair } (i,j) \in \mathcal{E}^k, \\ 0 & \text{else.} \end{cases}$$

That is, $Y_{ij}^k = 1$ iff nodes i and j are joined by an edge in the network associated to the k-th layer. For simplicity, we focus on the case of undirected networks, even though the extension to the directed case is straightforward.

Let $\{U_k\}_{k=(1,\ldots,K)}$ denote layer-specific, independent and identically distributed, latent variables defined over the support $\{1,\ldots,s\}$ and let $\eta_v = \Pr(U_k = v)$, for all $k \in 1,\ldots,K$. Furthermore, let $Z_i^k, i = 1,\ldots,n$, be a node-level latent variable, nested with respect to $U_k, k = 1,\ldots,K$, defined over the support $\{1,\ldots,m\}$ and let $\alpha_{qv} = \Pr(Z_i^k = q \mid U_k = v)$.

We assume that, conditional on the latent variables U_k, Z_i^k, and Z_j^k, the random variables Y_{ij}^k are independent each other and follow a Bernoulli distribu-

tion with tie probability only depending on the block membership of layers and nodes involved in the relation. That is,

$$Y_{ij}^k \mid Z_i^k = q, Z_j^k = l, U_k = v \overset{iid}{\sim} \mathcal{B}e(\pi_{qlv}).$$

Based on the above assumptions and denoting with θ the set of all free model parameters, the log-likelihood function can be written as

$$\begin{aligned} \ell(\theta) &= \log p(y) = \log \sum_u \sum_z p(y \mid u, z) p(z \mid u) p(u) \qquad (1) \\ &= \log \sum_u \sum_z \left\{ \left[\prod_{k=1}^{K} \prod_{i=1}^{n} \prod_{j>i} \mathcal{B}e(\pi_{z_i^k, z_j^k, u_k}) \right] \left[\prod_{k=1}^{K} \prod_{i=1}^{n} \alpha_{z_i^k, u_k} \right] \left[\prod_{k=1}^{K} \eta_{u_k} \right] \right\}, \end{aligned}$$

where y is a realization of $\mathcal{Y}$, and $\sum_u$ and $\sum_z$ are shorthands for $\sum_{u_1} \dots \sum_{u_K}$ and $\sum_{z_1^1} \sum_{z_1^2} \dots \sum_{z_n^{K-1}} \sum_{z_n^K}$, respectively.

As evident, deriving parameter estimates by either a direct or an indirect maximization of equation (1) is impractical. Indeed, this would require the computation of multiple summations, which is infeasible from a computational standpoint, even for networks of very limited size. To overcome the issue, an EM algorithm based on a variational approximation of the likelihood function may be employed as an effective alternative, as detailed in the following section.

3 Parameter estimation and inference

To derive parameter estimates, we extend the variational approach firstly introduced by Daudin *et al.*, 2008 in the SBM framework. Accordingly, starting from the likelihood function detailed in equation (1), estimates are derived by maximizing the following lower bound

$$\mathcal{F}(q(z,y), \theta) = \ell(\theta) - KL[q(z,u) \mid\mid p(z,u \mid \mathcal{Y}, \theta)], \qquad (2)$$

where $KL[\cdot \mid\mid \cdot]$ denotes the Kullback-Leibler divergence between the true, intractable, posterior distribution of the latent variables $p(z,u \mid y)$ and the corresponding approximating function $q(z,u)$. As we are not able to let KL vanish due to intractability of the likelihood, we look for the best approximation $q(z,u)$ in the class of completely factorized distributions

$$q(z,u) = q(u)q(z) = \prod_{k=1}^{K} \mathcal{M}ult(1, \tau_k) \prod_{i=1}^{n} \mathcal{M}ult(1, \phi_i).$$

The variational EM (VEM) algorithm alternates between two separate steps until convergence: (*i*) a VE-step, in which we maximize equation (2) with respect to the variational parameters τ_k and ϕ_i; (*i*) a VM-step, in which maximize (2) with respect to model parameters θ. Different works in the literature show the effectiveness of the variational approach in recovering the true value of model parameters in θ both with finite samples (see e.g., Mariadassou *et al.*, 2010) and asymptotically (see e.g., Celisse & Pierre, 2012).

To select the optimal number of blocks s and m, we may rely on an Integrated Classification Likelihood criterion (ICL - Biernacki *et al.*, 2000), as typically done in the SBM framework. Once the optimal model is selected, layer and node memberships are determined on the base of the parameter estimates $\hat{\tau}_k$ and $\hat{\phi}_i$, obtained at convergence of the estimation algorithm.

References

BARBILLON, PIERRE, DONNET, SOPHIE, LAZEGA, EMMANUEL, & BAR-HEN, AVNER. 2017. Stochastic block models for multiplex networks: an application to a multilevel network of researchers. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **180**, 295–314.

BIERNACKI, CHRISTOPHE, CELEUX, GILLES, & GOVAERT, GÉRARD. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, **22**(7), 719–725.

BLEI, DAVID M, KUCUKELBIR, ALP, & MCAULIFFE, JON D. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, **112**, 859–877.

CELISSE, ALAIN, & PIERRE, LAURENT. 2012. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, **6**, 1847–1899.

DAUDIN, J-J, PICARD, FRANCK, & ROBIN, STÉPHANE. 2008. A mixture model for random graphs. *Statistics and computing*, **18**(2), 173–183.

DEMPSTER, ARTHUR P, LAIRD, NAN M, & RUBIN, DONALD B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, **39**, 1–22.

MARIADASSOU, MAHENDRA, ROBIN, STEPHANE, & VACHER, CORINNE. 2010. Uncovering latent structure in valued graphs: a variational approach. *Annals of Applied Statistics*, **4**(2), 715–742.

THE MULTIVARIATE CLUSTER-WEIGHTED DISJOINT FACTOR ANALYZERS MODEL

Francesca Martella [1], Xiaoke Qin[2] and Wangshu Tu[2] and Sanjena Subedi[2]

[1] Department of Statistical Sciences, Sapienza University of Rome, (e-mail: francesca.martella@uniroma1.it)

[2] School of Mathematics and Statistics, Carleton University, (e-mail: XIAOKEQIN@cmail.carleton.ca, wangshu.tu@carleton.ca, Sanjeena.Dang@carleton.ca)

ABSTRACT: Cluster-weighted factor analyzers (CWFA) models are a flexible family of mixture models for fitting the joint distribution of a random vector constituted by a response variable and a set of explanatory variables. It is a useful tool especially when high-dimensionality and multicollinearity occurs. This paper extends CWFA models in two significant ways. Firstly, it allows to predict more than one response variable accounting for their potential interactions. Secondly, it identifies factors that relate to disjoint clusters of explanatory variables, simplifying their interpretatiblity. This leads to the multivariate cluster-weighted disjoint factor analyzers (MCWDFA) model. An alternating expectation-conditional maximization algorithm is used for parameter estimation. Application of the proposed approach to both simulated and real datasets is presented.

KEYWORDS: finite mixtures, factor regression model, disjoint factor analysis.

1 Introduction

Mixture models represent a powerful statistical tool for clustering observations which is an essential task in many fields, such as economics, engineering, and social sciences. In the context of media technology, Gershenfeld, 1997 proposed a particular family of Gaussian mixture models, called cluster-weighted models (CWMs), which has also been called saturated mixture regression models in Wedel, 2002. The context of interest is represented by data arising from a random vector $(\mathbf{X}, Y)'$, in which a functional dependence of Y on $\mathbf{X}$ is assumed for each mixture-component and the component-specific joint density of $(\mathbf{X}, Y)'$ is factorized into the product of the conditional density of $Y|\mathbf{X}$ and the marginal density of $\mathbf{X}$. Ingrassia *et al.*, 2012 reformulated the CWM in a statistical setting under the assumptions that both the component-specific conditional distributions of $Y|\mathbf{X}$ and the component-specific marginal

distributions of $\mathbf{X}$ are Gaussian. To allow the applicability of CWM in high dimensional $\mathbf{X}$-spaces or when multicollinearity occours, Subedi *et al.*, 2013 proposed the cluster-weighted factor analyzers (CWFA) model, which addressed the problem by assuming a latent structure for the explanatory variables in each mixture component. The aim of this paper is to propose a new model, called the multivariate cluster-weighted disjoint factor analyzers (MCWDFA) model, extending CWFA model in a two fold way. Firstly, it allows to predict more than one response variable accounting for their potential interactions. It leads to a more flexible model since it can capture the complexity and variability of real phenomena more accurately providing a more complete understanding of the underlying mechanisms of a case study. Secondly, it identifies factors that relate to disjoint clusters of explanatory variables which similarly predict the responses. In particular, following the idea of Martella *et al.*, 2008 and Vichi, 2017, we replace the factor loading matrix with the product of a binary row-stochastic matrix and a diagonal matrix in the factor analyzer structure. In this way, the explanatory variables that similarly predict the responses can be clustered into groups such that an explanatory variable loads only on one single factor, and thus, it is uniquely associated by a single factor only. This simplifies not only the interpretability of the resulting factors but also the interpretability of the (many) regression coefficients, especially when the explanatory variables matrices are in high-dimensional $\mathbf{X}$-spaces.

2 The cluster-weighted factor analyzers model

Briefly, the CWFA model (Subedi *et al.*, 2013) is a particular mixture model for fitting the joint distribution of a random vector composed of a response variable and a set of explanatory variables, where, within each Gaussian in the mixture, a single factor analysis regression (FAR) model (Basilevsky, 1981) is assumed. Let $y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$ be a response variable and a vector of explanatory variables, respectively, realizations of the pair $(\mathbf{X}, Y)$. Specifically, the CWFA model postulates that:

$$Y = \beta_{0g} + \beta'_{1g}\mathbf{X} + e_g \qquad \text{with} \qquad \mathbf{X} = \mu_g + \Lambda_g \mathbf{F}_g + \varepsilon_g \tag{1}$$

with probability π_g $(g = 1, \ldots, G)$. Terms μ_g represents the component-specific mean vectors of $\mathbf{X}$, Λ_g is a $p \times Q$ component-specific factor loadings matrix $(Q < p)$, $\mathbf{F}_g$ is a Q-dimensional vector of component-specific factors, which are assumed to be i.i.d. draws from a Gaussian distribution $N(0, \mathbf{I}_Q)$ and $\mathbf{I}_Q$ denotes the $Q \times Q$ identity matrix, ε_g are i.i.d. component-specific errors with

Gaussian distribution $N(0,\Psi_g)$, where $\Psi_g = \text{diag}(\psi_{1g},\ldots,\psi_{pg})$, that are assumed to be independent of $\mathbf{F}_g$. Furthermore, β_{0g} and β_{1g} are the component-specific intercept and the $(1\times p)$ component-specific vector of the regression coefficients, respectively; while e_g is a component-specific disturbances variable with Gaussian distribution $N(0,\sigma_g^2)$. Moreover, by assuming that Y is conditionally independent of $\mathbf{F}$ given $\mathbf{X}=\mathbf{x}$ in the generic g-th mixture component, we get that the joint density of $(\mathbf{X},Y)$ is given by:

$$p(\mathbf{x},y,\theta) = \sum_{g=1}^{G} \pi_g N(y|\mathbf{x};m(\mathbf{x};\beta_g),\sigma_g^2) N(\mathbf{x};\mu_g,\Lambda_g\Lambda_g'+\Psi_g) \tag{2}$$

where $m(\mathbf{x};\beta_g) = \beta_{0g}+\beta_{1g}'\mathbf{X}$ and $\theta = \{\pi_g,\beta_g,\sigma_g^2,\Lambda_g,\Psi_g; g=1,\ldots,G\}$. A collection of sixteen parsimonious CWFA models can be obtained by constraining or not $\sigma_g^2=\sigma^2$, $\Lambda_g=\Lambda$, $\Psi_g=\Psi$, and $\Psi_g=\psi_g\mathbf{I}_p$.

3 The multivariate cluster-weighted disjoint factor analyzers model

As mentioned previously, here we introduce the MCWDFA model that extends CWFA framework by considering more than one response variable and by identifying factors that relate to disjoint clusters of explanatory variables which similarly predict the responses. Let $\mathbf{X}$ be the p-dimensional vector of explanatory variables and $\mathbf{Y}$ be the M-dimensional vector of the response variables. For each component g ($g=1,\ldots,G$), the MCWDFA model is composed of two parts. The first extends the regression model in (1) with a multivariate regression model formalizing the relations between the M responses and the p explanatory variables, as follows:

$$\mathbf{Y} = \mathbf{B}_{0g} + \mathbf{B}_{1g}'\mathbf{X} + \mathbf{e}_g \tag{3}$$

where $\mathbf{B}_{0g}$ and $\mathbf{B}_{1g}$ are the $(M\times 1)$ component-specific vector of intercepts and the $(p\times M)$ component-specific matrix of the regression coefficients, respectively; $\mathbf{e}_g$ is the $(M\times 1)$ component-specific vector of disturbances variables with Gaussian distribution $N(0,\Sigma_{\mathbf{e}_g})$. On the other hand, the second part of the model assumes that the factor loading structure of the CWFA model holds except for the factor loading matrix Λ_g. In fact, to introduce explanatory variable clustering forming disjoint clusters which similarly predict the responses, Λ_g is replaced by the product of the specific matrices $\mathbf{V}_g$ and $\mathbf{W}_g$, where $\mathbf{V}_g = [v_{jqg}]$ is a $(p\times Q)$ component-specific binary row stochastic matrix

representing the membership matrix of the explanatory variables into Q clusters corresponding to Q factors, i.e. $v_{jqg} = 1$ if and only if, for observations in the g-th component, the j-th explanatory variable belongs to cluster q, 0 otherwise ($j = 1, \ldots, p$); while, $\mathbf{W}_g = \text{diag}(w_{1g}, \ldots, w_{pg})$ is a $(p \times p)$ component-specific diagonal matrix of weights for the explanatory variables. Constraint $\mathbf{V}_g' \mathbf{W}_g \mathbf{W}_g \mathbf{V}_g = \text{diag}(w^2_{\cdot 1g}, \ldots, w^2_{\cdot Qg})$, with $w^2_{\cdot qg} = \sum_{j=1}^{p} w^2_{jqg} > 0$ has to be satisfied, where the third index q added to w_{jg} indicates the factor associated with the j-th variable. Thus, the factor structure in (1) can be constrained in order to include the explanatory variables clustering as follows:

$$\mathbf{X} = \mu_g + \mathbf{W}_g \mathbf{V}_g \mathbf{F}_g + \varepsilon_g. \tag{4}$$

It is interesting observe that, recalling similar factor assumptions of the CWFA model, the component-specific covariance matrix of $\mathbf{X}$, after the proper permutation of explanatory variables, has a block diagonal form, where each block is the component-specific covariance matrix of the subset of the explanatory variables related to a specific factor. Maximum likelihood parameter estimates are derived using an alternating expectation-conditional maximization (AECM) algorithm. Application of the proposed approach to both simulated and real datasets is presented.

References

BASILEVSKY, A. 1981. Factor analysis regression. *Canadian Journal of Statistics.*, **9**(1), 109–117.

GERSHENFELD, N. 1997. Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences.*, **808**(1), 18–24.

INGRASSIA, S., MINOTTI, S.C., & VITTADINI, G. 2012. Local Statistical Modeling Via the Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification.*, **29**(3), 363–401.

MARTELLA, F., ALFO, M., & VICHI, M. 2008. Biclustering of gene expression data by an extension of mixtures of factor analyzers. *The international Journal of Biostatistics.*, **4**(1), 3.

SUBEDI, S., PUNZO, A, INGRASSIA, S., & MCNICHOLAS, P.D. 2013. Clustering and Classification Via Cluster-Weighted Factor Analyzers. *Advances in Data Analysis and Classification.*, **7**(1), 5–40.

VICHI, M. 2017. Disjoint factor analysis with cross-loadings. *Advances in Data Analysis and Classification.*, **11**(3), 563–591.

WEDEL, M. 2002. Concomitant Variables in Finite Mixture Models. *Statistica Neerlandica.*, **56**(3), 362–375.

SPATIAL MODELLING OF PYROCLASTIC COVER DEPOSIT THICKNESS WITH REMOTE SENSING DATA AND GROUND MEASUREMENTS: A FORECASTING COMBINATION APPROACH

Mattera, Raffaele[1], Scepi, Germana[2], Ebrahimi, Pooria[3] and Matano, Fabio [4]

[1] Department of Social and Economic Sciences, Sapienza University of Rome, Rome, Italy (email: raffaele.mattera@uniroma1.it)

[2] Department of Economics and Statistics, University of Naples "Federico II", Naples, Italy (email: scepi@unina.it)

[3] CNR-ISMAR, Naples, Italy (pooria.ebrahimi@na.ismar.cnr.it)

[4] CNR-ISMAR, Naples, Italy (fabio.matano@cnr.it)

ABSTRACT: Thickness of pyroclastic deposits governs various geomorphological and hydrological processes, but studies on the areas characterized by pyroclastic soil coverage are limited in the literature worldwide and the existing models predict thickness mainly based on morphological features of the slope. In this paper, additional variables are also derived from Digital Elevation Model (DEM) and satellite multispectral images to propose a spatial model for forecasting the thickness of pyroclastic deposits. For the prediction model, a two-step procedure is adopted: (1) the best subset of variables is selected; and (2) the predictions from different schemes are combined for deriving the final model. Predictive accuracy tests verify that the combination procedure provides a statistically significant improvement in predictions.

KEYWORDS: ensemble forecasting, remote sensing, environmental science, spatial modelling

1 Introduction

In an eruptive event, volcanic ash disperses in the atmosphere and deposits on the ground surface based on wind speed and direction. Because the geotechnical and hydraulic properties of the unconsolidated pyroclastic ash-fall deposits usually differ from bedrock, the spatial variation of their thickness significantly influences geomorphological and hydrological processes such as landscape evolution, hillslope hydrology, erosion, and landslides. Estimating the spatial distribution of thickness of pyroclastic ash-fall deposits is challenging because there might be more than one eruptive event, changes in wind characteristics during a single eruption can enhance complexity of the ash-dispersal pattern, and soil-forming and geomorphological processes continuously influence the expected spatial thickness.

In the literature, estimating thickness was mainly carried out for the areas covered by residual regolith (e.g., Saulnier et al., 1997; Saco et al., 2006; Tesfa et al., 2009; Segoni et al., 2013) and the approaches developed based on independent variables and applied to a specific site or in limited areas had a better performance (Del Soldato et al., 2018). There is, however, limited information on the thickness of pyroclastic ash-fall deposits under the influence of hillslope processes (De Vita et al., 2006).

Our paper proposes a new approach that considers additional variables for modelling and forecasting the thickness of pyroclastic ash-fall deposits. Combining the results provides more accurate forecasts, which are validated in terms of field measurements and compared with those obtained from three previously developed approaches: (1) Slope Angle Pyroclastic Thickness (SAPT; De Vita et al., 2006); (2) Geomorphological Pyroclastic Thickness (GPT; Del Soldato et al., 2016); and (3) Slope Exponential Pyroclastic Thickness (SEPT; Del Soldato et al., 2018). We apply the predictions for the area around Somma-Vesuvius, Phlegrean Fields and Roccamonfina volcanoes in southern Italy to evaluate the possibility of mapping thickness for a territory with complex geology and geomorphology.

2 Data and methodology

The literature on the tephra-producing eruptions of Somma-Vesuvius, Phlegrean Fields and Roccamonfina volcanoes was studied to prepare a database and compute the distance from eruptive vents along with the cumulative thickness of the ash layer deposited on the ground surface. The existing models like SAPT, GEPT and SEPT predict thickness mainly on the basis of morphological features of the slope. In this paper, we consider additional variables derived from Digital Elevation Model (DEM) and LANDSAT satellite multispectral images.

The following terrain features were then obtained from the DEM (resolution: 10×10m): altitude, slope degree, slope aspect, curvature, profile curvature, plan curvature, flow direction, flow accumulation, stream power index, stream transport index and topographic wetness index. Distance from the hydrographic network was also computed. The imageries of LANDSAT 8 Operational Land Imager (acquired in August 2017 and 2019), Collection 1 Level-1, were finally implemented to obtain four additional variables, i.e. Normalized Difference Vegetation Index (NDVI), Modified Secondary Soil-Adjusted Vegetation Index ($MSAVI_2$) and Normalized Clay Index (NCI) as proposed in the literature.

Following splitting a dataset of 7000 units (70% for training and 30% for testing), a stepwise regression (STPW) is applied to the training dataset for choosing the best subset of variables and for estimating the coefficients of the predictive model. Then, we use the resulting model for forecasting the thickness in the testing dataset.

The final forecasting model is obtained by combining the predictions of the GPT ($\hat{y}_{i,GPT}$), the SAPT ($\hat{y}_{i,SAPT}$), the SEPT ($\hat{y}_{i,SEPT}$) and the STPW ($\hat{y}_{i,STPW}$) approaches:

$$\hat{y}_i = w_1\,\hat{y}_{i,GPT} + w_2\,\hat{y}_{i,SAPT} + w_3\,\hat{y}_{i,SEPT} + w_4\,\hat{y}_{i,STPW} \quad (1)$$

where w_1, w_2, w_3, w_4 are the combination weights. We choose combination weights by evaluating the performance of five different schemes in out-of-sample. The first one is the Sample Average (SA) combination scheme. The second criterion is the Minimum Variance (MV, Hsiao and Wang, 2014):

$$w: \min_{w} w'\Sigma w \tag{2}$$

where $w = (w_1, w_2, w_3, w_4)'$ refers to the vector of unknown weights and Σ is the covariance matrix between forecasts of alternative models. The third scheme considers the inverse ranking (InvRank, Ailofi and Timmermann, 2006) of the alternative forecasting models in terms of Root Mean Square Error (RMSE):

$$w_k = Rank_k^{-1} / \sum Rank_k^{-1} \tag{3}$$

where k=1,2,3,4 is the index associated with the k-th forecasting model and $Rank_k^{-1}$ is the inverse ranking of the models in terms of RMSE. The last two considered combination schemes are the Ordinary Least Squares (OLS, Granger and Ramanathan, 1989) and the shrinkage approach (Shrink) of Bodnar et al. (2019).

3 Main results and final remarks

Tab. 1 shows results of the single forecasting models and the combination procedures. The GPT model is the best approach available in the literature, but the stepwise approach provides lower RMSE and MAE values (84.58 and 60.53, respectively). It indicates that including additional data derived from DEM and satellite imageries improves the accuracy of thickness predictions.

Tab. 1: Forecasting accuracy results

Category	Models	RMSE	MAE	Best performance
Single Fforecasting model	GPT	94.21843	62.64876	
	SEPT	110.2305	72.32251	
	SAPT	167.8572	141.7064	
	STPW	84.57994	60.52539	*
Combination procedure	SA	91.45242	67.57764	
	MV	84.56758	60.51947	
	InvRank	83.48263	60.24573	*
	OLS	84.54387	60.45999	
	Shrink	84.56749	60.92980	

In the next step, the predictions of single models are combined and it is revealed that performance of most combination approaches is better. The Inverse Ranking presents the best weighing system because the RMSE and MAE of the predicted thickness values are the lowest. Thus, we obtain a more representative pyroclastic

cover thickness distribution map for the areas affected by natural hazards such as landslides and floods.

References

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. Journal of Econometrics, 135(1-2), 31-53.

Bodnar, T., Dmytriv, S., Parolya, N., & Schmid, W. (2019). Tests for the weights of the global minimum variance portfolio in a high-dimensional setting. IEEE Transactions on Signal Processing, 67(17), 4479-4493.

De Vita, P., Agrello, D., & Ambrosino, F. (2006). Landslide susceptibility assessment in ash-fall pyroclastic deposits surrounding Mount Somma-Vesuvius: Application of geophysical surveys for soil thickness mapping. *Journal of Applied Geophysics*, 59(2), 126-139.

Del Soldato, M., Pazzi, V., Segoni, S., De Vita, P., Tofani, V., & Moretti, S. (2018). Spatial modeling of pyroclastic cover deposit thickness (depth to bedrock) in peri-volcanic areas of Campania (southern Italy). *Earth Surface Processes and Landforms*, 43(9), 1757-1767.

Del Soldato, M., Segoni, S., De Vita, P., Pazzi, V., Tofani, V., & Moretti, S. (2016). Thickness model of pyroclastic soils along mountain slopes of Campania (southern Italy). In Landslides and Engineered Slopes. Experience, Theory and Practice: Proceedings of the 12th International Symposium on Landslides, Napoli, Italy, 12-19 June 2016. CRC Press: Boca Raton, FL; 797-804.

Granger, C. W., & Ramanathan, R. (1984). Improved methods of combining forecasts. Journal of forecasting, 3(2), 197-204.

Hsiao, C., & Wan, S. K. (2014). Is there an optimal forecast combination?. Journal of Econometrics, 178, 294-309.

Saco, P. M., Willgoose, G. R., & Hancock, G. R. (2006). Spatial Organization of Soil Depths using a Landform Evolution Model/ NOVA. The University of Newcastle's Digital Repository: Newcastle.

Saulnier, G. M., Beven, K., & Obled, C. (1997). Including spatially variable effective soil depths in TOPMODEL. *Journal of Hydrology*, 202(1-4), 158-172.

Segoni, S., Martelloni, G., & Catani, F. (2013). Different methods to produce distributed soil thickness maps and their impact on the reliability of shallow landslide modeling at catchment scale. In Margottini, C., Canuti, P., Sassa, K. (Eds). Landslide Science and Practice. Springer: Berlin; 127-133.

Tesfa, T. K., Tarboton, D. G., Chandler, D. G., & McNamara, J. P. (2009). Modeling soil depth from topographic and land cover attributes. *Water Resources Research*, 45(10).

GRANGER NETWORK ON SANTA MARIA DEL FIORE DOME

Fiammetta Menchetti [1]

[1] DiSIA, University of Florence, (e-mail: fiammetta.menchetti@unifi.it)

ABSTRACT: The paper investigates the dynamic relationships between cracks and environmental variables, including temperature, humidity, and seismic activity, in the Santa Maria del Fiore Dome. Using Vector Autoregression (VAR) models and Granger causality tests, the study aims to understand the response of cracks to shocks on neighboring cracks. *

KEYWORDS: architectural heritage preservation, monument monitoring system, vector autoregressive model, Granger causality, impulse response functions

1 Introduction

The Santa Maria del Fiore Dome is a masterpiece of engineering and a symbol of Florence, Italy. Filippo Brunelleschi's design was revolutionary, and his innovative approach to construction enabled the Dome to be built without any scaffolding or temporary support structures. However, the first cracks on the Dome appeared soon after its construction in the 15th century and have progressively increased, giving rise to concerns about the stability of the monument (Ottoni & Blasi, 2015, Bertaccini, 2015, Bertaccini *et al.* , 2020). To address this issue, a monitoring system consisting of over 160 instruments was installed in the Dome starting from 1955. The present study is part of a long-term project aimed at monitoring the stability of the monument and predicting its future response to distressing phenomena. The objective of this work is to investigate the dynamic relationships between the cracks and the influence of environmental variables. To this aim, Vector Autoregression (VAR) models, Granger causality tests and Impulse Response Functions (IRF) are employed. The paper is structured as follows: Section 2 presents the data and the methodology used for the analysis; Section 3 describes the results.

*Funding for this study was provided by the National Centre for HPC, Big Data and Quantum Computing (project num. CN00000013). The author would also like to thank Silvia Bacci, Bruno Bertaccini and Fabrizio Cipollini for for their constructive feedback and suggestions, which helped to improve the quality of this article.

2 Data & Methodology

The data, provided by the Opera di Santa Maria del Fiore, consists in daily recordings of cracks width performed by the electronic system installed on the 8 webs of Brunelleschi's Dome by the ISMES (Istituto Sperimentale Modelli E Strutture) in 1987. In this analysis, we focus on the 13 deformometers located on web 4 and we restrict our attention to the period from January 1, 2001 to February 28, 2017. [†] Beside the wall temperature, the data have been supplemented with weather information such as air temperature and humidity, as well as information on earthquakes that occurred within a 50km radius of Florence during the analysis period.[‡]. Given the "breathing" mechanism of the Dome, we suspect that neighboring cracks could affect each other. To investigate the dynamic relationships between cracks as well as the influence of exogenous regressors, we fit the following VARX(p) model (Lütkepohl, 2005),

$$\Phi_p(L)Y_t = c + B_j(L)X_t + \varepsilon_t \tag{1}$$

where: $Y_t = (DF401, \ldots, DF413)$ is a vector containing the web crack measurements of all deformometers located on web 4; X_t is a vector of explanatory variables (namely, wall temperature, daily variation of air temperature, humidity and two dummy variables of earthquakes strength); c is a constant term; $\Phi_p(L) = I - \Phi_1 L - \cdots - \Phi_p L^p$ and $B_j(L) = B_0 + B_1 L + \cdots + B_j L^j$ are matrix polynomials in the lag operator L; $\Phi_1, \ldots, \Phi_p$ and $B_0, \ldots, B_j$ are coefficient matrices for lags 1 to p; and $\varepsilon_t \sim (0, \Sigma)$ is a multivariate white noise.

3 Results

All the variables included in the analysis show a yearly seasonal pattern that is highly persistent over time and there are some deformometers exhibiting non-linear trends (e.g., DF404). Before incorporating time series into a VAR model, they must be made stationary. To achieve this, Fourier terms are used

[†]This choice is motivated by empirical facts: major cracks are concentrated on the even webs, web 4 and 6 in particular (Ottoni & Blasi, 2015); moreover, early measurements evidence irregular patterns in the web crack evolution that may be due to the instruments' break-in period.

[‡]The city is located close to two fault lines, namely Mugello's composite seismogenic source and the (debated) Prato-Fiesole fault system. Earthquake data was sourced from the website of the Italian National Institute of Geophysics and Volcanology (INGV) and the map of fault lines can be found at https://diss.ingv.it/diss330/dissmap.html. Historical recordings of weather information for the city of Florence were obtained from the website "Il meteo", https://www.ilmeteo.it/meteo/Firenze

to remove the yearly seasonal pattern, and non-linear trends are removed with a natural cubic spline. This results in the estimation of the following model,

$$Y_t = c + g(t) + \beta_1 sin(2\pi t/365) + \beta_2 cos(2\pi t/365) + \nu_t \quad (2)$$

where: c is the intercept; β_1 and β_2 are the coefficients of the Fourier terms, representing the magnitude and the phase of the seasonal patterns; $g(t)$ is the natural cubic spline function capturing the non-linear trend in the data; and ν_t is the error term. After fitting model (2) separately to each variable, residuals are retained, as they represent the de-seasonalized and de-trended versions of the time series. To provide a snapshot of the results, Figure 1 plots the evolution of DF406 before and after the transformation, showing also the comparison between original and fitted values. The results of the VARX fit (available upon request) evidence that all the exogenous variables included in the model are significant predictors of the web cracks evolution in web 4. In particular, the lagged air temperature variation and lagged wall temperature are generally associated with a reduction in the crack width, whereas lagged humidity and the earthquakes are associated with an increase. Based on the VARX model above, it is also possible to explore the dynamic relationships between the cracks employing Granger causality by testing the pairwise combinations of the deformometers. The resulting Granger network is displayed in Figure 2.[§] Interestingly, DF404 seems the main driver for the evolution of several web cracks and its own dynamics does not appear to be Granger-caused by any other web crack; on the contrary, DF401 appears to be driven by several web cracks and doesn't seem to exert any impact on others. Finally, DF409 seems to be substantially isolated from the remaining deformometers. These results are supported by the IRFs, for which we report a snapshot in Figure 3 below.

[§]For ease of interpretation, the graph only shows unidirectional relationships, i.e., a directed edge is drawn from Y_1 to Y_2 only if past lags of Y_1 predict future values of Y_2 and not the reverse.

Figure 1. *Starting from the left: i) original vs. fitted values; iii) residuals.*

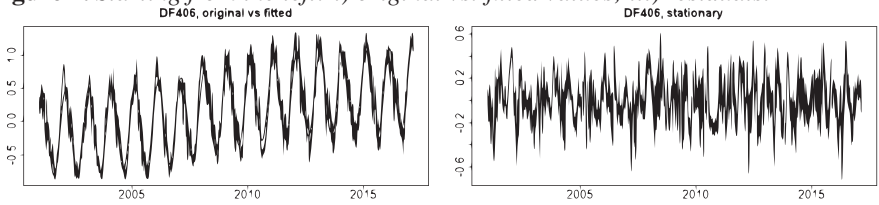


Figure 2. *Graphical representation of the Granger causality network originating from model (1).*

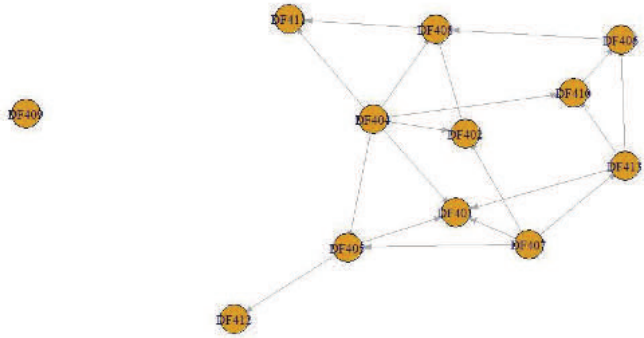
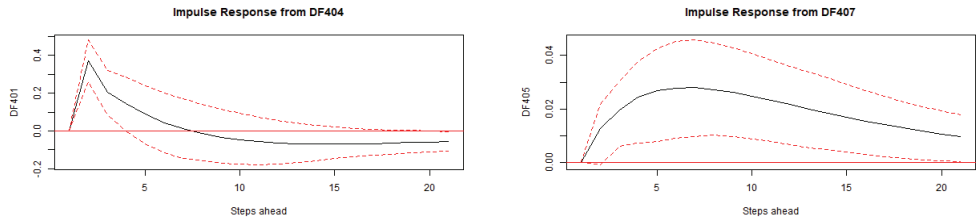


Figure 3. *IRFs for (a) a shock on DF401 from DF404 and (b) a shock on DF405 from DF407. Dashed lines indicate 95% bootstrap-based confidence intervals.*

(a) (b)



References

BERTACCINI, BRUNO. 2015. Santa Maria del Fiore dome behavior: Statistical models for monitoring stability. *International Journal of Architectural Heritage*, **9**(1), 25–37.

BERTACCINI, BRUNO, BACCI, SILVIA, & CRESCENZI, FEDERICO. 2020. A Dynamic Latent Variable Model for Monitoring the Santa Maria del Fiore Dome Behavior. *Pages 47–58 of: International Conference on Computational Science and Its Applications*.

LÜTKEPOHL, HELMUT. 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.

OTTONI, FEDERICA, & BLASI, CARLO. 2015. Results of a 60-year monitoring system for Santa Maria del Fiore Dome in Florence. *International Journal of Architectural Heritage*, **9**(1), 7–24.

Group's heterogeneity in rating tasks: a Bayesian semi-parametric approach

Giuseppe Mignemi[1], Ioanna Manolopoulou[2], Antonio Calcagnì[1]
[1]Università di Padova (email: giuseppe.mignemi@phd.unipd.it, antonio.calcagni@unipd.it)
[2]University College London (email:i.manolopoulou@ucl.ac.u)

Abstract In several observational contexts where different raters evaluate a set of items, it is common to assume that all raters draw their scores from the same underlying distribution. A common distributional assumption in this setting is that hierarchical effects as independent and identically distributed from a normal with the mean parameter fixed to zero and unknown variance. The present work aims to overcome this strong assumption in the inter-rater agreement estimation by assigning a Dirichlet Process (DP) mixture as the hierarchical effects' prior distribution. A new semi-parametric index λ is proposed to quantify raters polarization in presence of group heterogeneity. The model is applied to a real context.

Key words: rating process, inter-rater agreement, Dirichlet mixture process, Bayesian nonparametrics

1 A semi-parametric model proposal

Several methods and statistical models that aim to account for inter-rater variability have appeared in the literature [3]. Despite the popularity of work on this issue, less attention has been paid to possible latent dissimilarities among raters within inter-rater agreement studies[4]. From a psychometric point of view, it can be appealing to assess the extent to which different raters could have different latent opinions for specific rating processes.
To this aim, Hierarchical Generalized Linear Models (HGLM) are a natural choice, since they can account for the individual-variability specifying the effect of m covariates. The HGLM assumption regarding the distribution of the hierarchical effects is crucial in characterising different possible clusters or latent patterns of heterogeneity among raters. To this aim a Dirichlet Process Prior is specified over the hierarchical effects and the model is specified as follow.
The rating $y_{ij} \in \{0,1\}$ of the item $j \in \{1,..,J\}$ carried out by rater $i \in \{1,..,I\}$,

considering a set $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ of covariates for the different effects, respectively, is modelled as follows:

$$\begin{aligned} P(y_{ij}=1) &= F(\mathbf{x}_{ij}\beta + \mathbf{z}_i\mathbf{u}_i + \varepsilon_{ij}),\\ \mathbf{u}_i|\mu_c,\mathbf{Q} &\sim N_q(\mu_c,\mathbf{Q}),\\ \mu_c|G &\sim G,\\ G &\sim DP(\alpha,G_0). \end{aligned}$$

Here $F(\cdot)$ is a cumulative distribution function (e.g., Normal or Logistic), $N_q(\cdot)$ stands for a q-variate normal distribution, β is a $p\times 1$ vector of non hierarchical effects and $\mathbf{u}_i$ is a q vector of hierarchical effects. Here, $DP(\alpha,G_0)$ is a Dirichlet Process Mixture with $\alpha>0$ *precision parameter* and *base measure* G_0. It is assumed that $\mathbf{u}_i$ and ε_{ij} are independent.
The hierarchical effects distribution considering a stick breaking construction of the DPM might be then specified the as follow:

$$\mathbf{u}_i|\mu_c,\mathbf{Q},\alpha \overset{iid}{\sim} \sum_{c=1}^{R}\pi_c N_q(\mu_c,\mathbf{Q}),\quad i=i,\ldots,I$$

$$\mu_c \overset{iid}{\sim} G_0,$$

$$\pi_c = v_c\textstyle\prod_{l<c}(1-v_l),$$

$$v_c \overset{iid}{\sim} Be(1,\alpha),\ \ c=1\ldots,R.$$

Where $R\in\mathbf{N}$ and large enough [1].

1.1 The λ index

The marginal posterior distribution of the hierarchical effects in the model outlined above captures information about the dissimilarity or disagreement among raters (on the assumption that the model captures the data adequately). To this end the full estimated distribution of $\mathbf{u}$ resulting from the model might be useful. At each iteration t, the density of $\mathbf{u}$ is given by the corresponding mixture model given the parameters at iteration t. Following the formulation of [1] , the set of modes and antimodes (i.e., the least frequent values between two consecutive modes) is identified; the latent disagreement λ is then defined as the log ratio between the mean density of the modes and the that of the antimodes:

$$\lambda = \ln\left(\frac{\frac{1}{M}\sum_{m=1}^{M} f_{\mathbf{u}}(\gamma_m)}{\frac{1}{A}\sum_{a=1}^{A} f_{\mathbf{u}}(\zeta_a)}\right)$$

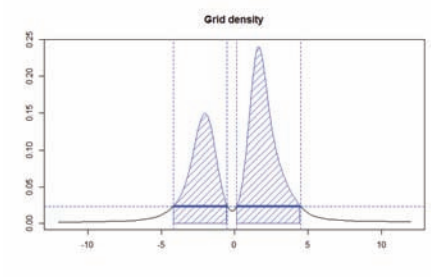
where M is the number of modes γ_m and A the number of antimodes ζ_a of $f(\mathbf{u})$. Larger values of λ indicate strongly multimodal distribution of the hierarchical effects, whereas smaller values are evidence of weak multimodality, thus the estimated hierarchical effects are less concentrated. In this sense this index is informative about the latent group polarization. Which in this context is assumed as a way of disagreement.

2 Posterior sampling and numerical example

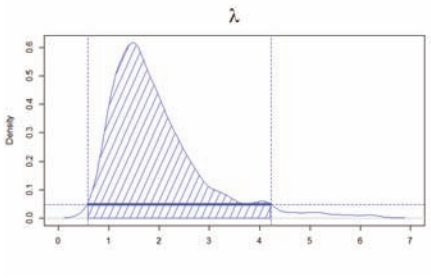
As a numerical example a real data set from the social sciences context was analysed. Fifty-two personnel selectors were asked to rate 40 different applicants per rater on a binary scale (0=not selected, 1=selected). In this case, y_{ij} is the binary score given to applicant i by selector j. Selectors' years of experience and applicants' age were two covariate considered in the model. The effect of the latter was specified as hierarchical with the distributional assumption outlined in the previous section.

Since most of the parameters in the model have conjugate prior distributions a blocked Gibbs sampling algorithm was used for the posterior sampling. An underline latent variable approach accounting for the probit link function of the HGLM was adopted. Weakly informative priors were elicited following [2]. As suggested by [1], in order to estimate the density of $\mathbf{u}$ the approach of monitoring $\mathbf{u} \overset{iid}{\sim} \sum_{c=1}^{R} \pi_c N_q(\mu_c, Q)$ at each iteration over a dense grid of u values was adopted. At each iteration t, the density of the parametric mixture was computed at each point of the grid. As result of some prior predictive check, a dense grid of 481 equally-spaced values from -12 to 12 (i.e., with a fixed interval of 0.05) was used to monitoring the mixture density of the hierarchical effects. The maximum number of mixture component R through the stick-breaking construction was fixed to 25. In all the computations 80.000 iteration with 8.000 burn-in were used, the Markov chains were thinned the by a factor of 80, resulting in samples of size 1000.

As shown in table 1 selector's years of experience has a positive effect on th probability of being selected. The marginal posterior distribution of the hierarchical effect of applicant's age showed a bimodal distribution. More precisely the effect of this predictor is positive for a subgroup of the overall sample, whereas it is negative in the other one. The presence of this heterogeneity is shown also by the λ- index which HPD interval is far from zero and includes large values.



(a)



(b)

Fig. 1 95% HPD intervals of the grid density (a) and λ-index.

Table 1 95% HPD intervals

95% HPD intervals	
β	$(1.58, 3.33)$
b_β	$(-0.56, 3.41)$
σ_{B_β}	$(0.16, 4.17)$
μ_0	$(-0.28, 0.75)$
σ_{D_0}	$(2.14, 5.25)$
Q	$(0.09, 0.29)$
σ_ε	$(0.86, 3.08)$
α	$(6.69, 16.06)$
Grid density	$(-4.15, -0.5) \cup (0.15, 4.50)$

References

1. GELMAN, A., CARLIN, J., STERN, H., DUNSON, D., AND VEHTARI, A.AND RUBIN, D. *Bayesian Data Analysis*. Chapman and Hall/CRC, 11 2013.
2. HEINZL, F., KNEIB, T., AND FAHRMEIR, L. Additive mixed models with dirichlet process mixture and p-spline priors. *AStA Advances in Statistical Analysis 96* (05 2012).
3. NELSON, K., AND EDWARDS, D. Measures of agreement between many raters for ordinal classifications. *Statistics in medicine 34* (06 2015).
4. WIRTZ, M. A. *Interrater Reliability*. Springer International Publishing, Cham, 2020, pp. 2396–2399.

LATTICE OF GAUSSIAN GRAPHICAL MODELS FOR PAIRED DATA WITH COMMON UNDIRECTED STRUCTURE

Dung Ngoc Nguyen [1] and Alberto Roverato[1]

[1] Department of Statistical Sciences, University of Padova, (e-mail: ngocdung.nguyen@unipd.it, alberto.roverato@unipd.it)

ABSTRACT: Typically, a model space embedded with a submodel order relationship has a lattice structure, called the model inclusion lattice. Recent works are related to the problems of joint learning of Gaussian graphical models suited for paired data, with exactly two dependent groups of variables. In this framework, it was shown that the model inclusion lattice does not satisfy the distributivity property, and this increases the complexity of procedures for the exploration of the search space. We consider a relevant subfamily of Gaussian graphical models for paired data represented by coloured graphs with common uncoloured structure. We show that this subfamily forms a proper sublattice of the family of Gaussian graphical models for paired data and that, within this sublattice, the distributivity property is satisfied. This can be exploited to improve efficiency in model search procedures.

KEYWORDS: coloured Gaussian graphical model, RCON model, distributivity.

1 Introduction

In the joint learning of multiple networks, recent works have considered the case of paired data, where the observations come from two dependent groups with the same variables, and every variable in the first group has a *homologous* variable in the second group. In this context, it is of interest to learn the similarities and differences between groups (Xie *et al.*, 2016; Ranciati *et al.*, 2021; Roverato & Nguyen, 2022; Zhang *et al.*, 2022; Roverato & Nguyen, 2023).

Let Y_V be a multivariate Gaussian random vector indexed by $V = \{1, \ldots, p\}$ with covariance matrix Σ and concentration matrix $\Sigma^{-1} = \Theta = (\theta_{ij})_{i,j \in V}$. An undirected graph $G = (V, E)$ consists of a set V of vertices and a set E of edges, which are unordered pairs of elements of V. In a *Gaussian graphical model* (GGM) for Y_V every missing edge of G implies that the corresponding entry of Θ is equal to zero (Lauritzen, 1996). A coloured version of G, denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, consists of a partition $\mathcal{V} = \{V_1, \ldots, V_v\}$ of V and a partition

$\mathcal{E} = \{E_1, \ldots, E_e\}$ of E, into colour classes. A coloured GGM (Højsgaard & Lauritzen, 2008) with coloured graph $\mathcal{G}$ is a GGM with additional symmetry restrictions on the parameters implied by the colouring of $\mathcal{G}$. More specifically, the parameters associated with vertices or edges belonging to a same colour class are restricted to be identical. One type of such restrictions are equalities between elements of the concentration matrix Θ, thereby identifying the family of RCON models.

In paired data problems, the random vector Y_V is partitioned into $Y_V = (Y_L, Y_R)^T$ with $|L| = |R| = p/2 = q$, and it is assumed, without loss of generality, that $L = \{1, \ldots, q\}$, $R = \{1', \ldots, q'\}$ with $i' = q + i$ for $i \in L$, and that for every $i \in L$, Y_i and $Y_{i'}$ form a pair of homologous variables.

Roverato & Nguyen, 2022 introduced a subfamily of RCON models specifically suited for paired data (PD-RCON models) where symmetries are implemented as equality constraints on the diagonal entries $\theta_{ii} = \theta_{i'i'}$ implied by the colour class $\{i, i'\}$ that we call a *twin-pairing* class. Similarly, for the symmetries of the off-diagonal entries that can be either $\theta_{ij} = \theta_{i'j'}$ implied by the twin-pairing class of edges $\{(i, j), (i', j')\}$ between groups, or $\theta_{ij'} = \theta_{ji'}$ implied by the twin-pairing class of edges $\{(i, j'), (j, i')\}$ across groups. Therefore, in the coloured graphs for paired data, $\mathcal{V}$ is divided into $\mathcal{V} = \mathcal{V}^{(t)} \cup \mathcal{V}^{(a)}$ that contains either twin-pairing classes of $\mathcal{V}^{(t)}$ consisting of a pair of homologous vertices, or atomic classes of $\mathcal{V}^{(a)}$ consisting of a single vertex. This is similar to colouring of edges with $\mathcal{E} = \mathcal{E}^{(t)} \cup \mathcal{E}^{(a)}$ where $\mathcal{E}^{(t)}$ contains twin-pairing classes made up of a pair of homologous edges between or across groups and $\mathcal{E}^{(a)}$ contains atomic classes made up of a single edge present on the graph.

2 Exploration of the search space

Typically, a model space is embedded with the *model inclusion*, or *submodel*, relationship resulting in a lattice structure, which is obtained by specifying the meet $\wedge$ and join $\vee$ operations between two models. These operations are used in structure learning of graphical models for the identification of neighbouring models, and it is important that they can be efficiently computed. This is the case of GGMs where the model inclusion coincides with the subset relation between edge sets. Formally, for two GGMs $G = (V, E_G)$ and $H = (V, E_H)$, G is a submodel of H, denoted by $G \preceq H$, if and only if $E_G \subseteq E_H$. Therefore, the family of GGMs is a lattice where the meet $G \wedge H$ and the join $G \vee H$ take particularly simple forms represented by graphs with the edge sets $E_G \cap E_H$ and $E_G \cup E_H$, respectively. The distributivity between these operations is thus satisfied, and we recall that distributivity is a useful property that facilitates the

implementation of efficient procedures in lattices and has also been exploited in model selection (Edwards & Havránek, 1987; Davey & Priestley, 2002; Gehrmann, 2011).

Roverato & Nguyen, 2022 considered the family of PD-RCON models, denoted by $\mathcal{P}$, and showed that $\mathcal{P}$ forms a proper sublattice of RCON models under model inclusion and that, also for this sublattice, the distributivity property does not hold. Here, we notice that the family of PD-RCON models can be naturally split into equivalence classes. All the models in a same class have the same underlying uncoloured undirected graph, that is, they are obtained by imposing additional equality restrictions to a common GGM. For an undirected graph $G = (V,E)$ we denote by $\mathcal{P}_E$ the family of PD-RCON models represented by coloured graphs with a common uncoloured structure G. In the following, we show that such equivalence classes form a proper sublattice of $\mathcal{P}$ that is distributive.

Theorem 1 *Let $G = (V,E)$ be an undirected graph. The class of PD-RCON models with a common uncoloured structure $\mathcal{P}_E$, equipped with the model inclusion order $\preceq$, forms a distributive lattice where if $\mathcal{G},\mathcal{H} \in \mathcal{P}_E$,*

(i) $\mathcal{G} \preceq \mathcal{H}$ if and only if $\mathcal{V}_{\mathcal{G}}^{(a)} \subseteq \mathcal{V}_{\mathcal{H}}^{(a)}$ and $\mathcal{E}_{\mathcal{G}}^{(a)} \subseteq \mathcal{E}_{\mathcal{H}}^{(a)}$,

(ii) the meet $(\mathcal{V}_\wedge, \mathcal{E}_\wedge) \in \mathcal{P}_E$ can be computed as

- *atomic classes:* $\mathcal{V}_\wedge^{(a)} = \mathcal{V}_{\mathcal{G}}^{(a)} \cap \mathcal{V}_{\mathcal{H}}^{(a)}, \quad \mathcal{E}_\wedge^{(a)} = \mathcal{E}_{\mathcal{G}}^{(a)} \cap \mathcal{E}_{\mathcal{H}}^{(a)}$
- *twin-pairing classes:* $\mathcal{V}_\wedge^{(t)} = \mathcal{V}_{\mathcal{G}}^{(t)} \cup \mathcal{V}_{\mathcal{H}}^{(t)}, \quad \mathcal{E}_\wedge^{(t)} = \mathcal{E}_{\mathcal{G}}^{(t)} \cup \mathcal{E}_{\mathcal{H}}^{(t)}$;

(iii) the join $(\mathcal{V}_\vee, \mathcal{E}_\vee) \in \mathcal{P}_E$ can be computed as

- *atomic classes:* $\mathcal{V}_\vee^{(a)} = \mathcal{V}_{\mathcal{G}}^{(a)} \cup \mathcal{V}_{\mathcal{H}}^{(a)}, \quad \mathcal{E}_\vee^{(a)} = \mathcal{E}_{\mathcal{G}}^{(a)} \cup \mathcal{E}_{\mathcal{H}}^{(a)}$
- *twin-pairing classes:* $\mathcal{V}_\vee^{(t)} = \mathcal{V}_{\mathcal{G}}^{(t)} \cap \mathcal{V}_{\mathcal{H}}^{(t)}, \quad \mathcal{E}_\vee^{(t)} = \mathcal{E}_{\mathcal{G}}^{(t)} \cap \mathcal{E}_{\mathcal{H}}^{(t)}$.

Proof. Point (i) follows from Proposition 2 of Roverato & Nguyen, 2022 because $E_{\mathcal{G}} = E_{\mathcal{H}} = E$. Furthermore, the meet and the join between $\mathcal{G}$ and $\mathcal{H}$ in (ii) and (iii) can be computed as described in Theorem 4 of Roverato & Nguyen, 2022, with $\tilde{\mathcal{E}}_{\mathcal{G}}^{(a)} = \mathcal{E}_{\mathcal{G}}^{(a)}$, $\tilde{\mathcal{E}}_{\mathcal{G}}^{(t)} = \mathcal{E}_{\mathcal{G}}^{(t)}$, $\tilde{\mathcal{E}}_{\mathcal{H}}^{(a)} = \mathcal{E}_{\mathcal{H}}^{(a)}$ and $\tilde{\mathcal{E}}_{\mathcal{H}}^{(t)} = \mathcal{E}_{\mathcal{H}}^{(t)}$; moreover, $E^* = \emptyset$ with $\mathcal{E}^{(a)}(E) \subseteq (\mathcal{E}_{\mathcal{G}}^{(a)} \cap \mathcal{E}_{\mathcal{H}}^{(a)})$, $\mathcal{E}_{\mathcal{G}}^{(t)}(E) = \mathcal{E}_{\mathcal{G}}^{(t)}$, $\mathcal{E}_{\mathcal{H}}^{(t)}(E) = \mathcal{E}_{\mathcal{H}}^{(t)}$. All notations $\tilde{\mathcal{E}}_{\mathcal{G}}^{(\cdot)}$, $\tilde{\mathcal{E}}_{\mathcal{H}}^{(\cdot)}$, E^*, $\mathcal{E}^{(a)}(E)$, $\mathcal{E}_{\mathcal{G}}^{(t)}(E)$, and $\mathcal{E}_{\mathcal{H}}^{(t)}(E)$ are defined in Section 3 of Roverato & Nguyen, 2022.

3 Conclusions

We have shown that the family of PD-RCON models can be split into equivalence classes which form distributive lattices with respect to model inclusion. Future research work will concern the exploitation of this property to achieve more efficiency in model search procedures.

Acknowledgement. Financial support has been provided by the MUR – Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) grant 2022 SMNNKY.

References

DAVEY, BRIAN A, & PRIESTLEY, HILARY A. 2002. *Introduction to lattices and order*. Cambridge University Press.

EDWARDS, DAVID, & HAVRÁNEK, TOMÁŠ. 1987. A fast model selection procedure for large families of models. *Journal of the American Statistical Association*, **82**(397), 205–213.

GEHRMANN, HELENE. 2011. Lattices of graphical Gaussian models with symmetries. *Symmetry*, **3**(3), 653–679.

HØJSGAARD, SØREN, & LAURITZEN, STEFFEN L. 2008. Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society: Series B*, **70**(5), 1005–1027.

LAURITZEN, STEFFEN L. 1996. *Graphical models*. Oxford University Press.

RANCIATI, SAVERIO, ROVERATO, ALBERTO, & LUATI, ALESSANDRA. 2021. Fused graphical lasso for brain networks with symmetries. *Journal of the Royal Statistical Society: Series C*, **70**(5), 1299–1322.

ROVERATO, ALBERTO, & NGUYEN, DUNG NGOC. 2022. Model inclusion lattice of coloured Gaussian graphical models for paired data. *Pages 133–144 of:* SALMERÓN, ANTONIO, & RUMÍ, RAFAEL (eds), *Proceedings of the 11th International Conference on Probabilistic Graphical Models*. Proceedings of Machine Learning Research, vol. 186. PMLR.

ROVERATO, ALBERTO, & NGUYEN, DUNG NGOC. 2023. Exploration of the search space of Gaussian graphical models for paired data. *arXiv preprint arXiv:2303.05561*.

XIE, YUYING, LIU, YUFENG, & VALDAR, WILLIAM. 2016. Joint estimation of multiple dependent Gaussian graphical models with applications to mouse genomics. *Biometrika*, **103**(3), 493–511.

ZHANG, HONGMEI, HUANG, XIANZHENG, & ARSHAD, HASAN. 2022. Comparing Dependent Undirected Gaussian Networks. *Bayesian Analysis*, 1–26.

MULTIVARIATE REGRESSION TREE TOPIC MODELING

Marco Ortu [1], Giulia Contu[1] and Luca Frigau[1]

[1] Dept. of Economics and Business Sciences, University of Cagliari, (e-mail: marco.ortu@unica.it, giulia.contu@unica.it, frigau@unica.it)

ABSTRACT: In this paper we propose Multivariate Tree Topic Modeling methodology, a general purpose approach to Topic Detection, which aims to refine the general results of a Topic Modeling methodology using Multivariate Trees in order to obtain consistent document groups. Topic modeling is defined as a mechanism for discovering low-dimensional, multi-faceted summaries of textual documents, typically by discovering hidden or latent topics in a corpus of documents. Given these hidden topics, we exploit the Multivariate Trees to obtain more homogeneous document groups with respect to the Topic Modeling output alone. We applied our model to a standard corpus of documents generally used in this kind of study to show that, when the aim of Topic Modeling is to generate coherent clusters of documents, the use of Multivariate Trees improves the overall coherence of these clusters for a wide range of Multivariate Trees' size.

KEYWORDS: Multivariate Analysis, Decision Trees, Topic Detection

1 Introduction

Topic modeling (TM) is a method for detecting latent structures in a collection of text documents. From the mathematical perspective, it can be seen as a dimensional reduction problem, where the vector space of a text document is often greater than several tens or hundreds of thousands, while the output vectorial space of topic modeling is typically in the order of tens and seldom hundreds. The typical use case of topic modeling is to represent themes and topics that are present in a large corpus of text data. This article proposes a method, based on Multivariate-Trees (MT), to refine the topic modeling output. In the literature, there are three main families of Topic Modeling methods: i) Matrix factorization based methods, ii) Probabilistic Methods, and more recently iii) Deep-learning based approaches. Matrix factorization-based topic modeling methods rely on a linear algebraic technique that factorizes a term-document matrix into two non-negative matrices, where one matrix represents

the topic-word distribution and the other matrix represents the document-topic distribution. The constraints used to solve the linear algebraic problem lead to the specific implementation, a widely used topic model is the Non-negative Matrix Factorization (NMF) (Lee & Seung, 2000). Probabilistic-based topic modeling methods rely on the hypothesis that documents are a mixture of topics guided by, typically, two hidden distributions of words in a collection of documents, one that models the distribution of words in hidden topics and one that models the distribution of topics in documents. One of the most popular approaches to topic modeling is Latent Dirichlet Allocation (Blei *et al.*, 2003) (LDA). It is a generative probabilistic model that assumes that each document in a corpus is generated by a mixture of latent topics, where a distribution over words characterizes each topic. When used for clustering, documents are grouped together by their dominant topic. However, these groups might be incoherent, since the assignment of the dominant topic is arbitrary. We show that, when the main goal of topic modeling is to generate a coherent clustering of documents, the use of multivariate trees improves the overall coherence of these clusters as measured by heterogeneous indexes such as Gini's Index.

2 Methodology And Data

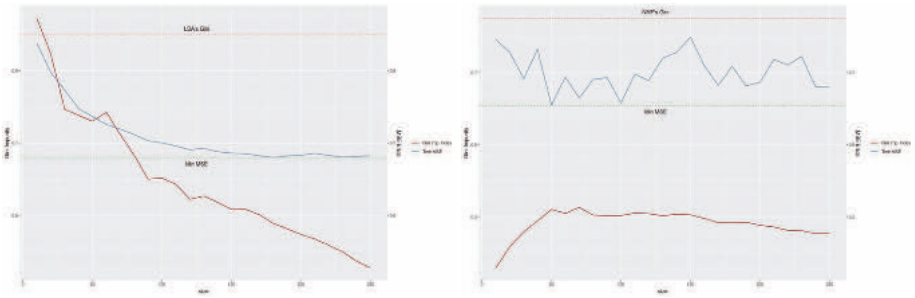
Multivariate-Tree Topic Modelling De'Ath, 2002 is a general purpose approach to Topic Modeling, which aims to refine the general results of a TM methodology using Multivariate Trees in order to obtain consistent documents groups. Algorithm 1 illustrates MTTM method. It is composed of two phases. In the first phase, a topic model is used to obtain the topic distributions of each document. The topic model leads to a topic distribution over the documents, as often in these applications, we considered the dominant topic, namely ϕ_t, and grouped the documents according to it. We finally evaluate the average Gini's index of each group considering the true category, namely y_t, obtaining our baseline measure of the groups' homogeneity ($\bar{G}_{topic}$). Our goal is to maximize the homogeneity of the groups (thus minimizing the average Gini's index). In the second phase, we apply a multivariate regression tree using the topic distributions as dependent variables and the words' frequencies as predictors. We evaluate the average Gini's index $\bar{G}_{tree}$ as a function of tree's size, and compare it with the topic baseline ($\bar{G}_{topic}$).

Figure 1 shows the results of MTTM for the 20 Newsgroups dataset using MT (De'Ath, 2002). Figure 1a reports our results using LDA topic modeling. The dashed red line is the baseline of the Gini's index obtained by the topic modeling alone, it is the average Gini's index of the topic modeling groups

Algorithm 1 MTTM Algorithm Definition

Require $\mathcal{D}$: Training documents of size N, each with a categorical response variable y_t and a set of quantitative variables X_t;
Step 1:
Input: $\mathcal{D} = \{d_1, \ldots, d_N\}$
Output: $\phi : \mathcal{W} \rightarrow \Theta$
$X \leftarrow W$
$Y \leftarrow \phi$
$\bar{G}_{topic} = \frac{1}{T} \sum_{i=1}^{T} G_t(\phi_t)$
Step 2:
set $(s_{min}; s_{max})$
for $s = s_{min}$, $s{+}{+}$, $s < s_{max}$ **do**
fit MT: $Y \sim X$
$\bar{G}_{tree} = \frac{1}{s} \sum_{i=1}^{s} G_s(y_t)$
end for

considering the true category. The green dashed line represents the Gini's index of the groups obtained by the multivariate tree obtained for the tree size that minimizes the tree MSE. The blue continuous line represents the tree MSE over the tree size, and the red continuous line represents the average Gini's Index over the tree size. We can observe that the average Gini's Index of the multivariate tree decreases as the number of splits increases, at the limit case when the number of splits equals the number of documents, each split contains only one document and the Gini's index is zero. In this particular case, we knew that there were 20 topics, thus we run the LDA algorithm using 20 as the number of topics. Analyzing the output of LDA, we could identify only nine out of 20 topics, namely, only nine topics were assigned with a probability greater than zero. On the other hand, the multivariate tree's output showed an optimal number of splits around about 150. In this range of the number of splits, we can observe that the MTTM output always yields more homogeneous groups. Figure 1b shows the results using NMF topic modelling. In this case, it can be observed that the average Gini's Index over the tree size exhibit a different behaviour, it starts with a lower value of average Gini's index, and it increases as the tree's size increases, then it stabilizes for a wide range of tree's size and finally start a decreasing trend for extreme values. In general, from Figure 1 we can observe that, after the application of the MT, the overall Gini's index is improved for a wide range of the tree's size.



(a) LDA algorithm. (b) NMF algorithm.

Figure 1: Average Gini's impurity index for the 20 Newsgroups dataset using Multivariate Tree.

3 Conclusion

This study proposes a methodology to enhance the ability of a topic detection method to create coherent groups leveraging decision trees. We presented the Multivariate-Tree Topic Modelling framework MTTM. MTTM is constructed by combining topic modeling and multivariate-trees methodologies. We applied our model to a standard corpus of documents (the 20 Newsgroup) and two topics models (LDA and NMF) generally used in this kind of studies, and we found that, when the aim of TM is to generate coherent clusters of documents, the use of a MT improves the overall coherence of these clusters for a wide range of the MT's size.

References

BLEI, DAVID M, NG, ANDREW Y, & JORDAN, MICHAEL I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, **3**, 993–1022.

DE'ATH, GLENN. 2002. Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*, **83**(4), 1105–1117.

LEE, DANIEL, & SEUNG, H SEBASTIAN. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, **13**.

DENDROGRAM SLICING THROUGH A PERMUTATION TEST APPROACH RECONSIDERED

L. Palazzo, A. Iodice D'Enza, F. Palumbo, and D. Vistocco
Department of Political Sciences, University of Naples Federico II
(e-mail: [lucio.palazzo, iodicede, domenico.vistocco, fpalumbo]@unina.it)

ABSTRACT: DESPOTA is a clustering method that cuts the tree branches at various heights to find the best division among those that can be achieved from the hierarchical clustering tree through the use of a permutation test at each node. In order to reduce the computational cost and increase the applicability of DESPOTA to huge data sets, the present study suggests two improvements to the DESPOTA original implementation that combine aggregation with either splitting or partitioning approaches. A dataset of the Italian universities' five-year periodical accreditation by the Italian national agency (ANVUR) is used to test the suggested approach.

KEYWORDS: hierarchical clustering, permutation test, top-down splitting

1 Introduction

Hierarchical algorithms represent excellent solutions for data clustering when one aims to get nested partitions in the data that can be easily visualized through tree-like representations, also referred to as dendrograms (from the Greek word δὲνδρον= tree). Cutting the tree at a given level defines data partitioning into disjoint clusters. Nonetheless, the optimal cutting level (corresponding to the optimal number of clusters) remains a ticklish problem, and the choice is generally left to the user's heuristic criteria. DESPOTA (Bruzzese & Vistocco, 2015, DEndrogram Slicing through a PermutatiOn Test Approach) seeks the best partition among the possible ones achievable from a hierarchical clustering tree, cutting the tree branches at different heterogeneity levels. DESPOTA performs a permutation test at each node under the null hypothesis that the two descending branches sustain only one cluster. It ensures that the optimal number of clusters is based on the decision made using independent permutation tests, considering the minimum cost required for joining two branches and the cost incurred in the merging process. DESPOTA does not require any distributional assumption and works in a purely data-driven approach. The use of permutations to test for clusteredness in abundance/species data has been proposed by Greenacre (2011). DESPOTA needs a considerable

computational burden, even for moderately large data sets. At each node of the dendrogram, an agglomerative procedure is applied on each branch and for each permutation.

This paper proposes two modifications of the DESPOTA original implementation, aiming to limit the computational effort and favor the applicability of DESPOTA to large data sets. In particular, while the DESPOTA original procedure is purely agglomerative, we propose two variations combining the agglomerative with divisive and partitioning approaches. The divisive approach-based proposal is based only on distances and is suitable for categorical and mixed data. The partitioning-based approach provides further computational efficiency, yet it requires continuous data.

The paper presents some main results concerning a dataset containing some variables that refer to the efficiency and effectiveness of education at Italian universities. These variables are a subset of those that are considered for the five-year periodical accreditation by the Italian national agency (ANVUR).

The paper is organized as follows: Section 2 recalls the DESPOTA test statistic while Section 3 describes the proposed enhancements; Section 4 provides an example and concludes the paper.

2 DESPOTA: general idea and test statistics

Any indexed hierarchy defines a sequence of nested partitions, and at each partitioning, it corresponds to a level of heterogeneity $h(\cdot)$ dictating whether observations/groups are clustered. The choice of $h(\cdot)$ and the corresponding cluster solution is up to the user's expertise and knowledge of the domain.

In order to provide a data-driven choice, Bruzzese & Vistocco (2015) provided a test statistic that evaluates whether two subgroups should be kept separated or merged together. Under the null hypothesis, it is assumed there is no gain in splitting the subgroups at hand. Let us consider a generic dendrogram and let $h(L_k)$ and $h(R_k)$ be, respectively, the left and right branch heterogeneity levels at the node k; then, the test statistic is obtained through the ratio of the minimum cost to the actual cost. Hence, for a generic node k the quantity $h(L_k \cup R_k)$ indicates the heterogeneity level merging the nodes L_k and R_k, and the test statistics is defined as:

$$rc_k = \frac{|h(L_k) - h(R_k)|}{h(L_k \cup R_k) - \min\{h(L_k), h(R_k)\}}, \tag{1}$$

is the ratio between the minimum and actual merging costs, which ranges in $[0,1]$. If rc_k is close to 1 means that L_k and R_k should be kept together.

The null hypothesis distribution is obtained via permutation: at each node k of the original hierarchy, M (usually M=999) permutations of the L_k-vs-R_k membership are considered and the corresponding rc_k values computed.

3 Using permutations to compute the null hypothesis distribution

The computation of quantities in 1 of the shuffled sets under the null hypothesis is a critical point in DESPOTA. In fact, an agglomerative procedure is applied on L_k and R_k to obtain $h(L_k)$ and $h(R_k)$. Finally, the M obtained values for rc_k (see Formula 1) will define the null distribution of the test statistics. For the general node k, the computation of rc_k only involves the second- and third-last aggregation levels. Since the agglomerative approach is bottom-up, the whole hierarchy is needed to compute the second- and third-last aggregation levels. When the complete linkage is considered, given a set $\mathcal{A}$ of observations, the following relation holds: $h(\mathcal{A}) = max(d(i,i'))$, $i,i' \in \mathcal{A}$. In this case, to compute the second- and third-last aggregation levels of the hierarchy, a top-down approach can be used, doing just the first split.

A classic implementation of divisive clustering (see, e.g., DIANA, Kaufman & Rousseeuw, 2009) has a complexity of $O(n^4)$ as opposed to the $O(n^3)$ of agglomerative procedures. Several proposals in the literature enhance the computational performance of divisive approaches, making them substantially more efficient than agglomerative procedures (seeRoux (2018) for a comparative review). To compute the rc_k null distribution, a single step of a divisive approach is used at each permuted node. A further enhancement to split up the permuted nodes is using a partitioning procedure like k-means with careful seeding (Arthur & Vassilvitskii, 2006) to avoid random starts and ensure quality bi-partitions.

4 Example and Conclusions

The considered data for the application consist of three standardized indicators: *iC13* (credits earned at the first year), *iC17* (students graduating up to one year late), and *iC28* (first-year students/faculty members ratio) measured over 68 Italian universities.

Both the agglomerative and the DESPOTA procedures are applied by using the Euclidean metric and the complete linkage aggregation. In Fig1 the results of the two clustering approaches are summarized.

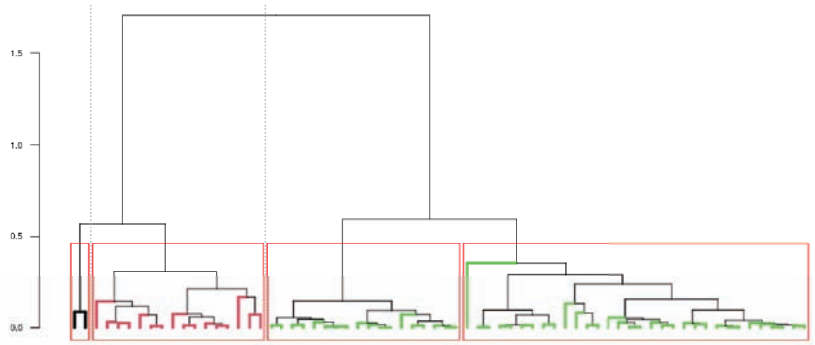


Figure 1. *Comparison between dendrogram cutting rules. The boxes depict the four clusters detected by the classical horizontal rule, while the colored leaves show the clusters selected by DESPOTA.*

DESPOTA and classical hierarchical clustering solutions disagree in the choice of the lower levels of the hierarchy: the horizontal cut splits the large group on the right-hand side of Figure 1, albeit there is no substantial difference between the two groups. DESPOTA sets Bicocca and Bocconi Universities in the same group as they present high values in all the indicators. While the best clustering solution is better interpretable, having a non-subjective procedure to pick a clustering solution is valid, even as a baseline.

References

ARTHUR, DAVID, & VASSILVITSKII, SERGEI. 2006. *k-means++: The advantages of careful seeding*. Tech. rept. Stanford.

BRUZZESE, DARIO, & VISTOCCO, DOMENICO. 2015. DESPOTA: DEndrogram slicing through a pemutation test approach. *Journal of classification*, **32**, 285–304.

GREENACRE, MICHAEL. 2011. A simple permutation test for clusteredness.

KAUFMAN, LEONARD, & ROUSSEEUW, PETER J. 2009. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

ROUX, MAURICE. 2018. A comparative study of divisive and agglomerative hierarchical clustering algorithms. *Journal of Classification*, **35**, 345–366.

MARKOV SWITCHING AUTOREGRESSIVE MODELS FOR THE ANALYSIS OF HYDROLOGICAL TIME SERIES

Roberta Paroli [1] and Luigi Spezia [2]

[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)

ABSTRACT: Markov switching autoregressive models (MSARMs) are proposed here in order to tackle the non-linearity, non-Normality, non-stationarity, and long memory of time series in hydrology. Bayesian inference, model choice, and stochastic variable selection are performed numerically by Markov chain Monte Carlo algorithms. Hence, it is possible to efficiently fit the data, reconstruct the sequence of hidden states, restore the missing values, classify the observations into a few regimes, and select the covariates. The efficiency of MSARMs is demonstrated by applications to isotope signatures, turbidity measurements, and river temperature. Our proposal is very general and flexible and can be applied to any kind of environmental time series.

KEYWORDS: marginal likelihood, non-linearity, non-Normality, non-stationarity, variable selection

1 Introduction and Data

Hydrological time series are realisations of complex stochastic systems. A few issues need to be taken into account by the modellers: non-Normality, non-linearity, non-stationarity, and long memory. These issues can be analysed by Markov switching autoregressive models (MSARMs): a class of models that is a popular tool within the econometrics community to model complex time series but has been considered quite rarely in other disciplines, including environmental sciences. Among the few applications in hydrology, Birkel et al. (2012) modelled isotope signatures; Spezia et al. (2021) turbidity measurements and Spezia et al. (2023) water temperature. In this work, we investigate the dynamic variability of water temperature by analysing an hourly water temperature time series automatically recorded in the Gairn catchment, in the North-East of Scotland, for more than five years, along with some covariates affecting both the latent process (i.e. the time-varying transition probabilities

of the hidden Markov chain) and the observed process. The water temperatures is recorded hourly from 16th August 2012 to 23rd November 2017; the length of the series is 46224 points (i.e. 1926 days; more than five years), with 328 missing values (0.71% of the total number of observations). The range of the series is between -0.02°C and 22.41°C. The contemporary series of the hourly river flows is also available. We also studied an intermediate series of water temperature from 13th June 2014 to 31st August 2016 (19440 observations; 810 days; more than two years) with 209 missing values (1.08%) along with three covariates (flow, air temperature, rainfall). Finally, a short series was considered: 1200 observations (50 days) with no missing values recorded from 18th August to 6th October 2012 along with seven covariates (flow, air temperature, rainfall, wind speed, wind direction, radiation, soil temperature). The length of series of the exogenous variables was limited by the need to not have missing values in these deterministic sequences. This because missing values within the covariates might bias the results of our analyses.

We propose MSARMs within the Bayesian framework: inference, model choice, and variable selection are performed numerically by Markov chain Monte Carlo (MCMC) algorithms.

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.

Seven covariates were also incorporated into the model through both the hidden Markov chain (the transition probabilities are time-varying and dependent on the dynamics of these exogenous variables) and the observed process (the state-dependent exogenous variables are added to the past observations). Thus, we have time-varying means and autocovariances, and hence, a non-stationary model. The covariates are: river flow, air temperature, rainfall, wind speed, wind direction, radiation, and soil temperature. The data set is also

characterised by periodicities: the hourly temperatures vary according to the dynamics of the year and of the 24 hours of the day. Hence, both an annual and a state-dependent daily harmonic component are added to the observed process.

In the Bayesian framework, inference, model choice, and variable selection are performed numerically by MCMC algorithms. The basic scheme for parameter estimation in the observed process is Gibbs sampling which also allows both restoration of the missing values occurring within the series of observations and reconstruction of the sequence of hidden states. Two random walk Metropolis moves are used to estimate the parameters of the hidden Markov chain. Adding extra-steps to the basic Metropolis-within-Gibbs scheme we can also compute the marginal likelihood of the various competing models through the MCMC sample. This procedure enables us to select the best model within a set of models varying for the number of hidden states and the order of the autoregressive processes. The exogenous, deterministic variables appearing in the observed process may be different in any state and they may be different from those affecting the transition probabilities. The transition matrix is affected by two sets of covariates (possibly different from each other and different from those in the observed process), one for the transitions from a lower to a higher state, and another for the transition from a higher to a lower state. The selection of the covariates appearing in each state-dependent autoregression and in the transition matrix is performed stochastically through the Metropolised-Kuo-MallicK (MKMK) method, proposed by Paroli and Spezia (2008). In the case of non-homogeneous hidden Markov models and MSARMs with covariates, the MKMK method improves the performance of the competing techniques, especially when the explanatory variables are strongly correlated, and/or when the complexity of the model is high.

3 Results

The flexibility of the MSARMs is demonstrated by the three applications we considered. For the whole series with a single covariate, the best model has three hidden states and autoregressions of the fifth order. Thus, the non-linear model (three hidden states) worked better than the corresponding linear model (no hidden states). Flow is relevant in the observed process for two states only, while it is not selected in the hidden process and the Markov chain is homogeneous. For the intermediate series with three covariates, the best model has three hidden states and autoregressions of the sixth order. Again, the non-linear model (three hidden states) works better than the corresponding linear

model (no hidden states). Flow is relevant in the observed process for one state only, while air temperature is always selected both in the observed and the hidden process. For the short series with seven covariates, we obtain that the best model is the linear autoregression of the sixth order, with no hidden Markov chain behind. Air temperature, solar radiation, and soil temperature are the relevant variables to explain the water temperature dynamics. Thus, discharge is a proxy for water temperature modelling, when no other more directly related variables are available. In those situations, the latent states will help to model the long-term dynamics, in the absence of true predictors with a physical meaning. As we saw in the first two applications, the hidden regimes can have an interpretation related to the seasonality. In fact, the Markov chain shows an annual dynamics which anticipates the annual dynamics of the water temperatures. It is not surprising that for the short series (50 days, i.e. no annual periodicity) the model is not multi-state. It would be interesting to see what happens when considering the seven covariates on longer series, that is if the same covariates are selected in a non-linear model (i.e., with a multi-state hidden Markov chain). Our study provides a novel application of the suitability of the MSARMs in hydrological time series analysis and environmental sciences in general. We hope our work can motivate other scientists to approach MSARMs and give their highly structured time series a valuable interpretation.

References

BIRKEL C., PAROLI, R. SPEZIA L. DUNN S.M. TETZLAFF D., & SOULSBY, C. 2012. A new approach to simulating stream isotope dynamics using Markov switching autoregressive models. *Advances in Water Resources*, **26**, 308–316.

PAROLI, R., & SPEZIA, L. 2008. Bayesian variable selection in Markov mixture models. *Communications in Statistics - Simulation and Computation*, **37**, 25–47.

SPEZIA, L., GIBBS S. GLENDELL M. HELLIWELL R. PAROLI R., & POHLE, I. 2023. Bayesian analysis of high frequency water temperature time series through Markov switching autoregressive models. *Under second revision for Environmental Modelling & Software*.

SPEZIA, L., VINTEN A. PAROLI R., & STUTTER, M. 2021. An evolutionary Monte Carlo method for the analysis of turbidity high-frequency time series through Markov switching autoregressive models. *Environmetrics*, **32**, e2695.

A CASE STUDY OF ELECTRONIC MEDICAL RECORDS USE FOR PREDICTING KIDNEY INJURY

Davide Passaro [1], Luca Tardella [1], Giovanna Jona Lasinio [1], Tiziana Fragasso [2], Valeria Raggi [2] and Zaccaria Ricci [3]

[1] Department of Statistical Sciences, Sapienza University of Rome, (e-mail: davide.passaro@uniroma1.it)

[2] Department of Cardiology and Cardiac Surgery, Pediatric Cardiac Intensive Care Unit, Bambino Gesù Children's Hospital, IRCCS, Rome, Italy

[3] Meyer University Hospital, University of Florence, Italy

ABSTRACT: We present a case study concerning the use of electronic medical records (EMRs) acquired in an intensive care Unit (ICU). In particular, we focus on the problem of exploiting this emerging new type of data for predicting Acute Kidney Injury (AKI), a frequent complication in hospitalized patients during patient stay using data collected in the Pediatric Cardiac Intensive Care Unit of Bambino Gesù Childen's Hospital. We discuss the methodological issues related to pre-processing the available EMR data, analyze the possible alternative ways of defining the outcome and use different tools for making predictions.

KEYWORDS: 'EMR', 'classification', 'forecast', 'predictive model'.

1 The challenges of electronic medical records

In the last thirty years, the development of technologies has favored the development of Electronic Medical Records (EMRs). EMRs are the digital version of a patient's paper chart. EMRs are real–time, patient–centered records that make information available instantly and securely to authorized users. A database with Electronic Health Records contains patient data recorded to varying levels of granularity.

The trend of adoption of digital health record systems in hospitals seems to be clear and no longer deferrable (Collins & Tabak, 2014). The increasingly widespread presence of this new type of data has involved the development of research with the aim of using this data to support doctors' decisions.

Hodgson *et al.*, 2019 observe that, within health care, clinical decision support systems (CDSS) are increasingly being introduced with the aim to provide pertinent information, intelligently filtered or presented at appropriate times, to enhance care and potentially improve outcomes.

Indeed, Electronic health records contain valuable data for identifying health outcomes, but these data also present numerous challenges. In fact, despite the progress realized in recent years, the EMRs data suffer yet of no standardization problem in measurements acquisition in the particular case of Intensive Care Unit (ICU).

Statistics and Machine learning methods could help with some of these challenges (Wong *et al.*, 2018). As highlighted by Shafaf & Malek, 2019, the use of statistical methods as well as artificial intelligence and machine learning techniques in different medical fields are rapidly growing, in particular for the case of prediction and early detection of disease.

We describe a case study of use of EMRs using the data collected by the Pediatric Cardiac Intensive Care Unit (PCICU) of Bambino Gesù Childen's Hospital focusing on the problem of predicting Acute Kidney Injury (AKI) beforehand. Our study involved patient records extracted from January 2018 to February 2020. All the data extracted by the EMR have been anonymized.

2 The case of acute Kidney injury prediction

AKI is an increasingly common clinical problem associated with mortality, length of stay, and healthcare cost. In light of the impact of AKI on short and long-term outcomes, it is of high importance to develop methods to identify when patients are at risk for AKI and to diagnose subclinical AKI in order to improve patient outcomes.

For these reasons, we focus our work on the objective of predicting the AKI defined according to the AKI stage criteria (described in Khwaja, 2012). We adopt a continuous forecasting approach of the state of AKI throughout the hospital stay with a time advance of 48 hours.

In the initial phase, we work on data selection, extraction, and management of missing data. In particular, according to the literature and the clinicians, we use a selection of objectively collected variables available in the EMR data grouped into the following groups: admission and post-surgical data, vital signs, fluids, blood gas analysis, laboratory analysis, and therapies administered. Since a pediatric patient admitted to intensive care can be subjected, although not frequently, to more than one surgical and/or hemodynamic procedure during the same hospitalization we decided to select the following subset of the dataset:

- only patient in pediatric age ($\leq$ 18 years) with a length of hospitalization greater than 48h

	RF (all var)	RF using RFE	GAM (all var)	BN (all var)	BN (MMPC)
AUC bin AKI	0.93 (0.92-0.94)	0.95 (0.94-0.96)	0.87 (0.85-0.89)	0.90 (0.88-0.92)	0.90 (0.88-0.92)
AUC severe AKI	0.99 (0.98-1)	0.98 (0.97-0.99)	0.94 (0.92-0.96)	0.97 (0.96-0.98)	0.97 (0.96-0.98)
Accuracy Max AKI	0.92 (0.91-0.93)	0.93 (0.92-0.93)	0.87 (0.86, 0.89)	0.88 (0.87-0.89)	0.87 (0.86-0.88)
Accuracy Mode AKI	0.95 (0.94-0.96)	0.96 (0.95-0.97)	0.90 (0.89, 0.91)	0.92 (0.91-0.93)	0.92 (0.91-0.92)

Table 1. *Summary of results of AKI prediction using RF, GAM, BN.*

- only the temporal data between admission and discharge date from PCICU or before the start of a second surgery.

For groups of variables for which there was missing data (blood gas analysis and vital signs) we assume the origin of the missing data is missing at random (MAR). Starting this assumption we use a nonparametric missing value imputation using Random Forest provided by MissForest R Package (Stekhoven & Bühlmann, 2011). We discretize all the different acquisition frequencies in a common sample frequency of $\Delta t = 6$ hours. Finally, we define different types of outcome grouping the AKI stage in the binary and multiclass way.

In the second phase, we develop different classification models: random forest (RF), Generalized Additive Method (GAM), and Bayesian Network (BN). In all the cases we split the dataset into train (70%) and test (30%) sets. The former is used to fit the classification model, whereas the latter is employed to evaluate its performance. In splitting the data, we preserve the percentages of each class in train and test sets.

The overall performances reported in Table 1 are evaluated using the standard measures as Area under the ROC Curve (AUC-ROC) for binary cases and accuracy and kappa for the multiclass cases. The obtained results are always good compared with other recent attempts in the literature (Gameiro *et al.*, 2020).

We use, furthermore, different techniques of variable selection. In the case of RF, we applied Recursive Feature Elimination RF as described in Chen *et al.*, 2020. In BN cases we use the Max-Min Parents and Children algorithm (MMPC) as described in Lagani *et al.*, 2017. The list of the most important variables obtained in the various classifications confirm the importance of some of the variables (such as creatinine) reported in other studies in the literature but also highlights the presence of variables that are specific to pediatric patients under examination (such as Pediatric Index of Mortality).

All implemented models confirm the possibility of making an accurate pre-

diction of the AKI stage using the PCICU. These models can be potentially included in a web interface and, in perspective, be integrated into the EMR of PCICU. This tool would allow the doctors to predict prospectively the patient's stage of AKI and evaluate how to intervene if necessary. In order to proceed with this, it would be necessary for the future to implement the export of a larger dataset adding new data acquired in the meantime in PCICU.

References

CHEN, RUNG-CHING, DEWI, CHRISTINE, HUANG, SU, & CARAKA, REZZY. 2020. Selecting critical features for data classification based on machine learning methods. *Journal Of Big Data*, **7**(07), 26.

COLLINS, F., & TABAK, L. 2014. Using machine learning to identify health outcomes from electronic health record data. *Nature*, **505**, 612–613.

GAMEIRO, JOANA, BRANCO, TIAGO, & LOPES, JOSÉ. 2020. Artificial Intelligence in Acute Kidney Injury Risk Prediction. *Journal of Clinical Medicine*, **9**(03), 678.

HODGSON, LUKE, SELBY, NICHOLAS, HUANG, TAO-MIN, & FORNI, LUI. 2019. The Role of Risk Prediction Models in Prevention and Management of AKI. *Seminars in Nephrology*, **39**(09), 421–430.

KHWAJA, A. 2012. KDIGO Clinical Practice Guidelines for Acute Kidney Injury. *Nephron Clin Pract*, 179–184.

LAGANI, VINCENZO, ATHINEOU, GIORGOS, FARCOMENI, ALESSIO, TSAGRIS, MICHAIL, & TSAMARDINOS, IOANNIS. 2017. Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. *Journal of Statistical Software*, **80**(7), 1–25.

MAKOUL, G., CURRY, H., R., & TANG, P. C. 2001. The use of electronic medical records: communication patterns in outpatient encounters. *Journal of the American Medical Informatics Association : JAMIA*, **8**(6), 610–615.

SHAFAF, NEGIN, & MALEK, HAMED. 2019. Applications of Machine Learning Approaches in Emergency Medicine; a Review Article. *Archives of academic emergency medicine*, **7**(06), 34.

STEKHOVEN, DANIEL J., & BÜHLMANN, PETER. 2011. MissForest—nonparametric missing value imputation for mixed-type data. *Bioinformatics*, **28**(1), 112–118.

WONG, J., M., HORWITZ M., ZHOU, L., & TOH, S. 2018. Using machine learning to identify health outcomes from electronic health record data. *Current epidemiology reports*, **5**(4), 331–342.

PERSONALIZED TREATMENT SELECTION MODEL FOR SURVIVAL OUTCOMES

Matteo Pedone [1], Raffaele Argiento[2] and Francesco C. Stingo[1]

[1] Department Statistics, Computer Science and Applications, University of Florence, (e-mail: matteo.pedone@unifi.it, francescoclaudio.stingo@unifi.it)

[2] Department of Economics, University of Bergamo, (e-mail: raffaele.argiento@unibg.it)

ABSTRACT: Precision medicine, a patient-centric approach to disease treatment, has attracted considerable interest in recent years. Building on a prior method focused on short-term outcomes, we introduce a model that clusters patients based on similar predictive characteristics and treatment responses, enabling optimal therapeutic strategy selection via predictive inference for new patients, incorporating long-term survival outcomes.

KEYWORDS: Nonparametric Bayes, personalized medicine, predictive probability, product partition models, time-to-failure endpoints.

1 Introduction

The field of oncology has shifted towards personalized treatments that take into account the heterogeneity of cancer pathogenesis. This is driven by the recognition that cancer molecular mechanisms are complex and multifactorial, involving multiple biomarkers and pathways. To address this complexity, the focus has shifted towards developing therapies that are based on multiple biomarkers.

Statistical methods for personalized treatment selection need to consider the uniqueness of each tumor and individual patient characteristics. The common assumption of statistical exchangeability among patients should be relaxed, and patients should only be treated as exchangeable to the extent to which their tumors are molecularly similar. By leveraging individual patient

Matteo Pedone gratefully acknowledges the support of the European Union - Next GenerationEU, UNIFI Young Independent Researchers Call - BayesMeCOS Grant no. B008-P00634.

characteristics, these personalized treatment strategies have the potential to improve treatment efficacy and patient outcomes.

? proposed a hybrid two-step approach for accurate treatment selection that integrates a Bayesian predictive model with prognostic and predictive biomarkers. Patients are grouped based on molecular similarity using heuristic clustering algorithms, and a Bayesian model predicts treatment response probabilities for each competing treatment. This approach improves upon existing methods by relaxing the assumption of full exchangeability among patients and utilizing complementary sources of information.

? developed a fully Bayesian method that builds upon **?**'s approach and improves upon it by jointly performing clustering and prediction using a nonparametric approach. By combining the two tasks into a single model, the need for multi-step procedures is eliminated. The nonparametric approach provides a sound framework for both clustering and prediction, accounting for the uncertainty in all modeling steps. Ultimately, the individualized treatment rule fully accounts for patients' heterogeneity, as confirmed by improved prediction performances compared to **?**'s method (**?**).

? used a categorical outcome to evaluate treatment effectiveness after a post-therapy follow-up period. However, this approach may be limited since it only considers short-term outcomes. To address this limitation, we suggest using time-to-event analysis to base treatment selection on long-term outcomes such as disease progression, relapse, or death. This approach can better capture treatment effectiveness.

2 Survival model

We examine a group of n patients from past clinical studies who were treated with T different treatments. The patients' predictive and prognostic biomarkers were measured, along with the survival times. The treatments are indexed by $a = 1, \ldots, T$, and the total number of treated patients is denoted by $n = \sum_{a=1}^{T} n^a$, where n^a is the number of patients receiving therapy a. The observed survival times of patients are represented as vectors $\boldsymbol{t}^a,\ a = 1, \ldots, T$. However, due to limited study duration, not all patients experience the event of interest, resulting in "right-censored" data. To account for this, binary indicator vectors $\boldsymbol{d}^a,\ a = 1, \ldots, T$ are introduced to identify patients whose event was observed during follow-up and those who were censored. In the case of patients who received treatment a and experienced an observed event or censored time ($d_i^a = 1$), their time to event is represented by t_i^a. On the other hand, if $d_i^a = 0$, meaning that the patient did not experience an event or was not cen-

sored during the study period, then t_i^a represents the length of their follow-up.

We use an accelerated failure time (AFT) model to analyse right-censored survival data, that takes into account the $p-$ and $q-$ dimensional vector of prognostic and predictive features. Prognostic and predictive markers are denoted as $\boldsymbol{z}_i^a$ and $\boldsymbol{x}_i^a$, respectively, measured on the $i-$th patient who received treatment a. It is assumed that patients with similar genetic profiles are likely to have similar responses to a given treatment. We assume that $\Pi_{n^a}^a = S_1^a, \ldots, S_{C_{n^a}^a}^a$ is a given treatment-specific partition of the indices $1, \ldots, n^a$, where $C_{n^a}^a$ is the number of clusters among patients treated with therapy a, and $n_j^a = |S_j^a|$ is the number of patients in cluster j, for $j = 1, \ldots, C_{n^a}^a$.

$$\log(t_i^a) = \mu_j^{a\star} + \boldsymbol{z}_i\boldsymbol{\beta} + \sigma\varepsilon_i,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the vector of regression coefficients and $\boldsymbol{\varepsilon}$ is the error vector. Assuming a minimum value Gumbel distribution for the error terms $\varepsilon_1, \ldots, \varepsilon_n \sim \text{Gumbel}(0,1)$, gives rise to the Gumbel AFT model. Moreover, we assume $\beta_k \overset{\text{iid}}{\sim} N(0, \lambda_k^2\tau^2), \lambda_k, \tau \overset{\text{iid}}{\sim} HC(0, 1/p)$ (namely, a horseshoe prior), and $\sigma \sim U(a_\sigma, b_\sigma)$. Moreover, $\mu_j^{a\star}$s are cluster-specific parameters. We assume a product partition model with covariates (PPMx, **?**) for the joint distribution of the clustering and the cluster-specific parameters $(\Pi_{n^a}^a,\ \mu_j^{a\star})$, to induce independence across clusters and conditional independence within clusters. The joint law of $(\Pi_n, \mu_j^{a\star})$ is assigned hierarchically as:

$$\begin{aligned} \mu_j^{a\star} \mid \boldsymbol{\zeta}, \Pi_{n^a}^a &\overset{\text{ind}}{\sim} N(\theta, \boldsymbol{\Sigma}) \\ \Pi_{n^a}^a &\sim PPMx(\boldsymbol{x}). \end{aligned}$$

All the details pertaining to the specification, construction, and posterior inference of PPMx can be found in **?**.

3 Example

We conduct a simulation study to evaluate the properties of the proposed method on finite samples. We consider 200 patients assigned to two competing treatments and use piecewise constant exponential distributions to generate survival outcomes, that is we do not simulate from our model. Our simulation design is inspired by the work of **?**. To evaluate the performance of the clustering procedure, we generate synthetic covariates (5 predictive and 5 prognostic biomarkers) with a known clustering structure (a two-component mixture of normals). The validation set comprised 100 patients. In Table **??** we report the

Adjusted Rand Index (ARI), the Mis-assigned Optimal Treatment (MOT), and Proportion of Treatment Utility (PTU) averaged over 10 replications (standard deviation in parenthesis).

ARI	MOT	PTU
1.00	7.50	0.93
(0.00)	(5.92)	(0.06)

Table 1: Results simulation study.

In terms of clustering, the model demonstrates a remarkable level of effectiveness. The quantity of non-optimal treatment assignments is approximately 8 per 100 patients, with a considerable standard deviation. However, the high PTU value suggests that the misassigned patients may belong to a subgroup with similar treatment benefits across therapies.

4 Conclusion

Overall, our study's preliminary results are promising and suggest that the proposed method has potential for accurately assigning treatments using long-term outcomes.

VARIABLE RANKING IN BIVARIATE COPULA SURVIVAL MODELS

Danilo Petti [1], Marcella Niglio[2] and Marialuisa Restaino[2]

[1] Department of Mathematical Sciences, University of Essex, (e-mail: d.petti@essex.ac.uk)

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

ABSTRACT: We propose a variable ranking procedure based on copula bivariate time-to-event margins under a general censoring scheme. The procedure identifies the important variables influencing the two time-to-events in a high dimensional setting introducing a proper metric able to take into account the probabilistic copula structure. The proposal is the first attempt to apply a variable selection method to a copula bivariate time-to-event domain. The advantages of the proposed approach are illustrated in a case study based on AREDS dataset.

KEYWORDS: bivariate survival, copula, variable ranking, ultrahigh dimensionality.

1 Introduction

Technologies have had a deep impact on society and on data collection in a wide range of scientific areas. With a relatively low cost, we are able to collect massive amounts of information (and noise). This has led to the high dimensional data phenomenon where the variable selection plays a central role. This is even more true in the case of bivariate copula survival models under a censoring scheme (presence of two outcomes and missing information). Under this domain, we are interested in identifying two sets of relevant covariates for two random times to event (T_1 and T_2). This can be achieved by ranking the covariates in order of importance through a given metric ω to assess the contribution of each feature in the dataset. As far as the authors are aware, there is no valuable variable selection or variable ranking method nor implementation available in the literature for Bivariate Copula Survival models. In Sect. 2 we shortly present the model under analysis, and in Sect. 3 we sketch the algorithm of variable ranking. The application to AREDS data is presented in Sect. 4.

2 The model

Let us consider the pair of survival times (T_{1i}, T_{2i}), a vector of covariates $\mathbf{x}_i$, for $i = 1, 2, \ldots, n$, and an associated generic parameter vector $\delta \in \mathbb{R}^w$ of dimension w. We assume that T_{1i} and T_{2i} have marginal survival functions given by $S_v(t_{vi}|\mathbf{x}_{vi};\beta_v) = P(T_{vi} > t_{vi}|\mathbf{x}_{vi};\beta_v) \in (0,1)$, for $v = 1,2$, and a joint survival function expressed as follows $S(t_{1i}, t_{2i}|\mathbf{x}_i;\delta) = C(S_1(t_{1i}|\mathbf{x}_{1i};\beta_1), S_2(t_{2i}|\mathbf{x}_{2i};\beta_2)$ $;m\{\eta_{3i}(\mathbf{x}_{3i};\gamma)\})$, where $\delta^T = (\beta_1^T, \beta_2^T, \gamma^T)$, $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$ and $\mathbf{x}_{3i}$ are vectors of covariates, with associated coefficient vectors $\beta_1 \in \mathbb{R}^{w_1}$, $\beta_2 \in \mathbb{R}^{w_2}$ and $\gamma \in \mathbb{R}^{w_3}$ such that $w = w_1 + w_2 + w_3$, $C : (0,1)^2 \to (0,1)$ is a uniquely defined 2-dimensional copula function with coefficient $\theta_i = m\{\eta_{3i}(\mathbf{x}_{3i};\gamma)\}$ modelling the potentially varying dependence of (T_{1i}, T_{2i}) across observations, $\eta_{3i}(\mathbf{x}_{3i};\gamma) \in \mathbb{R}$ is a predictor which includes generic additive covariate effects, and m is a monotonic and differentiable one-to-one transformation function. The marginal survival functions can be written as

$$g_v\left[S(t_{vi}|\mathbf{x}_{vi};\beta_\nu)\right] = \eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \mathbf{f}_v(\beta_v)),\ \nu = 1,2 \tag{1}$$

where $g_v : (0,1) \to \mathbb{R}$ is a monotone and twice continuously differentiable link function with bounded derivatives, $\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \mathbf{f}_v(\beta_v)) \in \mathbb{R}$ is an additive predictor which models the baseline hazard and several types of covariate effects, and $\mathbf{f}_v(\beta_v)$ has the role of imposing a monotonicity constraint. Equation 1 can be written as $S(t_{\nu i}|\mathbf{x}_{\nu i};\beta_\nu) = G_\nu(\eta_{\nu i}(t_{\nu i}, \mathbf{x}_{\nu i}, \mathbf{f}(\beta_\nu)))$ where G_ν is an inverse link function. The key difference between $\eta_{vi}(t_{vi}, \mathbf{x}_{vi}; \mathbf{f}_v(\beta_v))$, for $v = 1,2$, and $\eta_{3i}(\mathbf{x}_{3i};\gamma)$ is that the two former predictors must include smooth functions of times t_{vi} which can be treated as regressors. We, therefore, consider a generic $\eta_{\nu i}(\nu = 1,2,3)$, where the dependence on the covariates and parameters is momentarily dropped, an overall covariate vector $\mathbf{z}_{\nu i}$ containing $\mathbf{x}_{\nu i}$ and $t_{\nu i}$ when $\nu = 1,2$, and $\mathbf{z}_{3i} = \mathbf{x}_{3i}$. For simplicity, the dimensions of $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ are assumed to be W_1 and W_2. A generic additive predictor is specified as follows

$$\eta_{\nu i} = \beta_{\nu 0} + \sum_{k_\nu=1}^{K_\nu} s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu i}),\ \nu = 1,2,3 \tag{2}$$

where $\beta_{\nu 0} \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{\nu k_\nu i}$ denotes the k_ν^{th} sub-vector of the complete vector $\mathbf{z}_{\nu i}$ and the K_ν functions $s_{\nu k_\nu}(\mathbf{z}_{\nu k_\nu i})$ represent generic effects which are chosen according to the type of covariate(s) considered (Wood, 2017). The above formulation allows for many types of flexible covariate effects. For more details see Marra, 2020.

3 The Variable Selection Algorithm

We extend the variable selection procedure proposed by Baranowski, 2020 to the bivariate survival data. Given the set of w covariates, the variables with higher influence on $\eta_{\nu i}(x_{\nu i}, \beta_{\nu i})$, $(\nu = 1,2)$ are those that even in presence of randomly selected sub-samples exhibit consistent relationship to explain the dependence of the two survival functions.

Let $Z_i = \{T_{1i}, T_{2i}, X_{i1}, X_{i2}, \ldots, X_{iw}\}$, for $i = 1,2,\ldots,n$ and with w that grows with n, be the observed dataset used to select the subset of covariates $\{X_1, \ldots, X_w\}$. Further, let $\mathcal{A}^\nu \subset (1,\ldots,w_\nu)$ for $\nu = 1,2$ be the indices that identify a subset of covariates for the ν-th margin and let $|\mathcal{A}^\nu| = k$ be the cardinality of $\mathcal{A}^\nu$, for $k = 0,1,\ldots,w_\nu$. Let $R^s_{nj}(Z_1,\ldots,Z_n)$ be the ranking of the j-th covariate, based on a metric $\hat{\omega}^\nu_j = \hat{\omega}^\nu_j(Z_1,\ldots,Z_n)$ assessing the importance of each covariate of each margin, such that $\omega^\nu_{R^s_{n1}} \geq \cdots \geq \omega^\nu_{R^s_{n|\mathcal{A}^\nu|}}$. The probability of the set of $|\mathcal{A}^\nu|$ top-ranked variables in $\mathcal{A}^\nu$ is:

$$\pi_{n,m}(\mathcal{A}^\nu) = \mathbb{P}\left(\left\{R^\nu_{n1}(Z_1,\ldots,Z_m),\ldots,R^\nu_{n|\mathcal{A}^\nu|}(Z_1,\ldots,Z_m)\right\} = \mathcal{A}^\nu\right), \nu = 1,2 \tag{3}$$

that is obtained from a subset of m observations, with $1 \leq m \leq n$. To estimate 3 a bootstrap approach is proposed in Baranowski, 2020. It follows that $\pi_{n,m}(\mathcal{A}^\nu)$ is the probability that the covariates in $\mathcal{A}^\nu$ are ranked at the top, using a subset of m observations. The selection can be then performed on the set of top-ranked variables $\mathcal{A}^\nu$ from which the number of terms $\hat{s}^\nu$ can be determined using equation (2.5) in Baranowski, 2020. In practice, given the estimated probabilities of $\hat{\pi}_{n,m}\left(\hat{\mathcal{A}}^\nu_{k,m}\right)$, for $k = 0,\ldots,k_{\max}-1$, with $k_{\max}$ a fixed large integer, the number of relevant variables is related to the evaluation of the magnitude of the estimated probability.

4 Application to AREDS dataset

The performance of the algorithm in Sect. 3 is assessed using the AREDS data (available in the `R` package `CopulaCenR`). The dataset includes 629 Caucasian participants. The event of interest is late-AMD progression, which is a disease affecting both eyes. Less than 50% of the subjects had late-AMD in both eyes (bivariate interval-censored). Around 20% had late-AMD in one eye but not the other by the study end (mixed interval- and right-censored). More than 33% did not develop late-AMD in either eye (bivariate right-censored). The variables are Severity score, values that reflect the progression of the disease in the eyes (`SevScale1E` `SevScale2E`), enrollment age (`Age`), and a

genetic variant (`rs2284665`), factor variable with levels 0 (`GG`), 1 (`GT`) and 2 (`TT`), respectively). Furthermore, the AREDS dataset has been perturbed by adding 100 independent realizations of a standard Gaussian distribution. For sake of completeness, the algorithm has also been evaluated through a Monte Carlo study (not included in the paper), which confirmed the effectiveness of the method returning false positives and negatives close to zero.

We carried out some preliminary fitting from which emerged that {`C0`, `POPO`} is the combination with the lowest BIC (4330.08) considering a full model specification (all features included in all three margins). The procedure has been applied on a standardized version of the dataset, where `rs2284665` has been encoded as $0/1$, resulting in three new covariates. The tuning parameters has been specified as follows: $k_{max} = 10$, $m = \lfloor n/2 \rfloor$, $\tau = 0.5$, 50 bootstrap replicates, Clayton copula (`C0`) and Proportional odds (`PO,PO`). We have considered two metrics: $\omega_\nu = \beta_j^2 i(\beta_j)$ (with $i(\beta_j)$ be the associated element of the Fisher information matrix) and $\psi_\nu = |\beta_j|$. In pseudo code (ignoring the smooth functions of times t_ν) $\eta_\nu = \beta_{\nu 0} + \beta_{\nu j} x_j$ for $j = 1, \ldots, w$. The former metric is proposed specifically for the class of Bivariate Copula Survival models while the latter is the absolute value of the coefficient.

Comparing the variable selected with the two metrics, the selection with $\beta_j^2 i(\beta)$ has greater cardinality and is able to select those characteristics considered relevant for the event of interest in the literature (see Sun, 2021), giving empirical evidence of its goodness.

Table 1. *Results of the algorithm in Sect. 3 using Clayton copula, proportional hazard margins and using* $\omega_\nu = \beta_j^2 i(\beta)$ *and* $\psi_\nu = |\beta_j|$*, for* $j = 1, \ldots, w$*, as metrics. The covariates are ordered according to their importance. The BIC and AIC are obtained by applying* `gjrm()` *function to a non-standardized AREDS.*

	$\hat{\mathcal{A}}^1$	$\hat{\mathcal{A}}^2$	BIC	AIC
$\mathcal{M}_\omega$	{SevScale1E,SevScale2E,GG,TT}	{GG,SevScale2E,TT,SevScale1E,GT}	4325.849	4225.385
$\mathcal{M}_\psi$	{SevScale1E,SevScale2E}	{SevScale2E,SevScale1E,GG,TT,GT}	4324.518	4220.659

References

BARANOWSKI, R., CHEN Y. FRYZLEWICZ P. 2020. Ranking-based variable selection for high-dimensional data. *Stat. Sinica*, **30**(3), 1485–1516.

MARRA, G., RADICE R. 2020. Copula link-based additive models for right-censored event time data. *J. of the Am. Stat. Ass.*, **115**(530), 886–895.

SUN, T., DING Y. 2021. Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostat.*, **22**(2), 315–330.

WOOD, S. N. 2017. *Generalized additive models: an introduction with R*. CRC Press.

ROBUST PENALIZED MULTIVARIATE ANALYSIS FOR HIGH-DIMENSIONAL DATA

Pia Pfeiffer [1] and Peter Filzmoser[1]

[1] Institute of Statistics and Mathematical Methods in Economics, TU Wien, (e-mail: pia.pfeiffer@tuwien.ac.at, peter.filzmoser@tuwien.ac.at)

ABSTRACT: High-dimensional data sets, with fewer observations than variables, pose a challenge for statistical methods, particularly if outlying observations are present. Several proposals for robust and sparse estimation in the context of multivariate statistical methods are available, together with algorithms for the computation. We present a unified computational approach based on reformulating the problem as a constrained optimization problem, also incorporating sparsity constraints. Recent developments with adaptive gradient descent algorithms can efficiently solve such problems, and they are also scalable with data dimensionality. The procedures are illustrated in the example of canonical correlation analysis, where also higher-order directions can be directly computed, and the sparsity can be controlled easily. Extensions to other multivariate methods are possible.

KEYWORDS: high-dimensional data, robust multivariate analysis, sparse multivariate analysis.

1 Introduction

Classical methods for multivariate analyses, such as PCA (Principal Component Analysis), CCA (Canonical Correlation Analysis), and LDA (Linear Discriminant Analysis), are based on covariance estimation and aim to find projection directions in the data according to some criteria. This estimation procedure is not suitable for high-dimensional data sets, and therefore sparse methods have been proposed, e.g. by applying elastic net type penalties (Zou & Hastie, 2005) for the projection directions. Such methods are sensitive to outlying observations, and therefore methods combining sparsity with robust estimation have been proposed. In the context of CCA, for example, Wilms & Croux, 2015 suggest using alternating regressions with sparse and robust regression estimators. A disadvantage of this approach is that higher-order directions cannot be derived directly.

2 Methodology

In the example of CCA, we show how the objective can be reformulated as an optimization problem, directly stating the optimization conditions and offering a flexible choice of covariance estimator and penalty function. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote a p- and q-dimensional random variable, respectively, and $\boldsymbol{\Sigma}_{xx}, \boldsymbol{\Sigma}_{yy}$ and $\boldsymbol{\Sigma}_{xy}$ the corresponding covariance matrices. The first canonical correlation coefficient ρ_1 and the first pair of canonical vectors $(\boldsymbol{a}_1, \boldsymbol{b}_1)$ are given as a solution of the optimization problem

$$\max_{\boldsymbol{a}\in\mathbb{R}^p, \boldsymbol{b}\in\mathbb{R}^q} \boldsymbol{a}'\boldsymbol{\Sigma}_{xy}\boldsymbol{b} \tag{1}$$

under the constraints

$$\boldsymbol{a}'\boldsymbol{\Sigma}_{xx}\boldsymbol{a} = 1 \quad \textit{and} \quad \boldsymbol{b}'\boldsymbol{\Sigma}_{yy}\boldsymbol{b} = 1. \tag{2}$$

The k-th canonical correlation coefficient ρ_k and the respective pair of canonical vectors $(\boldsymbol{a}_k, \boldsymbol{b}_k)$ maximize (1) under the condition that they are uncorrelated with the previous $k-1$ directions, denoted as the constraints

$$\boldsymbol{a}'\boldsymbol{\Sigma}_{xx}\boldsymbol{a}_i = 0 \quad \textit{and} \quad \boldsymbol{b}'\boldsymbol{\Sigma}_{yy}\boldsymbol{b}_i = 0, \text{ for } i = 1, \ldots, k-1. \tag{3}$$

Penalty terms are added as further constraints for a sparse setting,

$$P_{\alpha_1}(\boldsymbol{a}) \leq c_1 \quad \textit{and} \quad P_{\alpha_2}(\boldsymbol{b}) \leq c_2 \tag{4}$$

where c_1 and c_2 denote positive constants, and the penalty terms (4) are given as elastic net (Zou & Hastie, 2005) penalties with mixing parameters α_1, α_2.

The augmented Lagrangian with $\boldsymbol{\lambda}$ denoting the Lagrange multipliers, and H summarizing the constraints, is then given as

$$\mathcal{L}_\rho(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\lambda}) = -|\boldsymbol{a}'\boldsymbol{\Sigma}_{xy}\boldsymbol{b}| + \boldsymbol{\lambda}' \cdot H(\boldsymbol{a}, \boldsymbol{b}) + \frac{\rho}{2}\|H(\boldsymbol{a}, \boldsymbol{b})\|_2^2. \tag{5}$$

Then, a solution to (1)-(4) can be found by minimizing (5). For the optimization algorithm, the method of multipliers (see e.g. Bertsekas, 1982) is combined with an adaptive gradient descent algorithm as described by Reddi *et al.*, 2018 for an alternating update of $(\boldsymbol{a}, \boldsymbol{b})$ and $\boldsymbol{\lambda}$.

Our approach is not only flexible in the choice of covariance estimator and penalty type, but we can also directly state the necessary conditions for higher-order canonical correlations. The robustness of the resulting canonical

correlations can be controlled by an appropriate choice of covariance estimators for $\mathbf{\Sigma}_{xx}, \mathbf{\Sigma}_{yy}$ and $\mathbf{\Sigma}_{xy}$. The penalty terms (4) induce sparsity in the resulting canonical directions. Conditions (3) ensure that higher-order directions are uncorrelated to lower-order canonical vectors. For the higher-order directions, again, a suitable level of sparsity can be chosen.

In a simulation study, we show the robustness and suitability of our approach for high-dimensional data in different simulation scenarios. Empirical applications from tribology underline the usefulness of this approach.

3 Outlook

The methodology can be adapted to other robust multivariate methods such as LDA or PCA for high-dimensional data. It is sufficient to formulate the optimization problem and the constraints in a joint Lagrangian problem. The advantage of using an adaptive gradient descent algorithm is its scalability to higher dimension, and it also leads to highly precise parameter estimates, especially for higher-order components.

References

BERTSEKAS, DIMITRI P. 1982. Constrained Optimization and Lagrange Multiplier Methods.

REDDI, SASHANK J., KALE, SATYEN, & KUMAR, SANJIV. 2018. On the Convergence of Adam and Beyond. *In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

WILMS, INES, & CROUX, CHRISTOPHE. 2015. Robust Sparse Canonical Correlation Analysis. *BMC Systems Biology*, **10**(01).

ZOU, HUI, & HASTIE, TREVOR. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B*, **67(2)**, 301–320.

STRUCTURAL ZEROS IN REGRESSION MODELS WITH COMPOSITIONAL EXPLANATORY VARIABLES

Francesco Porro [1]

[1] Dipartimento di Matematica, Università degli Studi di Genova,
e-mail: francesco.porro@unige.it

ABSTRACT: In many real-life situations it may happen to consider a regression model with compositional explanatory variables. Compositional data describe parts of some whole, having the feature to sum to a fixed value, so they are commonly presented as vectors of proportions, percentages, or frequencies. In the compositional framework, the presence of structural zeros in the regressors is problematic, since a composition is not allowed to have a part equal to zero. In the recent years, a few techniques have been introduced in the literature to adress this issue. In this paper a description and a comparison of the most interesting proposals are provided.

KEYWORDS: Compositional data, regression models, structural zeros, logratio transformation.

1 The compositional data framework

During the last decades Compositional Data (CoDa) have gained more attention in the literature. The relevant information in compositional data is in the ratios between the parts and not in their absolute values or in their sum. Different examples of compositional data can be easily found in every field: physics, chemistry, finance, social sciences, and economics, just to mention some of them (cf. Pawlowsky-Glahn *et al.* , 2015). The CoDa methodology has been developed to deal with the compositions.

Definition 1 *Let $D \in \mathbb{N}$. Consider the real-valued vectors $\mathbb{R}^D$, with all (strictly) positive components. Two of such vectors $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_D)$ are compositionally equivalent whether there exists a positive constant $c \in \mathbb{R}$ such that $\mathbf{x} = c \cdot \mathbf{y}$. A D-part composition is then a class of equivalence containing all the compositionally equivalent vectors in $\mathbb{R}^D$.*

Since a D-part composition is an equivalence class, a representative one has to be selected: it is usually the vector of proportions that sum to 1. The sample

space for the D-part compositions is the simplex $\mathbb{S}^D$, defined by:

$$\mathbb{S}^D = \{(x_1, x_2, \ldots, x_D) \in \mathbb{R}^D \ : \ x_i > 0 \, \forall i; \ \sum_{i=1}^{D} x_i = c\}, \tag{1}$$

where the arbitrary constant c is usually set to 1. For further details, see Pawlowsky-Glahn *et al.* , 2015, Filzmoser *et al.* , 2018, and references therein. Starting from the definition of composition, the so-called *Aitchison geometry on the simplex* can be defined: it is the suited framework to analyze compositional data, and it can be equipped with a coherent distance, norm, and inner product. In CoDa analysis, a dataset $\mathbf{X}$ is a sample of n observations, each one being a D-part composition $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$, with $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, $i = 1, 2, \ldots, n$. The standard statistical descriptive measures, based on the real Euclidean structure, should be used with attention in such a dataset, since they can lead to erroneous conclusions (see Pawlowsky-Glahn *et al.* , 2015). To overcome this issue, the compositional approach proposes alternative statistical tools and methods, based on the Aitchison geometry.

A usual practice in handling compositions is the application of transformations, mapping them into real vectors (belonging to suitable spaces) for exploiting the usual Euclidean structure. Several transformations based on logratios have been proposed: the additive (*alr*), the centered (*clr*) and the isometric (*ilr*) logratio transformations. Their features can be found in Pawlowsky-Glahn *et al.* , 2015 and Filzmoser *et al.* , 2018. The definition of *ilr*-transformation is the following one.

Definition 2 *For a D-part composition* $\mathbf{x} = (x_1, x_2, \ldots, x_D)$*, the isometric logratio (ilr) transformation associated to an Aitchison-orthonormal basis in* $\mathbb{S}^D$*,* $\{\mathbf{e}_i\}$*,* $(i = 1, 2, ..., D-1)$*, is the mapping from* $\mathbb{S}^D$ *to* $\mathbb{R}^{D-1}$ *given by:*

$$ilr(\mathbf{x}) = [\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \langle \mathbf{x}, \mathbf{e}_2 \rangle_a, ..., \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a],$$

where $\langle \cdot, \cdot \rangle_a$ *denotes the Aitchison inner product in* $\mathbb{S}^D$*, defined by:*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \right).$$

For the remainder of this paper, it is worth just mentioning that the *ilr*-transformation is characterized by two important features: (i) it reduces the number of parts, since a D-part composition is mapped into a vector in $\mathbb{R}^{D-1}$; (ii) it preserves both the distances and the angles: in the simplex the Aitchison distance of two compositions is equal to the distance of the corresponding *ilr*-transformed vectors in $\mathbb{R}^{D-1}$ (see Pawlowsky-Glahn *et al.* , 2015 for details).

2 Regression models with compositional regressors

Many examples of regression models with (at least some) compositional explanatory variables can be easily found. In such a case, the regressors can not be directly used since compositional data are by definition singular: the constraint about their sum provides the linear dependency of the regressors and a singular covariance matrix.

The standard approach is to apply the *ilr*-transformation to the original explanatory variables and to consider the corresponding *ilr*-transformed variables as new regressors (Hron *et al.* , 2012). In this way, the linear dependence of the compositional regressors can be discarded: the new obtained model can be easily handled, and then parameter estimation can be done as in usual linear regression. This approach cannot be applied whether there are zeros, since in this case, no logratio transformation can be carried out. It follows that in case of *structural zeros* in the regressors, a different procedure has to be undertaken. It is worth recalling that a structural zero is a value that is certain to be zero, and it is not due to imprecise or insufficient measurements.

3 Three approaches dealing with structural zeros

For facing the issue of structural zeros in regression models with compositional regressors a few approaches have been proposed, quite recently. In the following, three of them are briefly presented: the first one is due to Aitchison, 1986, while the other two are described in Verbelen *et al.* , 2018. In the presentation more details will be provided to characterize and compare the three methods.

3.1 A naive approach: the replacement

The replacement strategy is the first method proposed in the literature, and it is the most intuitive one. The idea is to take all those values giving problems (since, for example, they are zeros) and replace them with a nonproblematic value (for example, a value very close to zero). This approach can be very easily implemented, and it can also be used to remove missing values. The most relevant drawbacks are the arbitrary nature of the replaced values, and the inconsistency in case of structural zeros, as they are *true* zeros.

3.2 The conditioning approach

The conditioning approach consists in treating the observations with different structural 0 patterns as different subgroups within the data, so that the regression coefficients are modeled conditionally on the 0 patterns. This method requires to compute for each compositional observation with at least one 0 part, the *ilr*-transformation of the corresponding subcomposition with non-zero parts (obtained by removing the zero parts) and to model the compositional predictor effect separately by 0 pattern. The regression coefficients obtained by this method are different for each structural 0 pattern and hence only estimated by using observations with that particular 0 pattern.

3.3 The projection approach

The projection approach is more parsimonious than the conditional one, since the regression parameters are shared across the different 0 patterns. In this method, a *generalized isometric logratio transformation* from the simplex $\mathbb{S}^D$ to $\mathbb{R}^{D-1}$ is proposed as an extension of the usual one. This new transformation can be applied also to a compositions with one or more zero parts, since the logratios are calculated using the projections onto the orthogonal complement of the structural 0 parts.

References

AITCHISON, JOHN. 1986. *The Statistical Analysis of Compositional Data*. Chapman and Hall.

FILZMOSER, PETER, HRON, KAREL, & TEMPL, MATTHIAS. 2018. *Applied compositional data analysis*. Springer.

HRON, KAREL, FILZMOSER, PETER, & THOMPSON, K. 2012. Linear regression with compositional explanatory variables. *J.Appl.Stat.*, **39**(5), 1115–1128.

PAWLOWSKY-GLAHN, VERA, EGOZCUE, JUAN JOSÉ, & TOLOSANA-DELGADO, RAIMON. 2015. *Modeling and analysis of compositional data*. John Wiley & Sons.

VERBELEN, ROEL, ANTONIO, KATRIEN, & CLAESKENS, GERDA. 2018. Unravelling the predictive power of telematics data in car insurance pricing. *J. R. Stat. Soc., C: Appl. Stat.*, **67**(5), 1275–1304.

ONE-DIMENSIONAL MIXTURE-BASED CLUSTERING FOR ORDINAL RESPONSES

Kemmawadee Preedalikit[1], Daniel Fernández[2], Ivy Liu[3] , Louise McMillan[3] , Marta Nai Ruscone[4] and Roy Costilla[5]

[1] University of Phayao, (e-mail: kemmawadee@gmail.com)

[2] Universitat Politècnica de Catalunya - BarcelonaTech, (e-mail: daniel.fernandez.martinez@upc.edu)

[3] Victoria University Wellington, (e-mail: ivy.liu@vuw.ac.nz, mcmilllo@ecs.vuw.ac.nz)

[4] University of Genoa, (e-mail: marta.nairuscone@unige.it)

[5] AgResearch NZ, (e-mail: roy.costilla@agresearch.co.nz)

ABSTRACT: Existing methods can perform likelihood-based clustering on a multivariate data matrix of ordinal responses, using finite mixtures to cluster the rows and columns of the matrix. Those models can incorporate the main effects of individual rows and columns and the cluster effects to model the matrix of responses. However, many real-world applications also include available covariates. In this study, we have extended mixture-based models to include covariates and test what effect this has on the resulting clustering structures. We focus on clustering the rows of the data matrix, using the proportional odds cumulative logit model for ordinal data. We fit the models using the Expectation-Maximization (EM) algorithm and assess their performance. Finally, we also illustrate an application of the models to the well-known arthritis clinical trial data set.

KEYWORDS: cluster analysis, mixture models , EM algorithm, ordinal responses, proportional odds mode.

1 Introduction

A well-known definition of an ordinal variable says it is one characterized by a categorical data scale, which describes an order showing differing degrees of dissimilarity (Agresti, 2010). Thus, although ordinal variables are affected by the distances among their ordinal categories, those distances are not known. In this work our approach to mixture-based clustering involves constructing an additive linear model of parameters, connected to the response data via a link function. Additional terms such as covariates may easily be added to the linear

predictor. To the best of our knowledge, (Fernández *et al.*, 2019) introduced this formulation of model-based clustering for ordinal data with covariates, but the performance of these covariate methods and, more importantly, their influence on the resulting clustering structures, have not been documented so far. The main purpose of this article is to extend such models to include covariates and allow them to affect the detection of cluster structures. Moreover, we are also interested in comparing how the resulting clustering structures compare to those obtained without covariates, and how these changes may affect the interpretation of the results. We will focus on extending the one-dimensional clustering approach proposed in (Matechou *et al.*, 2016). This approach models ordinal response data using the proportional odds assumption of the cumulative logit model (from now on "proportional odds model"). We will include covariates directly in the linear predictor.

2 Model formulation

When the data are in matrix form, clustering of rows is called row clustering. We present the row clustering formulation for finite mixtures based on the proportional odds model. This closely follows the model formulations in (Matechou *et al.*, 2016 , Fernández *et al.*, 2019). We decided to focus on row clustering because it is more common to have covariates linked to observations (rows) than to variables (columns). We consider a set of n subjects and m ordinal response variables, each with q possible ordinal response categories. Thus, data can be represented by an $n \times m$ matrix $\mathbf{Y}$ with ordinal entries y_{ij}. The row cluster index r ($r = 1, \ldots, R$) represents the number of the row cluster and the symbol $i \in r$ indicates that row i is allocated to row cluster r. We shall assume that all rows belonging to the same row cluster r have ordinal responses driven by the same row cluster effect, i.e. that there are no individual row effects. In a simpler model with clustering of rows, the rows (observations/subjects) will tend to be clustered if they have similar patterns of responses, without taking into account the information present in the covariates.

Having in mind that R and C are the numbers of row clusters and column clusters, respectively, we will deal with the possible values of $C = m$ (when column effects are different and therefore they are included within the model, without clustering). $C = 1$ when the column effect is the same and it is not included into the model.

Considering the simplest row clustering model, without column effects,

the proportional odds model without covariates can be expressed as

$$logit\left(\sum_{h=1}^{k}\theta_{ijrh}\right)=\eta_{ijrk}=\mu_k-\alpha_r, \tag{1}$$

where the parameters μ_k are the cutpoints and α_r indicates the effects of row cluster r. Adding p covariates into Model 1, we obtain

$$logit\left(\sum_{h=1}^{k}\theta_{ijrh}\right)=\eta_{ijrk}=\mu_k-(\alpha_r+x_i^T\delta_r), \tag{2}$$

where δ_r represent the effects of the covariates Models 1 and 2 will be used in the simulation and application section to compare the clustering structure.

3 Application

We applied the models proposed in this article to the *arthritis clinical trial* data set (Lipsitz *et al.*, 1996), which compares the drug auranofin and placebo therapy for the treatment of rheumatoid arthritis. The data set is obtained from the **R** package *multgee* (Touloumis, 2015). In this application, the covariate-dependent clustering could help to identify subsets of patients with similar covariate information patterns. This insight would be important because it would provide a flexible approach for identifying potential heterogeneous gender, age, and auranofin treatment effects on the arthritis scores. After fitting the models without covariates Eq.(1) and with covariates Eq.(2), with different number of row clusters, we compared them using the information criteria AIC and BIC (see results in Table 1). AIC indicates that the best model is the version of the row clustering model including age and treatment covariates ($\mu_k-(\alpha_r+x_{i1}\delta_{1r}+x_{i2}\delta_{2r})$) with $R=4$ row clusters (AIC = 2136.78), which is better than its counterpart in the model without covariates (AIC=2154.40). However, BIC shows that the model without covariates ($\mu_k-\alpha_r$) and $R=4$ is the best model (BIC=2202.05). A possible reason is that BIC penalizes higher numbers of parameters more strongly than AIC does, leading to a preference for more parsimonious models.

References

AGRESTI, ALAN. 2010. *Analysis of Ordinal Categorical Data, Second Edition*. Wiley Series in Probability and Statistics: John Wiley and Sons, Inc.

Table 1. *Results of row clustering models fitted to the arthritis data set. The best model in each group of models (no covariates, one, two, or three covariates), based on AIC, is shown in bold.*

Model		R	number of parameter	Log-like	AIC	BIC
$\mu_k - \alpha_r$		2	6	-1096.99	2205.99	2234.58
		3	8	-1077.73	2171.46	2209.59
		4	10	-1067.20	**2154.40**	**2202.05**
		5	12	-1067.20	2158.40	2215.58
$\mu_k - (\alpha_r + x_i\delta_r)$	x= age	2	8	-1138.18	2292.37	2330.49
		3	11	-1071.88	2165.75	2218.17
		4	14	-1065.18	2158.37	2225.08
		5	17	-1060.84	**2155.68**	2236.68
	x=treatment	2	8	-1082.28	2180.57	2218.69
		3	11	-1067.93	2157.87	2210.28
		4	14	-1057.70	**2143.40**	2210.11
		5	17	-1056.23	2146.46	2227.46
	x= gender	2	8	-1096.89	2209.77	2247.89
		3	11	-1079.51	2181.02	2233.44
		4	14	-1066.92	**2161.84**	2228.55
		5	17	-1066.37	2166.74	2247.74
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r})$	x_1 = age, x_2= treatment	2	10	-1072.54	2165.07	2212.72
		3	14	-1059.23	2146.46	2213.17
		4	18	-1050.39	**2136.78**	2222.55
		5	22	-1048.53	2141.05	2245.88
	x_1 = age, x_2= gender	2	10	-1085.83	2191.67	2239.32
		3	14	-1068.97	2165.95	2232.66
		4	18	-1061.29	**2158.58**	2244.35
		5	22	-1059.26	2162.52	2267.35
	x_1 = treatment, x_2= gender	2	10	-1081.82	2183.64	2231.29
		3	14	-1065.99	2159.99	2226.71
		4	18	-1056.73	**2149.45**	2235.22
		5	22	-1055.06	2154.13	2258.96
$\mu_k - (\alpha_r + x_{i1}\delta_{1r} + x_{i2}\delta_{2r} + x_{i3}\delta_{3r})$	x_1 = age, x_2= treatment, x_3= gender	2	12	-1071.60	2167.21	2224.39
		3	17	-1060.50	2155.00	2236.01
		4	22	-1050.35	**2144.71**	2249.54
		5	27	-1052.14	2158.35	2287.00

FERNÁNDEZ, DANIEL, ARNOLD, RICHARD, PLEDGER, SHIRLEY, LIU, IVY, & COSTILLA, ROY. 2019. Finite mixture biclustering of discrete type multivariate data. *Advances in Data Analysis and Classification*, **13**, 117–143.

LIPSITZ, STUART R., FITZMAURICE, GARRETT M., & MOLENBERGHS, GEERT. 1996. Goodness-of-Fit Tests for Ordinal Response Regression Models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **45**(2), 175–190.

MATECHOU, ELENI, LIU, IVY, FERNÁNDEZ, DANIEL, FARIAS, MIGUEL, & GJELSVIK, BERGLJOT. 2016. Biclustering Models for Two-Mode Ordinal Data. *Psychometrika*, **81**(3), 611–624.

TOULOUMIS, ANESTIS. 2015. R Package multgee: A Generalized Estimating Equations Solver for Multinomial Responses. *Journal of Statistical Software*, **64**(8), 1–14.

A COMPOSITIONAL STOCHASTIC BLOCK MODEL FOR THE ANALYSIS OF THE ERASMUS PROGRAMME NETWORK

Iuliia Promskaia[1,2], Adrian O'Hagan[1,2] and Michael Fop[1]

[1] School of Mathematics and Statistics, University College Dublin,
(e-mail: iuliia.promskaia@ucdconnect.ie, adrian.ohagan@ucd.ie, michael.fop@ucd.ie)

[2] SFI Insight Centre for Data Analytics

ABSTRACT: The Erasmus Programme is one of the most well-known student exchange programmes in the world, with over 200,000 higher education students availing of its benefits each year. We explore the Erasmus exchange data, aiming to identify clustering structure in the mobility patterns of students between countries. A directed weighted network of country-to-country student exchanges is constructed, with edge weights representing the percentage of students travelling from one country to any other. These edge weights are compositional in nature, so they cannot be assumed independent. We propose an extension of the stochastic block model for clustering network data with compositional edge weights, and compare its performance to that of other models.

KEYWORDS: clustering, networks, compositional data, stochastic block model.

References

GADÁR, L., KOSZTYÁN, Z.T., TELCS, A. ET AL. 2020. A multilayer and special description of the Erasmus mobility network. *Scientific Data*, **7**, 41.

LEE, C., WILKINSON, D.J. 2019. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, **4**, 122.

GREENACRE, M. 2021. Compositional Data Analysis. *Annual Review of Statistics and Its Applications,* **8:1**, 271-299.

A PROPOSAL OF DEEP FUZZY CLUSTERING BY MEANS OF THE SIMULTANEOUS APPROACH

Claudia Rampichini [1] and Maria Brigida Ferraro[1]

[1] Department of Statistical Sciences, University of Rome "La Sapienza", (e-mail: claudia.rampichini@uniroma1.it, mariabrigida.ferraro@uniroma1.it)

ABSTRACT: Classical clustering methods may suffer from the presence of high dimensional or complex data. In this scenario, deep clustering can be useful to overcome such problems. The main idea is to use a neural network to reduce the input's complexity and apply a clustering algorithm to the reduced space. Our method consists in combining a neural network with the fuzzy *k*-means clustering algorithm. In particular, the proposal links the encoder part of an autoencoder neural network to a new layer, in which the membership degree values are calculated, and jointly optimizes the method by minimizing the fuzzy *k*-means objective function. Furthermore, to avoid the problem of collapsing centers, a penalization term is added. The adequacy of the proposal is evaluated by means of benchmark datasets.

KEYWORDS: deep clustering, neural networks, fuzzy *k*-means.

1 Introduction and background

Recent improvements in deep learning techniques have led to a new research field called deep clustering that shows new opportunities for conventional clustering to overcome problems with high-dimensional data. The idea of deep clustering is to learn latent features of training data using a deep neural network (DNN) and apply clustering methods to the resulting data representation. There exist two different deep clustering approaches: sequential and simultaneous. In the former, clustering algorithms are applied to the learned DNN representation, while in the latter, deep representation learning and clustering objectives are jointly optimized. Clustering approaches that are combined with deep learning models include *k*-means, graph clustering, spectral clustering, Gaussian mixture model, and many others, however, few studies focus on deep fuzzy clustering. One of the most famous models in the simultaneous approach is the deep embedded clustering method (DEC) that was proposed by Xie *et al.* (2016). This method simultaneously learns feature representations with stacked autoencoders and cluster assignments with soft *k*-means,

minimizing a joint loss function. Later, some more complex deep fuzzy clustering methods have been proposed. The main differences are in the use of convolutional networks and more complex structures for loss functions (see, for example, Feng *et al.*, 2020, and Zhang *et al.*, 2020).

Starting by considering the problem of clustering a set of n points $\{\mathbf{x}_i \in X\}_{i=i}^{n}$ into k clusters, each represented by a centroid $\boldsymbol{\mu}_g$, $g = 1, ..., k$, the DEC model consists in transforming the input data by a non-linear mapping $f_\theta : X \to Z$, where θ are learnable parameters and Z is the latent feature space where clustering is performed. Moreover, the Kullback–Leibler (KL) divergence loss between a centroid-based probability distribution and an auxiliary target distribution is used as the objective function:

$$L = KL(P||Q) = \sum_{i=1}^{n}\sum_{g=1}^{k} p_{ig} log\left(\frac{p_{ig}}{q_{ig}}\right). \tag{1}$$

In (1), q_{ig} is a Student's t-distribution used as a kernel to measure the similarity between embedded point $\mathbf{z}_i$ and centroid $\boldsymbol{\mu}_g$:

$$q_{ig} = \frac{(1 + ||\mathbf{z}_i - \boldsymbol{\mu}_g||^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{g'}(1 + ||\mathbf{z}_i - \boldsymbol{\mu}_{g'}||^2/\alpha)^{-\frac{\alpha+1}{2}}} \tag{2}$$

and it can be interpreted as the probability of assigning sample i to cluster g. Moreover, $\mathbf{z}_i = f_\theta(\mathbf{x}_i) \in Z$ corresponds to $\mathbf{x}_i \in X$ after embedding, α represents the degrees of freedom of the Student's t distribution. The auxiliary target distribution p_{ig} is calculated as the ratio between $\frac{q_{ig}^2}{f_g}$ and $\sum_{g'=1}^{k'} \frac{q_{ig'}^2}{f_{g'}}$ where $f_g = \sum_{i=1}^{n} q_{ig}$ are soft cluster frequencies; k-means is used only to initialize cluster centers. Their final model consists of the encoder part of the DNN and an additional layer in which the probability of assigning sample i to cluster g is calculated.

Starting from this model, we propose a simultaneous deep fuzzy clustering method in which the fuzzy k-means algorithm is involved. The idea for this proposal stems from noticing that the use of fuzzy clustering algorithms is often related to the initialization of cluster centers only, moreover, few works in the literature deal with these clustering algorithms. Additionally, we test our method on images improving upon traditional clustering methods which often poorly cluster this kind of data. The main reason is the difficulty of obtaining reliable similarity measures in high-dimensionality space but deep clustering methods have shown impressive performance in image clustering tasks.

2 Proposed method

The main idea is to replace the KL divergence loss with the fuzzy k-means objective function, hence

$$\arg\min_{U,C} \sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^{m} \|\mathbf{z}_i - \boldsymbol{\mu}_g\|^2, \tag{3}$$

$$\text{s.t. } u_{ig} \in [0,1], \quad i = 1,...,n \quad \textit{and} \quad g = 1,...,k; \quad \textstyle\sum_{g=1}^{k} u_{ig} = 1, \quad i = 1,...,n.$$

Similar to the work of Xie *et al.* (2016), we create a deep autoencoder neural network and keep only the encoder part. Then we link a new layer to this part which computes the membership degrees values u_{ig} as follows

$$u_{ig} = \frac{1}{\sum_{j=1}^{k} \left(\frac{\|\mathbf{z}_i - \boldsymbol{\mu}_g\|}{\|\mathbf{z}_i - \boldsymbol{\mu}_j\|} \right)^{\frac{2}{m-1}}}. \tag{4}$$

In this way, by minimizing the fuzzy k-means loss function, we jointly optimize the cluster centers and DNN parameters. Since in the optimization process, the encoder part may lead, in an attempt to reduce the initial data, to the collapse of all the points into a single cluster, we introduce a penalization term, which is the absolute value of the sum of the pairwise differences. Additionally, to ensure that the two terms are on the same scale, we normalized them by dividing the first term by the product of the number of training examples used in one iteration hence batch size (b) and the number of cluster centers (k), and the second term by the product of the batch size and itself. Hence the new loss function takes the following form

$$\arg\min_{C,U} \frac{1}{bk} \sum_{i=1}^{n} \sum_{g=1}^{k} u_{ig}^{m} \|\mathbf{z}_i - \boldsymbol{\mu}_g\|^2 - \frac{1}{b^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \overline{u}_{ig}^{m} |\mathbf{z}_i - \mathbf{z}_j|. \tag{5}$$

3 Results

We evaluate the proposed method on different benchmark datasets: Mnist, Fashion-Mnist and Cifar10. The first dataset consists of 70.000 black and white images of handwritten digits of 28$\times$ 28 pixel size, the second is a dataset of 70.000 Zalando's article images of 28$\times$ 28 pixel size and the last consists of 60.000 different colour images of 32$\times$32 pixel size. The accuracy results

achieved by the standard k-means, k-means in the embedding space (AE+k-means), DEC and our method are reported in Table 1.

Table 1. *Comparison of the accuracy level achieved by different methods on Mnist, Fashion-Mnist and Cifar10 datasets.*

Method	Mnist	Fashion-Mnist	Cifar10
k-means	53.5	47.4	22.9
AE+k-means	81.8	57.9	80.1
DEC	84.3	51.7	30.1
Our method	93.4	62.3	31.3

The results show the potential of the proposed method. In particular, on the Mnist dataset, we achieve an accuracy of 93.4% against 53.5% obtained with k-means and 84.3% with DEC; also on the Fashion-Mnist the accuracy of our method is higher than the others. On the Cifar10, the accuracy of our proposal is in line with the value reached by DEC but far from that of AE+k-means; this is probably related to the more complex dataset with colour images.

4 Concluding remarks

The new deep clustering method jointly learns feature representations with a deep autoencoder neural network and clusters assignments with fuzzy k-means by minimizing a loss function constructed in accordance with the chosen fuzzy algorithm. The results show margins of improvement with respect to the classical clustering methods and DEC model. Our future research will focus on the study of the different segmentation techniques for colour images.

References

FENG Q., CHEN L., CHEN P.C.L., & L., GUO. 2020. Deep Fuzzy Clustering-A Representation Learning Approach. *IEEE Transaction on Fuzzy Systems.*, **28**.

XIE, J., GIRSHICK R., & FARHADI, A. 2016. Unsupervised Deep Embedding for Clustering Analysis. *Proc. 33rd International Conference on Machine Learning.*, **48**, 478–487.

ZHANG R., LI X., ZHANG H., & F., NI. 2020. Deep Fuzzy K-Means With Adaptive Loss and Entropy Regularization. *IEEE Transaction on Fuzzy Systems.*, **28**.

WHEN NONRESPONSE MAKES ESTIMATES FROM A CENSUS A SMALL AREA ESTIMATION PROBLEM: THE CASE OF THE SURVEY ON GRADUATES' EMPLOYMENT STATUS IN ITALY*

Maria Giovanna Ranalli [1], Fulvia Pennoni[2], Francesco Bartolucci[3], and Antonietta Mira[4]

[1] Department of Political Science, University of Perugia, IT (e-mail: maria.ranalli@unipg.it)

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

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

[4] Università della Svizzera italiana, CH, and Department of Economics, University of Insubria, IT (e-mail: antonietta.mira@uninsubria.it)

ABSTRACT: In this paper we frame the problem of obtaining estimates from the survey on the employment status of graduates in Italy as a Small Area Estimation problem because of unit nonresponse. We propose to use generalized linear mixed models and to include two variables that can be considered proxies of the response propensity among the set of covariates to make the MAR assumption more tenable. Estimates for degree programmes are obtained as (semi-parametric) empirical best predictions.

KEYWORDS: generalized linear mixed model, latent trait models, mixed-mode survey, nonparametric maximum likelihood, paradata.

1 Introduction

Since 1998 AlmaLaurea, a consortium of 80 Italian Universities, carries out an annual survey on the employment status of graduates. The survey is carried out one, three, and five years after graduation and provides a broad picture of graduates' job placement in the labour market. The 2022 edition has involved 660,000 first- and second-level graduates in 2020 (AlmaLaurea, 2022). The survey is a census and targets many variables of interest other than the

*We are grateful to AlmaLaurea for making the data available and to AlmaLaurea researchers for sharing their precious insights that motivated the research questions and helped with interpretation.

employment status, such as job characteristics, including type of contract and salary, and of the use of the skills gained at university.

As with all surveys, nonresponse occurs: the overall response rate for the graduates involved one year after graduation (the focus here) is 68.4%. This is the outcome of a two-fold process. First, a subset of graduates (approximately 92%) is identified as those who have given consent to be contacted according to the General Data Protection Regulation no. 2016/679. Then, these graduates are contacted using a dual survey technique: CAWI (Computer-Assisted Web Interviewing) and CATI (Computer-Assisted Telephone Interviewing). CATI is used to contact those who did not respond to the online questionnaire. This sequential mixed-mode CAWI-CATI methodology leads to a response rate of 74.2% among graduates contacted with their consent in accordance with the GDPR. Estimates for the overall population of graduates are adjusted for nonresponse by means of calibration on known population totals coming from administrative registers (AlmaLaurea, 2022; Kott, 2006).

The survey aims at providing estimates not only at the population level, but also for subpopulations (domains) of interest given by the degree programmes. In the last edition, there are almost 5,700 degree programmes for which unweighted count data are publicly released (AlmaLaurea, 2023). Some of these domains have a very small number of observations: this is due to a small number of observations in the population coupled with nonresponse. This setting resembles that of Small Area Estimation (SAE, Rao & Molina, 2015): a SAE problem arises when the sample size available in a domain (area) of interest is so small that direct estimates, albeit (approximately) unbiased, have unduly large variances. Here, re-weighting methods such as calibration are of little use. SAE methods, on the other hand, are indirect as they make use of observations coming from other areas and are model-based. In SAE, the small sample size is the outcome of a process (the sampling design) that is known to the researcher. Here, the SAE problem arises from a process (the response) that is unknown. Often, the (unverifiable) assumption that data is Missing At Random (MAR) given the covariates included in the model is made. In this paper we propose a modeling approach that tries to go beyond the classical MAR assumption by making use of all the available auxiliary information on the response behaviour of graduates from paradata and other survey data.

2 The proposed modeling approach

We adapt here the framework proposed in Marino *et al.* , 2019, and use their notation. Let U denote the finite population of AlmaLaurea graduates in 2020

of size N, which can be partitioned into m non-overlapping small areas (degree programmes), with U_i denoting the i-th small area with size N_i, $i = 1,\dots,m$. For a given degree programme i, population data consist of N_i measurements of a response variable Y_{ij} and a vector of covariates $\boldsymbol{x}_{ij} = (x_{ij1},\dots,x_{ijp})'$, with $j = 1,\dots,N_i$. For ease of notation, we consider here the case of one variable of interest. Covariates $\boldsymbol{x}$ come from administrative registers, as well as from previous surveys conducted by AlmaLaurea such as that on the Profile of Graduates. Also, let $\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_m$ be iid, q-dimensional, vectors of area-specific random effects ($q \leq p$) with density $f_\alpha(\cdot)$, $E_\alpha(\boldsymbol{\alpha}_i) = 0$, and $E_\alpha(\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i') = \boldsymbol{\Sigma}$ for all $i = 1,\dots,m$. Last, let $\boldsymbol{w}_{ij}$ denote a q-dimensional subset of $\boldsymbol{x}_{ij}$ associated to $\boldsymbol{\alpha}_i$. Then a sample of size n of respondents is obtained from the above population and we denote by r_i the set containing the n_i population indexes of sample units belonging to degree programme i, with $n = \sum_{i=1}^{m} n_i$. Therefore, values of Y_{ij} are known only for the sample ($i = 1,\dots,m$, $j \in r_i$), while the values of $\boldsymbol{x}_{ij}$ and of $\boldsymbol{w}_{ij}$, are known for all units in the population ($i = 1,\dots,m$, $j = 1,\dots,N_i$).

Usually, it is assumed that the response process is non-informative for the small area distribution of $Y_{ij} \mid \boldsymbol{x}_{ij}$, allowing to use population level models with sample data. In order to make this assumption more tenable, we propose to include in each vector $\boldsymbol{x}_{ij}$ two covariates obtained as follows. The first one comes from paradata and has the following categories: "Response with CAWI", "Response with CATI", "Response with CATI recall", "Nonresponse", "No consent to GDPR". It can be considered as a proxy of the response propensity as these categories can be ordered along a decreasing response propensity. The second one exploits information on item nonresponse of graduates in the survey and in previous surveys to build a latent variable in the spirit of Matei & Ranalli, 2015. A set of binary indicators taking value 1 if the item is not missing and 0 if it is missing can be used to derive a latent trait using Item Response Theory models that can be interpreted as a proxy of the response propensity. Nonrespondents have all zeros and the smallest value of the latent trait.

We assume that, conditional on $\boldsymbol{\alpha}_i$, responses Y_{ij} from the same area i are independent with density $f_{y|\alpha}(y_{ij} \mid \boldsymbol{\alpha}_i;\boldsymbol{x}_{ij})$ in the Exponential Family with canonical parameter θ_{ij} modeled as $\theta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{w}_{ij}'\boldsymbol{\alpha}_i$. The marginal distribution of $\boldsymbol{y}_i$ is obtained as $f_y(\boldsymbol{y}_i;\boldsymbol{X}_i) = \int_{\mathbb{R}^q} f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i;\boldsymbol{X}_i) f_\alpha(\boldsymbol{\alpha}_i) d\boldsymbol{\alpha}_i$, where $f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i;\boldsymbol{X}_i) = \prod_{j\in r_i} f_{y|\alpha}(y_{ij} \mid \boldsymbol{\alpha}_i;\boldsymbol{x}_{ij})$ and $\boldsymbol{X}_i$ is the matrix of covariates for units in the i-th area. Typically, a parametric specification for $f_\alpha(\boldsymbol{\alpha}_i)$ is adopted, with a common choice being the $N_q(\boldsymbol{0},\boldsymbol{\Sigma})$ distribution. We also consider the more flexible alternative proposed in Marino *et al.* , 2019, in which the distribution of $\boldsymbol{\alpha}_i$ is left unspecified and nonparametric ML is used.

We use respondents data on Y_{ij} ($i = 1,\dots,m, j \in r_i$) and population data

on covariates $\boldsymbol{x}_{ij}$ ($i = 1,\ldots,m, j = 1,\ldots,N_i$) to predict a (possibly) non-linear function of fixed and random effects, say $\zeta(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma})$. According to Jiang, 2003, the Best Predictor (BP) of ζ in terms of minimum MSE is given by $\tilde{\zeta}^{BP} = E_{\alpha|y}[\zeta(\boldsymbol{\beta},\boldsymbol{\alpha}_i,\boldsymbol{\Sigma}) \mid \boldsymbol{y}] = \int_{\mathbb{R}^{\upsilon}} \zeta(\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma}) f_{\alpha|y}(\boldsymbol{\alpha} \mid \boldsymbol{y}) d\boldsymbol{\alpha}$, where

$$f_{\alpha|y}(\boldsymbol{\alpha} \mid \boldsymbol{y}) = \frac{\prod_{i=1}^{m} f_{y|\alpha}(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i;\boldsymbol{X}_i) f_{\alpha}(\boldsymbol{\alpha}_i)}{\prod_{i=1}^{m} f_y(\boldsymbol{y}_i;\boldsymbol{X}_i)},$$

$\boldsymbol{y} = (\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m)$ and $\upsilon = m \times q$. Estimates of model parameters can be obtained by maximizing the observed data likelihood function: $L(\boldsymbol{\Phi}) = \prod_{i=1}^{m} f_y(\boldsymbol{y}_i;\boldsymbol{X}_i)$. To maximize $L(\boldsymbol{\Phi})$, numerical approximations (e.g., Gaussian quadrature techniques) or simulation based methods (e.g., Monte Carlo integration) may be required. Once parameters are estimated, we may compute the empirical BP of ζ, that is $\hat{\zeta}^{EBP} = \tilde{\zeta}^{BP}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\Sigma}})$. To evaluate the quality of such predictions, the second-order MSE estimator can be considered as in Jiang, 2003 and in Marino *et al.* , 2019.

References

ALMALAUREA. 2022. *24th Report Occupational Condition of Graduates, 2022 Summary Report*. https://www.almalaurea.it/sites/default/files/2022-09/sintesi_occupazione_rapporto_2022_en.pdf. Accessed: 2023-04-21.

ALMALAUREA. 2023. *Graduates' employment status, data*. https://www2.almalaurea.it/cgi-php/universita/statistiche/tendine.php?anno=2021&LANG=en&config=occupazione. Accessed: 2023-04-21.

JIANG, J. 2003. Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference*, **111**, 117–127.

KOTT, P. S. 2006. Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32**, 133–142.

MARINO, M. F., RANALLI, M. G., SALVATI, N., & ALFÒ, M. 2019. Semiparametric empirical best prediction for small area estimation of unemployment indicators. *The Annals of Applied Statistics*, **13**, 1166–1197.

MATEI, A., & RANALLI, M. G. 2015. Dealing with non-ignorable nonresponse in survey sampling: A latent modeling approach. *Survey Methodology*, **41**, 145–165.

RAO, J. N. K., & MOLINA, I. 2015. *Small Area Estimation*. John Wiley & Sons.

A SUPERVISED CLASSIFICATION STRATEGY BASED ON THE NOVEL DIRECTIONAL DISTRIBUTION DEPTH FUNCTION

Edoardo Redivo[1] Cinzia Viroli[1]

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

ABSTRACT: Statistical depth functions are a class of functions that provide a center-outward ordering of sample points in multidimensional space. In this work we introduce a novel depth function that is based on the cumulative distribution function along random directions, and is thus termed directional distribution depth. Some properties and a connection to the Mahalanobis depth when applied to sphered data are shown. The proposed depth is used as a basis for supervised classification using maximum depth classifiers and more flexible polynomial separators in the depth space. It is shown to be effective and competitive against other depth functions through simulated experiments and real data applications.

KEYWORDS: depth functions, random projection, supervised classification

1 Introduction

In multivariate analysis the identification of order statistics, quantiles and atypical patterns is very challenging due to the lack of an order among observations, which is instead natural in the real line $\mathbb{R}^1$ (Kong & Mizera, 2012; Serfling, 2002). To overcome this challenge, the most important line of research is rooted in the concept of statistical depth, which leads to a center-outward ordering of the sample points in $\mathbb{R}^p$ with $p \geq 2$. More specifically, a depth function is a function that can assign a real number to each point of in multivariate space, measuring the outlyingness of the point with respect to the barycenter, and can be used as a starting point for outlier detection, clustering, classification.

Popular depth functions are the Mahalanobis depth, which is based on the Mahalanobis distance (Mahalanobis, 1936), and the halfspace depth, which measures the depth of a point by the smallest probability of a halfspace that contains that same point. Liu *et al.* (1999) described different depth functions as valuable exploratory tools in multivariate analysis. Introducing some

notation, let $\mathbf{X}$ be a multivariate random variable of order p with a probability distribution F: a data depth measures how deep (or central) a given value $\mathbf{x}$ of $\mathbf{X}$ is with respect to the data cloud or a given distribution function and is usually denoted as $D(\mathbf{x}, F)$. A simple example is the Mahalanobis depth, which is inversely proportional to the Mahalanobis distance: $MD(\mathbf{x}, F) = \left[1 + (\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-1}$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and dispersion matrix of $\mathbf{X}$ and can be estimated from the data.

Zuo & Serfling (2000) reviewed some of the most popular depth functions and introduced some desirable properties that in their view can be define a proper depth function. More precisely, a depth function is a non-negative and bounded function, which is: (i) invariant to the coordinate system or to the scale of the underlying measurements (affine invariance); (ii) maximum at its center; (iii) monotonically decreasing when a point moves away from the deepest central point and (iv) it should approach zero as a point approaches infinity. Some other properties that can be attractive and that we will consider are consistency of the function based on sample data to a population counterpart, and computational feasibility, *i.e.*, it should be possible to compute the depth values of data points efficiently even for large p.

2 Directional Distribution Depth

Let $\mathbf{S}$ be a random vector of length p with a uniform distribution on the sphere, that is any of its realizations $\mathbf{s}$ is a direction belonging to the sphere ($\mathbb{S}^{p-1}$) and having unit norm ($\|\mathbf{s}\|_2 = 1$). The depth of a point is derived by projecting it along any direction and evaluating the cumulative distribution function of the univariate distribution of the projected data $\mathbf{S}^\top\mathbf{X}$. The resulting probability is transformed so that the depth is symmetric with respect to the median, defined as the deepest point. As a last step we take the expected value over all random direction. More precisely, the directional distribution depth is the mapping $\mathbb{R}^p \times \mathcal{F} \to [0,1]$ defined as

$$D(\mathbf{x}, F) = E_{\mathbf{S}}\left[1 - 2|F_{\mathbf{S}^\top\mathbf{X}}(\mathbf{S}^\top\mathbf{x}) - 0.5|\right], \tag{1}$$

where $E_{\mathbf{S}}$ is the expectation with respect to the random vector $\mathbf{S}$, F is the probability distribution of the multivariate data and $F_{\mathbf{S}^\top\mathbf{X}}$ is the marginal probability distribution of the transformation $\mathbf{S}^\top\mathbf{X}$ evaluated at $\mathbf{S}^\top\mathbf{x}$. $F_{\mathbf{S}^\top\mathbf{X}}$ can be any (probabilistic or nonparametric) univariate distribution function differently parameterized along each direction. In this work we will focus and compare the depth based on the Gaussian distribution, on the *fgld* quantile function due to

its large flexibility (Redivo *et al.*, 2023; Chakrabarty & Sharma, 2021) and the nonparametric kernel density estimation.

Theorem 1. *Given whatever model choice of $F_{\mathbf{S}}$, the depth defined in (1) is a proper depth function in the sense of the definition given by Zuo & Serfling (2000).*

An interesting property closely related to the proposed depth function is that the average squared distance of univariate projections from the mean, applied to sphered data, is proportional to the Mahalanobis distance in the original multivariate space:

$$E_{\mathbf{S}}\left[\left(\mathbf{S}^\top\tilde{\mathbf{x}}-\mathbf{S}^\top\tilde{\boldsymbol{\mu}}\right)^2\right]=\frac{1}{p}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}),$$

where $\tilde{\mathbf{x}}$ and $\tilde{\boldsymbol{\mu}}$ are respectively the point and center transformed via the sphering matrix. Next we adapt our depth definition to sample data. Let $\mathbf{X}_n$ be a sample of size n from $\mathbf{X}$, without loss of generality we assume it to sphered. Let $\mathbf{s}_B$ be a set of B random directions. Then the sample version of the directional distribution depth for a generic point $\mathbf{x}_i$ is

$$D_n(\mathbf{x}_i,F)=\frac{\sum_{b=1}^{B}\left[1-2|\hat{F}_{\mathbf{s}_b^\top\mathbf{X}_n}(\mathbf{s}_b^\top\mathbf{x}_i)-0.5|\right]}{B}, \tag{2}$$

This quantity is strongly consistent with respect to its population counterpart, that is as $n\to\infty$ and $B\to\infty$, $D_n(\mathbf{x},F)\xrightarrow{a.s.}D(\mathbf{x},F)$.

3 Application to Supervised Classification

We apply to proposed depth function to supervised classification by allocating a new observation to the class with the maximum depth among the K populations (Ghosh & Chaudhuri, 2005). The performance of the proposed depth (with its three distribution estimators) is evaluated through a simulation study, comparing it to maximum depth classifiers based on other depth definitions (Mahalanobis, projection, simplicial and halfspace) and to linear and quadratic discriminant analysis. The simulation comprises three distributional scenarios: with Gaussian data classifiers based on the directional distribution depth perform similarly well to those based on data generating normal model; with t-distributed data, linear discriminant analysis performs the best, being quite robust to the heavier tails, with the distributional depth classifiers lagging shortly behind; with skewed data our depth performs generally better

than the alternatives, being the only one that can accommodate non-elliptical data, which is assumed by the Mahalanobis depth and the discriminant analysis methods. Throughout the simulations classifiers based on the halfspace depth have substantially worse results, and this is probably due to the difficulty in computing the depth, with only an approximation being available in higher dimensions, where the resulting classifier suffers the most.

We also applied depth based classifiers to commonly used benchmark data sets. Here we have considered polynomial separators for the classes in the depth space, in contrast to the quadrant bisector line implicitly assumed by the maximum depth classifier. This method is called DD-classifier and has been introduced in Li *et al.* (2012). The DD-classifier based on the new depth is able to achieve competitive accuracies (measured through mean accuracy in repeated training-testing splits) even against K-nearest neighbours and SVM.

References

CHAKRABARTY, T. K., & SHARMA, D. 2021. A Generalization of the Quantile-Based Flattened Logistic Distribution. *Annals of Data Science*, **8**(3), 603–627.

GHOSH, A. K., & CHAUDHURI, P. 2005. On Maximum Depth and Related Classifiers. *Scandinavian Journal of Statistics*, **32**(2), 327–350.

KONG, L., & MIZERA, I. 2012. Quantile Tomography: Using Quantiles With Multivariate Data. *Statistica Sinica*, **22**(4), 1589–1610.

LI, J., CUESTA-ALBERTOS, J. A., & LIU, R. Y. 2012. DD-Classifier: Nonparametric Classification Procedure Based on DD-Plot. *Journal of the American Statistical Association*, **107**(498), 737–753.

LIU, R. Y., PARELIUS, J. M., & SINGH, K. 1999. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics*, **27**(3), 783–858.

MAHALANOBIS, P.C. 1936. On the Generalized Distance in Statistics. *Proceedings of the National Institute of Science of India*, **2**, 49–55.

REDIVO, E., VIROLI, C., & FARCOMENI, A. 2023. Quantile-based distribution functions and their use for classification, with application to naïve Bayes Classifiers. *Statistics and Computing*.

SERFLING, R. 2002. Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, **56**(2), 214–232.

ZUO, Y., & SERFLING, R. 2000. General notions of statistical depth function. *The Annals of Statistics*, **28**(2), 461 – 482.

AN APPLICATION OF CART ALGORITHM TO ADMINISTRATIVE DATA: ANALYSIS OF YOUTH INITIAL EMPLOYMENT TRAJECTORIES

Rocco Ilaria[1,2]

[1] Labour Market Observatory of Veneto Lavoro, Italy,
(e-mail: ilaria.rocco@venetolavoro.it)

[2] Department of Statistical Sciences, Sapienza University of Rome, Italy

ABSTRACT: This work presents an application of the classification and regression tree (CART) algorithm to administrative data on employment in the Veneto region, an area in northern Italy with a strong economy and a dynamic labour market. This data, derived from the national stream of declarations due by employers to notify each activation, termination, extension, or transformation of employment relationships, allows to investigate the occupational condition of young people who enter for the first time in the regional labour market: we classified their initial work trajectories and then we analysed the individual and working features characterising each identified class. In the current socio-economic context, an insight on youth working conditions is crucial to facilitate the successful design of labour and education policies.

KEYWORDS: classification and regression tree, administrative data, youth, labour market.

1 Introduction

Tree-based methods, that find application in many disciplinary fields, from economics (Williams et al. 1987; Keely and Tan 2008; Manasse and Roubini 2009; Galletta 2016; Bilton et al. 2017), to engineering, medicine, biology, and marketing (De'ath and Fabricius 2000; Dacko et al. 2016), are useful statistical techniques for exploring patterns in complicated datasets if assumptions of linear models are somewhat violated (De'ath and Fabricius 2000; Frisman et al. 2008) or if response or explanatory variables present outliers, missing and unbalanced values (Low and Lai 2016). The classification and regression tree (CART) algorithm, introduced by Breiman et al. (1984), is a non-parametric approach without distributional assumptions that allows to handle datasets containing variables of categorical, scale, and ordinal measurement types (Wałęga and Wałęga, 2021).

This work aims to present an application of the CART algorithm to administrative data on employment in the Veneto region, a territory in northern Italy with a strong economy and a particularly dynamic labour market. Using this method we will classify the initial trajectories of young people into the labour market, and then we

will explore the individual and working characteristics associated to each identified class (through multinomial regressions).

2 Data source

The data used come from the database derived from the Labour Information System of Veneto (SILV). It is an administrative archive that collects the stream of declarations ("Compulsory Communications") due by employers to notify the events of activation, termination, extension, or transformation of each employment relationship. This database, that ensures in a timely manner a constant updating of the information, has as reference universe all the subordinate and para-subordinate employments activated by regional enterprises, both public and private; it also allows to monitor the work experiences like the non-curricular internships that are particularly relevant for the young component of the labour supply.
Moreover, this data source offers the possibility to observe the dynamics of the regional labour market since at least 2008, the year of the computerization of the national information system (in the Veneto region this process started in the '90s and the data collected since 2000 onwards can be considered reliable).
The wealth of data analytical details (information is available for single worker and company) opens broad possibilities for exploring in depth the characteristics of young workers and, in a longitudinal perspective, their trajectories in the labour market.

3 Application

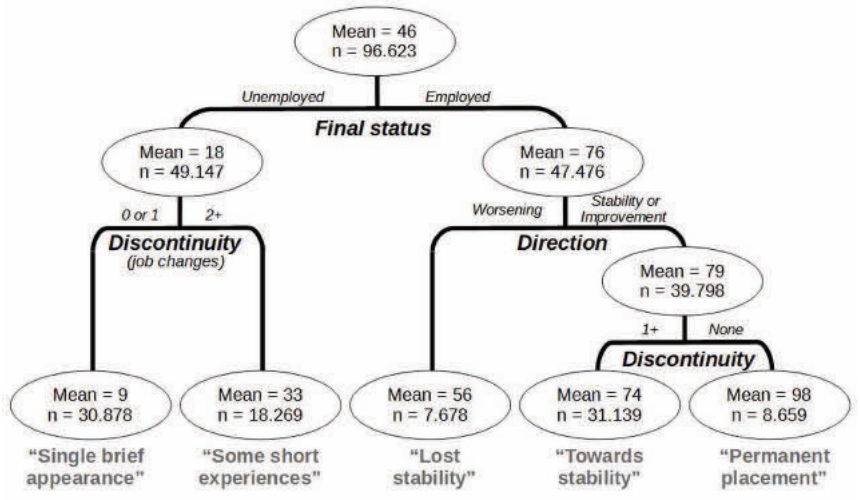
The population of interest includes the young people aged between 15 and 29 years old that were hired for the first time by a firm located in the Veneto region in 2007 (n=97,000). The working histories of these subjects were followed for 12 years after their first entrance into the labour market and five key indicators were selected to describe the main characteristics of their initial trajectories:

- the "initial status", i.e. the type of the first labour contract (stable vs fixed-term);
- the "final status", i.e. the employment condition after a 12-year follow-up, (employed vs unemployed);
- the career "direction", determined by comparing the prevalent contract in the first and in the last trimester (i.e. stability, improvement, or worsening);
- the "discontinuity" of the working paths (i.e. number of job changes);
- the "saturation rate" (i.e. the percentage of days worked in the observed period).

The CART model, built using R software, divides the entire sample (the initial parent node) into smaller, homogeneous groups (child nodes) based on a dependent variable, that in this case is the "saturation rate". The other four key indicators listed above were included in the model as predictor variables. As illustrated in Figure 1, the "final status" is the first variable that best splits data into homogeneous subgroups most

relevant to the outcome of interest. Also the "discontinuity" of the working paths has a crucial role, both among subjects employed at the end of the follow-up period and among the unemployed ones.

Figure 1: Tree plot for the "saturation rate"



The whole population was split into five classes that were named according to their main identifying characteristics as listed below:

- young people that have no employment contract open at the end of the follow-up period were classified in two groups according to the number of jobs they changed (often carrying out low-skilled professions):
  - *"Single brief appearance"* group has the lowest mean "saturation rate"; its members, mainly men and foreigners, were employed for a single short period;
  - *"Some short experiences"* group has a bit higher "saturation rate" and includes young people that were employed for short periods in at least two firms;
- young people that have a contract open at the end of the follow-up period were classified in three groups with an increasing "saturation rate":
  - *"Lost stability"* group is characterised by working careers that start with a stable contract and then move to a fixed-term job; this class shows a high percentage of apprentices and construction workers;
  - *"Towards stability"* group, on the contrary, comprises workers that start with a fixed-term job, in many cases before the age of 20, and then reach the contractual stability, usually changing firm;
  - *"Permanent placement"* includes careers that start with a stable contract which continues for the whole observed period; the subjects in this class, mainly aged

between 25 and 29 years old and often graduates, show the highest presence of qualified profiles, both in intellectual and technical professions.
The results of this analysis represent a preliminary exploration of the participation of young people in the labour market; a deep insight into their trajectories and conditions is crucial to facilitate the successful design of labour and education policies.

References

BILTON, P., JONES, G., GANESH, S., & HASLETT, S. 2017. Classification trees for poverty mapping. *Computational Statistics and Data Analysis*, **115**, 53–66.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R. A., & STONE, C.J. (Eds.). 1984. *Classification and regression trees (the wadsworth statistics/probability series)*. New York:: Chapman and Hall.

DACKO, M., ZAJAC, T., SYNOWIEC, A., OLEKSY, A., KLIMEK-KOPYR, A., & KULIG, B. 2016. New approach to determine biological and environmental factors influencing mass of a single pea (Pisum sativum L.) seed in Silesia region in Poland using a CART model. *European Journal of Agronomy*, **74**, 29–37.

DE' ATH, G., & FABRICIUS, K.E. 2000. ClassifIcation and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, **81**(11), 3178–3192.

FRISMAN, L., PRENDERGAST, M., LIN, H.J., RODIS, E., & GREENWELL, L. 2008. Applying classification and regression tree analysis to identify prisoners with high HIV risk behaviors. *Journal of Psychoactive Drugs*, **40**(4), 447–458.

GALLETTA, S. 2016. On the determinants of happiness: A classification and regression tree (CART) approach. *Applied Economics Letters*, **23**(2), 121–125.

KEELY, L.C., & TAN, C.M. 2008. Understanding preferences for income redistribution. *Journal of Public Economics*, **92**(5–6), 944–961.

LOW, C.T., & LAI, P.C. 2016. Personal factors influencing the perception of quality of life in Hong Kong–a classification tree approach. *Procedia Environmental Sciences*, **36**, 70–73.

MANASSE, P., & ROUBINI, N. 2009. "Rules of thumb" for sovereign debt crises. *Journal of International Economics*, **78**(2), 192–205.

WILLIAMS, M.A., JOSKOW, A.S., JOHSON, R.L., & HURDLE, G.J. 1987. Explaining and predicting airline yields with non-parametric regression trees. *Economics Letters*, **24**(1), 99–105.

WALEGA, G., WALEGA, A. 2021. Over-indebted Households in Poland: Classification Tree Analysis. *Social Indicator Research,* **153**, 561–584.

RESAMPLING FOR STABILITY ESTIMATION VS. CLUSTER VALIDATION VIA DATA SPLITTING AND SUBSAMPLING. WHICH APPROACH IS BETTER IN DETECTION OF CLUSTERS IN TAXONOMY?

Dorota Rozmus[1]

[1] Department of Economic and Financial Analysis, University of Economics in Katowice,
(e-mail: dorota.rozmus@ue.katowice.pl)

ABSTRACT: Due to the fact that there are no labels or gold standards by which performance of clustering can be measured, the problem of determining the right number of clusters (k) has not been solved to this day. However, new methods are proposed to ensure the best possible clustering performance.

KEYWORDS: Clustering, cluster stability, clustering performance.

1 Cluster stability

Clustering algorithms seek to partition data into groups, according to certain similarity measures. The overall goal is to place similar data points in the same cluster, and dissimilar data points in different clusters.

Due to the fact that there are no labels or gold standards by which performance can be measured, the problem of determining the right number of clusters (k) has not been solved to this day.

The concept of stability has emerged as a strategy for assessing the performance and reproducibility of data clustering. The underlying premise is that a good clustering of the data will be reproduced over perturbed datasets that are nearly identical to the original data.

Several methods have been developed for measuring of cluster stability. According to (Liu, Yu, Blair, 2021), these methods can be broken down into the following three categories: resampling for stability estimation, cluster validation via data splitting and subsampling, and alternative methods that do not adhere to these classic approaches. In this study only the first two approaches will be taken into consideration.

The aim of the research will be to compare these two approaches in the context of indicating the value of the *k* parameter (number of clusters). The study will be conducted on benchmark data sets, which are usually used in comparative studies.

References

DUDOIT, S., & FRIDLYAND, J. 2002. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, **3(7)**, 1–21.

FANG, Y., & WANG, J. 2012. Selection of the number of clusters via the bootstrap method. *Computational Statistics and Data Analysis*, **56**, 468–477.

HENNING, C. 2007. Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis*, **52**, 258–271.

LIU, T., YU, H., & BLAIR, R. H. 2022. Stability estimation for unsupervised clustering: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, **14(6)**.

ŞENBABAOĞLU, Y., MICHAILIDIS, G., & LI, J. Z. 2014. Critical limitations of consensus clustering in class discovery. *Scientific Reports 4*.

TIBSHIRANI, R., & WALTHER, G. 2005. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, **14(3)**, 511–528.

YU, H., CHAPMAN, B., DI FLORIO, A., EISCHEN, E., GOTZ, D., JACOB, M., & BLAIR, R. H. 2019. Bootstrapping estimates of stability for clusters, observations and model selection. *Computational Statistics*, **34(1)**, 349–372.

FUNCTIONAL DATA ANALYSIS APPROACH FOR IDENTIFYING REDUNDANCY IN AIR QUALITY MONITORING STATIONS

Annalina Sarra [1], Adelia Evangelista[1], Tonio Di Battista[1] and Sergio Palermi[2]

[1] Department of Philosophical, Pedagogical and Economic-Quantitative Sciences , University of Chieti-Pescara, (e-mail: annalina.sarra@unich.it, adelia.evangelista@unich.it, tonio.dibattista@unich.it)

[2] Agency of Environmental Protection of Abruzzo (ARTA), Pescara, Italy, (e-mail: s.palermi@artaabruzzo.it)

ABSTRACT: The assessment of air quality is of great importance for defining measures for pollution reduction and ensuring the public health protection. The monitoring stations are the tools established to measure and manage the compliance with national ambient air quality standards. Because these networks need considerable financial resources, many studies are aimed at identifying possible redundancy in air quality monitoring sites. Following these lines of research, we focus on ascertaining if the spatial distributions of NO_2, PM_{10}, $PM_{2.5}$ and benzene concentrations are homogenously distributed in the urban area of Pescara-Chieti (Central Italy). To this end we adopt a multivariate functional model-based clustering algorithm.

KEYWORDS: air quality, redundancy, meteorological normalization, FDA, model-based clustering.

1 Introduction

In recent decades there has been a growing interest in monitoring air pollution levels, especially in urban areas. Countries all over the world have set up air quality monitoring networks for collecting unbiased, accurate and comparable data on the air quality and supporting policies that lessen the impact on human health and the environment. In order to save money and avoid data duplication, it is preferable to use the fewest number of stations possible to meet monitoring goals. There are numerous studies in the literature that look for potential redundancy in air quality monitoring networks (see Wilson *et al.*, 2005 for a review). The majority of them concentrate on determining whether or not the pollutant is uniformly distributed throughout the area and on the intra-urban

variation of air pollutant concentrations. In this study, we address the problem of identifying possible redundancy in air quality monitoring stations using the FDA (Ramsay & Silverman, 2005) paradigm. FDA has gained considerable interest in the literature over the past two decades, and several benefits of using FDA over conventional vectorial approaches have been emphasized, such as the possibility to extract more information from the data (the smoothness of the data structure, rate of change, acceleration, and dynamic changes over a large-scale domain). In this work, we analyze the multivariate air pollution concentrations using a multivariate functional model-based clustering approach proposed by Schmutz *et al.*, 2020. The data set used is comprised of hourly measurements of air quality and weather data obtained from the automatic reporting platform operated by the Regional Agency for Environmental Protection of the Abruzzo Region (ARTA), in the urban area of Pescara-Chieti (Central Italy). We also implement a meteorological normalization to control for changes in the weather and lower the variability in air quality time series. The remainder of this paper is structured as follows. Section 2 describes the study area and the data used for the analysis, as well as the meteorological normalization procedure conducted. Section 3 provides background information on the functional clustering algorithm employed. Finally, Section 4 conveys the main findings of the analysis.

2 Study area and data

The study focused on the Chieti-Pescara urban area in the Abruzzo Region (Central Italy), which includes the conurbation of the major cities Pescara and Chieti, and the neighboring municipalities of Montesilvano and Francavilla al Mare. It is a nearly flat area located in the terminal stretch (about 15 km long) of the Pescara river valley, which flows into the Adriatic Sea. The valley industrial and vehicular traffic are the main contributors to air pollution, with domestic heating having a sizable impact during the winter. For this study, we take into account NO_2, PM_{10}, $PM_{2.5}$ and benzene measurements obtained from ARTA automatic reporting platform between January 2017 and December 2019 at 5 five monitoring sites divided into two categories: urban background (3 sites: Teatro d'Annunzio, Chieti, and Francavilla) and urban traffic (2 sites: Via Firenze and Montesilvano). The dataset also includes the following meteorological factors: wind speed, wind direction, temperature, relative humidity, solar radiation, air pressure and precipitation, measured on the ground at air quality monitoring stations. Since weather strongly influences pollutants formation and transport, in this paper we consider a meteorologi-

cal/weather normalization. More specifically, in our air quality data analysis over time, we control for changes of meteorology by means of boosted regression trees, as implemented in the R package deweather (Carslaw, 2021).

3 Model based clustering algorithm

The main steps involved by the model-based clustering algorithm for high-dimensional data (fun-HDDC) introduced by Schmutz *et al.*, 2020 can be summarized as follows. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are the observed multivariate curves, representing in our case the air quality data. The goal is to group them into K homogenous clusters, where K is fixed a priori. The core idea is to transform the high-dimensional data into group-specific subspaces. For each group k ($k = 1, \dots K$), let $d_k < R$ denote the intrinsic dimension of a low dimensional latent subspace in which the curve of each cluster could be described. Through a principal component analysis for multivariate functional data, curves are expressed into a group-specific basis

$$\varphi_r^k(t) = \sum_{l=1}^{R} q_{krl}\phi_l(t), 1 \leq r \leq R \tag{1}$$

obtained through a linear transformation from the matrix of principal factors $\left\{\phi_r^j\right\}_{1\leq j\leq p, 1\leq r\leq R}$ where q_{krl} are the basis expansion coefficients of the eigenfunctions, contained in an orthogonal matrix $R \times R$. Thus, each multivariate curve n_k, of cluster k, can be represented by its score $(\delta_i^k)_{1\leq i\leq n_k}$.
The scores are assumed to follow a Gaussian distribution $\delta_i^k \sim N(\mu_k, \Delta_K)$ with $\mu_k \in R^R$ the mean function and Δ_K the corresponding covariance matrix. Actually, the novel approach (fun-HDDC) is an extension of the work of Jacques & Preda, 2014, and it is advantageous from two perspectives: modeling all principal component scores with estimated variances that are not zero, and proposing a criterion for choosing the number of clusters using the expectation-maximization (EM) algorithm.

4 Results

In this section, we present the results obtained through the use of the algorithm illustrated in Section 3. All the analyses were performed using the R packages *fda* and *funHDDC* (R Core Team, 2022). The observed pollutant time series and the meteorological normalized pollutant time series have been

transformed into functional data with a process of smoothing, with 30 basis and cubic B-spline. In either instances, the model-based clustering algorithm applied here is the [AkjBQkDk] model (see, Schmutz *et al.*, 2020 for more details) and provided the partition of monitoring stations into two groups. We find out that the composition of the identified groups does not change after performing a meteorological normalization: the first cluster contains the monitoring stations of Chieti and Francavilla whereas the remaining monitoring sites of Via Firenze, Montesilvano and Teatro d'Annunzio are grouped in the second cluster. For NO_2, PM_{10} and $PM_{2.5}$, we observe that cluster 2 exhibits higher values throughout the period considered than cluster 1; conversely for benzene an opposite behaviour is recorded.

Interestingly, the functional multivariate clustering algorithm reveals a potential misclassification since the Pescara urban background station of "Teatro d'Annunzio" is grouped with two traffic stations. This result highlights the peculiarity of the municipality of the Pescara, characterized by a considerable population density and a capillary road network, with high volumes of traffic that insists on an area little extended. In this context, urban traffic emissions represent the dominant source of atmospheric pollution and make background stations similar to traffic ones.

References

CARSLAW, D.C. 2021. Deweather-An R Package to Remove Meteorological Variation from Air Quality Data. *Available online:https://github.com/davidcarslaw/deweather*.

JACQUES, J., & PREDA, C. 2014. Model based clustering for multivariate functional data. *Computational Statistics and Data Analysis.*, **71**, 92–106.

R CORE TEAM. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAMSAY, J.O, & SILVERMAN, B.W. 2005. *Functional data analysis, 2nd edn.* New York: Springer-Verlag.

SCHMUTZ, A., JACQUES, J., BOUVEYRON, C., CHÉZE, L., & MARTIN, P. 2020. Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics.*, **35**, 1101–1131.

WILSON, G., KINGHAM, S., PEARCE, J., & STURMAN, A. 2005. A review of intraurban variationsin particulate air pollution: implications for epidemiological research. *Atmospheric Environment.*, **34**, 6444–6462.

STUDENT MOBILITY IN HIGHER EDUCATION: A DESTINATION-SPECIFIC LOCAL ANALYSIS

Luca Scaffidi Domianello[1]

[1] Department of Statistics, Computer Science, Applications, University of Florence
(e-mail: luca.scaffididomianello@unifi.it)

ABSTRACT: Student mobility flows are usually analyzed through gravity models. However, researchers devote less attention to the potential spatial heterogeneity in the estimated parameters: indeed local analysis is a crucial task within the Italian territory, where, as a consequence of the decentralization process, there are universities with national or local vocation. Then, in the empirical analysis of our work, we estimate the parameters for each university to identify their catchment area: the results show different interaction behaviors among Italian universities.

KEYWORDS: gravity models, student mobility, higher education, Poisson regression.

1 Introduction

Mobility of students across a country in higher education has gained increasing attention in the last years, due to the socio-economic impact of such a migration. Flows between an origin and a destination are usually analyzed through gravity models, which rely on Newton's law of universal gravitation: the interactions among two areas are proportional to the product of their "masses" (attraction effect) and inversely proportional to their distance (deterrence effect). Generally, the employed gravity models in literature, in order to explain the flows between the area of origin and the university of destination, assume the same relationship for each origin and each destination (see e,g., Sa *et al.*, 2004, for the Dutch universities, and Bacci & Bertaccini, 2021, for the Italian ones), then they do not consider possibly different interaction dynamics.

In the present contribution, we consider a destination-specific gravity model to obtain disaggregated information for each university (Haynes & Fotheringham, 1984): allowing model parameters to vary across the space is a crucial task in heterogeneous countries like the Italian one, where many students coming from the South decide to study in the universities located in the North. Furthermore, we allow the distance parameter, reflecting the deterrence effect, to vary among three thresholds: less than 250 kilometers, between 250 and

500 kilometers, and more than 500 kilometers. The idea is that if we have an increasing value of the parameter as we move from the lowest classes to the highest ones, the university has a national vocation rather than local. Furthermore, identifying the catchment area is of interest for university administrators for their marketing strategies. More in detail, in this work we focus on students enrolled in Science & Technologies (S&T) courses: they are of particular interest for the policy makers, because they are directly related to the technological development of the area where the university is located (Dotti *et al.*, 2014).

The data employed for this work comes from the Italian National Student Registry (in Italian, Anagrafe Nazionale Studenti - ANS), the Italian administrative database that records the students, by their province of residence, enrolled in any degree program in a certain university located in Italy.

The work is structured as follows: Section 2 describes the model we employ for the empirical analysis, Section 3 analyzes the data and comments the estimation results, and Section 4 offers some concluding remarks.

2 Theoretical Model

In order to analyze student mobility in higher education we rely on gravity models, a useful tool to describe people flows over a geographic area. By assuming the flow T_{ij}, denoting the number of students moving from the province of residence i ($i = 1,\ldots,I$) to the university of destination j ($j = 1,\ldots,J$), as an outcome of a Poisson process, its conditional mean λ_{ij} can be expressed as follows (see, e.g., Flowerdew & Aitkin, 1982):

$$\lambda_{ij} = exp\left(k + \sum_{p=1}^{P} \alpha_p \log x_{ip} + \sum_{q=1}^{Q} \beta_q \log z_{jq} + \gamma \log d_{ij}\right) \tag{1}$$

where x_{ip} and z_{jq} are the explanatory variables measuring origin propulsiveness and destination attractiveness, respectively, d_{ij} is the road distance, expressed in kilometers, between each origin and destination (that we expect has a negative influence on the student flows), k is a constant of proportionality, while α_p, β_q and γ are the other parameters to be estimated. In the model specified above, we assume the same relationship for each origin and each destination. Then, we can obtain disaggregated information if we estimate the model for each university (see Haynes & Fotheringham, 1984), thus obtaining destination-specific parameters.

As opposed to log-normal models, the Poisson regression allows us to deal with the problem of zero-valued flows, while this is not the case when we have

to use the log-transformation of the dependent variable. Then, we estimate the parameters through the Poisson-Quasi-Maximum-Likelihood estimation (QMLE) technique, in order to obtain consistent estimates of the parameters even if the assumed distribution is no more valid, except the correct specification of the conditional mean, as it could be the case when dealing with a large amount of zero-valued flows.

3 Empirical Analysis

The data of this work come from the ANS, the Italian administrative database that records students' enrollment in Italian universities, by their province of residence. The analysis focuses on students enrolled in a bachelor or five-years degree program for the academic year 2011-2012. More specifically, we consider a subset of students, those attending S&T courses (ISCED 5, 6 and 7) due to their relevance for local technological development. As proxy for the origin propulsiveness, we use the total number of students resident in province i (O Mass), while for the distance, we allow its parameter to vary according to its belonging to one of the categories defined by the following thresholds: less than 250 kilometers, between 250 and 500 kilometers, and more than 500 kilometers. Table 1 reports the summary results of the estimated destination-specific gravity model: as we can see, there is a lot of variation in the value of the coefficients among the universities, thus supporting the hypothesis of spatial heterogeneity. For lake of space, we do not report the estimates for each destination, but we find that the universities offering very specialized degree programs (e.g., Polytechnic universities of Milan and Turin) show an increasing value of the deterrence effect (national vocation) as opposed to universities with a decreasing level (local vocation), this is the case for most of the universities of the South (Bacci & Bertaccini, 2021).

4 Conclusion

This work analyzes S&T student mobility flows in Italian higher education through a Poisson gravity model. More specifically, we allow the parameters to vary across institutions to detect some heterogeneity in the interaction behaviors. Empirical analysis supports our hypothesis thus allowing us to discriminate among universities with national vocation as opposed to universities with local vocation: this is of relevant interest for university administrators in implementing their orientation strategies aimed at high school students. As

Table 1. *Summary results of the estimated Poisson destination-specific gravity models for student mobility in higher education.*

	Coefficients					
	Min	Mean	Max	1Q	2Q	3Q
Constant	-6.604	3.476	12.591	0.831	3.395	5.749
d < 250	-2.156	-0.688	-0.261	-0.788	-0.656	-0.536
250<d<500	-4.479	-1.394	-0.568	-1.293	-1.123	-0.961
d>500	-3.954	-1.395	-0.417	-1.402	-1.04	-0.789
O Mass	-0.863	0.347	1.833	0.071	0.324	0.635
Pseudo-R^2	0.843	0.505	0.981	0.795	0.864	0.933

future research, it could be interesting to obtain deeper information by allowing the parameters to vary according to the origin through the Geographically Weighting Regression (GWR) technique (Fotheringham *et al.*, 1998).

References

BACCI, S, & BERTACCINI, B. 2021. Assessment of the university reputation through the analysis of the student mobility. *Social Indicators Research*, **156**, 363–388.

DOTTI, NICOLA FRANCESCO, FRATESI, UGO, LENZI, CAMILLA, & PERCOCO, MARCO. 2014. Local labour market conditions and the spatial mobility of science and technology university students: evidence from Italy. *Review of Regional Research*, **34**, 119–137.

FLOWERDEW, ROBIN, & AITKIN, MURRAY. 1982. A method of fitting the gravity model based on the Poisson distribution. *Journal of regional science*, **22**(2), 191–202.

FOTHERINGHAM, A STEWART, CHARLTON, MARTIN E, & BRUNSDON, CHRIS. 1998. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and planning A*, **30**(11), 1905–1927.

HAYNES, KINGSLEY E, & FOTHERINGHAM, A STEWART. 1984. Gravity and spatial interaction models. *CA: Sage Publications*.

SA, CARLA, FLORAX, RAYMOND JGM, & RIETVELD, PIET. 2004. Determinants of the regional demand for higher education in the Netherlands: A gravity model approach. *Regional Studies*, **38**(4), 375–392.

RESIDUALS DIAGNOSTICS FOR MODEL-BASED TREES FOR ORDERED RATING RESPONSES

Rosaria Simone [1]

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

ABSTRACT: The contribution illustrates how selection of model-based trees can be supplemented by local diagnostics on a necessary condition for the correct specification of the baseline model, based on surrogate residuals' analysis. The procedure can support the choice of the baseline model or the tuning of pre-pruning conditions. Examples are given for MOB trees based on ordinal logit models.

KEYWORDS: Model-based tree, ordered data, surrogate residuals

1 Motivating framework

The contribution discusses the advantages of performing residuals diagnostics for ordinal data models (Liu & Zhang, 2018) in the setting of model-based classification trees. Specifically, a necessary condition for a model to be correctly specified is that surrogate residuals are uniformly distributed. The paper shows the procedure for model-based trees (Zeileis *et al.*, 2008) with ordinal logit models to tune pre-pruning conditions, to identify the nodes that should be preferably pruned, or to select the best tree in terms of the maintained local model. For illustration, we consider data from the 5th European Working Condition Survey carried out in 2010 and focus on $N = 972$ responses for Italy to the question 'Do you experience stress in your work?' on a $m = 5$ wording-type scale: 'Always', 'Most of the time', 'Sometimes', 'Rarely', 'Never' [*].[†]

[*]Coded from 1 to 5 for convenience

[†]To avoid bias in favour of variables with many splits, we consider as covariates dichotomous factors *Gender* (G) experience of *Insomnia* (I), experience of *Fatigue* (F), experience of *Depression* (D), presence of *Risk* (R) connected to the job stability, being the Household Breadwinner (B). The only non dichotomous covariate is the size of the *Household* (H) as number of components.

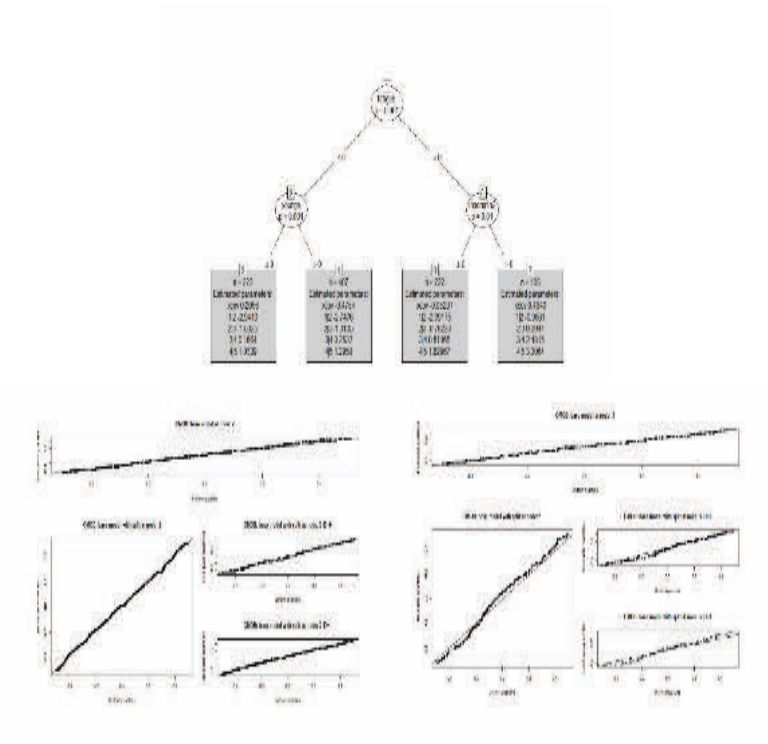


Figure 1. *MOB for M* : *Stress* $\sim$ *Gender (Top); Residuals' diagnostics for MOB based on M on perceived work-related stress (bottom)*

2 Residuals diagnostics of MOB trees for ordered responses

In the setting of MOB trees [‡], consider an ordered logit model $M: \ logit(Pr(R_i \leq j|x_i)) = \alpha_j - \beta_1 x_i, \ j = 1,\ldots,m$ as local maintained model. For instance, let M : *Stress* $\sim$ *Gender* (see the top panel in Figure 1). Then, Figure 1 (bottom) displays the uniform QQ plot of residuals at inner nodes and descendants, showing that the split at node 5 should be preferably pruned as M does not meet locally the necessary condition for being correctly specified.

Then, consider model M : *Stress* $\sim$ *Breadwinner* and the corresponding MOB with `minsplit=50`, `maxdepth=4` (see Figure 2 - left). Uniform QQ plots of residuals' at tree nodes are displayed in Figure 3, showing that - except for node 3 and its descendants - there is poor evidence for M being correctly specified locally. Then, modifying the pre-pruning condition on

[‡](`partykit` R package)

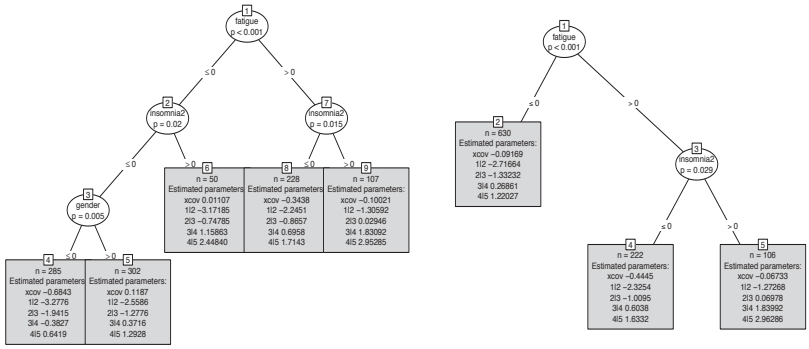


Figure 2. *MOB tree for M* : *Stress* $\sim$ *Breadwinner with* `minsplit = 50` *(left) and* `minsplit=100` *(right)*

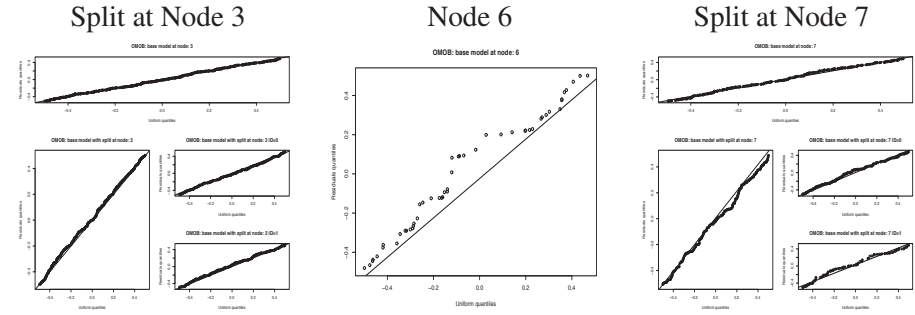


Figure 3. *QQ plot of surrogate residuals for a MOB tree based on M* : *Stress* $\sim$ *Breadwinner (*`minsplit=50`*)*

minimum sample size required to attempt a split (`minsplit = 100`) yields a reduced MOB tree (see Figure 2 - right), with evidence that the necessary condition for being correctly specified is overall satisfied (see Figure 4).

3 Concluding remarks

Residuals diagnostics in the setting of model-based trees can be successfully exploited also for trees based on CUB models (Cappelli *et al.*, 2019) to select the baseline model or the best performing partitioning criterion. The proposed procedure can be further integrated within model selection in order to focus only on models for which the necessary condition for being correctly specified

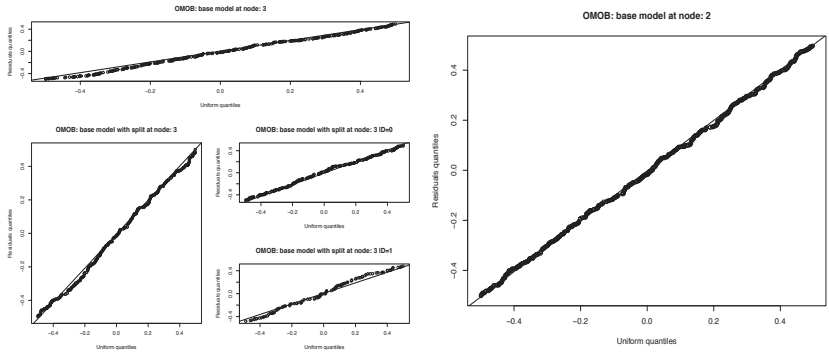


Figure 4. *Uniform QQ plot for local diagnostics on model M* : *Stress* $\sim$ *Breadwinner (minsplit=100)*

can be maintained. For instance, local uncertainty diagnostics of Binomial classification trees for rating data has been advanced in Simone, 2023. Further studies will investigate the impact of residual diagnostics on the derivation of variable importance measures from model-based tree ensembles.

References

CAPPELLI, C., SIMONE, R., & DI IORIO, F. 2019. CUBREMOT: a tool for building model-based trees for ordinal responses. *Expert Systems with Applications*, **124**, 39–49.

LIU, D., & ZHANG, H. 2018. Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach. *Journal of the American Statistical Association*, **113(522)**, 845–854.

SIMONE, R. 2023. Uncertainty diagnostics of Binomial Regression Trees for Ordered Rating Data. *Journal of Classification*, **40**, 79–105. 10.1007/s00357-022-09429-5.

ZEILEIS, A., HOTHORN, T., & HORNIK, K. 2008. Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, **17**, 492–514.

HIDDEN MARKOV MODELS FOR MULTIVARIATE LONGITUDINAL DATA

Alexa Sochaniwsky [1] and Paul D. McNicholas[1]

[1] Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada (e-mail: sochaa1@mcmaster.ca, paul@math.mcmaster.ca)

ABSTRACT: A method for handling the unique correlation structure that can occur in longitudinal data is introduced for hidden Markov models. This approach uses a family of independent mixture models that apply a variety of constraints to the covariance matrix, which is then used in hidden Markov models, i.e, dependent mixture models.

KEYWORDS: clustering, hidden Markov models, longitudinal data, EM algorithm.

1 Introduction

Longitudinal data is information that is collected on several subjects across several points in time. Longitudinal studies are often used in clinical or sociological research, but difficulties may arise as the correlation that can occur between subjects must be accounted for. For certain longitudinal studies, it would be useful not only to cluster the subjects but to model the transitions between states. The change in state can be modeled by hidden Markov models (HMMs). Efforts have been made in regression models, specifically AR and MA models (Hasan & Sneddon, 2009; Sutradhar, 2003), and in independent mixture models (McNicholas & Murphy, 2010) to account for the unique longitudinal correlation structure. This research modifies the EM algorithm for HMMs by using the covariance structures from the Cholesky-decomposed Gaussian mixture model (CDGMM) family (McNicholas & Murphy, 2010).

2 Background

A hidden Markov model comprises of two processes, an unobserved parameter process and an observed state-dependent process. The simplest HMM for longitudinal data can be defined as

$$P(C_{it}|\mathbf{C}^{(it-1)}) = P(C_{it}|C_{it-1}), \qquad \text{for } i = 1,...,n,\ t = 2,3,...,T$$
$$P(X_{it}|\mathbf{X}^{(it-1)},\mathbf{C}^{(t)}) = P(X_{it}|C_{it}), \qquad \text{for } i = 1,...,n,\ t = 1,...,T$$

where $\mathbf{C}^{(it)}$ represents the history of the unobserved parameter process $\{C_{it} : i = 1,...,n, t = 1,2,...,T\}$ with state space $C = 1,...,m$, and $\mathbf{X}^{(it)}$ represents the history of the state-dependent process $\{X_{it} : i = 1,...,n, t = 1,2,...,T\}$. The parameter process C_{it} satisfies the Markov property and is then used in the distribution of the state-dependent process X_{it}.

A common method for maximum likelihood estimation of an HMM is the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). An EM for HMMs is called the Baum-Welch algorithm (Baum *et al.*, 1970, 1972; Welch, 2003). Specifically, it is an EM for a hidden Markov model whose Markov chain is homogeneous. By assuming a homogeneous HMM, the parameter estimates have closed form solutions. The parameters are derived from the complete-data log-likelihood given by

$$l(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \left\{ \sum_{g=1}^{m} u_{i1g} \log \delta_i + \sum_{t=2}^{T} \sum_{g=1}^{m} \sum_{k=1}^{m} v_{itgk} \log \gamma_{gk} + \sum_{t=1}^{T} \sum_{g=1}^{m} u_{itg} \log f(x_{it}|S_{it} = g) \right\},$$

where $\boldsymbol{\vartheta}$ denotes the vector containing the model parameters, δ_i is the stationary distribution, γ_{gk} are the transition probabilities, the unknown labels $u_{itg} = 1$ if the observation i is in state g at time t and $u_{itg} = 0$ otherwise, and the other unknown labels $v_{itgk} = 1$ if the observation i is in state g at time $t-1$ and in state k at time t, and $v_{itgk} = 0$ otherwise.

For longitudinal data, McNicholas & Murphy (2010) use a Gaussian (independent) mixture model with a modified Cholesky decomposed covariance structure (Pourahmadi, 1999, 2000) such that the precision matrix $\boldsymbol{\Sigma}$ can be decomposed into $\boldsymbol{\Sigma}^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{T}$, where $\mathbf{T}$ is a unique unit lower triangular matrix and $\mathbf{D}$ is a unique diagonal matrix with strictly positive diagonal entries. For a p-dimensional random variable $\mathbf{X}$, the multivariate Gaussian mixture model with the modified-Cholesky decomposition, the gth component density is given by

$$f(\mathbf{x}|\boldsymbol{\mu}_g, (\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g)^{-1}) = \frac{1}{\sqrt{(2\pi)^p |\mathbf{D}_g|}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)'\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g(\mathbf{x} - \boldsymbol{\mu}_g) \right\}.$$

A family of eight Gaussian mixture models are constructed by constraining $\mathbf{T}_g$ and/or $\mathbf{D}_g$ with the option to impose the isotropic constraint $\mathbf{D}_g = \delta_g \mathbf{I}_g$. This family is called the Cholesky-decomposed Gaussian mixture model (CDGMM) family. The nomenclature, covariance structure, and number of free covariance parameters for all models are displayed in Table 1.

Table 1. *CDGMM Family*

Model	$\mathbf{T}_g$	$\mathbf{D}_g$	$\mathbf{D}_g$	Free Cov. Parameters
EEA	Equal	Equal	Anisotropic	$p(p-1)/2+p$
VVA	Variable	Variable	Anisotropic	$m[p(p-1)/2]+mp$
VEA	Variable	Equal	Anisotropic	$m[p(p-1)/2]+p$
EVA	Equal	Variable	Anisotropic	$p(p-1)/2+mp$
VVI	Variable	Variable	Isotropic	$m[p(p-1)/2]+m$
VEI	Variable	Equal	Isotropic	$m[p(p-1)/2]+1$
EVI	Equal	Variable	Isotropic	$p(p-1)/2+m$
EEI	Equal	Equal	Isotropic	$p(p-1)/2+1$

Constraining $\mathbf{T}_g$ such that $\mathbf{T}_g = \mathbf{T}$ suggests that all states have the same correlation structure. Constraining $\mathbf{D}_g$ such that $\mathbf{D}_g = \mathbf{D}$ suggests that all states have the same variability at each time point and the isotropic constraint $\mathbf{D}_g = \delta_g \mathbf{I}_p$ suggests that the variability at each time point is the same. All models would be fitted using an EM algorithm and then based on a model selection criterion, one would be selected.

3 Methodology

We propose modifying the M-step in the EM algorithm for a Gaussian HMM by substituting the 'traditional' covariance update, i.e.,

$$\mathbf{\Sigma}_g = \frac{1}{n_g} \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{u}_{itg} (x_{it} - \boldsymbol{\mu}_g)(x_{it} - \boldsymbol{\mu}_g)'$$

where $n_g = \sum_{i=1}^{n} \sum_{t=2}^{T} \hat{u}_{itg}$, with a member of the CDGMM family. This modified algorithm is outlined in Algorithm 1.

Algorithm 1 EM Algorithm for Gaussian HMM

initialize $\boldsymbol{\delta}$ and $\boldsymbol{\Gamma}$
initialize u_{itg} and v_{itgk}
while convergence criterion is not met **do**
    update $\hat{u}_{itg}$, $\hat{v}_{itgk}$
    update γ_{gk}, δ_g
    update $\hat{\boldsymbol{\mu}}_g$
    update $\hat{\mathbf{T}}_g$, $\hat{\mathbf{D}}_g$
    update $\hat{\boldsymbol{\Sigma}}_g^{-1} = \hat{\mathbf{T}}_g' \hat{\mathbf{D}}_g^{-1} \hat{\mathbf{T}}_g$
    check convergence criterion
end while

References

BAUM, LEONARD E, PETRIE, TED, SOULES, GEORGE, & WEISS, NORMAN. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, **41**(1), 164–171.

BAUM, LEONARD E, *et al.* 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, **3**(1), 1–8.

DEMPSTER, ARTHUR P, LAIRD, NAN M, & RUBIN, DONALD B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

HASAN, M TARIQUL, & SNEDDON, GARY. 2009. Zero-inflated Poisson regression for longitudinal data. *Communications in Statistics—Simulation and Computation®*, **38**(3), 638–653.

MCNICHOLAS, PAUL D, & MURPHY, T BRENDAN. 2010. Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, **38**(1), 153–168.

POURAHMADI, MOHSEN. 1999. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**(3), 677–690.

POURAHMADI, MOHSEN. 2000. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**(2), 425–435.

SUTRADHAR, BRAJENDRA C. 2003. An overview on regression models for discrete longitudinal responses. *Statistical Science*, **18**(3), 377–393.

WELCH, LLOYD R. 2003. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, **53**(4), 10–13.

K-MEANS CLUSTERING – NEW VARIATIONS

Andrzej Sokołowski[1], Małgorzata Markowska[1] and Maciej Laburda[2]

[1] Krakow University of Economics, Poland
(e-mail: sokolows@uek.krakow.pl)

[2] Wroclaw University of Economics and Business, Poland,

(e-mail: malgorzata.markowska@ue.wroc.pl)

3 Krakow University of Economics, Poland

(e-mail: s223060@student.uek.krakow.pl)

ABSTRACT: k-means is one of the most popular methods in cluster analysis. It can handle the large set of data since there is no need to store the distance matrix in the memory, and the algorithm converges very quickly to the situation when no object should be relocated (each one is closer to the mean of its "own" cluster, that to the other one). Two main drawbacks of the method are that the number of clusters should be defined properly and that the final partition tends to be formed by spherical clusters. In the literature, there are many variations, improvements, and new versions of k-means based on the original model.
In this contribution we discuss two new ideas. The first one can be called *n%-neighbors k-means*. When we have to decide whether an object should be relocated to another cluster, we consider only some percentage of the total set of objects, only points closest to the one which is considering at the moment. So partial means should be calculated and considered. It is possible that some distance clusters will not be taken into account if their members are not included in n% nearest neighbors of this point. The second new proposition can be called *local standardization k-means.* Standarization is performed separately for each cluster, using its mean and standard deviation, excluding point which is considered for relocation. Than this point is "standardized" using means and standard deviations of consecutive clusters and distances are calculated.
Simulation analysis is the main tool to evaluate the quality of the proposed approaches.

KEYWORDS: k-means, nearest neighbors, standardization.

References

MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. In: *Fifth Berkeley Symposium on Mathematics. Statistics and Probability.* University of California Press, 281-297

BOCK, H.-H. 2008. Origins and extentions of the k-means algorithm in cluster analysis. *Electronic Journal for History of Probability and Statistics*, **4** (2), 1-18

JAIN, A.K. 2009. Data Clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**(8), 651-666

A STATA IMPLEMENTATION OF CLUSTER WEIGHTED MODELS: THE CWMGLM PACKAGE

Daniele Spinelli [1], Salvatore Ingrassia [2] and Giorgio Vittadini [1]

[1] Department of Statistics and Quantitative Methods, University of Milano Bicocca (e-mail: daniele.spinelli@unimib.it, giorgio.vittadini@unimib.it)

[2] Department of Economics and Business ,University of Catania, (e-mail: s.ingrassia@unict.it)

ABSTRACT: The Cluster-Weighted Model (CWM) is a member of the family of the Mixtures of Regression Models and it is referred as Mixture of Regression with Random Covariates. Currently, the only procedure for estimating these models is R package **flexcwm**. The aim of this article is to introduce a new software component, the Stata package **cwmglm** which estimates CWMs based on the most common generalized linear models. Our software also extends to Stata users the possibility of estimating parsimonious models of Gaussian distributions with alternative specifications of the variance matrix. **cwmglm** also calculates the the generalized coefficients of determination and bootstrap standard errors that are not currently available in **flexcwm**. We illustrate the use of **cwmglm** with real data on Covid-19 admissions.

KEYWORDS: cluster weighted models, clustering, parsimonious models, Stata.

1 Introduction

The *Cluster-Weighted Model* (CWM) is a member of the family of the Mixtures of Regression Models and it is also referred as Mixture of Regression with Random Covariates. The model has been first proposed under Gaussian assumptions (Gershenfeld *et al.*, 1999). Assuming random covariates relaxes the assumption of assignment independence by allowing the component distribution of the covariates to affect the assignment of the observations to the mixture components (Mazza *et al.*, 2018). A CWM models parametrically the joint density $p(x,y)$ of response variable Y and covariates X using the conditional density $p(y|x)$ and the marginal density $p(x)$. In Ingrassia *et al.* (2012) the CWM has been formulated in the statistical framework under Gaussian assumptions and Ingrassia *et al.* (2015) introduced a broad family of CWMs modeling discrete responses in which the conditional densities are assumed to belong to the exponential family and the covariates are of mixed-type. For

such models, Di Mari *et al.* (2019) and Ingrassia & Punzo (2020) introduced local and overall coefficients of determination based on the decomposition of the deviance.

From the software point of view, Mazza *et al.* (2018) underlined the scarcity of packages aimed at estimating CWMs, the same authors developed **flexcwm** for R. To our knowledge, no other software is currently available. The aim of this article is to address such gap by introducing **cwmglm**, a Stata package focused on CMWs. Our software component is based on the framework of Ingrassia *et al.* (2012) and Ingrassia *et al.* (2015) and estimates mixtures of generalized linear models (GLMs) with random covariates. The supported families are Gaussian, Poisson and binomial. The supported marginalizations for the covariates are multivariate Gaussian, multinomial, binomial, and Poisson. The variance matrix of multivariate Gaussian covariates is parametrized according to Celeux & Govaert (1995). This feature is introduced in Stata for the first time with **cwmglm**. Other than extending the possibility of estimating CMWs to Stata users, **cwmglm** introduces new internal validity measures based on the generalized coefficient of determination and bootstrap-based inference, these features are not available in **flexcwm**.

2 Cluster Weighted Models

Assume a sample $(x_1, y_1), \ldots, (x_n, y_n)$ concerning a response variable Y and a set of covariates X. Assume that the sample comes from a heterogeneous population formed by K latent classes. The CWM models the density of (Y, X) as outlined by Equation 1.

$$p(x, y, \theta) = \sum_{j=1}^{K} \pi_j p(y|x; \zeta_j) q(x; \psi_j) \tag{1}$$

In Equation 1, π_j is the mixing proportion of latent class j, $p(y|x; \zeta_j)$ is the class j-specific conditional density of the response variable and $q(x; \phi_j)$ is the marginal density of X in class j. Densities are characterized by parameters ζ_j and ϕ_j to be estimated. In our framework, the conditional density belongs to the exponential family and it is modeled as a GLM, while the marginal density $q(x; \psi_j)$ is modeled according to the Gaussian, Bernoulli, multinomial and Poisson distributions. Parameters are obtained by maximizing the log-likelihood corresponding to the density of Equation 1 using the expectation-maximization (EM) algorithm. Assuming $p(y|x; \zeta_j) = 1$ in Equation 1 leads

to a mixture of distributions, while $q(x;\psi_j) = 1$ leads to a finite mixture of regressions (FMR).
In Equation 1, assuming multivariate Gaussian covariates implies that $q(x;\psi_j) = \phi(\boldsymbol{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ where $\boldsymbol{\mu}_j$ is the mean vector and $\boldsymbol{\Sigma}_j$ is the variance matrix for latent class j (to be estimated). The eigenvalue decomposition of the variance matrix $\boldsymbol{\Sigma}_j = \lambda_j \boldsymbol{D}_j \boldsymbol{A}_j \boldsymbol{D}_j'$ (Celeux & Govaert, 1995) can be used to model cluster volume, shape and orientation. Combining constraints on λ_j (class volume), $\boldsymbol{D}_j$ (orientation) and $\boldsymbol{A}_j$ (shape) define fourteen parsimonious models. Specifically, clusters may be constrained to have equal or variable volume, spherical, equal or variable shape and axis-aligned, equal or variable orientation. For example, possible specifications may be based on the assumption that clusters have equal volume, equal shape, equal orientation (EEE) or that cluster are characterized by variable volume, equal shape and variable orientation (VEV).

3 The cwmglm package

The **cwmglm** module is available in the Statistical Software Components (SSC) archive, can be installed by using the Stata command *ssc install cwmglm* and fits CWMs as mixtures of the most common GLMs with random covariates. To our knowledge, the features of **cwmglm** are completely new for Stata users as CWMs are not estimable with the current availability of Stata commands. Indeed, **gsem** and **fmm** are only capable to estimate FMR and mixtures of distributions, which are nested in CMWs and estimable using **cwmglm**. In **cwmglm**, the parametrization of the class j-specific variance matrix of multivariate Gaussian covariates is based on Celeux & Govaert (1995). Such models are available in R packages such as **mclust** (Fraley & Raftery, 2007) and **clustvarsel** (Scrucca & Raftery, 2018) but not in Stata. Estimation of models with variable orientation and equal shape is based on Browne & McNicholas (2014).
R users can estimate CWMs using **flexcwm**; our package is related to it by extending its capability to Stata. Further, **cwmglm** provides some new procedures based on novel deviance-based measures of model fit (Di Mari *et al.*, 2019) and bootstrap standard errors.
Moreover, besides controlling the number of EM iterations, **cwmglm** users can control the number of iterations of the maximization procedures occurring during each EM iteration. This option is useful when, during a single maximization step, models requiring iterative estimation such as GLMs fail to converge. Such feature is not available in **flexcwm**.

4 Empirical example

The dataset includes a random sample of 1000 hospital admissions during the first Covid-19 wave (Feb 2020 - May 2020) in the hospital of Brescia, Italy. The response variable is the length of stay in days. The covariates are the day of admission, the patient's demographic characteristics, procedures and comorbidities. The empirical strategy is concerned in estimating CWMs for different numbers of mixture components and compare their fit.

References

BROWNE, R., & MCNICHOLAS, P. 2014. Estimating Common Principal Components in High Dimensions. *Advances in Data Analysis and Classification*, **8**(2), 217–226.

CELEUX, G., & GOVAERT, G. 1995. Gaussian Parsimonious Clustering Models. *Pattern recognition*, **28**(5), 781–793.

DI MARI, R., INGRASSIA, S., & PUNZO, A. 2019. A Generalized Coefficient of Determination for Mixtures of Regressions. *Pages 27–35 of: Conference of the International Federation of Classification Societies*. Springer-Verlag.

FRALEY, C., & RAFTERY, A. 2007. Model-Based Methods of Classification: Using the mclust Software in Chemometrics. *Journal of Statistical Software*, **18**, 1–13.

GERSHENFELD, N., SCHÖNER, B., & METOIS, E. 1999. Cluster-Weighted Modelling for Time-Series Analysis. *Nature*, **397**, 329–332.

INGRASSIA, S., & PUNZO, A. 2020. Cluster Validation for Mixtures of Regressions Via the Total Sum of Squares Decomposition. *Journal of Classification*, **37**(2), 526–547.

INGRASSIA, S., MINOTTI, S.C., & VITTADINI, G. 2012. Local Statistical Modeling via the Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification*, **29**(3), 363–401.

INGRASSIA, S., PUNZO, A., VITTADINI, G., & MINOTTI, S. 2015. The Generalized Linear Mixed Cluster-Weighted Model. *Journal of Classification*, **32**(1), 85–113.

MAZZA, A., PUNZO, A., & INGRASSIA, S. 2018. flexCWM: a Flexible Framework for Cluster-Weighted Models. *Journal of Statistical Software*, **86**(2), 1–30.

SCRUCCA, L., & RAFTERY, A. 2018. clustvarsel: A Package Implementing Variable Selection for Gaussian Model-Based Clustering in R. *Journal of Statistical Software*, **84**.

MATRIX-VARIATE HIDDEN MARKOV REGRESSIONS

Salvatore D. Tomarchio [1], Antonio Punzo [1] and Antonello Maruotti [2]

[1] Department of Economics and Business, University of Catania, (e-mail: daniele.tomarchio@unict.it, antonio.punzo@unict.it)

[2] Department of Law, Economics, Political Sciences, and Modern Languages, LUMSA University, (e-mail: a.maruotti@lumsa.it)

ABSTRACT: We present two families of matrix-variate hidden Markov regression models, which differ in how they handle covariates (i.e., as fixed or random). The models achieve parsimony by using the eigen-decomposition of the components' covariance matrices. A two-step fitting strategy is implemented due to the high number of parsimonious models. These models are then investigated on a real dataset.

KEYWORDS: hidden Markov, matrix-variate, model-based clustering.

1 Introduction

Hidden Markov models (HMMs) are widely used for analyzing longitudinal data due to their mathematical flexibility. HMMs can also be modified to incorporate covariates, resulting in hidden Markov regression models (HMRMs), which are useful in regression settings (Bartolucci *et al*, 2012).

Broadly speaking, HMRMs can be divided into two main groups based on whether the covariates contribute to assigning observations to hidden states. The first group involves observed covariates that act as fixed effects shared by all units in the same hidden state, resulting in hidden Markov regression models with fixed covariates (HMRMFCs). Examples of this category can be found in studies by Bartolucci and Farcomeni (2015), and Maruotti and Punzo (2017). The second group, on the other hand, treats observed covariates as random and includes information about their distribution in the model to facilitate clustering. This approach leads to hidden Markov models with random covariates (HMRMRCs) as demonstrated in studies by Punzo *et al* (2018, 2021).

The focus of our study is to present and examine HMRMFCs and HMRMRCs as potential tools for analyzing matrix-variate longitudinal data. These models will be referred to as MV-HMRMFCs and MV-HMRMRCs, respectively. This type of data is typically obtained by observing $P \times R$ matrices of variables for I units over T periods. In essence, the data can be organized into a four-dimensional array with dimensions of $P \times R \times I \times T$.

To achieve parsimony, the two covariance matrices of each hidden state are subjected to eigen-decomposition. Because of the different formulations, the overall number of models is different between the two families. In the case of MV-HMRMFCs, only the covariance matrices of the response variables are available for each state, producing 98 MV-MRMFCs. On the other hand, for MV-HMRMRCs, both the response and covariate covariance matrices are available in each state, leading to 9604 MV-HMRMRCs. Therefore, a convenient approach for fitting the MV-HMRMRCs is employed to reduce the required computational effort.

We examine a dataset obtained from the Italian National Institute of Statistics to explore the relationship between unemployment and labor force participation in the Italian labor market. The data is structured in a two-factor design based on gender and age groups, and it covers four years at the provincial level.

2 Methodology

Let $\{\mathcal{Y}_{it}; i = 1,\ldots,I, t = 1,\ldots,T\}$ be a sequence of response variables, where each $\mathcal{Y}_{it}$ is a matrix of dimension $P \times R$ referring to the ith observation for the tth time point. The main assumption of an MV-HMM is that the random matrices in the above sequence are conditionally independent given a hidden process $\{S_{it}; i = 1,\ldots,I, t = 1,\ldots,T\}$ that follows a first-order Markov chain with state-space $\{1,\ldots,k,\ldots,K\}$. This process is governed by the initial probabilities $\pi_{ik} = \Pr(S_{i1} = k)$, $k = 1,\ldots,K$, and the transition probabilities $\pi_{ik|j} = \Pr(S_{it} = k|S_{it-1} = j)$, $t = 2,\ldots,T$ and $j,k = 1,\ldots,K$, where j refers to the state previously visited. We assume a matrix-variate normal distribution for the observations at every time occasion, that is, $f(\mathcal{Y}_{it} = \mathbf{Y}_{it}|S_{it} = s_{it}) \sim MVN_{P\times R}(\mathbf{M}_k, \mathbf{\Sigma}_k, \mathbf{\Psi}_k)$, where $\mathbf{M}_k$ is the $P \times R$ mean matrix, and $\mathbf{\Sigma}_k$ and $\mathbf{\Psi}_k$ are the $P \times P$ and $R \times R$ covariance matrices related to the P rows and R columns, respectively, for latent state k.

In numerous longitudinal studies, apart from the series of responses, there exists a series of covariates $\{\mathcal{X}_{it}; i = 1,\ldots,I, t = 1,\ldots,T\}$, being each $\mathcal{X}_{it}$ a matrix of dimension $Q \times R$, that we would like to functionally relate to the former. Thus, we have to extend MV-HMMs to the two regression-based categories introduced in Section 1. By starting with the fixed covariates approach (MV-HMRMFCs), in each latent state k, we are interested in modeling the conditional distribution

$$f(\mathcal{Y}_{it} = \mathbf{Y}_{it}|\mathcal{X}_{it} = \mathbf{X}_{it}, S_{it} = k), \tag{1}$$

by assuming a linear functional form for its expectation

$$\mathbb{E}(\mathcal{Y}_{it} = \mathbf{Y}_{it} | \mathcal{X}_{it} = \mathbf{X}_{it}, S_{it} = k; \mathbf{B}_k) = \mathbf{B}_k \mathbf{X}_{it}^*, \tag{2}$$

where $\mathbf{B}_k$ is a $P \times (1+Q)$ matrix of regression coefficients and $\mathbf{X}_{it}^*$ is a $(1+Q) \times R$ matrix having a vector of ones in the first row (to incorporate the intercept in the model) and the Q covariates from the second row onwards.

When the random covariates approach (MV-HMRMRCs) is considered, in each latent state k, we model the joint distribution

$$\begin{aligned} &f(\mathcal{Y}_{it} = \mathbf{Y}_{it}, \mathcal{X}_{it} = \mathbf{X}_{it} | S_{it} = k) = \\ &f(\mathcal{Y}_{it} = \mathbf{Y}_{it} | \mathcal{X}_{it} = \mathbf{X}_{it}, S_{it} = k) f(\mathcal{X}_{it} = \mathbf{X}_{it} | S_{it} = k), \end{aligned} \tag{3}$$

by also assuming (2).

To introduce parsimony in (1) and (3), we apply the eigen-decomposition to the covariance matrices, as commonly done in the model-based clustering literature (see, e.g. Tomarchio *et al*, 2022). This creates two families of models: 98 parsimonious MV-HMRMFCs and 9604 parsimonious MV-HMRMRCs.

Parameter estimation is implemented via a maximum likelihood approach based on the expectation conditional-maximization (ECM) algorithm (Meng and Rubin, 1993) and recursions widely used in the HMM literature (Baum *et al*, 1970). To make computationally affordable the fitting of 9604 parsimonious MV-HMRMRCs, a two-step fitting strategy (not discussed here for the sake of space) is implemented.

From a classification perspective, by using a maximum *a posteriori* probabilities approach (Punzo *et al*, 2021), each unit is classified to one of the K hidden states, at each time point. This information can be useful to track how the observations move between the hidden states as well as to identify which state is mainly sojourned by each observation.

3 Real data example

We examine the relationship between unemployment and the Labor Force Participation (LFP) of 106 Italian provinces, utilizing data from the Italian National Institute of Statistics (ISTAT). Our analysis focuses on the four years from 2018 to 2021. The unemployment and LFP for each province are recorded in a two-factor percentage format, categorized by gender (male and female) and age (15-24, 25-34, 35-49, 50-74). Therefore, both variables are presented in a four-way array format, with dimensions of $2 \times 4 \times 106 \times 4$.

By limiting here our discussion to the results obtained after the fitting of parsimonious MV-HMRMRCs, we found that the best solution according to the Bayesian information criterion (BIC) has $K = 5$ hidden states. The estimated regression coefficients (omitted here due to space constraints) indicate a negative sign in 80% of the cases. This suggests that the so-called discouraged worker effect is widespread across the provinces of Italy. The estimated mean matrices (omitted here due to space constraints) illustrate that the states can be sorted according to the levels of unemployment, both in gender and age factors. Specifically, the unemployment levels consistently decrease from the first state to the fifth state. Looking at the classification obtained by assigning each state to the province it mainly sojourns, it appears that there is a geographical pattern. The first two states seem to be predominantly composed of provinces located in the southern part of Italy, while the other three states appear to contain provinces located in the central and northern parts of the country.

References

BARTOLUCCI, F., & FARCOMENI, A. 2015. A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. *Biometrics*, **71**, 80–89.

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2012. *Latent Markov models for longitudinal data*. Boca Raton: CRC Press.

BAUM, L. E., PETRIE, T., SOULES, G., & WEISS, N. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41**, 164–171.

MARUOTTI, A., & PUNZO, A. 2017. Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. *Computational Statistics & Data Analysis*, **113**, 475–496.

MENG, X., & RUBIN, D. B. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267–278.

PUNZO, A., INGRASSIA, S., & MARUOTTI, A. 2018. Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population. *Statistics in Medicine*, **37**, 2797–2808.

PUNZO, A., INGRASSIA, S., & MARUOTTI, A. 2021. Multivariate hidden Markov regression models: random covariates and heavy-tailed distributions. *Statistical Papers*, **62**, 1519–1555.

TOMARCHIO, S. D., PUNZO, A., & MARUOTTI, A. 2022. Parsimonious hidden Markov models for matrix-variate longitudinal data. *Statistics and Computing*, **32**, 1–18.

INEQUALITIES AT ENTRANCE, LABOUR MARKET CONDITIONS AND UNIVERSITY DROPOUT: FIRST EVIDENCE FROM ITALY

Cristian Usala [1] , Isabella Sulis [1] and Mariano Porcu [1]

[1] Department of Political and Social Sciences, University of Cagliari, (e-mail: cristian.usala@unica.it, isulis@unica.it, mrporcu@unica.it)

ABSTRACT: This research investigates university students' dropout by focusing on the role of students' educational backgrounds and labour market conditions of both origin and destination areas. We exploit the administrative data from the Italian National Student Archive related to students enrolled in an Italian university between 2011 and 2018 by applying a two-step approach to separate students' career disruption effects related to their high school background, attended university, and degree program from those associated with labour market and socioeconomic conditions (e.g., unemployment, income, number of firms). In the first step, the effect of secondary schools, degree programs, and universities on students' performance in terms of earned formative credits is estimated using fixed regression models. In the second step, the estimated fixed effects are used in a multinomial logit model to account for the role played by high schools and universities in assessing the effect on students' career disruption of the labour market conditions of both origin and destination areas.

KEYWORDS: university dropout, multilevel models, university services, labor market conditions, school effect

1 Introduction and aims

According to the Human Capital Theory the investment in education is strictly affected by its expected returns (Becker, 1964). However, it is not straightforward how expectations related to the job market conditions affect the decision to invest in higher education (Di Pietro, 2006; Contini & Zotti, 2022). Indeed, these elements affect education decisions in two ways: fewer job opportunities may encourage students to have better credentials to spend in the labour market, while family income losses may push students to enter the labour market and abandon the university (Duncan, 1965; Rees & Mocan, 1997). Indeed, according to the main literature, better labour market conditions may increase the

cost of investment in education, determining higher dropout rates for students who experience difficulties in their university careers, come from disadvantaged backgrounds, or are enrolled at the university to avoid unemployment. However, better labour market conditions can provide more resources to support families' education investment and increase its expected returns. Indeed, several studies related to the Italian framework highlight a negative relationship between worse job market conditions and students' dropout (Di Pietro, 2006; **?**; Contini & Zotti, 2022; Contini & Cugnata, 2018; Meggiolaro *et al.*, 2017; Tedesco, 2022; Perchinunno *et al.*, 2021). However, Contini & Zotti, 2022 show that a general trend does not emerge if differences across disciplinary areas are considered.

This paper investigates the effects of labour market conditions, students' high school past experiences and universities' environment on students' dropout risk between the first and the second year of their university career. We combine administrative data on students' university careers in Italy with several complementary data sources regarding the socioeconomic conditions of both students' areas of residence and universities' hosting areas. This data is exploited using a two-step approach to disentangle the effect of students' educational background and individual characteristics from the one related to local labour market conditions. The paper is organised as follows: Section 2 describe the database used and the methods applied in the research; Section 3 discusses the preliminary results and concludes.

2 Data and methods

This research is based on the data from the MOBYSU.IT database regarding the careers of all students enrolled in an Italian university between 2011 and 2018.* MOBYSU.IT includes information on several students' characteristics such as their sex, high school background, municipality of residence, age, and the chosen university and degree programs. We have collected data on 1,668,882 first-year students.

MOBYSU.IT also contains the data on the formative credits (CFU) earned by each student within her/his first year of career.† This information helps

*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

†The CFU are similar to the European Credit Transfer and Accumulation System (ECTS),

measure the regularity of students' careers in terms of exams passed during their first year of career and to define students' risk of dropout. At this aim, we classify students into dropouts, at risk of dropout, and regulars. Dropouts are students who are not enrolled in any university in their second year of career and, therefore, that have abandoned the Italian university system. Students at risk of dropout are those that have obtained less than 25 CFU within their first year of career, while regulars are the residual category. The threshold of 25 CFU is chosen based on the observed values of CFU obtained among regulars and dropouts. Indeed, 95% of dropouts have obtained less than 25 CFU during the first year of their career.

The MOBYSU.IT database has been combined with the data obtained from ISTAT to assess the role of labour market conditions on students' career disruption. At this aim, we collected data on the provincial unemployment rate, total taxable income per capita at the municipal level, and the number of local firms or branches of firms in the municipality. These indicators have been obtained for students' areas of residence and universities' hosting areas.

This data is exploited by applying a two-step procedure to disentangle the effect related to students' educational backgrounds from the one related to origin and destination areas' labour market conditions. The first step consists of two fixed effects regression models that estimate the CFU obtained by students as a function of high school fixed effects and universities $\times$ degree programs fixed effects. The second step uses the fixed effects estimates to account for the average role of high schools and universitie' degree programs on students' dropout risk. More specifically, a multinomial logit model is estimated to assess the role of students' characteristics and labour market conditions on the probability of dropout, being at risk or being a regular student.

3 Preliminary Results

Preliminary results provide evidence that the inequalities at entrance related to the high school background and the attended university have a relevant role in affecting students' probability to experience an event of career disruption. Indeed, the average effect of high schools on the CFU obtained by students is one of the main predictors of the probability to dropout even when accounting for students' and areas' characteristics. Furthermore, high schools plays a relevant role also with respect to students' risk of dropout. However, in this case, the most important predictor is given by the set of university $\times$ degree

and they represent a measure of the workload associated with each exam.

programs fixed effects that measure the average performances of students in the institutions in terms of CFU. The results also show that labour market conditions have different effects depending on whether we consider dropouts or students' at risk of dropout and that these effects change when looking at origin or destination areas. Further research will study the effect of labour market conditions and how it changes depending on students' socioeconomic conditions. Moreover, the use of ad-hoc survey could provide valuable insight on the optimal policies to contrast university dropout.

References

BECKER, GARY S. 1964. *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago press.

CONTINI, D., & CUGNATA, F.AND SCAGNI, A. 2018. Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy. *Higher Education*, **75**, 785–808.

CONTINI, D., & ZOTTI, R. 2022. Do Financial Conditions Play a Role in University Dropout? New Evidence from Administrative Data. *In Checchi, D., Jappelli, T., and Uricchio, A. (eds): Teaching, Research and Academic Careers: An Analysis of the Interrelations and Impacts*, 39–70.

DI PIETRO, G. 2006. Regional labour market conditions and university dropout rates: Evidence from Italy. *Regional Studies*, **40**, 617–630.

DI PIETRO, GIORGIO, & CUTILLO, ANDREA. 2008. Degree flexibility and university drop-out: The Italian experience. *Economics of Education Review*, **27**(5), 546–555.

DUNCAN, B. 1965. Dropouts and the unemployed. *Journal of Political Economy*, **73**, 121–134.

MEGGIOLARO, SILVIA, GIRALDO, ANNA, & CLERICI, RENATA. 2017. A multilevel competing risks model for analysis of university students' careers in Italy. *Studies in Higher Education*, **42**(7), 1259–1274.

PERCHINUNNO, PAOLA, BILANCIA, MASSIMO, & VITALE, DOMENICO. 2021. A statistical analysis of factors affecting higher education dropouts. *Social Indicators Research*, **156**, 341–362.

REES, DANIEL I, & MOCAN, H NACI. 1997. Labor market conditions and the high school dropout rate: Evidence from New York State. *Economics of Education Review*, **16**(2), 103–109.

TEDESCO, N., SALARIS L. 2022. University drop out and mobility in Italy. First evidences on first level degrees. *In Pollice, A., Salvati, N., Schirripa Spagnolo, F. (eds): SIS 2020 - Book of Short Papers*, 1601–1606.

A CLUSTERING METHOD FOR DISTRIBUTIONAL DATA BASED ON A LDQ TRANSFORMATION

Rosanna Verde [1], Gianmarco Borrata [2] and Antonio Balzanella [1]

ABSTRACT: This work deals with a clustering method for distributional data. The set of objects to be clustered are described by p distributional variables. Each object is represented by p density probability functions (*dpf*'s), or empirical ones. In consideration of the most recent developments in distributional data analysis (DDA), we introduce a transformation of the quantile functions, *qf*'s, associated to the *dpf*'s, in Logarithm Derivative Quantiles (LDQ) functions, which allows to map density probability functions in an Hilbert space. Our proposal is based on a Dynamic Clustering Clustering type-algorithm, where the centroid of the clusters are represented by linear combination of LDQ functions; the objects are assigned to the clusters according to minimum sum of the squared distance from the centroid function. Applications on synthetic and real data have corroborated the new method.

KEYWORDS: symbolic data analysis, distributional data, quantile density functions

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

[2] Department of Social Science, University of Naples Federico II, (e-mail: gianmarco.borrata@studenti.unicampania.it)

Shrinkage of time-varying effects in panel data models

Helga Wagner[1], Roman Pfeiler[1]

[1] Department of Applied Statistics, Johannes Kepler University Linz, Austria, (e-mail: helga.wagner@jku.at, Roman.Pfeiler@jku.at)

Abstract: We consider regression models for panel data with time-varying effects in a Bayesian framework. We implement shrinkage of regression effects and the process variances of the effects to distinguish between effects that are practically zero, constant or time-varying via shrinkage priors. Longitudinal dependence is taken into account by including a subject specific random factor with weights that may also vary over time. The model is applied to analyse panel data on annual incomes of mothers returning to the job market after maternity leave.

Keywords: dynamic effects; factor model; shrinkage prior

1 Introduction

Panel data where subjects are observed at several time points provide richer information than cross sectional data but pose additional challenges as correlation of observations within subjects has to be taken into account. The multiple measurements per subject allow to model their heterogeneity and the longitudinal structure provides information on development over time. A standard way to take into account heterogeneity in panel data regression analysis is by including subject specific random effects in the linear predictor and development over time can be modelled by allowing for time-varying regression effects. However, modelling all regression effects as time-varying will result in an overspecified model if actually one or more effects are time-constant or even 0. In a Bayesian approach, based on an adaquate model formulation, appropriate prior distributions allow to identify constant or zero effects in time series regression models (Frühwirth-Schnatter & Wagner, 2010). In this paper we will use the shrinkage priors recently proposed in Bitto & Frühwirth-Schnatter, 2019 for time series and investigate their performance for panel data where the number of subjects is larger than 1

but time series are short, e.g. in our application we have individual time series of length 8.

2 Model specification and inference

2.1 Regression model with time-varying effects

To keep notation simple, we assume balanced panel data where $i = 1, ..., n$ subjects are observed at time points $t = 1, ..., T$. Let y_{it} denote the response of subject i at time t and $\mathbf{x}_{it}$ is the $p \times 1$ vector of covariates. We consider the following regression specification

$$y_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta}_t + \epsilon_{it}, \quad \epsilon_{it} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}) \tag{1}$$

where $\boldsymbol{\beta}_t$ is the $p \times 1$ vector of regression effects at time t and $\boldsymbol{\Omega}$ is a $T \times T$ covariance matrix.

To model time-varying parameters we assume that the regression effects follow a random walk

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$$

with independent increments, $\mathbf{Q} = \text{diag}(\theta_1^2, \ldots, \theta_p^2)$, and starting values

$$\boldsymbol{\beta}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_0).$$

The process variance θ_j^2, $j = 1, \ldots, p$ carries information on the evolvement of the regression effect β_{jt} over time.

To allow shrinkage to time-constant or zero effects we use shrinkage priors on the effects and process standard deviations in the non-centered parameterization (Frühwirth-Schnatter & Wagner, 2010), which is given as

$$\boldsymbol{\beta}_t = \boldsymbol{\beta} + \boldsymbol{\theta}\tilde{\boldsymbol{\beta}}_t.$$

Here $\boldsymbol{\theta} = \text{diag}(\theta_1, \ldots, \theta_p)$ is the vector of process standard deviations and $\tilde{\boldsymbol{\beta}}_t$ is defined as

$$\tilde{\boldsymbol{\beta}}_t = \tilde{\boldsymbol{\beta}}_{t-1} + \tilde{\boldsymbol{\omega}}_t, \quad \tilde{\boldsymbol{\omega}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Hence, the regression model (1) in its non-centered parameterization is given as

$$y_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{x}_{it}^T \boldsymbol{\theta} \tilde{\boldsymbol{\beta}}_t + \epsilon_{it}.$$

Shrinkage of elements of $\boldsymbol{\beta}$ as well as $\boldsymbol{\theta}$ is induced by appropriate prior distributions.

2.2 Modelling longitudinal association

To allow for longitudinal association within subjects we specify the error term ϵ_{it} in terms of a subject specific latent factor f_i and the idiosyncratic error ε_{it} as

$$\epsilon_{it} = \lambda_t f_i + \varepsilon_{it}, \quad \varepsilon_{it} \sim \mathcal{N}(0, \sigma_t^2)$$

and hence

$$\boldsymbol{\Omega} = \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Sigma}$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_T^2)$. This model encompasses as special case compound symmetry structure of $\boldsymbol{\Omega}$ when $\lambda_t = \lambda$. To model time-varying factor loadings we again model the evolvement of the factor loadings by a random walk

$$\lambda_t = \lambda_{t-1} + \nu_t, \quad \nu_t \sim \mathcal{N}(0, \psi^2).$$

parameterization.

2.3 Prior Distributions

To encourage shrinkage of constant effects β_j and their process variances θ_j^2, $j = 1, \ldots, p$, following Bitto & Frühwirth-Schnatter, 2019 we specify the priors on β_j as independent Normal-Gamma and on the process variances θ_j as independent double Gamma-priors. The same specification is used for the priors on the factor loading parameters in the noncentered parameterisation.

For the error variances σ_t^2 of the idiosyncratic errors we use independent uninformative Inverse Gamma priors.

2.4 Inference

Inference is performed by MCMC methods extending the Gibbs sampling proposed in Bitto & Frühwirth-Schnatter, 2019 by the additional steps to sample the subject specific factors and the factor loadings in the non-centered parameterization.

3 Application

We apply the developed methods to analyse earnings of mothers in Austria after their return to the labor market after their last maternity

leave. The data set comprises earnings for $n = 8877$ mothers after return to labour market observed for $T = 8$ panel periods.

Covariates in the regression model are categorical predictors of the number of children (baseline: 1 child, dummy for 2 children, dummy for 3 or more children), binary variables for type of contract (baseline: white collar), leave duration and working experience (for both the baseline is *below the median*) as well as the log-earnings before the maternity leave. All regression parameters, except the effect of 3 or more children and also the factor loadings vary over time. Figure 1 compares the estimated time-varying intercept and the effects of 3 or more children under the shrinkage priors to the estimated effects in a random intercept model with unstructured time-varying effects. The shrinkage prior results in smoother effects which can also be effectively reduced to zero, see the lower panel of Figure 1.

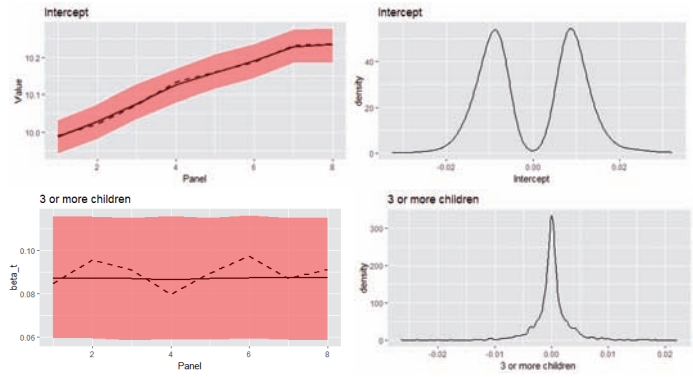


Figure 1. *Results for intercept and effect of 3 or more children. Left: Posterior mean estimates and 95%-HPD intervals of the regression effects. Dotted lines are the estimated time-varying regression effects from a random intercept model without smoothing. Right: Posterior of the process standard deviations.*

References

Bitto, A., & Frühwirth-Schnatter, S. 2019. Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics.*, **210**, 75–97.

Frühwirth-Schnatter, S., & Wagner, H. 2010. Stochastic model specification search for Gaussian and partially Non-Gaussian state space models. *Journal of Econometrics.*, **154**, 85–100.

A BAYESIAN SPATIO-TEMPORAL REGRESSION APPROACH FOR CONFOUNDING ADJUSTMENT

Carlo Zaccardi [1], Pasquale Valentini[1] and Luigi Ippoliti[1]

[1] University G. d'Annunzio, Chieti-Pescara, Department of Economics, Viale Pindaro 42, 65127 Pescara, Italy (e-mail: carlo.zaccardi@unich.it, pvalent@unich.it, luigi.ippoliti@unich.it)

ABSTRACT: For an accurate evaluation of the harmful impacts of pollution on human health, confounding variables must always be taken into account. Unfortunately, it oftentimes happens that some confounders might result unmeasured, hence, within a regression framework, the parameter that represents the exposure's effect might no longer be recoverable. In this paper, the unmeasured confounder is represented by a linear combination of basis functions, a technique that has been used in the spatial confounding literature, and that we expand to spatio-temporal designs. To reduce dimensionality and confounding bias, spike-and-slab priors are assumed on basis coefficients.

KEYWORDS: confounding, spatio-temporal, pollution, health, Bayesian.

1 Introduction

The principal objective in environmental epidemiology is to evaluate whether exposure to a pollutant has adverse health consequences. To this end, the relationship between exposure and outcome variables can be expressed in regression terms. An accurate evaluation of the relationship of interest requires that all variables correlated with both exposure and outcome (known as *confounders*), such as meteorological variables, should be included in the model as additional regressors (Dominici & Peng, 2008). However, data about some confounders could result not available because of, for example, budget constraints. If the model fails to account for confounding, it would be impossible to recover the parameter of interest. The estimator for the exposure's effect would then become biased, and its bias is known as *confounding bias* in the epidemiological literature (Dominici & Peng, 2008).

While smooth functions of calendar time are usually included in models for time-series data (e.g., see Dominici & Peng, 2008), in purely spatial settings, the simplest and more appealing remedy to the *spatial confounding* problem

is to add into the model a spatial random effect. However, Reich *et al.*, 2006 show that doing so distorts inference on the effect of interest and leads the practitioner to draw incorrect conclusions. Different other solutions are reviewed by Reich *et al.*, 2021 and Urdangarin *et al.*, 2022. To our knowledge, relatively few authors consider confounding adjustment in spatio-temporal designs. Reich *et al.*, 2021 reviews spatio-temporal methods as well to account for unmeasured confounding under causal inference hypotheses. More recently, two approaches in the spatial confounding literature are extended to account for temporal dependence as well (Adin *et al.*, 2023; Prates *et al.*, 2022). In the next Section, we discuss a different approach wherein, extending the work by Valentini *et al.*, 2022, unmeasured confounding is accounted for by including spatio-temporal basis functions into the regression model. We also impose a prior structure on the basis coefficients that encourages sparsity.

2 The Proposed Model

Consider a spatio-temporal process $\{Y(\mathbf{s},t) : \mathbf{s} \in \mathcal{D},\ t = 1,2,\ldots,T\}$, defined for every location, $\mathbf{s}$, over a continuous spatial domain $\mathcal{D} \subseteq \mathbb{R}^2$, and for discrete time periods $t = 1,2,\ldots,T$. Assume that it represents a health outcome observed at a finite set of locations, $\{\mathbf{s}_1,\ldots,\mathbf{s}_N\}$, for the entire study period. Moreover, suppose that $X(\mathbf{s},t)$ and $Z(\mathbf{s},t)$ are two correlated Gaussian spatio-temporal processes representing the exposure (observed at the same spatial locations and time instants as the outcome) and the unmeasured confounder, respectively. Assuming that the distribution F is a member of the exponential family, and that realizations are conditionally independent, it is possible to specify the following hierarchy, for $i = 1,\ldots,N$ and $t = 1,\ldots,T$:

$$Y(\mathbf{s}_i,t) \overset{ind}{\sim} F(\mu(\mathbf{s}_i,t),\phi) \tag{1}$$

$$g(\mu(\mathbf{s}_i,t)) = \beta_0 + \beta_x X(\mathbf{s}_i,t) + Z(\mathbf{s}_i,t) + \varepsilon(\mathbf{s}_i,t), \tag{2}$$

where $\mu(\mathbf{s}_i,t) = E[Y(\mathbf{s}_i,t)]$, ϕ is a scale parameter, $g(\cdot)$ is an opportune link function, and $\varepsilon(\mathbf{s}_i,t)$ represents a zero-mean stationary Gaussian process with realizations mutually independent in time but correlated in space such that the spatial covariance structure is defined by a parametric function with parameter vector $\boldsymbol{\theta}$, that is $Cov(\varepsilon(\mathbf{s}_i,t),\varepsilon(\mathbf{s}_j,t)) = C(|\mathbf{s}_i - \mathbf{s}_j|;\boldsymbol{\theta})$ for $i,j = 1,\ldots,N$. The primary aim of the analysis is to correctly recover the regression coefficient of the exposure, β_x, while controlling for confounding at the same time.

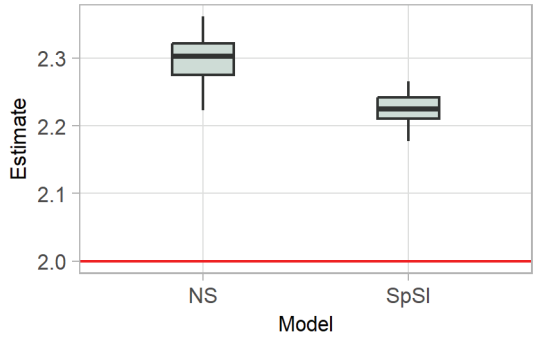


Figure 1. *Boxplots representing the estimated exposure effect in the simulation study. The red line represents the real value,* $\beta_x = 2$.

Thanks to the Karhunen-Loéve theorem (KLT, Banerjee *et al.*, 2014), the process $Z(\mathbf{s},t)$ can be represented as an infinite linear combination of pairwise orthogonal basis functions, but, operationally, a reduced-rank representation is given to it:

$$Z(\mathbf{s},t) \approx \sum_{m=1}^{M} \alpha_m \psi_m(\mathbf{s},t)\,, \tag{3}$$

where $\psi_m(\cdot,\cdot)$ are spatio-temporal basis functions, and α_m are expansion coefficients, for $m = 1,\ldots,M$. These bases are then introduced in Equation (2) in place of the unmeasured confounder. A necessary condition is that they must be correlated to both $X(\mathbf{s}_i,t)$ and $Z(\mathbf{s}_i,t)$, so the aforementioned drawbacks discussed by Reich *et al.*, 2006 could be overcome.

To select the most promising bases and hence obtain a parsimonious model, we assume spike-and-slab priors (Ishwaran & Rao, 2005) on the expansion coefficients. The Bayesian hierarchical specification is completed by assigning prior distributions to all the other parameters, and a Markov chain Monte Carlo (MCMC) algorithm is constructed for inferential purposes. To show whether our model is able to mitigate confounding issues, we set up a simulation study wherein $X(\mathbf{s},t)$ and $Z(\mathbf{s},t)$ are drawn from their joint distribution, under the assumptions that $Cor(X(\mathbf{s},t),Z(\mathbf{s},t)) = 0.5$, and that the second process varies at spatial and temporal scales coarser than those of the first process. The outcome is then generated using Eqs. 1–2, where F is the Gaussian distribution. We then fit a non-spatial (NS) model that does not account for confounding, and our proposal (denoted as SpSl). Figure 1 synthesize the main results: for each model, it depicts a boxplot of the posterior means for β_x obtained from fit-

ting 100 replicates. The red line represents its true value, $\beta_x = 2$. The proposed model can potentially reduce the confounding bias so it should be preferred to the non-spatial one.

Finally, a more extensive simulation study and real-data applications will be discussed in an extended version of this paper, wherein several types of basis functions will be examined as well.

References

ADIN, ARITZ, GOICOA, TOMÁS, HODGES, JAMES S, SCHNELL, PATRICK M, & UGARTE, MARÍA D. 2023. Alleviating confounding in spatio-temporal areal models with an application on crimes against women in India. *Statistical Modelling*, **23**(1), 9–30.

BANERJEE, SUDIPTO, CARLIN, BRADLEY P, & GELFAND, ALAN E. 2014. *Hierarchical modeling and analysis for spatial data*. CRC press.

DOMINICI, FRANCESCA, & PENG, ROGER D. 2008. *Statistical methods for environmental epidemiology with R: a case study in air pollution and health*. Springer.

ISHWARAN, HEMANT, & RAO, J. SUNIL. 2005. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, **33**(2), 730–773.

PRATES, MARCOS O, AZEVEDO, DOUGLAS RM, MACNAB, YING C, & WILLIG, MICHAEL R. 2022. Non-separable spatio-temporal models via transformed multivariate Gaussian Markov random fields. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **71**(5), 1116–1136.

REICH, BRIAN J, HODGES, JAMES S, & ZADNIK, VESNA. 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, **62**(4), 1197–1206.

REICH, BRIAN J, YANG, SHU, GUAN, YAWEN, GIFFIN, ANDREW B, MILLER, MATTHEW J, & RAPPOLD, ANA. 2021. A review of spatial causal inference methods for environmental and epidemiological applications. *International Statistical Review*, **89**(3), 605–634.

URDANGARIN, ARANTXA, GOICOA, TOMÁS, & UGARTE, MARÍA DOLORES. 2022. Evaluating recent methods to overcome spatial confounding. *Revista Matemática Complutense*, 1–28.

VALENTINI, PASQUALE, SCHMIDT, ALEXANDRA M., ZACCARDI, CARLO, & IPPOLITI, LUIGI. 2022. Adjusting for Unmeasured Spatial Confounding Through Shrinkage Methods. *In: Book of Short Papers of the 51st Scientific Meeting of the Italian Statistical Society*. Pearson.

LINEAR RANDOM FOREST TO PREDICT ENERGY CONSUMPTION

Gianpaolo Zammarchi [1]

[1] Department of Economics and Business Science, University of Cagliari, Via Sant'Ignazio da Laconi, 17, 09123, Cagliari (Italy). (e-mail: gp.zammarchi@unica.it)

ABSTRACT: Forecasting electricity consumption is a relevant task to ensure that the supply of energy fed into the grid always equals the demand. In this study we compare the performance of random forest and linear random forest in the prediction of daily electricity consumption in Italy. We show that both implementations reach a good performance in this task, with the best results obtained by linear random forest in a model including different features such as lags, difference variables and day - month variables.

KEYWORDS: linear random forest, time series, energy consumption

1 Introduction

Due to the rapid increase in world population and the global economic growth, the energy consumption is expected to increase in most countries. In particular, electricity is one of the main energy sources for homes, offices, factories and many other public and private places. A relevant problem is to ensure that the supply of energy fed into the grid always equals the demand or, in other words, to guarantee the equilibrium between the production of electricity and the consumption. For this reason, different companies and researchers have developed methods to forecast electricity daily consumption (Zhang *et al.*, 2021). In this study we assessed the performance of two different implementations of random forest in the prediction of energy consumption in Italy and compared their results with the effective consumption and with the prediction of Terna (the company that manages the Italian national transmission system).

The rest of the paper is organized as follows: first, we give an overview of the problem, the data collection and the methodology in Section 2. Then, we present the results in Section 3, and a brief summary and the future developments in Section 4.

2 Methods

In this section we will describe how data were collected, the features engineering process, and how these features were used to build the model used for predictions.

2.1 Data collection

The data related to the forecasts made by Terna's model, together with the actual consumption detected by the company (in Megawatt, MW), are published daily in the form of PDF files (Terna S.p.A., 2023). The files were downloaded and read in R (R Core Team, 2023). The data set included day-by-day hourly consumption values and forecasts for all days ranging from August 1, 2022 to March 31, 2023. Subsequently, these values were aggregated as follows: we computed $\upsilon_i = \{v_1, v_2, ..., v_j\}$ with $j = 1, ..., 24$, and a vector $\mathbf{V} = \{\upsilon_1, \upsilon_2, ..., \upsilon_i\}$ with $i = 1, ..., 243$ in order to obtain a vector with daily values obtained as the sum of individual hourly values.

2.2 Random forest

Random forest is a popular machine learning technique based on the combined use of decision trees, bootstrap, and ensemble methods (Breiman, 2001). It incorporates the output of several decision trees to produce a single evaluation. In this study we used the classical random forest implementation as well as a recently developed linear random forest variation based on the implementation of a ridge regression in the leaves (Künzel *et al.*, 2022). In this variation the returned value is computed using a linear aggregation function: $\hat{\mu}(x_{new}) := x^t_{new}(X^t_S X_S + \lambda I)^{-1} X^t_S Y_S$, where X_{new} is a new observation, S is a leaf, Y is the response variable, $\mathbf{X}$ the design matrix for the training set, and λ is a regularization parameter. The optimal splitting point is defined with a greedy strategy and the stopping criteria is based on an R^2 improvement threshold (Künzel *et al.*, 2022).

2.3 Feature engineering

We created lagged and difference variables to be used as predictors for the random forest. We therefore defined k as the number of lags that can be created starting from the response variable, the daily consumption of electricity. Difference variables were also created as in the following equation: $y_i - y_{i-t}$ where y_i is the energy consumpion during the day i. During the model

evaluation phase, various configurations were tested, using a number of lags $k \in \{k_1, ..., k_m\}$ with $m = 30$ and t equals to 7 and to 14. In addition, two variables relative to the day of the week (Monday-Sunday) and the month (from August 2022 to March 2023) have been included.

2.4 Models evaluation

The daily consumption values up to the end of February were used as the training set to predict daily consumption in March (test set). The predictions have been evaluated using two widely used metrics: root-mean-square error (RMSE) and mean absolute percentage error (MAPE). Terna's prediction has also been included as a benchmark to further compare the magnitude of the errors. Errors are computed using a moving window scheme.

3 Results

In this section we will present the results obtained using the two different implementations of random forest and compare these results with the effective consumption and with Terna's prediction. Figure 1 shows a comparison of the RMSE for both implementations of random forest compared with Terna's prediction. While both implementations of random forest showed a good performance in the prediction of the daily consumption of energy, Terna's model showed a lower error (RMSE: 8,796; MAPE: 1.05%). Linear random forest and classical random forest obtained an RMSE ranging from 11,853 to 16,886, and from 12,355 to 16,548, respectively, based on the different models we tested. As shown in Table 1, the best results were obtained by linear random forest in the configuration including 15 lags and the two difference variables. This configuration proved to be the best also for the classical implementation of random forest.

4 Conclusions

To conclude, we showed that random forest can provide accurate predictions even when used with time series. The two implementations of random forest used to forecast the energy consumption provides similar results and this might be due to, among other things, the specific properties of the time series used for the evaluation. As a future development we plan to further investigate the role of lags, differentiation and size of the training set.

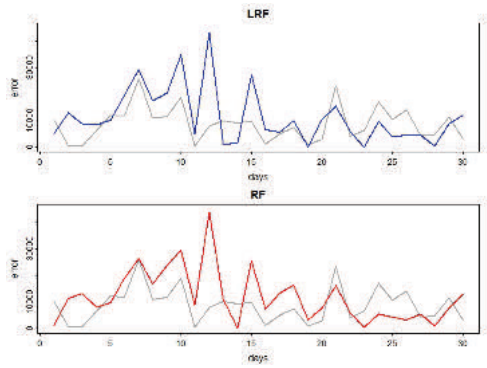


Figure 1. *Error of LRF (blue), RF (red) and Terna (grey) in the prediction of the electricity consumption. Abbreviations: LRF, linear random forest; RF, random forest*

Table 1. *Error of LRF and RF in the prediction of the electricity consumption*

Lags	Differences	RMSE LRF	RMSE RF	MAPE LRF	MAPE RF
5	-	14,740	14,400	1.78%	1.71%
15	-	15,241	15,643	1.80%	1.84%
30	-	16,886	16,548	1.99%	1.97%
5	2	15,585	13,076	1.88%	1.57%
15	2	**11,853**	**12,355**	**1.42%**	**1.48%**
30	2	12,763	13,236	1.54%	1.59%

In bold the best result (smallest error) for both models. Abbreviations: LRF, linear random forest; RF, random forest; RMSE, root-mean-square error; MAPE, mean absolute percentage error

References

BREIMAN, LEO. 2001. Random forests. *Machine learning*, **45**, 5–32.

KÜNZEL, SÖREN R, SAARINEN, THEO F, LIU, EDWARD W, & SEKHON, JASJEET S. 2022. Linear aggregation in tree-based estimators. *Journal of Computational and Graphical Statistics*, **31**(3), 917–934.

R CORE TEAM. 2023. R: A Language and Environment for Statistical Computing.

TERNA S.P.A. 2023. https://www.terna.it, Last access: April, 5 2023.

ZHANG, LIANG, WEN, JIN, LI, YANFEI, CHEN, JIANLI, YE, YUNYANG, FU, YANGYANG, & LIVINGOOD, WILLIAM. 2021. A review of machine learning in building load prediction. *Applied Energy*, **285**, 116452.