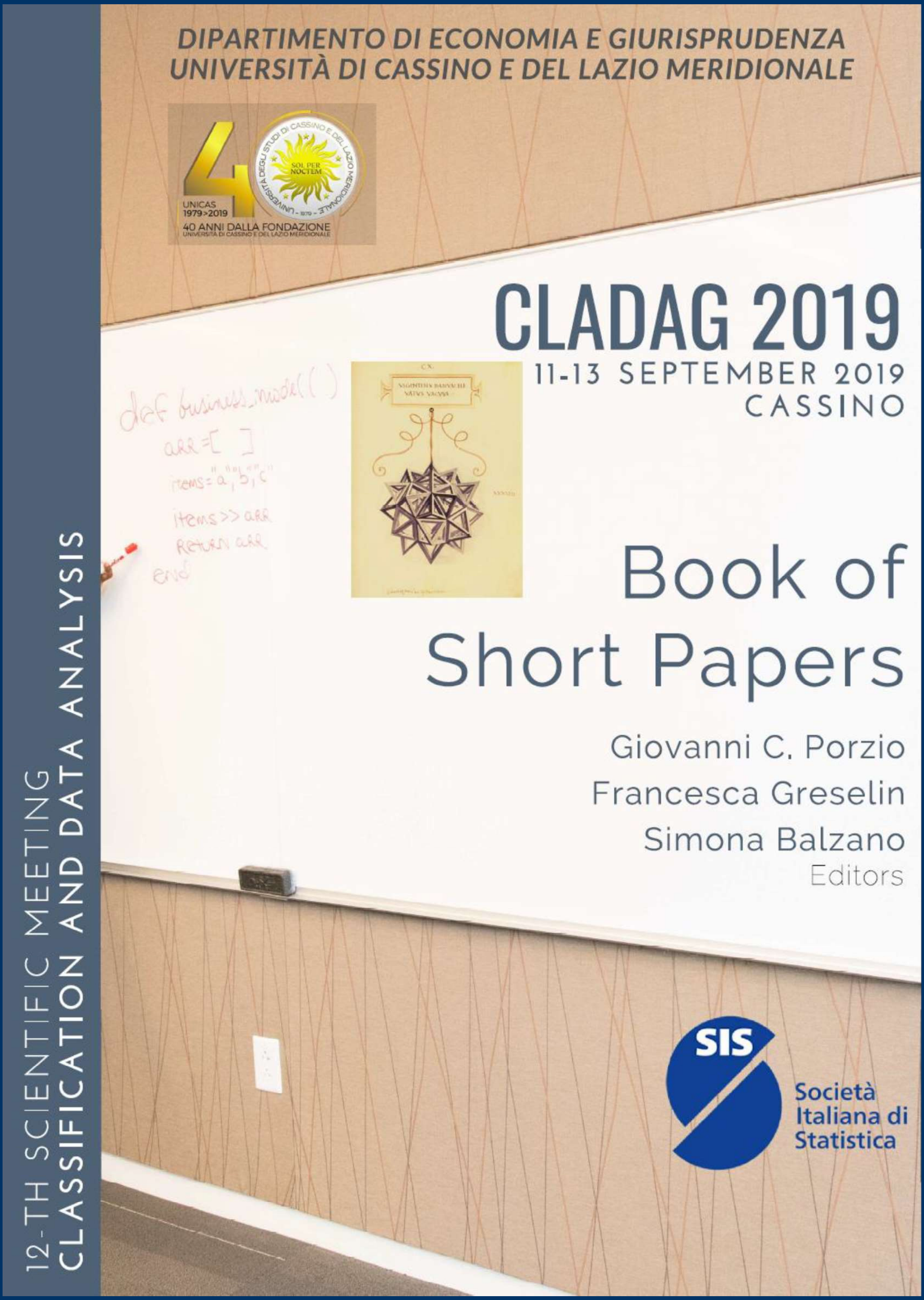
UNIVERSITÀ degli STUDI di CASSINO e del LAZIO MERIDIONALE

COLLANA SCIENTIFICA



DIPARTIMENTO DI ECONOMIA E GIURISPRUDENZA
UNIVERSITÀ DI CASSINO E DEL LAZIO MERIDIONALE

UNICAS 1979>2019
40 ANNI DALLA FONDAZIONE

CLADAG 2019

11-13 SEPTEMBER 2019
CASSINO

Book of Short Papers

Giovanni C. Porzio
Francesca Greselin
Simona Balzano
Editors

Società Italiana di Statistica

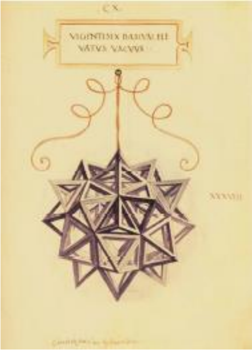
12-TH SCIENTIFIC MEETING
CLASSIFICATION AND DATA ANALYSIS

EDIZIONI UNIVERSITA' DI CASSINO

2019

Università di Cassino e del Lazio Meridionale
Centro Editoriale di Ateneo
Palazzo degli Studi Località Folcara, Cassino (FR), Italia

ISBN 978-88-8317-108-6



CLADAG 2019

Book of Short Papers

Giovanni C. Porzio

Francesca Greselin

Simona Balzano

Editors

2019

Contents

Keynotes lectures

Unifying data units and models in (co-)clustering 3
Christophe Biernacki

Statistics with a human face 4
Adrian Bowman

Bayesian model-based clustering with flexible and sparse priors 5
Bettina Grün

Grinding massive information into feasible statistics: current challenges and opportunities for data scientists 6
Francesco Mola

Statistical challenges in the analysis of complex responses in biomedicine 7
Sylvia Richardson

Invited and contributed sessions

Model-based clustering of time series data: a flexible approach using nonparametric state-switching quantile regression models 8
Timo Adam, Roland Langrock, Thomas Kneib

Some issues in generalized linear modeling 12
Alan Agresti

Assessing social interest in burnout using functional data analysis through google trends 16
Ana M. Aguilera, Francesca Fortuna, Manuel Escabias

Measuring equitable and sustainable well-being in Italian regions: a non-aggregative approach 20
Leonardo Salvatore Alaimo, Filomena Maggino

Bootstrap inference for missing data reconstruction 22
Giuseppina Albano, Michele La Rocca, Maria Lucia Parrella, Cira Perna

Archetypal contour shapes 26
Aleix Alcacer, Irene Epifanio, M. Victoria Ibáñez, Amelia Simó

Random projections of variables and units 30
Laura Anderlucci, Roberta Falcone, Angela Montanari

Sparse linear regression via random projections ensembles 34
Laura Anderlucci, Matteo Farnè, Giuliano Galimberti, Angela Montanari

High-dimensional model-based clustering via random projections 38
Laura Anderlucci, Francesca Fortunato, Angela Montanari

Multivariate outlier detection in high reliability standards fields using ICS 42
Aurore Archimbaud, Klaus Nordhausen, Anne Ruiz-Gazen

Evaluating the school effect: adjusting for pre-test or using gain scores? 45
Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto, Carla Rampichini

ACE, AVAS and robust data transformations 49
Anthony Atkinson

Mixtures of multivariate leptokurtic Normal distributions 53
Luca Bagnato, Antonio Punzo, Maria Grazia Zoia

Detecting and interpreting the consensus ranking based on the weighted Kemeny distance 57
Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio

Predictive principal components analysis 61
Simona Balzano, Maja Bozic, Laura Marcis, Renato Salvatore

Flexible model-based trees for count data 63
Federico Banchelli

Euclidean distance as a measure of conformity to Benford’s law in digital analysis for fraud detection 67
Mateusz Baryła, Józef Pociecha

The evolution of the purchase behavior of sparkling wines in the Italian market 71
Francesca Bassi, Fulvia Pennoni, Luca Rossetto

Modern likelihood-frequentist inference at work 75
Ruggero Bellio, Donald A. Pierce

Ontology-based classification of multilingual corpuses of documents 79
Sergey Belov, Salvatore Ingrassia, Zoran Kalinić, Paweł Lula

Modeling heterogeneity in clustered data using recursive partitioning 83
Moritz Berger, Gerhard Tutz

Mixtures of experts with flexible concomitant covariate effects: a bayesian solution 87
Marco Berrettini, Giuliano Galimberti, Thomas Brendan Murphy, Saverio Ranciati

Sampling properties of an ordinal measure of interrater absolute agreement 91
Giuseppe Bove, Pier Luigi Conti, Daniela Marella

Tensor analysis can give better insight 95
Rasmus Bro

A boxplot for spherical data 97
Davide Buttarazzi, Giuseppe Pandolfo, Giovanni C. Porzio, Christophe Ley

Machine learning models for forecasting stock trends 99
Giacomo Camba, Claudio Conversano

Tree modeling ordinal responses: CUBREMOT and its applications 103
Carmela Cappelli, Rosaria Simone, Francesca Di Iorio

Supervised learning in presence of outliers, label noise and unobserved classes 104
Andrea Cappozzo, Francesca Greselin, Thomas Brendan Murphy

Asymptotics for bandwidth selection in nonparametric clustering 108
Alessandro Casa, José E. Chacón, Giovanna Menardi

Foreign immigration and pull factors in Italy: a spatial approach 112
Oliviero Casacchia, Luisa Natale, Francesco Giovanni Truglia

Dimensionality reduction via hierarchical factorial structure 116
Carlo Cavicchia, Maurizio Vichi, Giorgia Zaccaria

Likelihood-type methods for comparing clustering solutions 120
Luca Coraggio, Pietro Coretto

Labour market analysis through transformations and robust multilevel models 124
Aldo Corbellini, Marco Magnani, Gianluca Morelli

Modelling consumers' qualitative perceptions of inflation 128
Marcella Corduas, Rosaria Simone, Domenico Piccolo

Noise resistant clustering of high-dimensional gene expression data 132
Pietro Coretto, Angela Serra, Roberto Tagliaferri

Classify X-ray images using convolutional neural networks 136
Federica Crobu, Agostino Di Ciaccio

A compositional analysis approach assessing the spatial distribution of trees in Guadalajara, Mexico 140
Marco Antonio Cruz, Maribel Ortego, Elisabet Roca

Joining factorial methods and blockmodeling for the analysis of affiliation networks 142
Daniela D'Ambrosio, Marco Serino, Giancarlo Ragozini

A latent space model for clustering in multiplex data 146
Silvia D'Angelo, Michael Fop

Post processing of two dimensional road profiles: variogram scheme application and sectioning procedure 150
Mauro D'Apuzzo, Rose-Line Spacagna, Azzurra Evangelisti, Daniela Santilli, Vittorio Nicolosi

A new approach to preference mapping through quantile regression 154
Cristina Davino, Tormod Naes, Rosaria Romano, Domenico Vistocco

On the robustness of the cosine distribution depth classifier 158
Houyem Demni, Amor Messaoud, Giovanni C. Porzio

Network effect on individual scientific performance: a longitudinal study on an Italian scientific community 162
Domenico De Stefano, Giuseppe Giordano, Susanna Zaccarin

Penalized vs constrained maximum likelihood approaches for clusterwise linear regression modelling 166
Roberto Di Mari, Stefano Antonio Gattone, Roberto Rocci

Local fitting of angular variables observed with error 170
Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor

Quantile composite-based path modeling to estimate the conditional quantiles of health indicators 174
Pasquale Dolce, Cristina Davino, Stefania Taralli, Domenico Vistocco

AUC-based gradient boosting for imbalanced classification 178
Martina Dossi, Giovanna Menardi

How to measure material deprivation? A latent Markov model based approach 182
Francesco Dotto

Decomposition of the interval based composite indicators by means of biclustering 186
Carlo Drago

Consensus clustering via pivotal methods 190
Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli

Robust model-based clustering with mild and gross outliers 194
Alessio Farcomeni, Antonio Punzo

Gaussian processes for curve prediction and classification 198
Sara Fontanella, Lara Fontanella, Rosalba Ignaccolo, Luigi Ippoliti, Pasquale Valentini

A new proposal for building immigrant integration composite indicator 199
Mario Fordellone, Venera Tomaselli, Maurizio Vichi

Biodiversity spatial clustering 203
Francesca Fortuna, Fabrizio Maturo, Tonio Di Battista

Skewed distributions or transformations? Incorporating skewness in a cluster analysis 207
Michael Gallaugher, Paul McNicholas, Volodymyr Melnykov, Xuwen Zhu

Robust parsimonious clustering models 208
Luis Angel Garcia-Escudero, Agustin Mayo-Iscar, Marco Riani

Projection-based uniformity tests for directional data 212
Eduardo García-Portugués, Paula Navarro-Esteban, Juan Antonio Cuesta-Albertos

Graph-based clustering of visitors' trajectories at exhibitions 214
Martina Gentilin, Pietro Lovato, Gloria Menegaz, Marco Cristani, Marco Minozzo

Symmetry in graph clustering 218
Andreas Geyer-Schulz, Fabian Ball

Bayesian networks for the analysis of entrepreneurial microcredit: evidence from Italy 222
Lorenzo Giammei, Paola Vicard

The PARAFAC model in the maximum likelihood approach 226
Paolo Giordani, Roberto Rocci, Giuseppe Bove

Structure discovering in nonparametric regression by the GRID procedure 230
Francesco Giordano, Soumendra Nath Lahiri, Maria Lucia Parrella

A microblog auxiliary part-of-speech tagger based on bayesian networks 234
Silvia Golia, Paola Zola

Recent advances in model-based clustering of high dimensional data 238
Isobel Claire Gormley

Tree embedded linear mixed models 239
Anna Gottard, Leonardo Grilli, Carla Rampichini, Giulia Vannucci

Weighted likelihood estimation of mixtures 243
Luca Greco, Claudio Agostinelli

A canonical representation for multiblock methods 247
Mohamed Hanafi

An adequacy approach to estimating the number of clusters 251
Christian Hennig

Classification with weighted compositions 255
Karel Hron, Julie Rendlova, Peter Filzmoser

MacroPCA: an all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers 256
Mia Hubert, Peter J. Rousseeuw, Wannes Van den Bossche

Marginal effects for comparing groups in regression models for ordinal outcome when uncertainty is present 258
Maria Iannario, Claudia Tarantola

A multi-criteria approach in a financial portfolio selection framework 262
Carmela Iorio, Giuseppe Pandolfo, Roberta Siciliano

Clustering of trajectories using adaptive distances and warping 266
Antonio Irpino, Antonio Balzanella

Sampling and learning Mallows and generalized Mallows models under the Cayley distance: short paper 270
Ekhine Irurozki, Borja Calvo, Jose A. Lozano

The gender parity index for the academic students progress 274
Aglaia Kalamatianou, Adele H. Marshall, Mariangela Zenga

Some asymptotic properties of model selection criteria in the latent block model 278
Christine Keribin

Invariant concept classes for transcriptome classification 282
Hans Kestler, Robin Szekely, Attila Klimmek, Ludwig Lausser

Clustering of ties defined as symbolic data 283
Luka Kronegger

Application of data mining in the housing affordability analysis 284
Viera Labudová, Ľubica Sipková

Cylindrical hidden Markov fields 288
Francesco Lagona

Comparing tree kernels performances in argumentative evidence classification 292
Davide Liga

Recent advancement in neural network analysis of biomedical big data 296
Pietro Liò, Giovanna Maria Dimitri, Chiara Sopegno

Bias reduction for estimating functions and pseudolikelihoods 297
Nicola Lunardon

Large scale social and multilayer networks 301
Matteo Magnani

Uncertainty in statistical matching by BNs 305
Daniela Marella, Paola Vicard, Vincenzina Vitale

Evaluating the recruiters' gender bias in graduate competencies 309
Paolo Mariani, Andrea Marletta

Dynamic clustering of network data: a hybrid maximum likelihood approach 313
Maria Francesca Marino, Silvia Pandolfi

Stability of joint dimension reduction and clustering 317
Angelos Markos, Michel Van de Velden, Alfonso Iodice D'Enza

Hidden Markov models for clustering functional data 321
Andrea Martino, Giuseppina Guatteri, Anna Maria Paganoni

Composite likelihood inference for simultaneous clustering and dimensionality reduction of mixed-type longitudinal data 325
Antonello Maruotti, Monia Ranalli, Roberto Rocci

Bivariate semi-parametric mixed-effects models for classifying the effects of Italian classes on multiple student achievements 329
Chiara Masci, Francesca Ieva, Tommaso Agasisti, Anna Maria Paganoni

Multivariate change-point analysis for climate time series 333
Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio, Carlo Blasi

A dynamic stochastic block model for longitudinal networks 337
Catherine Matias, Tabea Rebafka, Fanny Villers

Unsupervised fuzzy classification for detecting similar functional objects 339
Fabrizio Maturo, Francesca Fortuna, Tonio Di Battista

Mixture modelling with skew-symmetric component distributions 343
Geoffrey McLachlan

New developments in applications of pairwise overlap 344
Volodymyr Melnykov, Yana Melnykov, Domenico Perrotta, Marco Riani, Francesca Torti, Yang Wang

Modelling unobserved heterogeneity of ranking data with the bayesian mixture of extended Plackett-Luce models 346
Cristina Mollica, Luca Tardella

Issues in nonlinear time series modeling of European import volumes 350
Gianluca Morelli, Francesca Torti

Gaussian parsimonious clustering models with covariates and a noise component 352
Keefe Murphy, Thomas Brendan Murphy

Illumination in depth analysis 353
Stanislav Nagy, Jiří Dvořák

Copula-based non-metric unfolding on augmented data matrix 357
Marta Nai Ruscone, Antonio D'Ambrosio

A statistical model for software releases complexity prediction 361
Marco Ortu, Giuseppe Destefanis, Roberto Tonelli

Comparison of serious diseases mortality in regions of V4 365
Viera Pacáková, Lucie Kopecká

Price and product design strategies for manufacturers of electric vehicle batteries: inferences from latent class analysis 369
Friederike Paetz

A Mahalanobis-like distance for cylindrical data 373
Lucio Palazzo, Giovanni C. Porzio, Giuseppe Pandolfo

Archetypes, prototypes and other types 377
Francesco Palumbo, Giancarlo Ragozini, Domenico Vistocco

Generalizing the skew-t model using copulas 381
Antonio Parisi, Brunero Liseo

Contamination and manipulation of trade data: the two faces of customs fraud 385
Domenico Perrotta, Andrea Cerasa, Lucio Barabesi, Mario Menegatti, Andrea Cerioli

Bayesian clustering using non-negative matrix factorization 389
Michael Porter, Ketong Wang

Exploring gender gap in international mobility flows through a network analysis approach 393
Ilaria Primerano, Marialuisa Restaino

Clustering two-mode binary network data with overlapping mixture model and covariates information 395
Saverio Ranciati, Veronica Vinciotti, Ernst C. Wit, Giuliano Galimberti

A stochastic blockmodel for network interaction lengths over continuous time 399
Riccardo Rastelli, Michael Fop

Computationally efficient inference for latent position network models 403
Riccardo Rastelli, Florian Maire, Nial Friel

Clustering of complex data stream based on barycentric coordinates 407
Parisa Rastin, Basarab Matei, Guénaël Cabanes

An INDSCAL based mixture model to cluster mixed-type of data 411
Roberto Rocci, Monia Ranalli

Topological stochastic neighbor embedding 415
Nicoleta Rogovschi, Nistor Grozavu, Basarab Matei, Younès Bennani, Seiichi Ozawa

Functional data analysis for spatial aggregated point patterns in seismic science 419
Elvira Romano, Jonatan González Monsalve, Francisco Javier Rodríguez Cortés, Jorge Mateu

ROC curves with binary multivariate data 420
Lidia Sacchetto, Mauro Gasparini

Silhouette-based method for portfolio selection 424
Marco Scaglione, Carmela Iorio, Antonio D'Ambrosio

Item weighted Kemeny distance for preference data 428
Mariangela Sciandra, Simona Buscemi, Antonella Plaia

A fast and efficient modal EM algorithm for Gaussian mixtures 432
Luca Scrucca

Probabilistic archetypal analysis 436
Sohan Seth

Multilinear tests of association between networks 438
Daniel K. Sewell

Use of multi-state models to maximise information in pressure ulcer prevention trials 442
Linda Sharples, Isabelle Smith, Jane Nixon

Partial least squares for compositional canonical correlation 445
Violetta Simonacci Massimo Guarino, Michele Gallo

Dynamic modelling of price expectations 449
Rosaria Simone, Domenico Piccolo, Marcella Corduas

Towards axioms for hierarchical clustering of measures 453
Philipp Thomann, Ingo Steinwart, Nico Schmid

Influence of outliers on cluster correspondence analysis 454
Michel Van de Velden, Alfonso Iodice D'Enza, Lisa Schut

Earthquake clustering and centrality measures 458
Elisa Varini, Antonella Peresan, Jiancang Zhuang

Co-clustering high dimensional temporal sequences summarized by histograms 462
Rosanna Verde, Antonio Irpino, Antonio Balzanella

Statistical analysis of item pre-knowledge in educational tests: latent variable modelling and optimal statistical decision 466
Chen Yunxiao, Lu Yan, Irini Moustaki

Evaluation of the web usability of the University of Cagliari portal: an eye tracking study 468
Gianpaolo Zammarchi, Francesco Mola

Application of survival analysis to critical illness insurance data 472
David Zapletal, Lucie Kopecka

Preface

This book collects the short papers presented at CLADAG 2019, the 12th 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 Law of the University of Cassino and Southern Lazio, under the auspices of 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 a most important meeting point for people interested in classification and data analysis. One reason was certainly the fact that a selection of the presented papers is regularly published in (post-conference) proceedings, typically by Springer Verlag.

The Scientific Committee of CLADAG2019 conceived the Plenary and Invited Sessions to provide a fresh perspective on the state of the art of knowledge and research in the field. The scientific program of CLADAG 2019 is particularly rich. All in all, it comprises 5 Keynote Lectures, 32 Invited Sessions promoted by the members of the Scientific Program Committee, 16 Contributed Sessions, a Round Table and a Data Competition. We thank all the session organizers for inviting renowned speakers, coming from 28 countries. We are greatly indebted to the referees, for the time spent in a careful review.

The editors would like to express their gratitude to the Rector of the University of Cassino and Southern Lazio and the Director of the Department of Economics and Law for having hosted the meeting. Special thanks are finally due to the members of the Local Organizing Committee and all the people who with their abnegation and enthusiasm have worked for CLADAG 2019.

Special thanks go to Alfiero Klain and Livia Iannucci for the editorial and administrative support.

Last but not least, we thank all the authors and participants, without whom the conference would not have been possible.

Cassino, September 11, 2019

Giovanni C. Porzio
Francesca Greselin
Simona Balzano

UNIFYING DATA UNITS AND MODELS IN (CO-)CLUSTERING

Christophe Biernacki[1]

[1] Université Lille 1, FRANCE,
(e-mail: christophe.biernacki@math.univ-lille1.fr)

Statisticians are already aware that any modelling process issue (exploration, prediction) is wholly data unit dependent, to the extend that it should be impossible to provide a statistical outcome without specifying the couple (unit,model). In this talk, this general principle is formalized with a particular focus in model-based clustering and co-clustering in the case of possibly mixed data types (continuous and/or categorical and/or counting features), being also the opportunity to revisit what the related data units are.

Such a formalization allows to raise three important spots: (i) the couple (unit,model) is not identifiable so that different interpretations unit/model of the same whole modelling process are always possible; (ii) combining different "classical" units with different "classical" models should be an interesting opportunity for a cheap, wide and meaningful enlarging of the whole modelling process family designed by the couple (unit,model); (iii) if necessary, this couple, up to the non identifiability property, could be selected by any traditional model selection criterion. Some experiments on real data sets illustrate in detail practical beneficits from the previous three spots.

It is a joint work with Alexandre Lourme (University of Bordeaux).

STATISTICS WITH A HUMAN FACE

Adrian Bowman[1]

[1] University of Glasgow, (e-mail: adrian.bowman@glasgow.ac.uk)

Three-dimensional surface imaging, through laser–scanning or stereo–photogrammetry, provides high-resolution data defining the surface shape of objects.

Human faces are of particular interest and there are many biological and anatomical applications, including assessing the success of facial surgery and investigating the possible developmental origins of some adult conditions.

An initial challenge is to structure the raw images by identifying features of the face. Ridge and valley curves provide a very good intermediate level at which to approach this, as these provide a good compromise between informative representations of shape and simplicity of structure.

Some of the issues involved in analysing data of this type will be discussed and illustrated. Modelling issues include simple comparison of groups, the measurement of asymmetry and longitudinal patterns of shape change. This last topic is relevant at short scale in facial animation, medium scale in individual growth patterns, and very long scale in phylogenetic studies.

BAYESIAN MODEL-BASED CLUSTERING WITH FLEXIBLE AND SPARSE PRIORS

Bettina Grün[1]

[1] Johannes Kepler Universitat Linz, (e-mail: bettina.gruen@jku.at)

Finite mixtures are a standard tool for clustering observations. However, selecting the suitable number of clusters, identifying cluster-relevant variables as well as accounting for non-normal shapes of the clusters are still challenging issues in applications.

Within a Bayesian framework we indicate how suitable prior choices can help to solve these issues. We achieve this considering mainly prior distributions that have the characteristics that they are conditionally conjugate or can be reformulated as hierarchical priors, thus allowing for simple estimation using MCMC methods with data augmentation.

GRINDING MASSIVE INFORMATION INTO FEASIBLE STATISTICS: CURRENT CHALLENGES AND OPPORTUNITIES FOR DATA SCIENTISTS

Francesco Mola[1]

[1] University of Cagliari, (e-mail: mola@unica.it)

Massive amounts of data used to make quicker, better and more intelligent decisions to create business value are nowadays available for companies and organizations. Terms like big data, data science, analytics, artificial intelligence, machine learning etc., are very common in both academia and industry. All these areas of research are orientated towards answering the increasing demand for understanding trends and/or discovering patterns in data. Usually, collected data is massive and uncertain due to noise, incompleteness and inconsistency. The main goal of a statistician/data scientist is therefore to turn massive data into feasible information, the latter intended as able to describe efficiently an observed phenomenon, to gain indications about its future evolution as well as to provide useful insights for the ongoing decisional process. All these considerations lead towards arguing that the role of the statistician/data scientist considerably evolved in the latest years. In my presentation, after a brief description of the scenario summarized above, I will discuss three examples/case studies concerning image validation, hotels' reputation and social media popularity trying to give a contribution to the debate about turning the enormous amount of available data into feasible statistics. In all cases, adhoc but standard classification methods are used to obtain information that is extremely feasible and adds value to a decisional process.

STATISTICAL CHALLENGES IN THE ANALYSIS OF COMPLEX RESPONSES IN BIOMEDICINE

Sylvia Richardson[1]

[1] University of Cambridge, (e-mail: sylvia.richardson@mrc-bsu.cam.ac.uk)

To exploit better the structure of the rich sets of characteristics, such as clinical biomarkers, molecular profiles or detailed ontology records, that are currently being collected on large samples of healthy or diseased individuals, statistical models of the variations within and the interplay between different layers of data can be constructed.

Generic Bayesian model building strategies and algorithms have been tailored for this purpose. In this talk, I will discuss three areas: implementing joint hierarchical modelling of a large number of responses and a large number of features to discover features associated with many responses; analysing tree structured ontology data with application for finding the underlying genetic origin of rare diseases; and characterising network structures using fast Bayesian inference in large Gaussian graphical models. Common statistical issues of accounting for model uncertainty, ability to borrow information for retaining power and scalability of Bayesian computations will be highlighted. Modelling strategies and computations will be illustrated on case studies.

Model-based clustering of time series data: a flexible approach using nonparametric state-switching quantile regression models

Timo Adam[1], Roland Langrock[1] and Thomas Kneib[2]

[1] Bielefeld University, (e-mail: timo.adam@uni-bielefeld.de, roland.langrock@uni-bielefeld.de)

[2] University of Göttingen, (e-mail: tkneib@uni-goettingen.de)

Abstract: We propose a model-based clustering approach for time series data applications where clusters are inferred from the conditional quantiles of the variable of interest given the current state of a hidden state process. The suggested methodology allows us to draw a detailed picture of i) the effect of some covariate on those quantiles within clusters, and ii) the entire response distribution in a flexible data-driven way without the need to specify a parametric family of distributions. As an illustrating example, we model Spanish energy prices to obtain clusters relating to periods of relatively calm and nervous market regimes, respectively.

Keywords: hidden Markov models, penalized B-splines, quantile regression.

1 Introduction

Quantile regression models (QMs, Koenker, 2005) are widely used for modeling the conditional quantiles of the variable of interest given some covariate. In this paper, we extend QMs to time series data applications where the quantile curves are subject to state switching controlled by a hidden Markov chain, which provides an essentially distribution-free alternative to Markov-switching generalized additive models for location, scale, and shape (MS-GAMLSS, Adam *et al.*, 2017, Langrock *et al.*, 2018). By decoding the hidden states underlying the observations, the resulting class of Markov-switching QMs (MS-QMs) can be used for model-based clustering of time series data.

2 Methodology

2.1 Model formulation and dependence structure

MS-QMs comprise two stochastic processes, a *hidden* state process, $\{S_t\}_{t=1,\dots,T}$, and an *observed* state-dependent process, $\{Y_t\}_{t=1,\dots,T}$. The state process is

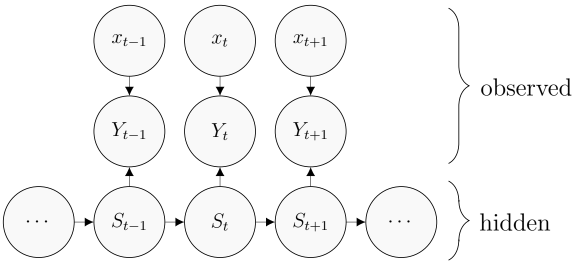


Figure 1. *Dependence structure of a Markov-switching quantile regression model.*

modeled as a discrete-time N-state Markov chain (where N determines the number of clusters) with transition probability matrix $\Gamma = (\gamma_{ij})$, where $\gamma_{ij} = \Pr(S_t = j | S_{t-1} = i)$, $i, j = 1, \dots, N$, and initial distribution (row) vector $\delta = (\delta_i)$, where $\delta_i = \Pr(S_1 = i)$, $i = 1, \dots, N$.

At each time t, the state-dependent process generates an observation from some (unspecified) distribution with state-dependent quantile functions $g_\tau^{(s_t)}(x_t)$, where $0 < \tau < 1$ denotes the quantile of interest. Using penalized B-splines (Eilers & Marx, 1996), the quantiles are modeled as functions of covariates,

$$g_\tau^{(s_t)}(x_t) = \beta_{\tau,0}^{(s_t)} + \sum_{k=1}^{K} \beta_{\tau,k}^{(s_t)} B_k^d(x_t), \tag{1}$$

where $\beta_{\tau,0}^{(s_t)}$ denotes the state-dependent intercept and $\beta_{\tau,k}^{(s_t)}$ the coefficient associated with the k-th B-spline basis function of degree d evaluated at the covariate value x_t, i.e. $B_k^d(x_t)$; we consider cubic basis functions, i.e. $d = 3$.

2.2 Model fitting and clustering

For a fixed quantile τ, quantile regression is commonly carried out by optimization with respect to the loss function $\rho_\tau(y_t - g_\tau(x_t)) = (y_t - g_\tau(x_t))\{\tau - \mathbb{1}_{(y_t - g_\tau(x_t))<0}\}$. This is equivalent to maximum likelihood assuming an asymmetric Laplace (AL) distribution, with density $f_{\text{AL}}(y_t; \mu_t, \sigma, \tau)$, which yields, in a Bayesian setup, posterior consistent estimators even if the observations are not AL-distributed (Sriram *et al.*, 2016). Defining the forward variables $\alpha_t(i) = f(y_1, \dots, y_t, S_t = i)$, which are summarized in the row vectors $\alpha_t = (\alpha_t(1), \dots, \alpha_t(N))$, the recursion

$$\alpha_1 = \delta \mathbf{P}(y_1); \alpha_t = \alpha_{t-1} \Gamma \mathbf{P}(y_t), \; t = 2, \dots, T, \tag{2}$$

can be applied to evaluate α_T, where $\mathbf{P}(y_t) = \text{diag}(f_{\text{AL}}(y_t; \mu_t^{(1)}, \sigma^{(1)}, \tau), \dots, f_{\text{AL}}(y_t; \mu_t^{(N)}, \sigma^{(N)}, \tau))$, with state-dependent quantile curves $\mu_t^{(i)} = g_\tau^{(i)}(x_t)$, scale parameters $\sigma^{(i)}$, and (fixed) quantile τ. From α_T, the likelihood is obtained as $\mathcal{L}(\theta) = f(y_1,\ldots,\, y_T|\theta) = \sum_{i=1}^{N} f(y_1,\ldots,y_T,s_T = i) = \alpha_T\mathbf{1}$. For simultaneously considering multiple quantiles τ_q, $q = 1,\ldots,Q$, we follow Sriram *et al.* (2016) and consider a pseudo-likelihood as the objective criterion to be maximized, where the quantile-specific state-dependent densities in (1) are replaced by $\prod_{q=1}^{Q} f_{\mathrm{AL}}(y_t;\,\ldots,\tau_q)$. To avoid i) overfitting and ii) quantile crossing, two penalties are added to the pseudo-log-likelihood, which leads to

$$\begin{aligned}\log\left(\mathcal{L}_{\text{pen.}}(\theta)\right) =& \log\left(\mathcal{L}(\theta)\right)\\ &\underbrace{-\sum_{i=1}^{N}\sum_{q=1}^{Q}\lambda_q^{(i)}\sum_{k=3}^{K}\left(\Delta^2\beta_{\tau_q,k}^{(i)}\right)^2}_{\text{roughness penalty}}\underbrace{-c\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{q=2}^{Q}\mathbb{1}_{(g_{\tau_q}^{(i)}(x_t)-g_{\tau_{q-1}}^{(i)}(x_t))\leq 0},}_{\text{quantile crossing penalty}}\end{aligned}$$

where $\lambda_q^{(i)}$ denotes some smoothing parameter, $\Delta^2\beta_{\tau_q,k}^{(i)}$ the squared second-order differences between adjacent coefficients, and c some (arbitrary) constant which ensures non-crossing quantile curves.

From some fitted MS-QM, clusters can be obtained by computing the most likely state sequence underlying the observations via the Viterbi algorithm. The state-dependent densities as required for Viterbi can be approximated based on the estimated quantile curves as

$$\hat{f}_Y^{(i)}(y_t) = \frac{\tau_{q^*} - \tau_{q^*-1}}{\hat{g}_{\tau_q^*}^{(i)}(x_t) - \hat{g}_{\tau_{q^*-1}}^{(i)}(x_t)},\; q^* = \min\left\{q \in 0,\ldots Q+1 : y_t \leq \hat{g}_{\tau_q}^{(i)}(x_t)\right\},$$

where the (not estimated) quantile curves associated with $\tau_0 = 0$ and $\tau_{Q+1} = 1$ are defined as $\min(y_1,\ldots,y_T)$ and $\max(y_1,\ldots,y_T)$, respectively.

3 Illustrating example

As an illustrating example, we model the conditional quantiles of daily energy prices in Spain, Y_t, given the oil price, x_t, over time. The data (Sanchez-Espigares & Lopez-Moreno, 2014) comprise 1761 daily observations between February 1, 2002, and October 31, 2008. For each of $N = 2$ states, we used $K = 30$ B-spline basis functions, where (for simplicity) all smoothing parameters were set to 1, and c was chosen to be 5.

The results are displayed in Figure 2. Within cluster 1, the energy prices are fairly low and exhibit a moderate volatility. Within cluster 2, the prices are generally higher and exhibit a considerably higher volatility. Overall, the energy price distribution is quite heavily affected by the oil price, where the corresponding effect substantially differs across both clusters and quantiles.

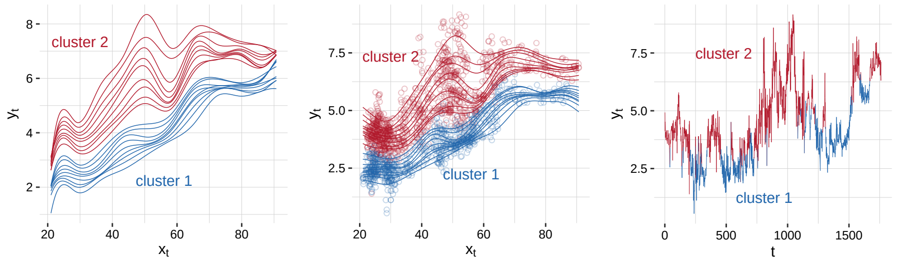


Figure 2. *Fitted state-dependent quantile curves for* $\tau = (0.1, 0.2, 0.3, \dots, 0.9)$ *without penalization (left), with penalization (center), and Viterbi-decoded time series as obtained under the model with penalization (right).*

4 Discussion

We have proposed MS-QMs as a model-based clustering approach for time series data applications. Key features of MS-QMs include i) their feasibility to infer cluster-specific covariate effects on various quantiles, and ii) the flexible, data-driven way in which the entire response distribution is modeled. The immense flexibility, however, comes at the cost of a potentially large set of tuning parameters. The development of efficient model selection techniques may therefore provide an important avenue for future research.

References

Adam, T., Mayr, A., & Kneib, T. 2017. Gradient boosting in Markov-switching generalized additive models for location, scale, and shape. *arXiv*, 1710.02385.

Eilers, P.H.C., & Marx, B.D. 1996. Flexible smoothing with B-splines and penalties. *Statistical Science*, 89–102.

Koenker, R. 2005. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Langrock, R., Adam, T., Leos-Barajas, V., Mews, S., Miller, D.L., & Papastamatiou, Y.P. 2018. Spline-based nonparametric inference in general state-switching models. *Statistica Neerlandica*, **72**(3), 179–200.

Sanchez-Espigares, J.A., & Lopez-Moreno, A. 2014. *MSwM: Fitting Markov-Switching Models. R package, version 1.2.*

Sriram, K., Ramamoorthi, R.V., & Ghosh, P. 2016. On Bayesian quantile regression using a pseudo-joint asymmetric Laplace likelihood. *Sankhya A*, **78**(1), 87–104.

SOME ISSUES IN GENERALIZED LINEAR MODELING

Alan Agresti[1]

[1] Distinguished Professor Emeritus, University of Florida, (e-mail: agresti@ufl.edu)

ABSTRACT: My talk discusses topics pertaining to generalized linear modeling, with focus on categorical data: (1) bias due to floor and ceiling effects in using ordinary linear models with ordinal response data, (2) interpreting effects with nonlinear link functions, (3) alternatives to logit and probit link functions with binary responses, (4) cautions in using Wald tests and confidence intervals when effects are large, and (5) the behavior and choice of residuals. In this accompanying paper, we discuss topics (2) and (3), which involve new and recent research.

KEYWORDS: Ordinal models, binary data, nonlinear link functions.

1 Introduction

We discuss some issues about generalized linear models that deserve more attention in terms of additional research or greater awareness of existing literature. I became increasingly aware of the issues years while writing a book on linear and generalized linear models (Agresti 2015) and while revising three books on categorical data analysis (Agresti 2010, 2013, 2019). This paper discusses two of five topics from my talk: Section 2 proposes a simple way to interpret effects in generalized linear models that use nonlinear link functions, by comparing groups using a probability summary about the higher response. Section 3 argues that for modeling binary responses, the identity link and log link functions can often supplement the logit and probit links.

2 Interpreting Effects in GLMs with Nonlinear Link Function

For many standard nonlinear link functions in generalized linear modeling, the interpretation of the model effects is difficult for non-statisticians and for methodologists who are mainly familiar with ordinary linear models. To illustrate, suppose y is ordinal with c outcome categories. For observation i, let x_{ik} denote the value of explanatory variable k. Consider the *cumulative link model*

$$\text{link}[P(y_i \leq j)] = \alpha_j + \sum_k \beta_k x_{ik}, \quad j = 1, \ldots, c-1,$$

for links such as the logit, probit, or complementary log-log,. For the probit link (i.e., the inverse of the standard normal cdf Φ), β_k represents the change in $\Phi^{-1}[P(y_i \leq j)]$ for a 1-unit increase in x_k, adjusting for the other explanatory variables. This is a rather obscure interpretation, as few people can make sense of effects on the scale of an inverse of a cdf.

One way used to interpret effects relies more on an underlying latent variable model (McKelvey and Zavoina 1975). For the observed ordinal response y and for a latent response y^*, suppose $y_i^* = \boldsymbol{\beta}^T \boldsymbol{x}_i + \varepsilon_i$, where ε_i has some parametric cdf G with mean 0. Suppose that thresholds (cutpoints) $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_c = \infty$ exist such that

$$y_i = j \text{ if } \alpha_{j-1} < y_i^* \leq \alpha_j.$$

Then, at a fixed value $\boldsymbol{x}$,

$$P(y_i \leq j) = P(y_i^* \leq \alpha_j) = P(y_i^* - \beta^T \boldsymbol{x}_i \leq \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i)$$

$$= P(\varepsilon_i \leq \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i) = G(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i).$$

This implies the model

$$G^{-1}[P(y_i \leq j \mid \boldsymbol{x}_i)] = \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}_i$$

with G^{-1} as the link function. In particular, one obtains the cumulative probit model when G is the standard normal cdf Φ; then Φ^{-1} is the probit link.

We suggest a simple interpretation that utilizes this latent variable model, formulated in terms of a summary for comparing two groups, adjusting for the other explanatory variables. Let z be an indicator variable for the two groups. At any potential setting $(x_1, \ldots, x_p)$ of p explanatory variables, let y_1^* and y_2^* denote independent latent variables when $z = 1$ and when $z = 0$, respectively. For the latent variable model that generates the cumulative probit model

$$\Phi^{-1}[P(y \leq j)] = \alpha_j - \beta z - \beta_1 x_1 - \cdots - \beta_p x_p,$$

the difference between the conditional means of y_1^* and y_2^* is β, and

$$P(y_1^* > y_2^*) = P[(y_1^* - y_2^*) > 0]$$

$$= P\left[\frac{(y_1^* - y_2^*) - \beta}{\sqrt{2}} > \frac{-\beta}{\sqrt{2}}\right] = 1 - \Phi(-\beta/\sqrt{2}) = \Phi(\beta/\sqrt{2}).$$

At any setting of the p explanatory variables, differences between the normal conditional means for the two groups of $\beta = (0, 0.5, 1, 2, 3)$ standard deviations correspond to $P(y_1^* > y_2^*)$ values of (0.50, 0.64, 0.76, 0.92, 0.98).

For details, including corresponding expressions with logit and complementary log-log links, see Agresti and Kateri (2016). That article also discusses related measures for the observed response scale. The idea can extend to other generalized linear models and to more complex models, such as generalized additive models.

3 Using Alternatives to the Logit and Probit Links with Binary Responses

For binary responses, the logit and probit links are used almost exclusively. Sometimes, however, we can also use the log and the identity links.

- The identity link provides similar fits as the logit or probit link when $P(y=1)$ falls mainly between about 0.2 and 0.8. It has simpler interpretations, as the model parameters relate to *differences of probabilities* instead of *ratios of odds*.
- The log link provides similar fits as the logit or probit link when $P(y=1)$ falls mainly below 0.5. It has simpler interpretations, as the model parameters relate to *ratios of probabilities* instead of *ratios of odds*.
- With uncorrelated explanatory variables, the effects with log and identity links are the same in the full model as in marginal models with sole predictors, which is not true with logit or probit links.

We illustrate the first two points with data from a recent Istat survey. For the binary response y = whether employed (i.e., $y=1$ means that the person is present in some administrative source), we use explanatory variables x_1 = gender (1 = female, 0 = male), x_2 = whether an Italian citizen (1 = yes, 0 = no), and x_3 = whether receiving a pension (1 = yes, 0 = no).

Consider first the 27,775 subjects in the survey having age over 65. For the 8 combinations of x_1, x_2, x_3, the sample proportions employed fall between 0.02 and 0.12. The main-effects logit and log-link model fits are

$$\text{logit}[\hat{P}(y=1)] = -1.8686 - 1.3236x_1 - 0.4295x_2 + 0.2162x_3,$$

$$\log[\hat{P}(y=1)] = -2.0374 - 1.2388x_1 - 0.3619x_2 + 0.2003x_3.$$

The absolute difference in fitted proportions, averaged over the 27,775 cases, is 0.0001. For the log-link model, the exponentiated coefficients estimate probability ratios; e.g., adjusting for x_2 and x_3, the probability a woman is employed is estimated to be $\exp(-1.2388) = 0.2897$ times the probability a man is employed.

Consider next the 72,225 subjects having age under 65. For the 8 combinations of x_1, x_2, x_3, the sample proportions employed fall between 0.18 and 0.74. The main-effects logit and identity-link model fits are

$$\text{logit}[\hat{P}(y=1)] = 0.3502 - 0.6440x_1 + 0.7017x_2 - 1.8737x_3,$$

$$\hat{P}(y=1) = 0.5875 - 0.1386x_1 + 0.1513x_2 - 0.4079x_3.$$

The absolute difference in fitted proportions, averaged over the 72,225 cases, is only 0.004. For the identity-link model, the coefficients estimate differences of probabilities. For instance, adjusting for x_2 and x_3, the probability that a woman is employed is estimated to be 0.1386 lower than the probability that a man is employed.

The effects in the models using log and identity links can be approximated by linearizations of logit-link models, such as by using *average marginal effects* measures that are available with software such as R and Stata. For details, see Agresti, Tarantola, and Varriale (2019). Such measures are also relevant for ordinal responses (Agresti and Tarantola 2018).

References

AGRESTI, A. 2010. *Analysis of Ordinal Categorical Data*, Wiley.

AGRESTI, A. 2010. *Analysis of Ordinal Categorical Data*, Wiley.

AGRESTI, A. 2013. *Categorical Data Analysis, 3rd ed.*, Wiley.

AGRESTI, A. 2015. *Foundations of Linear and Generalized Linear Models*, Wiley.

AGRESTI, A. 2019. *An Introduction to Categorical Data Analysis, 3rd ed.*, Wiley.

AGRESTI, A. & KATERI, M. 2016. Ordinal probability effect measures for group comparisons in multinomial cumulative link models, *Biometrics*, **73**, 214-219.

AGRESTI, A. & TARANTOLA, C. 2018. Simple effect measures for interpreting models for ordinal categorical data, *Statistica Neerlandica*, **72**, 210-223.

AGRESTI, A., TARANTOLA, C. & VARRIALE, R. 2019 (to appear). Simple ways to interpret effects in modeling binary responses, . in *Book of contributions to workshop on analyzing discrete data*, Aachen, Germany.

MCKELVEY, R.D. & ZAVOINA, W. 1975. A statistical model for the analysis of ordinal level dependent variables. *J. Math. Sociol.* **4**, 103-120.

ASSESSING SOCIAL INTEREST IN BURNOUT USING FUNCTIONAL DATA ANALYSIS THROUGH GOOGLE TRENDS

Ana Aguilera[1], Francesca Fortuna[2] and Escabias Manuel[1]

[1] Department of Statistics and O.R., University of Granada,
(e-mail: aaguiler@ugr.es, escabias@ugr.es)

[2] DISFPEQ, G. d'Annunzio University, Pescara,
(e-mail: francesca.fortuna@unich.it)

ABSTRACT: Burnout is a serious problem in modern society and early detection methods are needed to successfully handled its multiple effects. However, in many countries, official statistics on this topic are not available. For this reason, we propose to use Google Trends data as proxies for the interest in burnout and to analyze them through the functional data analysis (FDA) approach. Under this framework, the functional analysis of variance (FANOVA) model is used for testing a macro geographic area effect on search queries for the keyword "burnout" in Italy. The estimation of the FANOVA model is proposed in a finite dimensional space generated by a basis function representation. Thus, the functional model is reduced to a MANOVA model on the basis coefficients.

KEYWORDS: Burnout, Google Trends data, FDA, FANOVA model.

1 Introduction

Burnout is typically defined as a three dimensional syndrome characterised by emotional exhaustion, depersonalization and lack of professional efficacy (Maslach & Jackson, 1981). It has a strong impact not only on working well-being as it inevitably influences the private and social life of individuals. Indeed, burnout can affect health, giving rise to both physical and psychosomatic problems such as depression, anxiety, low self-esteem, guilt feelings, and low tolerance of frustration (Maslach *et al.*, 2001). In this context, the role of social support in reducing the negative effects of burnout becomes fundamental, especially under the current situation of crisis in the world of work. Although the importance of this phenomenon is now recognized, in many countires official statistics on the rates of burnout among workers are not available. For this reason, we propose the use of Google Trends data as proxies for assessing burnout. The basic idea is that internet searches may be considered indicators of the public interest. Indeed, people reveal information about their needs, wants, interests, moods and phycological problems through their Internet search histories, which are stored in the form of Google Trends data. More specifically, we propose to analyze Google Trends data through the functional data analysis (FDA) approach (Ramsay & Silverman, 2005) because data floowing from the web can be viewed as an infinite process, which continuously evolve over the time domain (Fortuna *et al.*, 2018). Since functional data are infinite-dimensional objects, they provide a more suitable representation of Google Trends search queries than traditional multivariate vectors. Moreover, FDA allows to address the so-called 'curse of dimensionality' of big data, enabling an effective statistical analysis when the number of variables exceeds the number of observations. Under this framework, the functional analysis of variance (FANOVA) model has been applied for studying the relationship between the functional queries and an explanatory categorical variable. In particular, the problem of testing the null hypothesis of equality of mean functions across different groups is addressed. In this paper, the estimation of the FANOVA model has been considered in a finite dimensional space generated by a basis. Then, the problem has been reduced to a finite multivariate ANOVA (MANOVA) model on the vector of basis coefficients.

2 The FANOVA model with regularized basis expansions for Google Trends data

Since Google Trends data continuously flow from the server of a web site, they can be seen as functions in a continuous domain, rather than scalar vectors (Fortuna *et al.*, 2018). Specifically, let $y_j(t) = \left\{ y_j(t_{jl}) \right\}_{l=1}^{L}$, $j = 1, 2, ..., n$, be a functional variable observed in a discrete set of sampling points, $l = 1, 2, ..., L$, in the temporal domain $\mathcal{T}$. Let us also assume that $y(t) \in L^2(\mathcal{T})$, where $L^2(\mathcal{T})$ is the Hilbert space of square integrable functions. One usual solution to reconstruct the functional form of the n samples starting from the discrete observations, is to assume that sample paths belong to a finite-dimension space spanned by a basis $\{\phi_1(t), \phi_2(t), \cdots, \phi_K(t)\}$, so that they can be expressed as follows:

$$\boldsymbol{y}(t) = \boldsymbol{A}\boldsymbol{\phi}(t) \tag{1}$$

where $\boldsymbol{y} = [y_1(t), ..., y_n(t)]^T$; $\boldsymbol{A} = (a_{jk})$ is the matrix of basis coefficient expansion; and $\boldsymbol{\phi}(t) = [\phi_1(t), \cdots, \phi_K(t)]^T$ is a K dimensional vector of basis functions.

Let $\{y_{ij}(t) : t \in \mathcal{T}, i = 1,...,I; j = 1,...,n_i\}$ be I independent samples of functions drawn from a second order stochastic process $Y = \{Y(t) : t \in \mathcal{T}\}$, continuous in quadratic mean, whose sample functions belong to $L^2(\mathcal{T})$. Assuming that there is a single factor with I different levels or groups ($i = 1,2,...,I$) and n_i observations within each group; the model for the j-th observation ($j = 1,2,...,n_i$) in the i-th group can be expressed as follows:

$$\boldsymbol{y}(t) = \boldsymbol{Z}\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t) \tag{2}$$

where $\boldsymbol{y}(t) = [y_1(t), y_2(t), ..., y_n(t)]^T$ is a vector of functional observations of length $n = \sum_{i=1}^{I} n_i$; $\boldsymbol{\beta}(t) = [\beta_1(t) = \mu(t), \beta_2(t) = \gamma_1(t), ..., \beta_Q(t) = \gamma_I(t)]^T$ is a vector of functional effects of length $Q = I + 1$; $\boldsymbol{\varepsilon}(t) = [\varepsilon_1(t), \varepsilon_2(t), ..., \varepsilon_n(t)]^T$ is a vector of n residual functions and $\boldsymbol{Z}$ is a $(n \times Q)$ design matrix, coding the group membership. The FANOVA model is equivalent to a standard ANOVA model, with the difference that the parameters $\boldsymbol{\beta}(t)$, and hence the predicted observations $\widehat{\boldsymbol{y}}(t) = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}(t)$, are vectors of functions rather than vectors of numbers.

The parameter vector $\boldsymbol{\beta}(t)$ in equation (2) can be estimated using the standard least squares criterion; thus, minimizing the residual sum of squares:

$$LMSSE(\boldsymbol{\beta}) = \int [\boldsymbol{y}(t) - \boldsymbol{Z}\boldsymbol{\beta}(t)]^T [\boldsymbol{y}(t) - \boldsymbol{Z}\boldsymbol{\beta}(t)]\, dt \tag{3}$$

To fit the model (2), it is usual to assume that the sample paths and the parameter functions belong to the same finite space generated by a basis of functions, so that the observed response functions are expressed as in (1) and the regression functions as follows:

$$\beta_q(t) = \sum_{k=1}^{K} b_{qk}\phi_k(t) = \boldsymbol{B}\boldsymbol{\phi}(t) \quad q = 1,...,Q; \tag{4}$$

where $\boldsymbol{B} = (b_{ik})$ is the matrix of basis function coefficients and $\boldsymbol{\phi}(t) = (\phi_1(t), ...,$ $..., \phi_K(t))^T$ is the K dimensional vector of basis functions. In this context, the least squares fitting criterion in (3) can be defined as follows:

$$LMSSE(\boldsymbol{\beta}) = \int \left[\boldsymbol{A}\boldsymbol{\phi}(t) - \boldsymbol{Z}\boldsymbol{B}\boldsymbol{\phi}(t)\right]^T \left[\boldsymbol{A}\boldsymbol{\phi}(t) - \boldsymbol{Z}\boldsymbol{B}\boldsymbol{\phi}(t)\right] dt \tag{5}$$

which leads to the following estimation of the functional effects:

$$\widehat{\boldsymbol{\beta}}(t) = (\boldsymbol{Z}^T\boldsymbol{\Psi}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{\Psi}\boldsymbol{A} \tag{6}$$

where $\boldsymbol{\Psi} = (\psi_{jq})_{K \times K}$ is the symmetric matrix of the inner products between basis functions, $\boldsymbol{\Psi} = \int_{\mathcal{T}} \boldsymbol{\phi}(t)^T \boldsymbol{\phi}(t)$, and $\boldsymbol{A}$ has an additional row of zeros to satisfy the constraint on the functional effects (Sayes *et al.*, 2008).

3 Conclusions

Burnout is a growing problem in the modern society. It is usually thought of as an individual response to prolonged work related stress, which in turn, impacts on job satisfaction and thereafter, can affect the phycological, physiological, affective and behavioral well-being of workers (Dyrbye *et al.*, 2011). The estimation of this phenomenon is essential to design social support for reducing its negative effects. However, in many countries, official statistics for the rates of burnout are not available. In this context, we propose the use of Google Trends data as proxies for the interest in burnout. In this scenario, we aim to provide an original methodological approach for the analysis of social indicators based on big data, through the FDA approach. The latter has the advantage of reducing the dimension of the huge amount of data with the conversion of vectors into functions. Under this framework, the FANOVA model can be used for testing a possible effect of different factors on the search queries.

References

DYRBYE, L.N., SHANAFELT, T.D., BALCH, C.M., SATELE, D., SLOAN, J., & FREISCHLAG, J. 2011. Relationship between work-home conflicts and burnout among American surgeons: A comparison by sex. *Archives of Surgery*, **146**, 211–217.

FORTUNA, F., MATURO, F., & DI BATTISTA, T. 2018. Clustering functional data streams: Unsupervised classification of soccer top players based on Google trends. *Quality and Reliability Engineering International*, **34**, 1448–1460.

MASLACH, C., & JACKSON, S.E. 1981. The measurement of experienced burnout. *Journal of Organizational Behavior*, **2**, 99–113.

MASLACH, C., SCHAUFELI, W.B., & LEITER, M.P. 2001. Job burnout. *Annual Review of Psychology*, **52**, 397–422.

RAMSAY, J. O., & SILVERMAN, B. W. 2005. *Functional Data Analysis*. New York: Springer. 2nd edition.

SAYES, W., DE KETELAEREA, B., & DARIUSA, P. 2008. Potential applications of functional data analysis in chemometrics. *Journal of Chemometrics*, **22**, 335–344.

MEASURING EQUITABLE AND SUSTAINABLE WELL-BEING IN ITALIAN REGIONS: A NON-AGGREGATIVE APPROACH

Leonardo Salvatore Alaimo[1] and Filomena Maggino[2]

[1] Department of Social Sciences and Economics, Sapienza University of Rome. Italian National Institute of Statistics, (e-mail: leonardo.alaimo@istat.it)

[2] Department of Statistics, Sapienza University of Rome,
(e-mail: filomena.maggino@uniroma1.it)

ABSTRACT: There are many attempts to measure well-being in various countries around the world. The Italian experience, conducted by the Italian National Institute of Statistics (Istat), is probably the most advanced (Equitable and Sustainable Well-being – BES): the selection of indicators involved many actors of civil, entrepreneurial and institutional society; there are twelve well-being domains and around 130 individual indicators drawn mainly from Istat surveys and archives. In this way, Istat generated a complex multi-indicator system, the understanding of which required the adoption of approaches that would allow for more concise views that could summarise the complexity. In this perspective, the guiding concept crossing all possible strategies is synthesis. In the last four BES reports, the Istat adopted the aggregative approach to synthesis and calculated composite indicators to provide one-dimensional measurements for each domain. Nowadays, in literature, the work paradigm adopted by Italian official statistics seems to be the most complete and imitated. The objective of our work is to provide, starting from the indicators of each domain, synthesis adopting a non-aggregative approach, namely the Partial Order Set Theory (Poset). In particular, the synthetic indicators in time series from 2010 to 2017 will be constructed for the Italian Regions (provided by Istat using an aggregative procedure) using the posets trying to analyse the phenomenon from a spatial and temporal perspective.

KEYWORDS: well-being, italian regions, synthesis, non-agggregative approach, poset.

References

ALAIMO, L. S., & CONIGLIARO, P. forthcoming. Assessing Subjective Well-being in Wide Populations. A Posetic Approach to Micro-data Analysis. *In:* R. BRÜGGEMANN, F. MAGGINO, C. SUTER, & BEYCAN, T. (eds), *Measuring and Understanding Complex Phenomena. Indicators and their Analysis in Different Scientific Fields*. Cham: Springer.

ALAIMO, L. S., & MAGGINO, F. 2019. Sustainable Development Goals Indicators at Territorial Level: Conceptual and Methodological Issues—The Italian Perspective. *Social Indicators Research*, 1–37.

ANNONI, PAOLA, & BRÜGGEMANN, RAINER. 2009. Exploring partial order of European countries. *Social indicators research*, **92**(3), 471.

FATTORE, M. 2016. Partially ordered sets and the measurement of multidimensional ordinal deprivation. *Social Indicators Research*, **128**(2), 835–858.

FATTORE, M. 2017. Synthesis of Indicators: The Non-aggregative Approach. *Pages 192–212 of:* MAGGINO, F. (ed), *Complexity in society: From indicators construction to their synthesis*. Cham: Springer.

FATTORE, M., MAGGINO, F., & ARCAGNI, A. 2015. Exploiting ordinal data for subjective well-being evaluation. *Statistics in Transition new series*, **3**(16), 409–428.

MAGGINO, F. 2017a. Dealing with syntheses in a system of indicators. *Pages 115–137 of:* MAGGINO, F. (ed), *Complexity in society: From indicators construction to their synthesis*. Cham: Springer.

MAGGINO, F. 2017b. Developing indicators and managing the complexity. *Pages 87–114 of:* MAGGINO, F. (ed), *Complexity in society: From indicators construction to their synthesis*. Cham: Springer.

BOOTSTRAP INFERENCE FOR MISSING DATA RECONSTRUCTION

Giuseppina Albano[1], Michele La Rocca[2], Maria Lucia Parrella[2] and Cira Perna[2]

[1] Department of Political and Social Studies, University of Salerno,
(e-mail: pialbano@unisa.it)

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

ABSTRACT: Imputing missing data from a data set is still a challenging issue both in theoretical and applied statistics. In the context of multivariate time series, the problem of missing data becomes even more challenging due to the dependence structure which is present in the data. Recently, a new imputation procedure for multivariate time series has been proposed in Parrella *et al.* , 2018, which uses the class of *Spatial Dynamic Panel Data* models (*SDPD*) to model serial correlation and cross-correlation simulanteously. This paper is aimed at discussing a residual bootstrap construction to approximate the sampling distribution of the missing value estimators.

KEYWORDS: multivariate time series, missing values, bootstrap.

1 Introduction: the model and the imputation procedure

Let $\mathbf{y}_t$ be a multivariate stationary process of order p, assumed for simplicity with zero mean value, collecting the observations at time t from p different variables. Following Dou *et al.* , 2016 and Parrella *et al.* , 2018, we assume that the process can be modeled by the following *SDPD* model

$$\mathbf{y}_t = D(\lambda_0)\mathbf{W}\mathbf{y}_t + D(\lambda_1)\mathbf{y}_{t-1} + D(\lambda_2)\mathbf{W}\mathbf{y}_{t-1} + \mathbf{u}_t, \qquad (1)$$

where $D(\cdot)$ are diagonal matrices with diagonal coefficients from the vectors λ_0, λ_1 and λ_2, and the error process $\mathbf{u}_t$ is serially uncorrelated. Model (1) belongs to the family of *spatial econometric models*, so it is particularly oriented to model spatio-temporal data. The matrix $\mathbf{W}$ is called *spatial matrix* and collects the weigths used in the *spatial regression* of each time series observation with simultaneous or delayed observations of neighboring data. However, if one uses a correlation based matrix $\mathbf{W}$ to measure variable distances, instead of using physical distances, one can use model (1) to analyse any kind of multivariate time series, not necessarily of strictly spatial nature.

In the following, we assume that $\mathbf{y}_1, \cdots, \mathbf{y}_T$ are realizations from the stationary process defined by (1). Then, we denote with $\Sigma_j = Cov(\mathbf{y}_t, \mathbf{y}_{t-j}) = E(\mathbf{y}_t \mathbf{y}'_{t-j})$ the autocovariance matrix of the process at lag j, where the prime subscript denotes the transpose operator. Let us assume that $\widetilde{\mathbf{y}}_1, \cdots, \widetilde{\mathbf{y}}_T$ are realizations from a stationary process as in (1), not necessarily with zero mean value. In case of processes with no zero mean, model (1) can be still used for parameter estimation after a pre-processing step which centers the observed time series. Let $\delta_t = (\delta_{t1}, \ldots, \delta_{tp})$ be a vector of zeroes/ones that identifies all the missing values in the observed vector $\widetilde{\mathbf{y}}_t$, so that $\delta_{ti} = 0$ if the observation $\widetilde{y}_{ti}$ is missing, otherwise it is $\delta_{ti} = 1$.

The imputation procedure for missing values and missing sequences has been proposed in Parrella *et al.* , 2018. It starts, at iteration 0, by initializing the mean centered vector $\mathbf{y}_t^{(0)}$, for $t = 1, \ldots, T$, as

$$\mathbf{y}_t^{(0)} = \delta_t \circ \left(\widetilde{\mathbf{y}}_t - \bar{\mathbf{y}}^{(0)} \right), \qquad \text{with } \bar{\mathbf{y}}^{(0)} = \sum_{t=1}^{T} \delta_t \circ \widetilde{\mathbf{y}}_t / \sum_{t=1}^{T} \delta_t, \tag{2}$$

where the operator $\circ$ denotes the Hadamard product and the ratio between the two vectors in the formula of $\bar{\mathbf{y}}^{(0)}$ is made component-wise.

Then, the generic iteration s of the procedure, with $s \geq 1$, requires that:

a) we estimate $(\widehat{\lambda}_0^{(s-1)}, \widehat{\lambda}_1^{(s-1)}, \widehat{\lambda}_2^{(s-1)})$ as in equation (8) reported in the appendix section, using the centered data $\{\mathbf{y}_1^{(s-1)}, \ldots, \mathbf{y}_T^{(s-1)}\}$;

b) we compute, for $t = 1, \ldots, T$,

$$\widehat{\mathbf{y}}_t^{(s)} = D(\widehat{\lambda}_0^{(s-1)}) \mathbf{W} \mathbf{y}_t^{(s-1)} + D(\widehat{\lambda}_1^{(s-1)}) \mathbf{y}_{t-1}^{(s-1)} + D(\widehat{\lambda}_2^{(s-1)}) \mathbf{W} \mathbf{y}_{t-1}^{(s-1)} \tag{3}$$

$$\bar{\mathbf{y}}^{(s)} = \frac{1}{T} \sum_{t=1}^{T} \left(\delta_t \circ \widetilde{\mathbf{y}}_t + (\mathbf{1} - \delta_t) \circ (\widehat{\mathbf{y}}_t^{(s)} + \bar{\mathbf{y}}^{(s-1)}) \right) \tag{4}$$

$$\mathbf{y}_t^{(s)} = \delta_t \circ (\widetilde{\mathbf{y}}_t - \bar{\mathbf{y}}^{(s)}) + (\mathbf{1} - \delta_t) \circ \widehat{\mathbf{y}}_t^{(s)}, \tag{5}$$

where **1** is a vector of ones.

c) We iterate steps a) and b) with increasing $s = 1, 2, \ldots$, until

$$\|\mathbf{y}_t^{(s)} - \mathbf{y}_t^{(s-1)}\|_2^2 \leq \gamma, \tag{6}$$

with γ sufficiently small.

At the end of the procedure, the reconstructed multivariate time series is given by $\widetilde{\mathbf{y}}_t^{(s)} = \mathbf{y}_t^{(s)} + \bar{\mathbf{y}}^{(s)}, t = 1, 2, \ldots, T$.

2 A Bootstrap construction for missing values estimation

The residual bootstrap approach can be effectively used to approximate the sampling distribution of the missing value estimators. The theoretical properties of the following residual bootstrap scheme for time series can be derived following Choi & Hall, 2000. The bootstrap algorithm can be implemented as follows.

Denote with $\mathcal{Y} = (\widetilde{\mathbf{y}}_1, \cdots, \widetilde{\mathbf{y}}_T)$ the observed time series. The bootstrap resampled time series $\mathcal{Y}^* = (\widetilde{\mathbf{y}}_1^*, \cdots, \widetilde{\mathbf{y}}_T^*)$ is built as follows.

1. Compute the residuals $\widehat{\varepsilon}_t^{(s)} = \mathbf{y}_t^{(s)} - \widehat{\mathbf{y}}_t^{(s)}$, where $\mathbf{y}_t^{(s)}$ is computed by the (5) and $\widehat{\mathbf{y}}_t^{(s)}$ is computed by the (3). The value for the index s is taken from the last iteration of the imputation procedure described in the previous section.
2. Obtain the bootstrap error series $\{\varepsilon_t^*\}$ by drawing T samples independently and uniformly, with replacement, from the centered residuals $\tilde{\varepsilon}_t^{(s)} = \widehat{\varepsilon}_t^{(s)} - \bar{\varepsilon}_t^{(s)}$.
3. Generate the bootstrap series $\widehat{\mathbf{y}}_t^*$, for $t = 1, \ldots, T$, as

$$\widehat{\mathbf{y}}_t^* = (\mathbf{I}_p - D(\widehat{\lambda}_0^{(s)})\mathbf{W})^{-1} \left[\left(D(\widehat{\lambda}_1^{(s)}) + D(\widehat{\lambda}_2^{(s)})\mathbf{W} \right) \mathbf{y}_{t-1}^{(s)} + \varepsilon_t^* \right].$$

This bootstrap construction induces a conditional probability P_*, given the sample $\mathcal{Y}$. As usual, the bootstrap distribution can be approximated by Monte Carlo simulation, by repeating the steps 1-3 for B times and by using the empirical distribution of the bootstrap replicates

$$\widetilde{\mathbf{y}}_t^{*(b)} = \widehat{\mathbf{y}}_t^{*(b)} + \bar{\mathbf{y}}^{(s)} \qquad b = 1, \ldots, B$$

Given the bootstrap distribution, a number of problems could be addressed effectively. For example, it can be used to approximate confidence intervals and confidence bands, of nominal level $1 - \alpha$ for missing value sequences. Moreover, when applying this model class to environmental pollution time series time series, such as PM_{10} and $PM_{2.5}$, the bootstrap distribution can be adequately used to estimate exceedance probability that the pollution levels exceed a specific threshold (Draghicescu & Ignaccolo, 2009), as defined by European law rules.

These lines of research are still under active developing.

3 Appendix: estimation of model parameters

The parameters of model (1) can be estimated following Dou *et al.* , 2016. In particular, given stationarity, from (1) we derive the Yule-Walker equations

$$(\mathbf{I} - D(\lambda_0)\mathbf{W})\Sigma_1 = (D(\lambda_1) + D(\lambda_2)\mathbf{W})\Sigma_0,$$

where $\mathbf{I}$ is the p-order identity matrix. The i-th row of the equation system is

$$(\mathbf{e}_i' - \lambda_{0i}\mathbf{w}_i')\Sigma_1 = (\lambda_{1i}\mathbf{e}_i' + \lambda_{2i}\mathbf{w}_i')\Sigma_0, \quad i = 1, \ldots, p, \tag{7}$$

with $\mathbf{w}_i$ the i-th row vector of $\mathbf{W}$ and $\mathbf{e}_i$ the i-th unit vector. The vector $(\lambda_{0i}, \lambda_{1i}, \lambda_{2i})'$ is estimated by the generalized Yule-Walker estimator, available in closed form,

$$(\widehat{\lambda}_{0i}, \widehat{\lambda}_{1i}, \widehat{\lambda}_{2i})' = (\widehat{\mathbf{X}}_i'\widehat{\mathbf{X}}_i)^{-1}\widehat{\mathbf{X}}_i'\widehat{\mathbf{Y}}_i, \quad i = 1, 2, \ldots, p, \tag{8}$$

where $\widehat{\mathbf{X}}_i = \left(\widehat{\Sigma}_1'\mathbf{w}_i, \widehat{\Sigma}_0\mathbf{e}_i, \widehat{\Sigma}_0\mathbf{w}_i\right)$, $\widehat{\mathbf{Y}}_i = \widehat{\Sigma}_1'\mathbf{e}_i$ and the estimated Σ_0 and Σ_1 are

$$\widehat{\Sigma}_1 = \frac{1}{T}\sum_{t=1}^{T-1}\mathbf{y}_{t+1}\mathbf{y}_t' \quad \text{and} \quad \widehat{\Sigma}_0 = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t'.$$

References

Choi, E., & Hall, P. 2000. Bootstrap confidence regions computed from autoregressions of arbitrary order. *J. R. Statist. Soc., series B*, **62**, 461–477.

Dou, B., Parrella, M.L., & Yao, Q. 2016. Generalized Yule-Walker Estimation for Spatio-Temporal Models with Unknown Diagonal Coefficients. *J. Econometrics*, **194**, 369–382.

Draghicescu, D., & Ignaccolo, R. 2009. Modeling threshold exceedance probabilities of spatially correlated time series. *Electronic Journal of Statistics*, **3**, 149–164.

Parrella, M.L., Albano, G., La Rocca, M., & Perna, C. 2018. Reconstructing missing data sequences in multivariate time series: an application to environmental data. *Statistical Methods & Applications*. https://doi.org/10.1007/s10260-018-00435-9.

ARCHETYPAL CONTOUR SHAPES

Aleix Alcacer[1], Irene Epifanio[1], M.Victoria Ibáñez[1] and Amelia Simó[1]

[1] Department of Mathematics, Jaume I University, (e-mail: aalacacer@uji.es, epifanio@uji.es, mibanez@uji.es, simo@uji.es)

ABSTRACT: Shapes are represented by contour functions from planar object outlines. Functional archetypal analysis is proposed to describe closed contour shapes. Each contour function is approximated by a convex combination of functional contour archetypes, which are a mixture of cases in the data set. Archetypes represent extreme shape patterns and improve the interpretability of highly complex distributions. The archetypal contours of feet from an anthropometric database of the adult Spanish population are extracted, which is useful for improving the fit in footwear.

KEYWORDS: shape analysis, archetype analysis, functional data analysis, footwear.

1 Introduction

Archetype Analysis (AA) (Cutler & Breiman, 1994) is an unsupervised technique that describes cases of a sample as a mixture of archetypes, which in turn, are mixtures of the cases in the sample. This multivariate technique was extended to functional data (Epifanio, 2016; Vinué & Epifanio, 2017).

Shape is all the geometrical information that remains after location, scale and rotational effects are removed from an object. Shapes can be analyzed from three approaches (Stoyan & Stoyan, 1994): objects can be treated as subsets of $\mathbb{R}^2$, they can be described by landmarks, or by using functions that represent their contours. Epifanio *et al.*, 2018 propose archetypal shapes based on landmarks. Here we propose archetypal shapes based on contour functions. In particular, we consider the natural parametrization of the contour, i.e. when the contour is parametrized by its arc length. This can be applied to any contour (other contour functions have limitations (Kindratenko, 2003)).

In Sect. 2 the methodology is introduced and it is applied on a foot shape data set in Sect. 3. The work ends with some conclusions in Sect. 4.

2 Methodology

Let $\mathbf{X}$ be an $n \times m$ matrix with n observations and m variables. AA seeks to find k archetypes, i.e a $k \times m$ matrix $\mathbf{Z}$, in such a way that $\mathbf{x}_i$ is approximated by a mixture of $\mathbf{z}_j$'s (archetypes): $\sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j$, with the mixture coefficients contained in the $n \times k$ matrix α. Additionally, $\mathbf{z}_j$'s is expressed as a mixture of the data through the mixture coefficients found in the $k \times n$ matrix β: $\mathbf{z}_j = \sum\limits_{l=1}^{n} \beta_{jl}\mathbf{x}_l$. To obtain the archetypes, AA computes two matrices α and β that minimize the following residual sum of squares (RSS): $\sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij}\mathbf{z}_j\|^2$ = $\sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{x}_l\|^2$, under the constraints 1) $\sum\limits_{j=1}^{k} \alpha_{ij} = 1$ with $\alpha_{ij} \geq 0$ for $i = 1, \ldots, n$ and 2) $\sum\limits_{l=1}^{n} \beta_{jl} = 1$ with $\beta_{jl} \geq 0$ for $j = 1, \ldots, k$.

2.1 Functional Archetype Analysis (FAA)

In the functional context, the values of the m variables in the standard multivariate context are replaced by function values with a continuous index t. Similarly, summations are replaced by integration to define the inner product. See Epifanio, 2016 for details about extension of AA to functional data.

In our problem, two functions characterize each contour, so FAA for bivariate functions must be considered. Let $f_i(t) = (x_i(t), y_i(t))$ be a bivariate function. Its squared norm is $\|f_i\|^2 = \int_a^b x_i(t)^2 dt + \int_a^b y_i(t)^2 dt$. Let $\mathbf{b}^{x_i}$ and $\mathbf{b}^{y_i}$ be the vectors of length m of the coefficients for x_i and y_i respectively for the basis functions B_h. Therefore, FAA is defined by $RSS = \sum\limits_{i=1}^{n} \|f_i - \sum\limits_{j=1}^{k} \alpha_{ij}z_j\|^2 =$ $\sum_{i=1}^{n} \|f_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} f_l\|^2 = \sum_{i=1}^{n} \|x_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} x_l\|^2 + \sum_{i=1}^{n} \|y_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl} y_l\|^2$ = $\sum_{i=1}^{n} \mathbf{a}^{x_i\prime}\mathbf{W}\mathbf{a}^{x_i} + \sum_{i=1}^{n} \mathbf{a}^{y_i\prime}\mathbf{W}\mathbf{a}^{y_i}$, where $\mathbf{a}^{x_i\prime}$ = $\mathbf{b}^{x_i\prime}$ - $\sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{b}^{x_l\prime}$ and $\mathbf{a}^{y_i\prime}$ = $\mathbf{b}^{y_i\prime} - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{jl}\mathbf{b}^{y_l\prime}$, with the corresponding constraints for α and β; and where $\mathbf{W}$ is the order m symmetric matrix with elements w_{m_1,m_2} = $\int_a^b B_{m_1} B_{m_2} dt$. In the case of an orthonormal basis, $\mathbf{W}$ is the order m identity matrix, and FAA is reduced to AA of the basis coefficients. But, in other cases, we may have to resort to numerical integration to evaluate $\mathbf{W}$, but once $\mathbf{W}$ is computed, no more numerical integrations are necessary.

3 Application

Knowledge of foot shape has a great relevance for the appropriate design of footwear. It is a main issue for manufacturing shoes, since a proper fit is a key factor in the buying decision, besides improper footwear can cause foot pain and deformity, especially in women. Therefore, the objective is to identify the shapes that represent the fitting problems of the population by means of archetypal shapes, which are extreme patterns. Then the shoe designer may adapt the design to the measurements of the extremes of a size.

Footprints have been extracted from an database of 775 3D right foot scans representing Spanish adult population. The anthropometric study was carried out by the Instituto de Biomecánica de Valencia. Data was collected in different regions across Spain using an INFOOT laser scanner. The binary images have been centered and scaled to remove the effects of translations and changes of scale as explained by Epifanio & Ventura-Campos, 2011.

In order to obtain the contour functions, the tracing begins counterclockwise in the most eastern outline point in the same row as the centroid, using *bwtraceboundary* of the image toolbox of MatLab. We normalize these functions in such a way that the perimeter length is eliminated, and the functions are defined on [0,1]. We approximate each curve by a linear combination of 51 Fourier basis (note that this basis system is periodic with period 1). All this work has been done by means of *fda* library (Ramsay & Silverman, 2005). We have therefore two pairs of functions (representing coordinates) $\{X(t),Y(t)\}$ for each foot, with $t \in [0,1]$.

3.1 Results

FAA is applied to the database. The screeplot is represented in Fig. 1, with the number of archetypes versus the respective RSS, and an elbow is found at $k = 3$. Fig. 1 also shows the contour of the 3 archetypes and the ternary plot (black circles and red triangles indicate women and men, respectively), where α values are displayed. The feet distribute more densely between archetype 2 and 3. The first archetype correspond with the solid black contour, the second one with the dashed red contour, while the third one is the dotted green contour.

4 Conclusions

AA for contour functions has been proposed. We have applied it to a novel data set of foot images. Knowing the extreme shapes can help shoe designers adjust their designs to a larger number of the population and be aware of the characteristics of the users that will not be comfortable to use them, whether to consider a line of special sizes or modify any shoe feature to cover more customers. As future work, we can extend AA to surface functions in order to analyze 3D foot shapes.

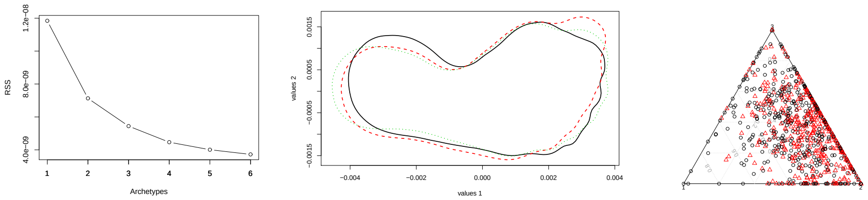


Figure 1. *Screeplot (left-handed). Archetypes (central panel) and ternary plot (right-handed). See text for details.*

5 Acknowledgments

This work is supported by the grant DPI2017-87333-R from the Spanish Ministry of Science, Innovation and Universities (AEI/FEDER, EU) and UJIB2017-13 from Universitat Jaume I.

References

CUTLER, A., & BREIMAN, L. 1994. Archetypal Analysis. *Technometrics*, **36**(4), 338–347.

EPIFANIO, I. 2016. Functional archetype and archetypoid analysis. *Computational Statistics & Data Analysis*, **104**, 24 – 34.

EPIFANIO, I., & VENTURA-CAMPOS, N. 2011. Functional data analysis in shape analysis. *Computational Statistics & Data Analysis*, **55**(9), 2758–2773.

EPIFANIO, I., IBÁÑEZ, M. V., & SIMÓ, A. 2018. Archetypal shapes based on landmarks and extension to handle missing data. *Advances in Data Analysis and Classification*, **12**(3), 705–735.

KINDRATENKO, V. V. 2003. On Using Functions to Describe the Shape. *Journal of Mathematical Imaging and Vision*, **18**, 225–245.

RAMSAY, J. O., & SILVERMAN, B. W. 2005. *Functional Data Analysis*. 2nd edn. Springer.

STOYAN, D., & STOYAN, H. 1994. *Fractals, Random Shapes and Point Fields. Methods of Geometrical Statistics*. Wiley.

VINUÉ, G., & EPIFANIO, I. 2017. Archetypoid Analysis for Sports Analytics. *Data Mining and Knowledge Discovery*, **31**(6), 1643–1677.

RANDOM PROJECTIONS OF VARIABLES AND UNITS

Laura Anderlucci[1], Roberta Falcone[1] and Angela Montanari[1]

[1] Department of Statistical Sciences - University of Bologna,
(e-mail: `laura.anderlucci@unibo.it`, `roberta.falcone3@unibo.it`, `angela.montanari@unibo.it`)

ABSTRACT: Random projections have recently emerged as a powerful tool to address the computational issues posed by high dimensional or very long data-sets. They can be applied to reduce either the number of columns or the number of rows. Besides facilitating computations they have also shown relevant statistical properties. In this work we highlight the main aspects and applications of random projections and present the use of matrix sketching in linear discriminant analysis with a focus on the issue of imbalanced classes.

KEYWORDS: random projections, sketching, supervised classification.

1 Introduction

High dimensional or very long datasets pose challenging issues for multivariate analysis. Let's consider an $n \times p$ data matrix X, related to p variables observed on n units. When the number of features is large compared to the number of units, estimates are rather unstable. When $p > n$ most of the classical multivariate methods can no longer be applied because the involved Gram matrix ($X^\top X$) cannot be inverted (as it is no longer full rank). When, on the contrary, it is the number of units to be very large, the Gram matrix can be easily inverted but its computation becomes heavily demanding.

Reduction in the number of columns (variables) or in the number of rows (units) can be dealt with in a unified framework resorting to random projections. Through random projections the columns of the data matrix are linearly combined with randomly generated weights and mapped to a d-dimensional subspace, with $d \ll p$, while approximately preserving interpoint distances. In the same way, the rows of the data matrix can be linearly combined with randomly generated coefficients, thus reducing the dataset size from n to k while approximately preserving the inner product, i.e. the Gram matrix.

The theoretical motivation for this is given by Johnson & Lindenstrauss, 1984's Lemma according to which given $\mathbf{u},\mathbf{v} \subset Q \subset \mathbb{R}^n$ and $k = \dfrac{20 \log p}{\varepsilon^2}$, where $\varepsilon \in (0, 1/2)$, there exists a Lipschitz mapping $f : \mathbb{R}^n \longrightarrow \mathbb{R}^k$ such that for all $\mathbf{u}$, $\mathbf{v} \in Q$:

$$(1-\varepsilon)\|\mathbf{u}-\mathbf{v}\|^2 \leq \|f(\mathbf{u})-f(\mathbf{v})\|^2 \leq (1+\varepsilon)\|\mathbf{u}-\mathbf{v}\|^2$$

This means that, after mapping, distances and hence scalar product are preserved up to a constant ε.
Linear combinations generated by suitably chosen random matrices have been proved to satisfy the lemma. When applied to the columns they are referred to as *random projections* (RPs); when applied to the rows the term *matrix sketching* is used instead.

Besides these common features, random projections and matrix sketching present specific characteristics which are mirrored in the algorithms based on them. In this work we highlight the main aspects and applications of random projections and present the use of matrix sketching in linear discriminant analysis with a focus on the issue of imbalanced classes.

2 Random projection based multivariate methods

All the most successful methods based on random projections entail the following steps: (i) map at random the original high-dimensional data onto a lower subspace, (ii) apply the chosen method to the dimensionally reduced data, (iii) combine the results on (selected) RPs via ensemble methods.

Analysis in the dimensionally reduced space have successfully been applied in order to perform large covariance estimation (Marzetta *et al.*, 2011), supervised (Cannings & Samworth, 2017) and unsupervised (Anderlucci *et al.*, 2019a) classification, sparse principal components (Gataric *et al.*, 2017) and multiple regression analysis (Anderlucci *et al.*, 2019b).

However useful in reducing dimensionality, random projections are highly unstable and, exactly because of their randomness, most of them can completely miss the relevant structure in the data. Moreover, they are run to obtain a final dimension d which is typically lower than the limit suggested by Johnson-Lindenstrauss' Lemma in order to preserve distances. Ensembles of the results obtained on suitably chosen random projections compensate for the instability and enlarge the dimension of the explored space, thus reaching the lemma's limit in an indirect way.

Ensembles have turned out to be very powerful, but they sometimes may be redundant and may hide the role of the original variables, which is also somehow masked by the recourse to random projections. These aspects have been addressed in the literature (see Fortunato *et al.*, 2017).

3 Matrix sketching for Multivariate Analysis

When n is too large to allow fast computations, matrix sketching has turned out to be an extremely useful and theoretically grounded device. It usually involves the following steps: (i) reduce the number of rows of the data matrix from n to k $(k \ll n)$ premultiplying it by the random Sketching Matrix $\{S\}$, (ii) apply the chosen method to the sketched data.

The similarity with the random projection strategy is striking, but one immediately notices the lack of the ensemble step. Indeed ensembles are no longer required for matrix sketching for practical and theoretical reasons. They definitely increase the computational burden without providing better results as the sketched dimension k is well within the Johnson-Lindenstrauss limit and provides a sufficiently good approximation in itself.

Sketching methods have been theoretically developed and successfully applied in the context of multiple linear regression (Ahfock *et al.*, 2017) and, recently, of linear discriminant analysis (Falcone *et al.*, 2019). In this paper we propose to adopt matrix sketching in a rather unconventional setting.

In many practical contexts, observations have to be classified into two classes of remarkably distinct size. In such cases, many established classifiers often trivially classify instances into the majority class achieving an optimal overall misclassification error rate. This leads to poor performance in classifying the minority class.

To tackle this problem, researchers often first rebalance the class sizes in the training dataset, through oversampling the minority class or undersampling the majority class, and then use the rebalanced data to train the classifiers. It is well known however that undersampling may lose some relevant information while oversampling may lead to overfitting. As previously stressed, the main feature of matrix sketching is that it preserves the scalar product. We propose to use this relevant property in order to rebalance class sizes by performing what we called *Group-wise Sketching*. It can be used in the classical sense to reduce the size of the largest class, thus keeping most of the information since all the original units are linearly combined (GwPS); but, by choosing a sketching dimension larger than the class size, it can also be used to increase the size of the small class thus preventing overfitting thanks to the randomness involved in the linear combination of the units (GwPOS).

The following empirical results reported here as an example show the good performances of the proposed method, definitely in line with the oversampling (OverS) and undersampling (UnderS) results and with a slight improvement in the ability to detect the small class.

Table 1. *Dataset* `mammography` *- Median values (over 200 reps)*

	Accuracy	Sensibility	Specificity	AUC
LDA	0.98	0.99	0.55	0.90
OverS	0.83	0.83	0.89	0.93
UnderS	0.83	0.83	0.89	0.93
GwPOS	0.83	0.83	0.90	0.93
GwPS	0.83	0.83	0.90	0.93

Empirical results The Mammography dataset (`https://www.openml.org/d/310`) has $p = 6$ and $n = 11,183$ labeled as noncalcification ($\pi_0 = 97.68\%$) and calcifications ($\pi_1 = 2.32\%$). Data have been split into training (75%) and test (25%) sets. Median values are reported in Table 1.

References

AHFOCK, D., ASTLE, W.J., & RICHARDSON, S. 2017. Statistical properties of sketching algorithms. *arXiv preprint arXiv:1706.03665.*

ANDERLUCCI, L., FORTUNATO, F., & MONTANARI, A. 2019a. High-dimensional model-based clustering via Random Projections. *In: Book of abstracts - CLADAG 2019.*

ANDERLUCCI, L., FARNÈ, M., GALIMBERTI, G., & MONTANARI, A. 2019b. Sparse linear regression via Random Projections Ensembles. *In: Book of abstracts - CLADAG 2019.*

CANNINGS, T.I., & SAMWORTH, R.J. 2017. Random-projection ensemble classification. *JRSS-B*, **79**(4), 959–1035.

FALCONE, R., ANDERLUCCI, L., & MONTANARI, A. 2019. Matrix sketching and supervised classification. *In: Book of abstracts - ERCIM 2018.*

FORTUNATO, F., ANDERLUCCI, L., & MONTANARI, A. 2017. Classifier selection and Variable Importance in RPE classificationn. *In: Book of abstracts - IFCS 2017.*

GATARIC, M., WANG, T., & SAMWORTH, R. J. 2017. Sparse principal component analysis via random projections. *arXiv:1712.05630.*

JOHNSON, WILLIAM B, & LINDENSTRAUSS, JORAM. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, **26**(189-206), 1.

MARZETTA, T.L., TUCCI, G. H., & SIMON, S.H. 2011. A random matrix-theoretic approach to handling singular covariance estimates. *IEEE Transactions on Information Theory*, **57**(9), 6256–6271.

SPARSE LINEAR REGRESSION VIA RANDOM PROJECTIONS ENSEMBLES

Laura Anderlucci[1], Matteo Farnè[1], Giuliano Galimberti[1]
and Angela Montanari[1]

[1] Department of Statistical Sciences, University of Bologna,
(e-mail: laura.anderlucci@unibo.it, matteo.farne2@unibo.it,
giuliano.galimberti@unibo.it, angela.montanari@unibo.it)

ABSTRACT: In this paper we propose a variable selection method for multiple linear regression which is based on axis-aligned random projections and accounts for partial correlation between each predictor and the response. Performances of the proposed method are evaluated on simulated data.

KEYWORDS: variable screening, sparsity, high-dimensional data.

1 Introduction

It is well known that, when dealing with high dimensional data, most of the classical multivariate methods cannot be applied or give unreliable results and it is known as well that when the number of observed variables p is large the relevant information may be contained in an s-dimensional subset of the observed variables.

In the context of multiple linear regression this means that the vector of regression coefficients for the model involving all the p variables is sparse. The ordinary approach for variable selection based on stepwise methods has turned out to produce very unstable results and new alternative solutions have recently appeared in the literature. The problem, for instance, has been addressed by either directly applying l_1 norm regularization to the original data (Tibshirani, 1996) or by screening the variables to identify the most relevant ones and then applying an l_1 penalty to the selected subset (Fan & Lv, 2008). The reasons for this two-step approach lie in the high computational load inherent in the penalized approach.

In this paper we propose a new method for variable selection in multiple linear regression which is based on random projections. The use of random projections to reduce the dimensionality of a data set is becoming increasingly popular in the multivariate statistical literature. The common trait of the most effective solutions consists in randomly combining the p columns of the data matrix X, thus mapping the data onto a random d-dimensional (with $d \ll p$) subspace on which classical analyses can be performed. The results obtained on different random projections are then summarized by ensemble methods in order to obtain the final estimates. Successful applications include supervised classification (Cannings & Samworth, 2017), large covariance estimation (Marzetta *et al.*, 2011), large-scale regression (Thanei *et al.*, 2017) and sparse principal components (Gataric *et al.*, 2017).

2 Predictor selection via Random Projections

In our proposal we exploit the special feature of axis aligned random projections, which represent a fast and analytically tractable way to perform random variable selection. Given a data matrix X we consider XA where A is a $p \times d$ axis aligned random matrix. The least squares problem is than rephrased in terms of XA as $b_A = argmin_b||y - XAb||$ and many different A matrices are considered. In particular we consider B_1 sets composed by B_2 random projections each and within each block of B_2 projections we chose the one for which the fitted regression model shows the largest R^2. As the matrix A is axis aligned only a few variables will contribute to b_A in each selected projection but combining the models fitted in all the B_1 top projections we can obtain a ranking of the p variables and after cutting the ranking at the assumed sparsity level s we identify the most relevant predictors for y.

3 Simulation Study

To study and to evaluate the performance of the proposed method, we partially reproduce the numerical study of Fan & Lv, 2008. In particular, Fan and Lv consider two main scenarios to validate their Sure Independence Screening (SIS) method: independent and correlated features.

Simulation I: 'independent features'. The first scenario considers a linear model with IID standard Gaussian predictors and Gaussian noise with standard deviation σ=1.5. Two settings with (n,p)=(200,1000) and (n,p)=(500,2000) are considered. The number s of relevant predictors is 8 and 18, and the corresponding non-zero coefficients are randomly chosen as follows. Let's set $a = 4 \cdot log(n)/n^{(1/2)}$ and $5 \cdot log(n)/n^{(1/2)}$ respectively; the non-zero coefficients are of the form $(-1)^u a|z|$ for each model, where u is drawn from a Bernoulli distribution with parameter 0.4 and z is drawn from the standard Gaussian distribution. In particular, the l_2-norms β of the two simulated models are 6.695

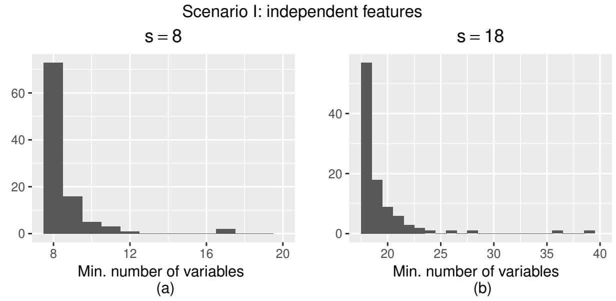


Figure 1. *Scenario 1, Distribution of the minimum number of selected variables that is required to include the true model when (a) $n = 200$ and $p = 1000$ and (b) $n = 800$ and $p = 2000$.*

and 9.582. For each model 100 data sets are simulated; the size of the projected space d is set to 10 and 500 blocks of 50 axis-aligned projections each are considered. In order to facilitate the comparison with the results of Fan and Lv, Figure 1 reports the distribution of the minimum number of variables to be selected in order to include the true model. More than the 70% of the datasets ranked the relevant variables as first. Such results clearly outperform those of SIS reported in Figure 5 (a), page 862 of Fan & Lv, 2008.

Simulation II: 'dependent features'. The scenario with dependent features considers three settings with (n,p,s) equal to (200,1000,5), (200,1000,8) and (800,2000,14), s denoting the number of non-zero coefficients. The three p-vectors β are generated in the same way as in simulation I. Let's set (σ,a)=(1, $2 \cdot log(n)/n^{(1/2)}$), (1.5, $4 \cdot log(n)/n^{(1/2)}$), (2, $4 \cdot log(n)/n^{(1/2)}$). In particular, the l_2-norms $||\beta||$ of the three simulated models are 3.618, 6.696 and 6.788. To introduce correlation between predictors, an $s \times s$ symmetric positive definite matrix C was generated with condition number about $n^{(1/2)}/log(n)$; samples of s predictors $X_1, \ldots, X_s$ are then generated from $\mathcal{N}(0,C)$. The remaining predictors are taken as $X_i = Z_i + (1-r)X_1$, $i = 2s+1, \ldots, p$, with $r = 1 - 4 \cdot log(n)/p$, $1 - 5 \cdot log(n)/p$ and $1 - 5 \cdot log(n)/p$, being $Z_{s+1}, \ldots, Z_p \sim \mathcal{N}(0, I_{p-s})$. For each model 100 data sets are simulated; the size of the projected space d is set to 10, B_1=500, B_2=50. Figure 2 includes the distribution of the minimum number of selected variables that is required to include the true model: compared with the independent case, the algorithm requires a larger model size; however, such number is still very limited, particularly if compared with that of SIS (see Figure 6 (a)-(b), page 863 of Fan & Lv, 2008).

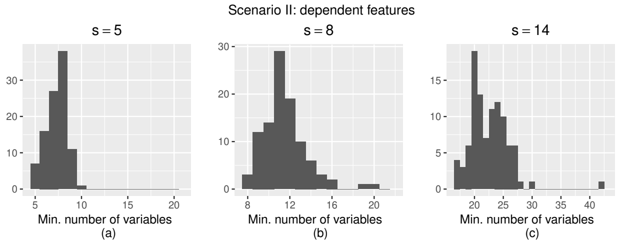


Figure 2. *Scenario 2, Distribution of the minimum number of selected variables that is required to include the true model when (a)-(b) $n = 200$ and $p = 1000$ and (c) $n = 800$ and $p = 2000$.*

4 Conclusions

This paper present a novel approach to sparse linear regression via Random Projections that accounts for partial correlation between predictors; as the simulation studies highlight, the proposed method improves upon SIS which only considers marginal correlations. The optimal choice of the tuning parameters, B_1, B_2, d, and the estimation of s are object of ongoing research.

References

CANNINGS, T.I., & SAMWORTH, R.J. 2017. Random-projection ensemble classification. *JRSS-B*, **79**(4), 959–1035.

FAN, J., & LV, J. 2008. Sure independence screening for ultrahigh dimensional feature space. *JRSS-B*, **70**(5), 849–911.

GATARIC, M., WANG, T., & SAMWORTH, R.J. 2017. Sparse principal component analysis via random projections. *arXiv preprint arXiv:1712.05630.*

MARZETTA, T.L., TUCCI, G.H., & SIMON, S.H. 2011. A random matrix-theoretic approach to handling singular covariance estimates. *IEEE Transactions on Information Theory*, **57**(9), 6256–6271.

THANEI, G., HEINZE, C., & MEINSHAUSEN, N. 2017. Random projections for large-scale regression. *Pages 51–68 of: Big and complex data analysis*. Springer.

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the lasso. *JRSS-B*, **58**(1), 267–288.

HIGH-DIMENSIONAL MODEL-BASED CLUSTERING VIA RANDOM PROJECTIONS

Laura Anderlucci[1], Francesca Fortunato[1] and Angela Montanari[1]

[1] Department of Statistical Sciences, University of Bologna,
(e-mail: laura.anderlucci@unibo.it, francesca.fortunato3@unibo.it,
angela.montanari@unibo.it)

ABSTRACT: Random projections (RPs) have shown to provide promising results in the context of high-dimensional supervised classification. In this work, we address the unsupervised classification issue by exploiting the general idea of RP ensemble. Specifically, we generate a set of low dimensional independent random projections and we perform a model-based clustering on each of them. The top B* projections, i.e. the projections which show the best grouping structure, are then retained. The final partition is obtained by aggregating the chosen classifiers via consensus. The performances of the method are assessed on a set of both real and simulated data.

KEYWORDS: high-dimensional clustering, random projections, model-based clustering.

1 Introduction

It is well known that, when dealing with high dimensional data, most of the classical multivariate methods for unsupervised learning cannot be applied or give unreliable results; in order to overcome this problem, often dimension reduction procedures are applied before carrying out any clustering.

A recent method for dimension reduction that has been gaining increasing attention is based on Random Projections (RPs) and consists in mapping at random the original high-dimensional data onto a lower subspace by using a matrix with orthogonal columns of unit length. Regardless of the original data dimension, the final solution preserves the global information almost perfectly; such a result is guaranteed by the Johnson and Lindenstrauss' Lemma (1984).

Specifically, in the context of supervised classification, Cannings and Samworth (2017) proposed a very general method for high dimensional classification, based on careful combination of the results of applying an arbitrary base classifier (like Linear Discriminant Analysis, *k*-NN, ...) to random projections of the feature vectors into a lower dimensional space. Such combination refers to the aggregation of the results of the base classifiers that yielded the smallest estimate of the test errors. Inspired by their original idea for supervised classification, we propose to extend the procedure to the context of unsupervised learning. Our idea is to generate a set of B low dimensional independent random projections and to apply a Gaussian Mixture Model (GMM) on each of them. Our Random Projection Ensemble Clustering (RPE Clu) algorithm then obtains the final partition by combining via consensus the clustering results from the top B^* projections, i.e. the projections which show the best grouping structure according to a given criterion.

In this work, we exploit the general idea of RP ensemble for high dimensional clustering. In particular, our novel proposal consists in applying a Gaussian Mixture Model (GMM) to *carefully chosen* random projections of the original data, and in using the GMM properties for both projection selection and consensus aggregation.

2 Random projection ensemble clustering

Random projections have shown to provide promising results for the analysis of high-dimensional data. The main inconvenience is that they are highly unstable; as a consequence of that, results from distinct configurations of the same data can be dramatically different: some projections indeed can induce a clear group structure in the lowered data, whilst some others can derail any hope of learning by confusing all the groups together. That is the reason why, in order to address this issue, the most successful proposals on RPs resort to ensembles.

In principle, we search for the solution that maximizes the log-likelihood of the GMM fitted on the original data, penalized by the number of free parameters. In practice, in order to avoid the drawbacks associated with the high-dimensional spaces, a feasible solution consists in considering the following variable partition

$$Y^* = [Y, \bar{Y}] = [XA|X\bar{A}],$$

where $X \in \mathbb{R}^{n \times p}$ is the original high-dimensional data matrix, $A \in \mathbb{R}^{p \times d}$ is the random projection matrix and $\bar{A} \in \mathbb{R}^{p \times (p-d)}$ is its orthogonal complement. The basic idea is to perform model-based clustering on the reduced data $Y = XA$, assuming that the underlying group structure may be well approximated by the one in the d dimensions of the block matrix Y^*. The projected solutions are then ranked according to the goodness of the partition they induce, measured by a specific transformation of the BIC, say BIC*.

The final partition is obtained through the following steps:

1. Generate B independent d-dimensional random projections A_b, $b = 1, \dots, B$, according to a specific measure, e.g. the Haar measure;
2. Compute the BIC* for the partition C_b induced by the GMM fitted on the projected data $Y = XA$;
3. Among the B possible solutions, select the B^* projections that exhibit the highest values for the BIC*: $A = [A_1; A_2; \dots; A_{B^*}]$;
4. Aggregate the cluster membership vector of the best B^* projections via consensus (Hornik, 2005).

On the basis of the numerical evidence, we suggest $B = 1000$ and $B^* = 100$ as good choices. A value for d equal to $O(10 \log G)$ works pretty well; higher values of d do dot noticeably improve the final performance.

3 Numerical Study on Gene Expression

The performances of the proposed method have been assessed on a set of both real and simulated data; in order to validate the results, we compare them with those of other clustering algorithms, such as the 'standard' Gaussian Mixture Model, the K-means algorithm, the Ward's method agglomerative hierarchical clustering, the Partition Around Medoids, the Spectral clustering and the Affinity Propagation algorithm. A further comparison is with the variable selection methodology for Gaussian model-based clustering (see Raftery & Dean, 2006). Due to space constraints, we illustrate the performance on real data only.

The lymphoma dataset (taken from the `R` package `spls`) contains the expression levels of $p = 4026$ genes for $n = 62$ patients. The study reports that 42 subjects have diffuse large B-cell lymphoma (DLBCL), 9 follicular lymphoma (FL), and 11 chronic lymphocytic leukemia (CLL).

The objective of the analysis is to group patients according to the corresponding lymphoma diagnosis, by using the information on their gene expression levels. RPE Clu procedure is performed with $B = 1000$, $B^* = 100$ and $d = \lfloor 10 \log 3 + 0.5 \rfloor + 1 = 12$; the number of groups is taken as known and set equal to 3 for all the methods. Clustering results in terms of ARI are reported in Table 1. As can be seen, the random projection ensemble clustering algorithm is capable to perfectly detect the grouping structure identified by the diagnosis. Mixture of Gaussians, K-means and hierarchical agglomerative clustering with Ward's method provide exactly the same (good) result, up to a label switching.

Table 1. *ARI for the Gene Expression Data.*

Method	ARI
RPEClu	1.00
GMM	0.95
k-means	0.95
h-ward	0.95
pam	0.84
Clust VarSel	0.49
Specc	0.95
AClust	0.85

4 Conclusions

This paper present a novel approach to cluster high-dimensional data via Random Projections; as the numerical results highlight, the proposed method improves upon the other clustering methods. Estimating the number of clusters is left for future work.

References

CANNINGS, T.I., & SAMWORTH, R.J. 2017. Random-projection ensemble classification. *JRSS-B*, **79**(4), 959–1035.

HORNIK, K. 2005. A CLUE for CLUster ensembles. *Journal of Statistical Software*, **14**(12), 1–25.

JOHNSON, W. B., & LINDENSTRAUSS, J. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, **26**(189-206), 1.

MCLACHLAN, G., & PEEL, D. 2004. *Finite mixture models*. John Wiley & Sons.

RAFTERY, A. E., & DEAN, N. 2006. Variable selection for model-based clustering. *Journal of the American Statistical Association*, **101**(473), 168–178.

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), 205–233.

MULTIVARIATE OUTLIER DETECTION IN HIGH RELIABILITY STANDARDS FIELDS USING ICS

Aurore Archimbaud[1], Klaus Nordhausen[2] and Anne Ruiz-Gazen[1]

[1] Toulouse School of Economics, University of Toulouse 1 Capitole,
(e-mail: aurore.archimbaud@tse-fr.eu, anne.ruiz-gazen@tse-fr.eu)

[2] Computational Statistics, Institute of Statistics & Mathematical Methods in Economics,
Vienna University of Technology, (e-mail: klaus.nordhausen@tuwien.ac.at)

ABSTRACT: In high reliability standards fields such as automotive or avionics, the detection of anomalies is crucial. An efficient methodology for automatically detecting multivariate outliers is detailed. It takes advantage of the remarkable properties of the Invariant Coordinate Selection method.

KEYWORDS: affine invariance, dimension reduction, unsupervised outlier identification.

1 Mahalanobis distance and PCA

Detecting outliers in multivariate data sets is of particular interest in industrial, medical and financial applications. Among the many existing method, some classical detection methods are based on the Mahalanobis distance and its robust counterpart (Rousseeuw & Van Zomeren, 1990), or on robust principal component analysis (Hubert *et al.*, 2005). One advantage of the Mahalanobis distance (MD) is its affine invariance while Principal Component Analysis (PCA) is only invariant under orthogonal transformations. For its part, PCA allows some components selection and facilitates the interpretation of the detected outliers.

2 Invariant Coordinate Selection

We propose an alternative to MD and PCA in a casewise contamination context when the number of observations is larger than the number of variables. The method we consider is the Invariant Coordinate Selection (ICS) as proposed by Tyler *et al.*, 2009. The principle of ICS is quite similar to Principal Component Analysis (PCA) with coordinates or components derived from an eigendecomposition followed by a projection of the data on selected eigenvectors.

However, ICS differs in many respects from PCA. It relies on the simultaneous spectral decomposition of two scatter matrices instead of one for PCA. While principal components are orthogonally invariant but scale dependent, the invariant components are affine invariant for affine equivariant scatter matrices. Moreover, under some elliptical mixture models, the Fisher's linear discriminant subspace coincides with a subset of invariant components in the case where group identifications are unknown (see Theorem 4 in Tyler *et al.*, 2009). This remarkable property is of interest for outlier detection since outliers can be viewed as data observations that differ from the remaining data and form separate clusters.

Compared to the MD which has some limitations in a context where the dimension of the data is large, ICS makes it possible to select relevant components which removes the limitations. Owing to the resulting dimension reduction, the method is expected to improve the power of outlier detection rules such as MD-based criteria. It also greatly simplifes outliers interpretation.

3 Practical guidelines for using ICS

We propose practical guidelines for using ICS in the context of a small proportion of outliers which is relevant in high reliability standards fields. The choice of scatter matrices together with the selection of relevant invariant components through parallel analysis and normality tests are addressed. The use of the regular covariance matrix and the so called matrix of fourth moments as the scatter pair is recommended. This choice combines the simplicity of implementation together with the possibility to derive theoretical results. Further details and results can be found in Archimbaud *et al.*, 2018.

References

ARCHIMBAUD, A., NORDHAUSEN, K., & RUIZ-GAZEN, A. 2018. ICS for multivariate outlier detection with application to quality control. *Computational Statistics & Data Analysis*, **128**, 184–199.

HUBERT, M., ROUSSEEUW, P. J., & VANDEN BRANDEN, K. 2005. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, **47**(1), 64–79.

ROUSSEEUW, P. J., & VAN ZOMEREN, B. C. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, **85**(411), 633–639.

TYLER, D. E., CRITCHLEY, F., DÜMBGEN, L., & OJA, H. 2009. Invariant co-ordinate selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(3), 549–592.

EVALUATING THE SCHOOL EFFECT: ADJUSTING FOR PRE-TEST OR USING GAIN SCORES?

Bruno Arpino[1], Silvia Bacci[1], Leonardo Grilli[1], Raffaele Guetto[1]
and Carla Rampichini[1]

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

ABSTRACT: We consider the issue of estimating the effect of schools on student achievement when a pre-test is available. Based on Invalsi data, our focus is on the causal effect of the lower secondary school type (public versus private) on test scores at the 8th grade (post-test), accounting for the student test scores at the 5th grade (pre-test). The causal effect can be estimated by either adjusting for the pre-test score (i.e. conditioning) or by using the difference between post-test and pre-test scores (gain score) as response variable. The performance of the two approaches, in terms of bias and efficiency, depends on several factors, such as pre-test reliability and validity of the common trend assumption. We compare the two approaches by an application using Invalsi data and by a simulation study.

KEYWORDS: causal effect, Invalsi achievement tests, multilevel model, random effects model.

1 Introduction

We consider the problem of estimating the school effect on student achievement, when a pre-test is available. Our work is inspired by Invalsi achievement tests implemented at the 5th grade (end of primary school) and 8th grade (end of lower secondary school) in Italy. We merge students with scores on these two grades to assess the school value added based on the progress from grade 5th to grade 8th. Specifically, we aim to evaluate if the school effect is different between public and private schools.

Two main methodological approaches have been considered in the literature to deal with the estimation of a causal effect when pre-test measures of the outcome are available (Kim & Steiner, 2019). The first approach consists in estimating the effect of the variable of interest on the post-test score, conditionally on the pre-test score (*conditioning* approach). In the second approach, the analysis is conducted on the gain score, namely the difference between the post-test and the pre-test scores (*gain score* approach).

In the causal inference literature, the conditioning approach is implemented via regression models or matching on the pre-test score, which can be regarded as methods to remove confounding, when conditioning on the pre-test score is sufficient to make the unconfoundedness assumption plausible (Arpino & Aassve, 2013). On the other hand, the gain score approach is related to difference-in-difference methods, which are devised to remove the effect of unobservable confounders under the assumption that such confounders have a time invariant effect, known as *common trend assumption*. In such a case, taking the first difference of the outcome removes confounding (e.g., Lechner, 2011).

Recently, Kim & Steiner, 2019 reconsidered the choice between the conditioning and gain score approaches. They consider a linear data generating model with constant effects across units. The treatment variable Z affects the post-test score Y, while an unobservable ability A affects both Z and Y. Thus, A is an unobserved confounder. In addition, the ability A affects the pre-test score P. If P is a reliable measure of A (i.e. the Cronbach alpha is high), conditioning on P removes most of the confounding effect of A. On the other hand, a low pre-test reliability suggest to consider the gain score approach, which is not affected by the reliability. However, the gain score approach is based on the *common trend assumption*. The authors derive formulas for the bias of the causal effects estimators under the two approaches, highlighting the assumptions required for unbiasedness. They also consider other scenarios, in particular a direct effect of the pre-test score on the treatment variable, which makes more problematic the assessment of the bias under the gain score approach.

In this contribution, we compare the two approaches, based on conditioning and gain scores, in a more complex setting with hierarchical data. Specifically, we consider students (level 1 units) nested within schools (level 2 units), where ability, pre-test and post-test scores are level 1 variables, while the treatment is a level 2 binary variable (public vs private school). Moreover, we investigate through a simulation study the performances of the estimators in terms of both bias and efficiency.

2 Case study

We aim at evaluating the effect of the Italian lower secondary schools on student achievement measured by Invalsi tests, focusing on the differences between public and private schools. To this end, we alternatively apply the conditioning and the gain score approaches, outlined in Section 1.

The data set collects information on a cohort of students that participated to the Italian language and mathematics Invalsi tests at grades 5th and 8th (i.e., the last year of the primary school and the last year of the lower secondary school, respectively). The data set has been obtained by merging data on students who attended the 5th grade in school year 2013-2014 with data on students who attended the 8th grade in school year 2016-2017. We retain data on students present in both occasions. The resulting data set consists of 436889 students who took part on both occasions: 427950 participated to both occasions of the language test, 427256 participated to both occasions of the math test. A subset of 418330 students participated to both occasions of both tests.

The students are nested in 5777 Italian schools. The average number of tested students per school is 103.91 with a standard deviation of 54.97 (min = 1; max = 334).

Each of the two achievement tests is composed of a set of items measuring the unobservable ability in language and mathematics, respectively. Items are dichotomously scored, with value 1 for a correct answer and value 0 for a wrong answer. The selection of the set of items relies on internationally validated methods based on the Rasch model (Rasch, 1960). For this reason, the ability level of a student is measured by the raw score (i.e., the total number of correct answers to the test items). As the number of items is different across subject areas (language and mathematics) and grades, we divide the raw scores by their maximum so that they are normalised in the range 0-100.

Several background variables are available both at student and school levels. Student covariates include gender, citizenship, and marks in language and mathematics resulting from the school reports. Data also include information about the parents educational level and job condition, which are exploited by Invalsi to define an index of the socio-economic status. In addition, a wide set of indicators measured at the end of the 5th grade provides information on student material deprivation, motivation and interest in learning, and relations with the class mates. School characteristics include information on the geographical location (municipality, urban area, altimetric area, and population density), the average number of students per class and the type of school (public vs private). Other school level variables are obtained averaging the student level characteristics (e.g., proportion of immigrants per school).

We specify a multilevel model (Goldstein, 2010) with students at level 1 and schools at level 2. In order to compare the conditioning and the gain score approaches, we specify two versions of the model. In the first version, the response variable is the post-test score (8th grade test), while the pre-test score enters as a covariate. In the second version, the response variable is the gain score (difference between the 8th and 5th grade tests), while the pre-test score is omitted from the covariates. Both versions of the model include the treatment variable, that is the indicator of the type of school (public vs private), as well as student and school characteristics.

3 Simulation study

The results of Kim & Steiner, 2019, described in Section 1, are based on a very simple setting that may be unrealistic in some circumstances. For example, in our application on Invalsi data (Section 2) the treatment variable is binary rather than continuous and the data have a hierarchical structure requiring random effects modelling. In such type of setting, it is not possible to obtain analytical results in closed form, thus we perform a simulation study to investigate the properties of the causal effect estimators under the conditioning and gain score approaches. The performance of the two approaches in terms of bias and efficiency of estimators is evaluated under different conditions depending on: pre-test reliability, validity of the common trend assumption, heterogeneity of the causal effects, and direct effect of the pre-test on the treatment variable. The simulation set-up mimics our case study on Invalsi data.

References

ARPINO, B., & AASSVE, A. 2013. Estimation of causal effects of fertility on economic wellbeing: Data requirements, identifying assumptions and estimation methods. *Empirical Economics*, **44**, 335–385.

GOLDSTEIN, H. 2010. *Multilevel Statistical Models, 4th ed.* Wiley.

KIM, Y., & STEINER, P. M. 2019. Gain scores revisited: a graphical models perspective. *Sociological Methods & Research, DOI: 10.1177/0049124119826155.*

LECHNER, M. 2011. The Estimation of Causal Effects by Difference-in-difference Methods. *Foundations and Trends in Econometrics*, **4**, 165–224.

MARTINI, A. (A CURA DI). 2018. *L'effetto scuola (valore aggiunto) nelle prove Invalsi 2018.* Tech. rept. Invalsi.

RASCH, G. 1960. *Probabilistic Models for some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

ACE, AVAS AND ROBUST DATA TRANSFORMATIONS

Anthony C. Atkinson[1]

[1] Department of Statistics, London School of Economics, London WC2A 2AE,
(e-mail: a.c.atkinson@lse.ac.uk)

ABSTRACT: The paper presents a series of robust parametric and non-parametric procedures for the transformation of positive and negative observations. The methods are also to be used for determining the relationship between the two transformation families.

KEYWORDS: constructed variable; diagnostic plots; extended Yeo-Johnson transformation; forward search; linked plots; robustness.

1 Introduction

The parametric family of power transformations analysed by Box & Cox, 1964 is widely used for the transformation of non-negative responses to approximate normality. Advantages of such transformation include the availability of software to analyse data from a wide range of models and the simplicity of inferences based on the normal distribution. Non-parametric alternatives use smoothing to find a transformation. Neither procedure is robust; the estimated transformation can be strongly affected by outliers and influential observations. The purposes of the work of which this is an extended abstract are:

1. To describe extensions of the Box-Cox transformation to responses which can be positive or negative.
2. To use the forward search, Atkinson *et al.*, 2010, to provide a robust method of data analysis in which outliers are detected.
3. To use a graphical display, the fan plot, to detect observations influential for the estimated transformation.
4. To investigate two non-parametric methods: ACE - Alternating Conditional Expectations, Breiman & Friedman, 1985 and AVAS - transformations for Additivity And Variance Stabilisation, Tibshirani, 1988.
5. To show the extension of the fan plot to investigating transformations of positive and negative responses and to illustrate its use in checking proposed transformations.

6. To provide a robust analysis of ACE and AVAS by comparing them with parametric transformations over values of the transformation parameter.
7. To illustrate these methods on a set of well-behaved investment fund data and on the data with appreciable contamination.

2 Extended Parametric Transformations

The normalized form of the Box-Cox transformation is

$$(z^{\lambda}-1)/(\lambda\dot{y}^{\lambda-1}) \qquad (\lambda \neq 0); \qquad \dot{y}\log y \qquad (\lambda = 0), \tag{1}$$

where $\dot{y}$ is the geometric mean of y and J, the Jacobian of the transfroamtion is given by $\log J = n(\lambda-1)\log\dot{y}$. The linear models is

$$z(\lambda) = X\beta(\lambda) + \varepsilon, \tag{2}$$

where X is $n \times p$, β is a $p \times 1$ vector of unknown parameters and the variance of ε is σ^2. For comparisons of estimates of parameters for different values of λ, many authors, starting with Box & Cox, 1964, stress the importance of working with $z(\lambda)$.

Yeo & Johnson, 2000 extended the Box-Cox transformation to observations that can be positive or negative by using different Box-Cox transformations for the two classes of response. The normalized transformation for their single parameter family is given by Atkinson *et al.*, 2020, where the Jacobian is now a more complicated function of the observations.

3 Robustness and the Fan Plot

We use a robust procedure, the Forward Search Atkinson *et al.*, 2010 to order the data by closeness to the fitted model. The procedure starts from a carefully chosen subset of $m_0 = p+1$ observations and moves forward increasing the subset size m by introducing the observation, not used in fitting, that is closest to the fitted model, until all observations have been fitted. Outliers, if any, enter at the end of the search. The outliers detection procedure is described, for multivariate data, by Riani *et al.*, 2009. The understanding of outliers is helped by brushing linked plots.

Outliers in one value of λ may not be so for some other values. We therefore need to repeat the forward search for a grid of values of λ. For each resultant ordering of the data we monitor evidence for the correctness of the transformation as m increases. We avoid repeated calculation of $\hat{\lambda}$ by use of an approximate score statistic. Taylor series expansion of the linear model (2) about the value λ_0 leads to the approximate model

$$z(\lambda_0) = x^T\beta + \gamma\, w(\lambda_0) + \varepsilon, \tag{3}$$

where the constructed variable $w(\lambda) = \partial z(\lambda)/\partial\lambda$. The approximate score statistic for testing the transformation is the t statistic for regression on $w(\lambda_0)$ in (3) in the presence of all other variables.

Atkinson *et al.*, 2020 derive constructed variables for the one-parameter Yeo-Johnson transformation. They further derive constructed variables for testing whether positive and negative observations require the same transformation. These come from the extended transformation in which one kind of response has parameter $\lambda + \alpha$ and the other λ. The test is for $\alpha = 0$.

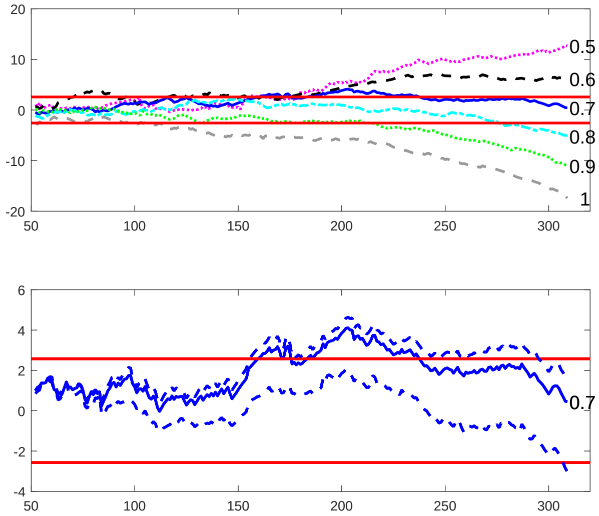


Figure 1. *Investment fund data, fan plots from Yeo-Johnson transformation. Upper panel, fan plot for single parameter distribution indicating the overall transformation* $\lambda = 0.7$*; lower panel, extended fan plot for* $\lambda = 0.7$ *suggesting different transformations for positive (upper trajectory) and negative responses*

4 Some Data Analysis

As a brief illustration of the fan plot and its extension we give a small part of the analysis of the performance of 309 investment funds, 99 of which have negative performance. The purpose is to relate the medium term performance to two indicators. The upper panel of Figure 1 suggests an overall transformation with $\lambda = 0.7$. The horizontal bands are the 99% confidence intervals for the score test. Although the value of 0.7 is acceptable at the end of the search it is rejected around $m = 200$. The lower panel of the figure indicates that different transformations are needed for positive and negative observations.

The strategy now is to try sets of pairs of values of the parameters for transformation of the positive and negative values, λ_P and λ_N. When we have found the correct transformation, the fan plot of the transformed data indicates that no further transformation is required; that is we accept the value $\lambda = 1$ in this fan plot. The analysis of extended fan plots with this strategy led to the values $\lambda_P = 1$ and $\lambda_N = 0$, which is not the log transformation for negative variables. With this transformation all three trajectories in the extended fan plot are close together, lying within the bounds throughout the search. This is also the strategy we apply to evaluating the transformations from ACE and AVAS.

References

ATKINSON, A. C., RIANI, M., & CERIOLI, A. 2010. The Forward Search: theory and data analysis (with discussion). *Journal of the Korean Statistical Society*, **39**, 117–134. doi:10.1016/j.jkss.2010.02.007.

ATKINSON, A. C., RIANI, M., & CORBELLINI, A. 2020. The transformation of profit and loss data. (Submitted).

BOX, G. E. P., & COX, D. R. 1964. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, **26**, 211–246.

BREIMAN, L., & FRIEDMAN, J. H. 1985. Estimating optimal transformations for multiple regression and transformation (with discussion). *Journal of the American Statistical Association*, **80**, 580–619.

RIANI, M., ATKINSON, A. C., & CERIOLI, A. 2009. Finding an Unknown Number of Multivariate Outliers. *Journal of the Royal Statistical Society, Series B*, **71**, 447–466.

TIBSHIRANI, R. 1988. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, **83**, 394–405.

YEO, I.-K., & JOHNSON, R. A. 2000. A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959.

MIXTURES OF MULTIVARIATE LEPTOKURTIC NORMAL DISTRIBUTIONS

Luca Bagnato[1], Antonio Punzo[2] and Maria Grazia Zoia[3]

[1] Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore,
(e-mail: luca.bagnato@unicatt.it)

[2] Dipartimento di Economia e Impresa, Università di Catania,
(e-mail: antonio.punzo@unict.it)

[3] Dipartimento di Politica economica, Università Cattolica del Sacro Cuore,
(e-mail: maria.zoia@unicatt.it)

ABSTRACT: We introduce mixtures of multivariate leptokurtic normal (LN) distributions as a tool for robust model-based clustering in the presence of mild outliers. Compared to the normal distribution, the LN has an additional parameter and, advantageously with respect to the existing elliptical heavy-tailed distributions, the additional parameter directly corresponds to the quantity of interest, namely, the excess kurtosis. We outline an EM algorithm for maximum likelihood estimation of the parameters of the mixture. As an illustration, we analyze the well-known Old Faithful geyser data.

KEYWORDS: Leptokurtik normal distribution, mixture models, EM algorithm.

1 The model

A d-variate random vector $\boldsymbol{X}$ follows a leptokurtic normal distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and excess kurtosis β, in symbols $\boldsymbol{X} \sim \mathcal{LN}_d(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$, if its density is given by

$$f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\beta) = q(t;\beta)\,\phi(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}), \qquad \boldsymbol{x} \in \mathbb{R}^d, \tag{1}$$

where $\phi(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the density of a d-variate normal random vector with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and $q(t;\beta)$ is defined as follows

$$q(t;\beta) = 1 + \frac{\beta}{8d(d+2)}\left[t^2 - 2(d+2)t + d(d+2)\right], \quad t = (\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}). \tag{2}$$

The kurtosis of $\boldsymbol{X} \sim \mathcal{LN}_d(\boldsymbol{\mu},\boldsymbol{\Sigma},\beta)$ is $d(d+2)+\beta$. So, β directly represents the excess kurtosis. Such a parameter must satisfy the constraint $\beta \in [0, \min(4d, 4d(d+2)/5)]$, which is the intersection of two constraints:

i) $\beta \in [0, 4d]$, which assures that $f_{\boldsymbol{X}}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta)$ is a positive elliptical density;
ii) $\beta \in [0, 4d(d+2)/5]$, which guarantees that $f_{\boldsymbol{X}}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta)$ is unimodal.

For a d-variate random vector $\boldsymbol{X}$, a finite mixture of MLN distributions can be written as

$$p(\boldsymbol{x}; \boldsymbol{\vartheta}) = \sum_{j=1}^{k} \pi_j f\left(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \beta_j\right), \tag{3}$$

where π_j is the mixing proportion of the jth component, with $\pi_j > 0$ and $\sum_{j=1}^{k} \pi_j = 1$, f is defined as in (1), and $\boldsymbol{\vartheta}$ contains all the parameters of the mixture. As a special case, when $\beta_j = 0$ for each $j = 1, \ldots, k$, we obtain classical mixtures of multivariate normal distributions.

2 An EM algorithm for maximum likelihood estimation

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be a random sample from model (3). To find maximum likelihood (ML) estimates for the parameters of our model, we adopt the classical expectation-maximization (EM) algorithm. We need to introduce an indicator vector $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ik})'$, where $z_{i1} = 1$ if $\boldsymbol{x}_i$ comes from component j and $z_{ij} = 0$ otherwise. The values of z_{ij} are used for the definition of the following complete-data log-likelihood

$$l_c(\boldsymbol{\vartheta}) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \ln(\pi_j) + \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \ln\left[f\left(\boldsymbol{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \beta_j\right)\right], \tag{4}$$

which is the core of the EM algorithm. The EM algorithm iterates between two steps, one E-step and one M-step, until convergence.

The E-step on the $(q+1)$th iteration requires the calculation of

$$E_{\boldsymbol{\vartheta}^{(q)}}\left[Z_{ij} | \boldsymbol{x}_i\right] = z_{ij}^{(q)} = \pi_j^{(q)} f\left(\boldsymbol{x}_i; \boldsymbol{\mu}_j^{(q)}, \boldsymbol{\Sigma}_j^{(q)}, \beta_j^{(q)}\right) \Big/ p\left(\boldsymbol{x}_i; \boldsymbol{\vartheta}^{(q)}\right). \tag{5}$$

Then, by substituting z_{ij} with $z_{ij}^{(q)}$ in (4), we obtain the conditional expectation of the complete-data log-likelihood, say $Q(\boldsymbol{\vartheta}) = Q_1(\boldsymbol{\pi}) + Q_2(\boldsymbol{\psi})$, where the two terms on the right-hand side are ordered as the two terms on the right-hand side of (4), being $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ and $\boldsymbol{\psi} = \boldsymbol{\vartheta} \setminus \boldsymbol{\pi}$.

The M-step on the same iteration requires the calculation of $\boldsymbol{\vartheta}^{(q+1)}$ as the value of $\boldsymbol{\vartheta}$ that maximizes $Q(\boldsymbol{\vartheta})$. As $Q_1(\boldsymbol{\pi})$ and $Q_2(\boldsymbol{\psi})$ have zero cross-derivatives, they can be maximized separately. Maximizing $Q_1(\boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$, subject to the constraints on those parameters, yields

$$\pi_j^{(q+1)} = \sum_{i=1}^{n} z_{ij}^{(q)} \Big/ n, \quad j = 1, \ldots, k. \tag{6}$$

Maximizing $Q_2(\boldsymbol{\psi})$ with respect to $\boldsymbol{\psi}$ is equivalent to independently maximize each of the k weighted log-likelihood functions

$$Q_{2j}\left(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j,\beta_j\right)=\sum_{i=1}^{n} z_{ij}^{(q)}\ln\left[f\left(\boldsymbol{x}_i;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j,\beta_j\right)\right], \tag{7}$$

with respect to $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$, and β_j, $j=1,\ldots,k$. Details about the maximization of Q_{2j} can be found in Bagnato *et al.* (2017).

3 Application: Old Faithful Geyser

We analyze the `geyser2` data set accompanying the **tclust** package for R, a bivariate ($d=2$) data set containing the eruption lengths and the corresponding previous eruption lengths for $n=271$ eruptions of the Old Faithful Geyser.

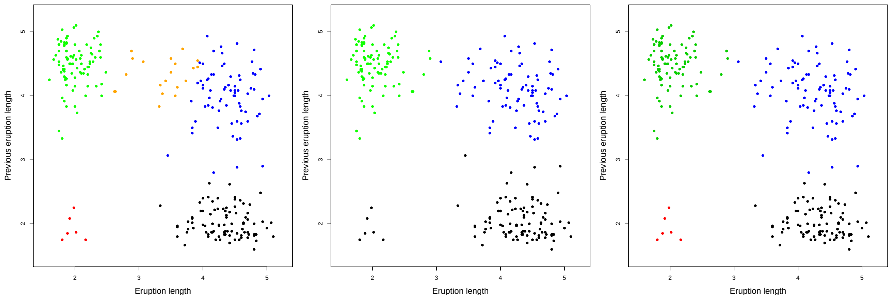
We provide a comparison with (unconstrained) finite mixtures of some well-established multivariate elliptically contoured distributions. In particular, for $k=1,\ldots,6$, we estimate: 1) mixtures of multivariate normal distributions (MNMs), 2) mixtures of multivariate t distributions (M*t*Ms; Peel & McLachlan, 2000), 3) mixtures of multivariate contaminated normal distributions (MCNMs; Punzo & McNicholas, 2016), 4) mixtures of multivariate power exponential distributions (MPEMs; Zhang & Liang, 2010), and 5) mixtures of multivariate leptokurtic normal distributions (MLNMs).

Table 1 compares the best BIC value, and the associated value of k, for each of the competing models. The best model is the MNM with $k=5$ components, while the worst is the MPEM with $k=3$ components. However, the clustering provided by the former model (see Figure 1(a)) is not as expected: the orange group seems to be composed by two well-separated subgroups, being one of them very overlapped with the blue group. Motivated by these results, we look for a different model. The second best MNM, having BIC $=-1128.001$, has $k=3$ components; compared with the BIC values in Table 1, this MNM is no more the best one. So, the overall second best model is the MLNM with $k=4$ components (see Figure 1(c) for the obtained clustering). For the selected MLNM, the estimates of the excess kurtosis for the four components are $\widehat{\beta}_1=$

Table 1. *Best BIC values, and associated value of k, for the fitted mixtures.*

	MNM	M*t*M	MCN	MPEM	MLNM
k	5	4	3	3	4
BIC	-1113.080	-1118.659	-1139.531	-1145.911	-1115.995

$9.768 \cdot 10^{-8}$ (refer to the black bullets in Figure 1(c)), $\widehat{\beta}_2 = 1.282 \cdot 10^{-6}$ (red bullets), $\widehat{\beta}_3 = 2.339$ (green bullets), and $\widehat{\beta}_4 = 0.972$ (blue bullets); therefore, it seems that two of the obtained clusters need heavier tails than the normal ones. For completeness, Figure 1(b) displays the clustering results obtained for the MLNM with $k = 3$ components (BIC $= -1119.837$). As we can note by the green bullets in Figure 1(b), the small cluster on the left-down corner is captured by the tail of the MLN distribution located on the cluster on the right-down corner; this is confirmed by the estimated excess kurtosis of such a component which is almost 6.4, which is the maximum excess kurtosis the MLN distribution can reach in the bivariate case.



(a) MNM: $k = 3$ (b) MLNM: $k = 3$ (c) MLNM: $k = 4$

Figure 1. *Clustering results for some MNM and MLNM models.*

References

Bagnato, L., Punzo, A., & Zoia, M. G. 2017. The multivariate leptokurtic-normal distribution and its application in model-based clustering. *Canadian Journal of Statistics*, **45**(1), 95–119.

Peel, D., & McLachlan, G. J. 2000. Robust mixture modelling using the *t* distribution. *Statistics and Computing*, **10**(4), 339–348.

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

Zhang, J., & Liang, F. 2010. Robust clustering using exponential power mixtures. *Biometrics*, **66**(4), 1078–1086.

DETECTING AND INTERPRETING THE CONSENSUS RANKING BASED ON THE WEIGHTED KEMENY DISTANCE

Alessio Baldassarre[1], Claudio Conversano[1] and Antonio D'Ambrosio[2]

[1] Department of Business and Economics, University of Cagliari, (e-mail: al.baldassarre1@gmail.com, conversa@unica.it)

[2] Department of Economics and Statistics, University of Naples Federico II, (e-mail: antdambr@unina.it)

ABSTRACT: This paper outlines a way for finding the consensus ranking minimizing the sum of the weighted Kemeny distance, using positional weights. The weighted Kemeny distance, introduced by García-Lapresta and Pérez-Román, meets the original Kemeny-Snell axioms and it is fully applicable in treating weak orderings. A differential evolution algorithm is ad-hoc defined in order to detect the consensus ranking, namely that ranking that best represents the preferences expressed by a set of individuals.

KEYWORDS: preference rankings, genetic algorithms, consensus ranking, weighted distance.

1 Introduction

Preference data are analyzed in several fields, such as political and social sciences, behavioral sciences, economics, and computer science. They are generally expressed through either ordering (when a person places in order a set of items according to his/her preference), or rank vectors (when an individual assigns a rank to each item). Even though their meaning is different, these terms can be used interchangeably (Marden, 1996).

In the specific, m judges could express their preference on n items by assigning values from 1 to n, where 1 represents the most preferred item and n the object in the last position. If the whole item set is judged and a judge assigns a different rank to the items, a full ranking is furnished. When a judge assigns the same value to two or more items, the resulting ordering is called tied (or weak) ranking. Lastly, if judges express their preference for an items subset only, the ordering is called partial ranking (D'Ambrosio *et al.* , 2017).

Often, the goal is to find the ranking that best represents the preferences stated by the individuals. This goal is known as consensus ranking problem, or Kemeny problem, or rank aggregation problem. When there is a large number of objects to be ranked, the solution of the stated problem can be really complex, falling in fact into the category of Non-deterministic Polynomial-time (NP) hard problems (Bartholdi III *et al.* , 1989). The solution is indeed carried out in a space of dimensions equal at least to the number of all possible permutations of items. Note that if there are tied rankings, the searching space is larger and larger than the space of permutations (D'Ambrosio *et al.* , 2019).

In order to carry out the search for the ranking that is most in agreement with the others, it is possible to follow two aggregation approaches: the ad hoc methods (de Condorcet, 1785) and distance-based methods.

This paper focuses on the distance-based approach to find the consensus ranking that is the ranking that minimizes the sum of a given distance between itself and all the orderings in a data matrix. Several distance measures for rankings have been defined. The most used measures are based on the geometric representation of the permutation set, that is called permutation polytope. It is a convex hull of a finite set of points in $\mathbb{R}^n$, whose coordinates are the permutations of n distinct numbers (Thompson, 1993). The permutation polytope is an $(n-1)$ dimensional object and for such reason it can only be represented for $n = 3$ or with $n = 4$ (D'Ambrosio *et al.* , 2015). The distance between two vertices corresponds to the minimum number of transpositions of adjacent objects needed to transform one ranking into another (Heiser, 2004).

Probably, the most known distance for rankings is the Kendall distance (Kendall, 1938), which is the natural measure defined on the permutation polytope. For two rankings, it is equal to the total number of steps to migrate from the first to the second ordering by reversing adjacent pairs of objects (Heiser, 2004). If ties are allowed, it is better to use the Kemeny distance (Kemeny & Snell, 1962), which counts the number of interchanges of couples of elements that are required to transform one (partial) ranking into another (Emond & Mason, 2002). Kemeny and Snell defined the median ranking as that ranking $\hat{S}$ that minimizes the sum of the distances between itself and all the other m rankings R:

$$\hat{S} = \underset{S \in Z^n}{argmin} \sum_{i=1}^{m} d(S, R_i), \tag{1}$$

where Z^n represents the universe of rankings with n items.

2 Weighted Kemeny distance and Differential Evolution algorithm

García-Lapresta and Pérez-Román (2010) introduced the possibility of weighting the discrepancies between weak orderings. They demonstrated that the Kemeny distance doesn't consider whether the judges' choices diverge in relation to the objects classified in the first positions rather than in the last ones.

Let A and B be two rankings of n items. Let w be a set of $(n-1)$ positional weights, which equal for both rankings with the restriction that $w_1 \geq w_2 \geq \ldots \geq w_{n-1}$ and $\sum_i w_i = 1$. Let a_{ij} and b_{ij} be the elements of the score matrices (Kemeny & Snell, 1962) associated with the rankings A and B. García-Lapresta & Pérez-Román, 2010 defined the weighted Kemeny distance that can be formulated as follows:

$$d^{w}(A,B) = \frac{1}{2}\left(\sum_{i<j=1}^{n} w_i \left|a_{ij}^{(A)} - b_{ij}^{(A)}\right| + \sum_{i<j=1}^{n} w_i \left|a_{ij}^{(B)} - b_{ij}^{(B)}\right|\right), \tag{2}$$

where the subscripts (A) and (B) mean that the orderings B and A are ordered with respect to rankings A and B, respectively.

Recently, D'Ambrosio *et al.* , 2017, proposed a Differential Evolution algorithm aimed at the detection of the consensus ranking for complex problems (i.e., with $n > 200$) called DECoR. Here, we modify the DECoR algorithm so that we can find the solution by minimizing the weighted Kemeny distance. The behavior of the algorithm is checked through both a simulation study and applications to well-known data set.

The goal of the experimental evaluation is understanding the role of the positional weights in detecting the consensus ranking and, at the same time, providing a flexible tool for complex problems in which the (weighted) consensus ranking detection is the starting point for other kinds of analysis, as in recursive partitioning methods (Plaia & Sciandra, 2017).

References

BARTHOLDI III, J., TOVEY, C.A., & TRICK, M.A. 1989. Voting schemes for which it can be difficult to tell who won the election. *Social Choice and welfare*, **6**(2), 157–165.

D'AMBROSIO, A., AMODIO, S., & IORIO, C. 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.

D'AMBROSIO, A., MAZZEO, G., IORIO, C., & SICILIANO, R. 2017. A differential evolution algorithm for finding the median ranking under the Kemeny axiomatic approach. *Computers & Operations Research*, **82**, 126–138.

D'AMBROSIO, A., IORIO, C., STAIANO, M., & SICILIANO, R. 2019. Median constrained bucket order rank aggregation. *Computational Statistics*. `https://doi.org/10.1007/s00180-018-0858-z`.

DE CONDORCET, M. 1785. *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix (Essay on the Application of Analysis to the Probability of Majority Decisions)*.

EMOND, E.J., & MASON, D.W. 2002. A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, **11**(1), 17–28.

GARCÍA-LAPRESTA, J. L., & PÉREZ-ROMÁN, D. 2010. Consensus measures generated by weighted Kemeny distances on weak orders. *Pages 463–468 of: 2010 10th International Conference on Intelligent Systems Design and Applications*. IEEE.

HEISER, W.J. 2004. Geometric representation of association between categories. *Psychometrika*, **69**(4), 513–545.

KEMENY, J.G., & SNELL, J.L. 1962. *Mathematical models in the social sciences*. Vol. 9. Ginn Boston.

KENDALL, M.G. 1938. A new measure of rank correlation. *Biometrika*, 81–93.

MARDEN, J.I. 1996. *Analyzing and modeling rank data*. CRC Press.

PLAIA, A., & SCIANDRA, M. 2017. Weighted distance-based trees for ranking data. *Advances in Data Analysis and Classification*. `https://doi.org/10.1007/s11634-017-0306-x`.

THOMPSON, G. 1993. *Generalized permutation polytope and exploratory graphical methods for ranked data*.

PREDICTIVE PRINCIPAL COMPONENT ANALYSIS

Simona Balzano[1], Maja Bozic[1], Laura Marcis[1] and Renato Salvatore[1]

[1] Department of Economics and Law, University of Cassino,
(e-mail: s.balzano@unicas.it, m.bozic@unicas.it,
laura.marcis@gmail.com, rsalvatore@unicas.it)

ABSTRACT: This work introduces a multi-group Principal Component Analysis, in analogy with the linear predictor as in the general linear mixed model approach.

Estimating PCs simultaneously in different groups provides a joint dimension reduction solution (Flury, 1988, Härdle and Simar, 2015), representing the so-called Common Principal Components (CPC). The literature proposes two types of CPC - one for independent groups (Flury, 1984), and the other for dependent groups (Neuenschwander and Flury, 2000).

The CPC basic assumption is that the space spanned by the eigenvectors, that leads to a joint eigenstructure across the structure, is identical across groups, but in practice variances associated with the components are allowed to vary. Some recent approaches address this issue incorporating the analysis of the differences among groups in the Structural Equation Modeling (SEM) framework (Bechger et al., 2014). Gu and Wu ()2016) propose to exploit a state-space model analysis (Dolan et al, 1999).

We present a model-based solution to some of the issues of the multi-group PCA. We refer to this approach as *Predictive* PC (PPC) as the PC loadings and scores are based on the results of a Singular Value Decomposition of the matrices of a linear model predicted values. The empirical predictor is given by an extension of the distribution-free variance least squares method to an iterative multivariate response algorithm.

KEYWORDS: Principal components, linear mixed model, empirical best linear unbiased predictor, variance least squares.

References

BECHGER, T.M., BLANCA M.J. & MARIS G. 2014. The analysis of multivariate group differences using common principal components. *Structural Equation Modeling*, **21**, 577–587.

DOLAN, C.V. 1996. Principal component analysis using LISREL 8. *Structural Equation Modeling*, **2**, 307–322.

FLURY, B. 1988. *Common principal components related multivariate models*. John Wiley & Sons, Inc.

FLURY, B.N. 1984. Common principal components in k groups. *Journal of the American Statistical Association*, **79**, 892–898.

GU, F. & WU, H. 2016. Raw Data Maximum Likelihood Estimation for Common Principal Component Models: A State Space Approach. *Psychometrika*, **81**, 751–773.

HÄRDLE, W.K. & SIMAR L. 2015. Principal Components Analysis. In: *Applied Multivariate Statistical Analysis*. Springer Berlin Heidelberg.

NEUENSCHWANDER, B.E. & FLURY B.D. 2000. Common Principal Components for Dependent Random Vectors. *Journal of Multivariate Analysis*, **75**, 163–183.

FLEXIBLE MODEL-BASED TREES FOR COUNT DATA

Federico Banchelli[1, 2]

[1] Department of Medical and Surgical Sciences, University of Modena and Reggio Emilia,
(e-mail: federico.banchelli@unimore.it)

[2] DG Personal Care, Health and Welfare, Emilia-Romagna Region,
(e-mail: federico.banchelli@regione.emilia-romagna.it)

ABSTRACT: One of the major developments in the last two decades in the field of recursive partitioning was the use of "hybrid" tree models. These methods have the structure of a traditional non-parametric decision tree, but a parametric regression model is fitted within each node of the tree, instead of a constant value. The aim of the present work is to illustrate the use of a new class of flexible model-based trees for count data, where the novelty is to allow different regression models to be fitted in different nodes. In each node, partitioning is performed using the model which better locally fits, instead than the one which better globally fits. A performance-complexity assessment of this method on simulated data is reported, comparing the proposed flexible model-based tree for count data to the standard model-based tree that uses the same model in each node.

KEYWORDS: model-based trees, count regression, regression trees, model selection.

1 Rationale and aim of the research

Since their introduction in the mid 60's, but mostly following their diffusion in the mid 80's, recursive partitioning methods are widely used in applied research. A major advantage is the interpretability of their decision tree structure, which gives the potential to easily communicate the result of the statistical analysis.

One of the major developments in the last two decades in this field was the use of "hybrid" tree models (Loh, 2014). These methods have the structure of a traditional non-parametric decision tree, but a parametric regression model is fitted within each node of the tree, instead of a constant value (e.g. the arithmetic mean). This gave birth to a new class of recursive partitioning methods - which in this work are named model-based trees – aimed at finding subgroups that have different values of a model's parameters. Indubitably, this gives the possibility of a great customization of the splitting criteria, since researchers can specify a domain-specific regression model and perform recursive partitioning with respect to one or more of this model's parameters. The key point is setting a model's equation which leads to the estimation of parameters that have a straightforward and relevant interpretation, with respect to the scope of the analysis. Some recent proposal of this kind were made, among others, in the field of medical and health statistics. These were focused on the identification of subgroups of patients that have differential treatment effects in randomized and observational clinical trials (Seibold et al., 2016) (Loh et al., 2015) and also in individual-level meta-analysis of clinical trials (Fokkema et al., 2018).

Despite the fervent research activity in this field, the current available model-based trees are mainly based on fitting the same model in all nodes of the tree. In particular, the same distribution for the response variable is assumed, and the model's independent variables are often kept fixed. However, since to find different data patterns is a goal itself in recursive partitioning, the use of a fixed model in all nodes has not to be considered as the only possibility. Instead, it could be possible that different model, e.g. models that suppose different distribution for the response variable, can fit better in different subsets of data, during the recursive partitioning procedure. Hence the need for the study and assessment of the possibility of selecting different models in different nodes of a model-based tree.

Therefore, based on these considerations, the aim of the present work is to illustrate the use of a new class of flexible model-based trees, where the novelty is to allow different regression models to be fitted in different nodes. In particular, the work is focused on the study of these flexible model-based trees for a count response variable, as this latter can be described by several alternative statistical distributions.

2 Flexible model-based trees for count data

In the model-based tree literature, one of the most promising methods for exploratory analysis of subgroups which differ for the values of a model's parameters is the Model-Based Recursive Partitioning (MOB), which was exhaustively described in (Zeileis et al., 2008).

This general, unbiased and broadly applicable recursive partitioning method is based on a class of parameter instability tests - M-fluctuation tests - for detecting different values of a model's parameters. Given the flexibility and adaptability of this method, several researchers have proposed the use of particular classes of models for use within the MOB algorithm. Some examples are: the generalized linear model-based tree described in (Rusch & Zeileis, 2013); the generalized linear mixed model-based tree in (Fokkema et al., 2018); the beta model-based tree in (Grun et al., 2012); the Rasch model-based tree in (Strobl et al., 2015).

All of these implementations share a common principle, which is to use the same model across all nodes. Define $\mathbb{M}(Y, X, \theta)$ as the chosen regression model to be fitted in the nodes, where Y is a count dependent variable that follows a fixed parametric distribution, X is a vector of independent variables and θ is a vector of regression parameters. The traditional model-based tree seek for a partition of the covariate space, where each subgroup has an associated model and a node-specific vector of parameters. The result of such a model-based tree T can be seen as a segmented (or piecewise) model:

$$\mathbb{M}^{\tau}(Y, X, \theta^{\tau}), \quad \tau = 1, \dots , \left|\bar{\tilde{T}}\right|$$

where $\mathbb{M}^{\tau}(Y, X, \theta^{\tau})$ is the parametric regression model fitted in generic node τ, θ^{τ} is a vector of node-specific regression parameters in node τ, $\tilde{\tilde{T}}$ is the subset of terminal nodes of T and $|\tilde{\tilde{T}}|$ is the cardinality of the tree (the number of its terminal nodes).

Regarding the general structure of the newly proposed flexible model-based tree, this is the same of MOB, preceded by a preliminary model choice step in each node. It then consists in five steps, which are iteratively performed: 1) to fit $D > 1$ different regression models to all observations in the current node τ and to select the best fitting one among them; 2) to fit the selected regression model to all observations in the current node, in order to estimate θ^{τ}; 3) to assess whether the parameter estimates θ^{τ} show parameter instability; 4) to detect the partitioning variable which is associated to the maximum parameter instability; 5) to find the best binary split. The five steps are repeated recursively until stopping or pre-pruning criteria occur. The first step is the novelty proposed in this work, whereas those that follow are the original steps of MOB. The results of such a procedure is not a segmented model in its previously described form, because of the different underlying distributions. However, provided that the different models are expressed in terms of parameters with the same interpretation, the result of the flexible model-base tree is a segmented model of the form:

$$\mathbb{M}^{\tau, d_{\tau}}(Y, X, \theta^{\tau, d_{\tau}}), \quad \tau = 1, \dots, |\tilde{\tilde{T}}|$$

where $\mathbb{M}^{\tau, d_{\tau}}(Y, X, \theta^{\tau, d_{\tau}})$ is the best-fitting parametric regression model d_{τ} in the generic node τ and $\theta^{\tau, d_{\tau}}$ is a vector of node-specific regression parameters estimated from the best-fitting model d_{τ} in node τ.

In the present work, the attention was focused on a flexible model-based tree for count data. Four models $(D = 4)$ of common use in count data analysis are considered in each node: Poisson, Negative Binomial (NB), Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) models. Each of these models is nested within another one - with ZINB being the more general case. As a consequence, the splitting criterion always looks for differential values of θ, whatever model is used to estimate it. The partitioning criteria are therefore coherent across all nodes, even if the models are different. Recursive partitioning via this method is still based on the search for differential model's parameters, like the standard model-based trees; rather, these parameters are estimated in each node according to the model which better locally fits, instead than from the one which better globally fits.

3 Planning of an experimental design for performance-complexity comparison of model-based trees on simulated data

The proposed flexible model-based trees for count data are compared to the standard model-based trees that use a fixed count data model in each node. This is a performance-complexity comparison, where the statistical performance is assessed as a function of the complexity of the tree (the number of terminal nodes). In order to do that, a sequence of nested subtrees $T_0 \supset T_1 \supset \cdots \supset T_m$, $|T_0| > |T_1| > \cdots > |T_m|$, $m + 1 \leq |\tilde{T}|$ is identified via post-pruning of the model based-trees, following a bottom-up procedure.

The comparison is carried out by assessing performance-complexity curves on simulated datasets in the field of medical and health statistics.

References

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. 2018. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*., **50**(5), 2016-2034.

Grun, B., Kosmidis, I., & Zeileis, A. 2012. Extended Beta Regression in R: Shaken, Stirred, Mixed, and Partitioned. *Journal of Statistical Software*., **50**(5), 2016-2034.

Hothorn T., & Zeileis, A. 2015. Partykit: a modular toolkit for recursive partitioning in R. *Journal of Machine Learning Research*., **16**, 3905-3909.

Loh, W.-J., He, X., & Man, M. 2015. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*., **34**, 1818-1833.

Loh, W.-J. 2014. Fifty years of classification and regression trees (with discussion). *International Statistical Review*., **34**, 329-370.

Rusch, T, & Zeileis, A. 2013. Gaining insights with recursive partitioning of generalized linear models. *Journal of Statistical Computation and Simulation*., **83**(7), 1301-1315.

Seibold, H., Zeileis, A., & Hothorn, T. 2016. Model-based Recursive Partitioning for subgroup analyses. *The International Journal of Biostatistics*, **12**(1), 45-63.

Strobl, C., Kopf, J., & Zeileis, A. 2015. A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, **80**(2), 289-316.

Zeileis, A., & Hornik K. 2007. Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*., **61**(4), 488-508.

Zeileis, A., Hothorn T., & Hornik K. 2008. Model-based Recursive Partitioning. *Journal of Computational and Graphical Statistics*., **17**(2), 492-514.

EUCLIDEAN DISTANCE AS A MEASURE OF CONFORMITY TO BENFORD'S LAW IN DIGITAL ANALYSIS FOR FRAUD DETECTION

Mateusz Baryła[1] and Józef Pociecha[1]

[1] Department of Statistics, Cracow University of Economics
(e-mail: mateusz.baryla@uek.krakow.pl, jozef.pociecha@uek.krakow.pl)

ABSTRACT: The main goal of the article is to discuss methods based on the Euclidean distance which can be used during the identification of financial frauds. The methods enable assessing data conformity to Benford's Law, within the primary tests of this law. After discussing techniques based on the Euclidean distance, an example of accounting fraud detection is presented.

KEYWORDS: Euclidean distance, fraud detection, primary Benford's Law tests.

1 Introduction

Benford's Law deals with the probability of the occurrence of significant digits in numbers. The law was discovered by S. Newcomb (1881) who noticed that there are more numbers that start from lower digits than from the higher ones. Exactly the same observation was also made by F. Benford (1938).

Let D_1 be the first significant digit of a number. The probability that a number has the first significant digit d_1 is calculated in the following way:

$$P(D_1 = d_1) = \log_{10}(1 + d_1^{-1}), \tag{1}$$

where $d_1 \in \{1, 2, ..., 9\}$.

Similarly, let D_1D_2 denote the first two significant digits of a number. Then, the probability that the first two significant digits of a number equal d_1d_2 is calculated as follows:

$$P(D_1D_2 = d_1d_2) = \log_{10}(1 + (d_1d_2)^{-1}), \tag{2}$$

where $d_1d_2 \in \{10, 11, ..., 99\}$.

One of the applications of Benford's Law is to use it as a tool in fraud detection procedures. M. Nigrini (2012) classified Benford's Law tests into three main categories: the primary tests, the advanced tests, and the associated tests. The idea of the primary tests is to verify whether an empirical distribution of digits in numbers conforms to Benford's Law or not. Some of the methods employed in this kind of evaluation are very popular (e.g. the Kolmogorov-Smirnov test, the chi-square goodness of fit test), while other techniques (e.g. tests based on distance measures), are less known.

The main objective of the article is to present methods based on the Euclidean distance which can be applied during the assessment of data conformity to Benford's Law in the process of detecting financial frauds. Moreover, an example of detecting irregularities in accounting data is presented.

2 Methods based on Euclidean distance

W.K.T. Cho and B.J. Gaines (2007) proposed the following measure based on the Euclidean distance when checking data conformity to Benford's distribution:

$$d = \left[\sum_{d_1=1}^{9}\left(p_{d_1} - w_{d_1}\right)^2\right]^{1/2}, \tag{3}$$

where: p_{d_1} is the probability of appearance of digit d_1, resulting from Benford's Law (see: equation (1)), w_{d_1} denotes the observed relative frequency of digit d_1 in a data set consisting of n records.

Dividing d by the maximum possible distance between two distributions, one of which is Benford's distribution and the other is the distribution where only digit 9 appears (i.e. a digit which is expected to occur the least often according to Benford's Law), W.K.T Cho and B.J. Gaines obtained:

$$d^* = d/\left[\sum_{d_1=1}^{8} p_{d_1}^2 + (1-p_9)^2\right]^{1/2} \cong d/1{,}0363. \tag{4}$$

The measure d^* can take on any value from 0 to 1. The lower the value of d^*, the higher conformity to Benford's Law.

Deliberations presented in (Cho and Gaines (2007)) concentrate on the first significant digit. However, it is possible to extend them to the first two significant digits case. Thus, we have:

$$d = \left[\sum_{d_1d_2=10}^{99}\left(p_{d_1d_2} - w_{d_1d_2}\right)^2\right]^{1/2}, \tag{5}$$

and

$$d^* = d/\left[\sum_{d_1d_2=10}^{98} p_{d_1d_2}^2 + (1-p_{99})^2\right]^{1/2} \cong d/1{,}0041, \tag{6}$$

where $p_{d_1d_2}$ and $w_{d_1d_2}$ are the probability of appearance of digits d_1d_2 (resulting from Benford's Law; see: equation (2)) and the observed relative frequency of digits d_1d_2 in a data set consisting of n records, respectively.

J. Morrow (2014) introduced the measure (3) to a hypothesis-testing framework. So as to verify the null hypothesis that the first significant digit distribution stays in accordance with Benford's distribution, against the alternative hypothesis that the first significant digit distribution does not conform to Benford's distribution, one uses the statistic:

$$D_n = \sqrt{n}\sqrt{\sum_{d_1=1}^{9}\left(W_{d_1} - p_{d_1}\right)^2}. \tag{7}$$

In his paper, J. Morrow gives also critical values for the above written statistic.

An analogous statistic, but this time focused on the first two significant digits distribution analysis, takes the following form:

$$D_n = \sqrt{n}\sqrt{\sum_{d_1d_2=10}^{99}\left(W_{d_1d_2} - p_{d_1d_2}\right)^2}. \tag{8}$$

D.W. Joenssen and T. Muellerleile (2015) created in R environment the package 'BenfordTests' which enables, among others, the analysis of the first significant digit distribution and the first two significant digits distribution.

3 Example

In our study we used data set A which contained 12,104 foreign revenues from the sales of finished products in 2016 in a certain company. The conducted analysis was based on the examination of the first two significant digits distribution. In the case of the statistical test based on the Euclidean distance, the p-value was calculated by means of the bootstrap technique, assuming 10,000 replicates.

Since the company's financial statements for 2016 were accepted without any reservations by an auditor, the expectation was that the financial data, including the foreign revenues, was the result of the proper accounting process, and therefore the data should follow Benford's distribution. Indeed, at the 0.05 level of significance, the Euclidean distance test did not permit to reject the null hypothesis stating that the distribution of the first two significant digits of foreign revenues conforms to Benford's distribution (D_n = 1,0736, p-value = 0,1653).

Next, a certain accountant (who did not know Benford's Law) was asked to commit fraud by adding 121 falsified records to data set A. Thus, a new data set (data set B) contained 12,225 foreign revenues. The results of the Euclidean distance test (D_n = 1,1472, p-value = 0,0470) led to the conclusion that at the 0.05 level of significance, we accept the alternative hypothesis stating that the distribution of the first two significant digits of foreign revenues does not conform to Benford's Law. For this reason, the foreign revenues from data set B resulted from the improper accounting process.

Table 1 presents the results of checking data conformity to Benford's Law for both analyzed data sets, taking into account the first two significant digits and using the measure d^*. The analysis of foreign revenues was made for the whole data sets (the last row of the table), and for five subsets of the data.

Table 1. *Outcomes of the assessment of data conformity from data sets A and B to Benford's distribution.*

Foreign revenues (PLN)	A		B	
	n	d^*	n	d^*
10.00 to less than 100.00	51	0.2066	51	0.2066
100.00 to less than 1,000.00	2,201	0.0814	2,213	0.0820
1,000.00 to less than 10,000.00	6,340	0.0271	6,443	0.0280
10,000.00 to less than 100,000.00	3,336	0.0757	3,337	0.0757
100,000.00 to less than 1,000,000.00	176	0.2315	181	0.2304
10.00 to less than 1,000,000.00	12,104	0.0097	12,225	0.0103

The obtained results allow to formulate the following main conclusions. Firstly, data set A is characterized by a higher level of agreement with Benford's Law than data set B. Secondly, the poorest fit for both data sets is observed for [10.00, 100.00) and [100,000.00, 1,000,000.00) intervals. Such a situation is caused by a small amount of numbers in these two intervals. Thirdly, the introduction of 121 falsified records into data set A resulted in: (a) the decline in the level of conformity to Benford's Law in the case of five- and six-digit revenues, (b) the increase in the level of conformity to Benford's Law in the case of eight-digit revenues, (c) no change (or a very small change) in the level of conformity to Benford's Law in the case of four- and seven-digit revenues.

4 Conclusion

The article discussed methods based on the Euclidean distance which are employed when assessing data conformity to Benford's Law. The outcomes of the conducted study indicated that these techniques can be a useful tool in the process of financial fraud detection. Although J. Morrow described the statistical test based on the Euclidean distance, there is still an open problem which deals with determining the distribution of the statistic D_n.

References

BENFORD, F. 1938. The law of anomalous numbers. *Proceedings of the American Philosophical Society.*, **78**(4), 551-572.

CHO, W.K.T. & GAINES, B.J. 2007. Breaking the (Benford) law: statistical fraud detection in campaign finance. *The American Statistician.*, **61**(3), 218-223.

JOENSSEN, D.W. & MUELLERLEILE, T. 2015. BenfordTests: statistical tests for evaluating conformity to Benford's law, R package version 1.2.0. Available at: https://cran.r-project.org/web/packages/BenfordTests/index.html.

MORROW, J. 2014. Benford's Law, Families of Distributions and a Test Basis. CEP Discussion Paper No 1291, Centre for Economic Performance, London School of Economics and Political Science. Available at: https://cep.lse.ac.uk/pubs/download/dp1291.pdf.

NEWCOMB, S. 1881. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics.*, **4**(1), 39-40.

NIGRINI, M.J. 2012. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. New Jersey: John Wiley & Sons.

THE EVOLUTION OF THE PURCHASE BEHAVIOR OF SPARKLING WINES IN THE ITALIAN MARKET

Francesca Bassi[1], Fulvia Pennoni[2] and Luca Rossetto[3]

[1] Department of Statistical Sciences, University of Padova,
(e-mail: francesca.bassi@unipd.it)

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

[3] Department of Land. Environment, Agriculture and Forestry, University of Padova,
(e-mail: luca.rossetto@unipd.it)

ABSTRACT: A dynamic analysis of purchases in the Italian market of sparkling wines is conducted by using scanner data derived from a consumer panel. We propose a continuous-time hidden Markov model that allows the transitions between states at any point in time. Results identify consumers' profiles in terms of type of purchases and socio-economic characteristics and describe the dynamics, and its determinants, across market segments. The findings improve the understanding of the market and provide useful evidences to design successful marketing strategies.

KEYWORDS: consumers' profiles, hidden Markov model, market dynamics, consumer panel.

1 Introduction

We study the dynamics of consumers' behavior in the Italian market of sparkling wine. In the last decade, the strong increase in the sparkling wines market has been coupled by a growth in brands, appellations, price range as well as other attributes (e.g. packaging), to catch consumers' attention. While in many countries the market tends to be dominated by Champagne, Cava or Prosecco, in Italy there is a greater fragmentation due to the preponderance of numerous domestic products and their complex denomination of origin classification. The consumption occasions for drinking sparkling wines have changed. Italian drinkers have started to drink and buy sparkling wine throughout the year rather than at specific seasons (e.g. Christmas); for this reason, the market is growing and it is expected to grow, offering more opportunities for sparkling wine producers.

In 2017 in Italy 31.6 million people (64% of adults)[1] consumed sparkling wine at least once; the majority of purchases is made in supermarkets; some specific

[1] https://www.wineintelligence.com

appellations, especially Champagne, are bought also in wine shops. Brand awareness, promotional offers and friends and family recommendations are the most important drivers of choice. Other relevant wine attributes for preferences are the method of production (Charmat, like Prosecco vs. Classic or Champenoise), appellations, especially the Controlled Denomination of Origin (DOC) and the Guaranteed Controlled Denomination of Origin (DOCG), the producer brand, the label, and its location.

The market of sparkling wines is relatively young, therefore the literature focusing on this topic is quite scarce. It mainly reports works on technical and sensorial aspects (Culbert *et al*., 2017) or on consumer's behavior and preferences (Cohen *et al*., 2012). With the proposed study of the market dynamics and of the factors that favor it we provide important information for designing successful marketing strategies. By using information collected on a panel of Italian families with purchases in stores we aim at identifying typical customers' profiles and analysing if and how they change acquisition behavior within two years of time. We also evaluate the effects of the characteristics of the consumers and the families on purchases.

2 Data and method

Data concerns a panel of 9,000 Italian households who registered their purchases in 2015 and 2016. The sample is representative of the Italian population with reference to the area of residence, number of components, monthly per capita income, age of the person responsible for purchases, type of the family. The survey collects longitudinal data with continuous time, each household may perform multiple purchases in the reference period. We observe a total of 22,362 purchases in unspecialized stores, made by 5,155 households, they make from 1 to 230 purchases in the reference period.

The dynamics of consumers' behavior is analyzed by assuming that preferences can be represented by an underlying latent variable $\boldsymbol{U} = (U_{i1},\ldots,U_{iT})$ for each customer $i = 1,\ldots,n,$ at occasion t, $t = 1,\ldots,T,$ where each t refers to the purchase time period. The latent process follows a continuous-time hidden Markov chain with discrete states, initial and transition probabilities parameterized with covariates (Bartolucci *et al*., 2013). We propose a multivariate hidden Markov model (HMM) for the vector of categorical responses $\boldsymbol{Y}_{it} = (Y_{i1t},\ldots,Y_{irt})$ where the responses observed for every customer are: the value of each purchase in Euros, denomination and type of wine. The main assumption is that the latent process fully explains the observed customer behavior and the time-fixed and time-varying customer socio-demographic characteristics describe the dynamics of the underlying latent preferences. The conditional probabilities of the responses are assumed constant over time to stabilize the customer's profiles. The Expectation-Maximization algorithm is employed to maximize the log-likelihood. The suitable number of latent states is selected by using the Bayesian Information Criterion.

3 Results

Five clusters of homogeneous purchases are identified as reported in Table 1. The first latent state identifies consumers that spend no more than four Euros for an ordinary sparkling wine, with no specific appellation. We refer to this segment as that of customers with low quality purchases, concerning 19% of the population. Latent state 2 defines customers with low quality purchases preferring sweet wine, this concerns 34% of the customers. Latent state 3 defines customers with mainly purchases of Prosecco wine both with DOC and DOCG, of dry or extra dry type, and the amount spent per purchase is over three Euros; we define this profile as Prosecco (20% of the population). Latent state 4 denotes the profile of sophisticated customers, not choosing Prosecco since they select mainly prestigious denominations such as Franciacorta, Asti, Brachetto D'Aqui, Oltrepo Pavese (13% of the customers). Latent state 5 denotes the profiles of sparkling wine connoisseurs, since they show purchases with the highest purchasing power, over six Euros, for brut classic sparkling wine with appellations such as Franciacorta, Trento and Champagne (14% of the customers).

Table 2 lists the average transition probabilities among each segment. The percentage of customers who do not change purchase behavior can differ quite a lot across states. Purchases of types 1, 2 and 3 are more stable than those of type 4 and 5. However, a non-negligible percentage of customers, greater than 12%, tend to move towards segment 1. The state from which there is the highest mobility is 5, these are purchases with the highest amount of money spent and the most prestigious appellations: this reveals as an occasional consumption behavior.

Concerning the effects of the covariates we mention that purchases in segment 3 of Prosecco refer with higher probability to middle-age consumers, living mainly in the North-east of Italy and Lazio region, with a medium-level income.

4 Conclusions

We propose a dynamic analysis of the Italian market of sparkling wines estimating a hidden Markov model on scanner data from a consumer panel. Latent states identify five homogenous types of purchases according to prices, type of wine and appellation. A non-negligible proportion of consumers perform purchases of different types, the most unstable segment is that with the highest price. Consumers tend to move to a segment of lower quality wine for their subsequent purchase and consumers' characteristics act as drivers of preferences.

Table 1. Latent states' profiles

Response variables	*Estimated conditional probabilities*				
Average purchase in Euros					
<2.99	0.22	0.51	0.03	0.05	0.00
2.99-3.98	0.30	0.21	0.19	0.24	0.00
3.99-5.68	0.19	0.13	0.26	0.30	0.01
5.69-8.98	0.18	0.10	0.28	0.27	0.22
>8.98	0.11	0.05	0.24	0.14	0.76
Type of wine (sugar content)					
Brut	0.78	0.07	0.07	0.00	0.97
Extra dry	0.14	0.02	0.59	0.02	0.00
Dry	0.08	0.04	0.34	0.00	0.00
Sweet	0.01	0.88	0.00	0.98	0.00
Denomination					
No appellation	0.69	0.95	0.04	0.01	0.01
Prosecco DOCG	0.06	0.01	0.41	0.00	0.00
Prosecco DOC	0.11	0.00	0.58	0.00	0.01
Franciacorta DOCG	0.00	0.00	0.00	0.16	0.40
Asti DOCG	0.00	0.03	0.00	0.39	0.00
Trento DOC	0.00	0.00	0.00	0.00	0.42
Brachetto DOCG	0.00	0.00	0.00	0.28	0.00
Oltrepo DOCG	0.04	0.00	0.00	0.06	0.00
IGT	0.07	0.00	0.00	0.00	0.01
French appellation	0.00	0.00	0.00	0.00	0.14
Other	0.01	0.00	0.00	0.07	0.00

Table 2. Average transition matrix

	State 1	State 2	State 3	State 4	State 5
State 1	0.74	0.06	0.11	0.04	0.04
State 2	0.12	0.66	0.09	0.09	0.04
State 3	0.12	0.08	0.72	0.04	0.05
State 4	0.14	0.18	0.11	0.52	0.05
State 5	0.18	0.11	0.15	0.10	0.47

References

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

COHEN, J., LOCKSHIN, L., & SHARP, B. 2012. A better understanding of the structure of a wine market using the attribute of variety. *International Journal of Business and Globalisation,* **8**, 66-80.

CULBERT J.A., RISTIC R., OVINGTON A., SALIBA A.J. & WILKINSON K. 2017. Influence of production method on the sensory profile and consumer acceptance of Australian sparkling white wine styles. *Australian Journal of Grape and Wine Research*. **23**, 170-178.

MODERN LIKELIHOOD-FREQUENTIST INFERENCE AT WORK*

Ruggero Bellio[1] and Donald A. Pierce[2]

[1] Department of Economics and Statistics, University of Udine,
(e-mail: ruggero.bellio@uniud.it)

[2] Statistics Department, Oregon State University,
(e-mail: pierce.don.a@gmail.com)

ABSTRACT: Modern likelihood asymptotics make available several inferential methods that can be applied to a large class of statistical models. In this contribution we summarize some of such methods, and provide an illustration by means of an application to a hierarchical nonlinear regression model. The methodology presented can be readily applied by the R package likelihoodAsy.

KEYWORDS: likelihood asymptotics; nonlinear regression; statistical software.

1 Background and theory

Modern likelihood asymptotics is a well established theory for inferential methods in parametric statistical models. The relevant literature is large, with surveys of the main results given in Severini, 2000 and Skovgaard, 2001, among others. The recent expository paper by Pierce & Bellio, 2017 tracks some of the developments, with an effort to make them accessible to a wider audience. The paper has an accompanying R package, named likelihoodAsy (available at the CRAN repository), which implements some of the methods.

The starting point is a parametric statistical model for the sample $y = (y_1,\dots,y_n)$, given by a density $p(y;\theta)$, indexed by a p-dimensional parameter θ. Let $\ell(\theta;y) = \log L(\theta;y)$ the log likelihood function and $\widehat{\theta}$ the maximum likelihood estimate.

The main methodology concerns inference about a scalar smooth function of the parameter θ, defined as $\psi(\theta)$. For testing the hypothesis $\psi(\theta) = \psi$, the recommended approach relies on the directed deviance

$$r_\psi(y) = \mathrm{sgn}(\widehat{\psi} - \psi)\left[2\left\{\ell(\widehat{\theta};y) - \ell(\widehat{\theta}_\psi;y)\right\}\right]^{1/2},$$

*This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS_003.

where $\widehat{\theta}_\psi$ is the maximum likelihood estimate of θ for fixed ψ. The key theoretical result is the r^*-formula (see for example Severini, 2000)

$$\Pr\left\{r_\psi(Y) \leq r_\psi(y); \theta : \psi(\theta) = \psi\right\} = \Phi\{r^*_\psi(y)\}\left\{1 + O\left(n^{-1}\right)\right\}, \qquad (1)$$

where $r^* = r^*_\psi(y)$ is a modified directed deviance and $\Phi(\cdot)$ is the standard normal cdf. The formula improves on the usual first-order version employing $\Phi\{r_\psi(y)\}$ for the probability on the left-hand side, and it provides a fairly accurate approximation to the distribution of the directed deviance. This can be readily used for computing confidence intervals and p-values for $\psi(\theta)$.

Some remarks on (1) are in order.

(i) The computation of r^* is challenging, but code in `likelihoodAsy` provides a fairly accurate approximation to it, with protection for large deviations (see Skovgaard, 2001). The code requires the user to supply a function implementing the log likelihood function and a function to generate a data set from the model. The latter is used for computing certain expected values entering the r^*-formula by a Monte Carlo approach.

(ii) A remarkable feature of the code is that the interest parameter $\psi(\theta)$ need not be a coordinate of the parametrization employed for the model.

(iii) Pierce & Bellio, 2017 provides a detailed account on the nature of (1). Moreover, results cited in the article show that inferences based on (1) are quite close to those of the most accurate parametric bootstrap method, which simulates a large number of bootstrap samples from $p(y; \widehat{\theta}_\psi)$.

Code in the package includes also routines for computing the modified profile likelihood, which is the inferential tool suitable for inference on multidimensional parameter of interest ψ accounting for nuisance parameters. Namely, $\ell_{\mathrm{M}}(\psi) = \ell_{\mathrm{P}}(\psi) + \log M(\psi; y)$ is returned, where $\ell_{\mathrm{P}}(\psi)$ is the profile log likelihood for ψ and $M(\psi; y)$ an adjustment term, whose computation entails a task similar to that required for $r^*_\psi(y)$.

2 Application to a nonlinear regression model

For an illustration of the scope of the methodology, we summarize here the analysis of the theophylline data, already considered by several authors. In particular the data are thoroughly analyzed in the monograph by Pinhéiro & Bates, 2000, and made available in the `R` package `nlme` associated to the book. The data are about a longitudinal study on 12 patients, each observed on 10 time points. The response is the theophylline concentration, for which a nonlinear regression model is adopted (see Pinhéiro & Bates, 2000, p. 351)

$$y_{ij} = f(\phi_i; d_i, t_{ij}) + \varepsilon_{ij}. \qquad (2)$$

Here i is the index for subject and j for time point, d_i a time-invariant dose and t_{ij} the j-th time point for the i-th subject. The model assumed is

$$f(\phi_i; d_i, t_{ij}) = d_i \frac{\exp(\phi_{1i} + \phi_{2i} - \phi_3)}{\{\exp(\phi_{2i}) - \exp(\phi_{1i})\}} \left[\exp\{-\exp(\phi_{1i})t_{ij}\} - \exp\{-\exp(\phi_{2i})t_{ij}\}\right] .$$

This is a one-compartment model, with two subject-specific parameters ($\exp(\phi_{1i})$ and $\exp(\phi_{2i})$) and one common parameter ($\exp(\phi_3)$; the exponential form for them is introduced for numerical stability. In what follows, we take the *clearance* of the model as the parameter of interest, defined as $\psi = \exp(\phi_3)$.

At first, we follow a fixed-effects approach, treating the subject-specific coefficients as model parameters, and assuming a normal distribution for the error term with subject-specific standard deviation. Implementation in R is straightforward, due to the independence assumption, yet the resulting model has 37 parameters for a sample with 120 observations. This is a setting where the r^* statistic may be useful, since this typically occurs in settings with either very small sample size or with many nuisance parameters.

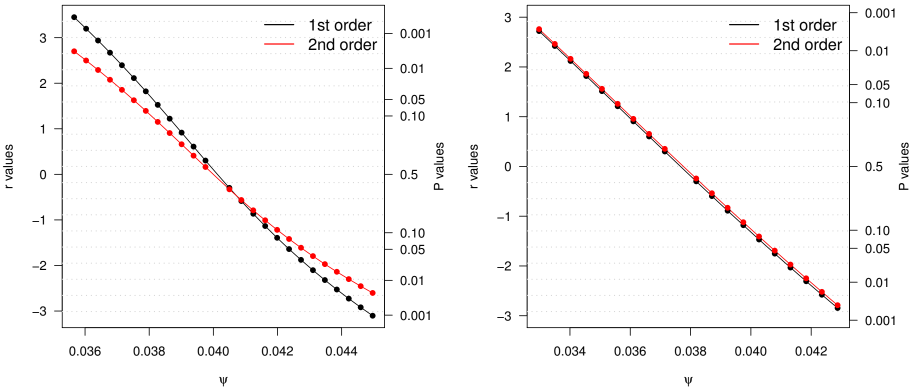


Figure 1. *$r_\psi(y)$ and $r^*_\psi(y)$ as a function of $\psi = \exp(\phi_3)$ for a fixed-effects model (left) and a random effects model (right). The P-values are the one-sided error rates given by $1 - \Phi\{|r_\psi(y)|\}$ or $1 - \Phi\{|r^*_\psi(y)|\}$.*

This fact is represented in the left panel of Figure 1, which shows the values of $r_\psi(y)$ and $r^*_\psi(y)$ as ψ varies along a grid of values around $\widehat{\psi}$=0.038. Confidence intervals are read off the plot, and correspond to ψ-points where the curves intersect the normal quantiles for the level of interest. The adjustment performed by the r^*-formula is noticeable, with the 95% confidence interval which is about 25% wider than that based on the first order solution.

A more customary modelling approach would treat ϕ_{1i} and ϕ_{2i} as normal random effects, and integrate them out to obtain the likelihood function. This entails a more challenging implementation, which is doable by recourse to the `TMB` package for automatic differentiation (Kristensen *et al.* , 2016). This has been done for a simpler version of the model, assuming homoscedasticity for ε_{ij}, obtaining $\widehat{\psi}$=0.040. A close agreement between $r_{\psi}(y)$ and $r^*_{\psi}(y)$ is found, as shown in the right panel of Figure 1. Indeed, the random effects models has only 6 parameters and the nuisance parameters adjustment is small. Finally, in the random effects model $\ell_{\mathrm{M}}(\psi)$ can be used to estimate the variance components in a REML-like fashion. This gives a 10% inflation for the estimates of random effects standard deviation.

3 Conclusion and ongoing research

Modern likelihood asymptotics has the potential to supplement standard analysis for models currently used in applications in several fields. The availability of suitable software appears to be the key factor. To this end, some developments of the `likelihoodAsy` package may involve the inclusion of further methods, such as the multidimensional tests (Skovgaard, 2001). Further extension would involve closer integration with the `TMB` package, which seems a promising route for targeting more realistic and complex models.

References

KRISTENSEN, K., NIELSEN, A., BERG, C. W., SKAUG, H., & BELL, B. M. 2016. TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, **70**, 1–21.

PIERCE, D. A., & BELLIO, R. 2017. Modern likelihood-frequentist inference. *International Statistical Review*, **85**, 519–541.

PINHÉIRO, J. C., & BATES, D. M. 2000. *Mixed-effects Models in S and S-PLUS*. New York: Springer-Verlag.

SEVERINI, T. A. 2000. *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

SKOVGAARD, I. M. 2001. Likelihood asymptotics. *Scandinavian Journal of Statistics*, **28**, 3–32.

ONTOLOGY-BASED CLASSIFICATION OF MULTILINGUAL CORPUSES OF DOCUMENTS

Sergey Belov[1, 2], Salvatore Ingrassia[3], Zoran Kalinić[4] and Paweł Lula[5]

[1] Laboratory of Information Technologies, Joint Institute for Nuclear Research, Dubna, (e-mail: sergey.belov@jinr.ru)

[2] Plekhanov Russian University of Economics,

[3] Department of Economics and Business, Catania University, (e-mail: s.ingrassia@unict.it)

[4] Faculty of Economics, University of Kragujevac, (e-mail: zkalinic@kg.ac.rs)

[5] Department of Computational Systems, Cracow University of Economics, (e-mail: pawel.lula@uek.krakow.pl)

ABSTRACT: The ontology-based classification of multilingual documents is the main problem discussed in the paper. The system proposed here is focused on the issue of cluster analysis of large sets of job offers prepared in various languages, but the method has universal character and can be used for analysis of documents related to a specific domain. Taking into account computational requirements, the authors propose to conduct all calculations in the cluster environment.

KEYWORDS: cluster analysis, exploratory analysis of multilingual documents, computational infrastructure, analysis of competencies.

1 Introduction

Ontology-based approach in computational text analysis allows to explore large corpuses of documents related to a specific domain. This technique appeared in the literature about twenty years ago (cf. Hotho et al, 2002) and has been used in different areas. Analysis of competencies expected by employers constitutes one of the most important field of application for this approach. However, it is worth emphasizing that all solutions developed in the area related to competencies have universal character and can be applied to other types of documents.

The issue of competencies appeared in the research literature in the 1960s (White, 1959) and has been being developed in many further publications (Boyatzis, 1982), (Levy-Leboyer, 1996), (Bengtsson, 1996). The problem of competency development is also widely discussed in the context of education, cf. (Lambrechts et al., 2013) or Oczkowska et al., 2017).

Employers' expectations towards competencies of candidates for employment are being considered by managers, experts responsible for education, individuals thinking about their future career, employees or politicians. This fact justifies the necessity of building IT solutions for labour market monitoring. Two systems belonging to this area were presented in (Lula et al., 2018) and (Belov et al., 2018).

However, it seems that globalization processes and swelling migration flows create the need for analysis of competencies in a context wider as national labour market which that takes into account the situation prevailing in many countries at the same time. Job offers published online on different labour markets can be treated as very important source of information on employees' competencies. Its universality and ease of access is a key advantage. Nevertheless, their analysis may cause many difficulties due to the lack of a uniform format of offers, the ambiguity of expressions used in their contents and multilingual nature of advertisements.

In the context presented above, the main goal of this study can be stated as the development of a system for analysis of job offers with particular reference to three issues: automatic exploration of offers prepared in various languages, cluster analysis of offers and performing required calculation in cluster environment.

2 The methodology

Let's assume that the $\boldsymbol{O}$ is a set of job offers:

$$\boldsymbol{O} = \{O_1, O_2, \dots, O_N\}$$

and $\boldsymbol{C}$ is a set of competencies:

$$\boldsymbol{C} = \{C_1, C_2, \dots, C_M\}.$$

The main goal of the analysis as performing a cluster analysis of objects belonging to the set $\boldsymbol{O}$ with respect to competencies taken from the set $\boldsymbol{C}$.

The process of a job offer analysis can be divided into two parts:

1. analysis of job offers' contents and their transformation into a form suitable for cluster analysis,
2. cluster analysis of objects representing job offers' contents.

2.1 The analysis of job offers' contents

It was assumed that the system proposed here should process job offers prepared in various languages (the current version can analyse offers prepared in English, Italian, Polish, Russian and Serbian language) and that modules designed for different language versions should recognise the same set of competencies. Otherwise the scope of the further analysis would be seriously limited.

Past experiences of the authors presented in (Lula et al., 2018) and (Belov et al., 2018) and literature survey indicates that the schema of analysis presented in the Figure 1 can be used.

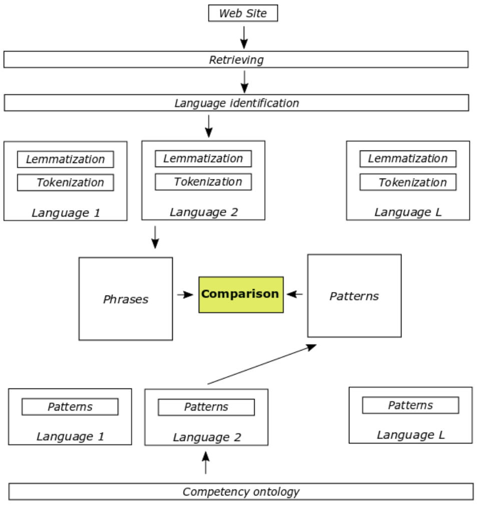


Figure 1. The schema of ontology-based analysis of job offers

The process of analysis of a given job offer allows to identify main competencies expected by an employer. During the analysis the significance of every competency is estimated for every phrase occurring in an offer. Next, significant coefficient calculated for every phrase can be aggregated to express the importance of a given competence in a whole job offer.

2.2 Cluster analysis of job offers

Cluster analysis of job offers should identify their homogenous groups. It seems that model-based approach can find formal description for every group (Ingrassia et al., 2015). The authors would like to find answers to research questions concerning similarities and differences between clusters identified in various countries. Also it may be interesting to compare formal descriptions of profiles of the most popular positions with respect of countries, employers, sectors and offer's language.

3 Implementation

All algorithms presented in the paper were implemented in R and Python languages. Raw data is being gathered from the publicly available sources (hunting agencies, job seeking sites, governmental organizations) using the following methods: web scraping, archives download, using sites' application programming interfaces. For the further processing to provide scalability and affordable speed, the technology stack of Big Data is used (Zaharia et al, 2012). Among the variety of products involved to build the processing pipeline, it is worth to mention Apache Spark which is used as a core platform to organize the computational part of the whole system (Armbrust et al, 2015). The infrastructure is based on the university cloud with application of Docker containerisation approach to ease the processing chain management.

References

ARMBRUST, M., XIN, R.S., LIAN, C., HUAI, Y., LIU, D., BRADLEY, J.K., MENG, X., KAFTAN, T., FRANKLIN, M.J., GHODSI. A., & ZAHARIA, M. 2015. Spark SQL: Relational Data Processing in Spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (SIGMOD '15). ACM, New York, NY, USA, 1383-1394. DOI: https://doi.org/10.1145/2723372.2742797

BELOV, S., FILOZOVA, I., KADOCHNIKOV, I., KORENKOV, V., SEMENOV, R., SMELOV, P., ZRELOV, P. 2018. Labour market monitoring system, *CEUR Workshop Proceedings*, ISSN:1613-0073, Изд:CEUR Workshop Proceedings, 2267, 528-532.

BENGTSSON, J. 1996. Rynki pracy przyszłości: wyzwania polityki edukacyjnej. *Nauka i Szkolnictwo Wyższe*, **7**, 24-45.

BOYATZIS, R.E. 1982. *The competent manager: a model for effective performance*. New York: John Wiley & Sons.

HOTHO A., MAEDCHE A., & STAAB S. 2002. Ontology-based text document clustering, University of Karlsruhe, https://www.researchgate.net/publication/220633810_Ontology-based_Text_Document_Clustering

INGRASSIA S., PUNZO A.,VITTADINI G., AND MINOTTI S.C. 2015. The Generalized Linear Mixed Cluster-Weighted Model, *Journal of Classification*, **32**(1): 85-113.

LAMBRECHTS, W., MULÀ I., CEULEMANS, K., MOLDEREZ, I., & GAEREMYNCK, V. 2013. The integration of competences for sustainable development in higher education: An analysis of bachelor programs in management. *Journal of Cleaner Production*, **48**, 65-73.

LEVY-LEBOYER, C. 1996. *La Gestion des competences*. Paris: Les Editions d'Organisation.

LULA, P., OCZKOWSKA, R., WIŚNIEWSKA, S., & WÓJCIK, K. 2018, Ontology-Based System for Automatic Analysis of Job Offers. *Information Technology for Practice*, 205-212

OCZKOWSKA, R., WIŚNIEWSKA, S., & LULA, P. 2017. Analysis of the competence gap among vocational school graduates in the area of smart specialization in Poland. *International Journal for Quality Research*, **11**(4), 945-966.

WHITE, R. 1959. Motivation reconsidered: The Concept of Competence. *Psychological Review*, **66**(5), 279-333.

ZAHARIA, M., CHOWDHURY, M., DAS, T., DAVE, A., MA, J., MCCAULEY, M., FRANKLIN M.J., SHENKER, S., & STOICA, I.. 2012. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation* (NSDI'12). USENIX Association, Berkeley, CA, USA, 2-2.

MODELING HETEROGENEITY IN CLUSTERED DATA USING RECURSIVE PARTITIONING

Moritz Berger[1] and Gerhard Tutz[2]

[1] Department of Medical Biometry, Informatics and Epidemiology, Rheinische Friedrich-Wilhelms-Universität München, (e-mail: moritz.berger@imbie.uni-bonn.de)

[2] Department of Statistics, Ludwig-Maximilians-Universität München, (e-mail: tutz@stat.uni-muenchen.de)

ABSTRACT: The analysis of longitudinal and cross-sectional data requires taking the dependence of observations and the heterogeneity of measurement units into account. A very flexible tool to account for unobserved heterogeneity are fixed effects models because they do not make assumptions on the distribution of effects. On the basis of a fixed effects model, we propose a recursive partitioning method that identifies clusters of units that share the same effect. The approach reduces the number of parameters to be estimated and is beneficial in particular if one is interested in identifying clusters with the same effect on the outcome variable. The usefulness of the approach is illustrated in an application using data from CTB/McGraw-Hill.

KEYWORDS: clustered data, fixed effects model, recursive partitioning, tree-structured regression.

1 Fixed Effects Models

Consider clustered data given by (y_{ij}, x_{ij}, z_{ij}), $i = 1, \dots, n$, $j = 1, \dots, n_i$, where y_{ij} denotes the response of measurement j for unit i. There are two sets of predictive variables $x_{ij}^\top = (1, x_{ij1}, \dots, x_{ijp})$ and $z_{ij}^\top = (1, z_{ij1}, \dots, z_{ijq})$ including p and q covariates, respectively. In a fixed effects model the mean response $\mu_{ij} = \mathbb{E}(y_{ij}|x_{ij}, z_{ij})$ is linked to the explanatory variables in the form

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^\top \beta + z_{ij}^\top \beta_i , \tag{1}$$

where x_{ij} is a vector of covariates that has the same effect β across all units and z_{ij} contains the covariates with effects that vary over units. With regard to z_{ij} each measurement unit has its own parameter vector $\beta_i^\top = (\beta_{i0}, \dots, \beta_{iq})$. The specification of one parameter vector per unit results in a very large number of parameters which can affect estimation accuracy. Moreover, typically there is not enough information to distinguish between all units. To address these issues, one can assume that there are groups of units (i.e. clusters) that share the same effect on the outcome.

2 Tree-Structured Models

Consider the fixed effects model with unit-specific intercepts, only. In the simplest case in which all intercepts are equal the linear predictor has the form $\eta_{ij} = x_{ij}^{\top}\beta + \beta_0$. If there are two clusters, the corresponding linear predictor is given by

$$\eta_{ij} = x_{ij}^{\top}\beta + \beta_{i0}^{(k)}, \quad k = 1,2, \tag{2}$$

where k denotes the membership to a group and $\beta_{i0}^{(k)}$ is the corresponding effect for the group. A simple test, for example a likelihood ratio test, for the hypothesis $H_0 : \beta_{i0}^{(1)} = \beta_{i0}^{(2)}$ can be used to determine if the model with two groups is more adequate for the data than the model in which all the intercepts are equal. By iterative splitting into subsets guided by test statistics one obtains a clustering of units that have to be distinguished with regard to their intercept.

In general, a tree is built by successively splitting one node A, that is already a subset of the predictor space, into two subsets A_1 and A_2 with the split being determined by only one variable. In a fixed effects model, when specifying intercepts for each unit, the unit number $i \in \{1,\dots,n\}$ itself can be seen as a nominal categorical covariate with n categories. The partition has the form $A \cap S_1$, $A \cap S_2$, where S_1 and S_2 are disjoint, non-empty subsets $S_1 \subset \{1,\dots,n\}$ and its complement $S_2 = \{1,\dots,n\} \setminus S_1$. Using this notation another representation of model (2) is given by

$$\eta_{ij} = x_{ij}^{\top}\beta + \beta_{i0}^{(1)} I(i \in S_{10}) + \beta_{i0}^{(2)} I(i \in S_{20}), \tag{3}$$

where $I(\cdot)$ denotes the indicator function with $I(a) = 1$, if a is true and $I(a) = 0$ otherwise. After several splits one obtains a clustering of the units $\{1,\dots,n\}$ and the predictor of the resulting model can be represented by

$$\eta_{ij} = x_{ij}^{\top}\beta + \sum_{k=1}^{m_0} \beta_{i0}^{(k)} I(i \in S_{k0}), \tag{4}$$

where $S_{10},\dots,S_{m_0 0}$ is a partition of $\{1,\dots,n\}$ consisting of m_0 clusters that have to be distinguished in terms of their individual intercepts. To determine the optimal number of splits (i.e. to decide when to stop) our strategy is to

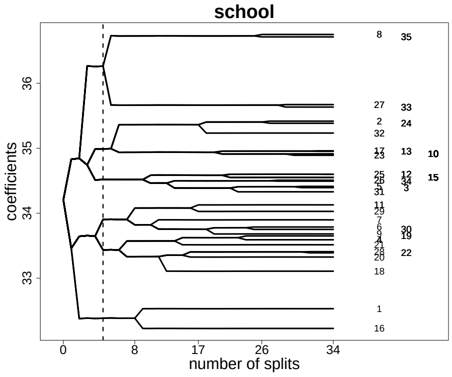


Figure 1. *Analyis of the CTB data. Paths of coefficients of school-specific intercepts against all splits. The paths build a tree that successively partitions the schools. The optimal number of splits is marked by a dashed line.*

check if the heterogeneity of measurement units is already modeled sufficiently in each step. To decide for the first split one has to examine the null hypothesis $H_0 : \beta_{10} = \beta_{20} = \ldots = \beta_{n0}$, which corresponds to the case of no heterogeneity. The hypothesis is tested by a likelihood-ratio test with significance level α and $n-1$ degrees of freedom. After several splits only differences of units within already built clusters are tested. In the $\ell - th$ step $n - \ell$ differences have to be tested because $\ell - 1$ splits are already performed. If a significant effect is found the selected split is performed, otherwise splitting is stopped.

3 Analysis of the CTB Data

We consider a data set from CTB/McGraw-Hill, a division of the Data Recognition Corporation (DRC). For a description of the original data, see De Boeck & Wilson, 2004. The data includes results of an achievement test that measures different objectives and subskills of subjects in mathematics and science. For our investigation we used the results of 1500 grade 8 students from 35 schools. They had to respond to 56 multiple-choice items (31 mathematics, 25 science). The outcome y_{ij} was the overall test score of student j in school i, defined as the number of correctly solved items. The main objective was to adequately describe the heterogeneity of the 35 schools. As additional covariate we included the gender of the students (male: 0, female: 1). There were 761 males and 739 females achieving an average test score of 34.

The coefficient paths of the school-specific intercepts obtained when fitting

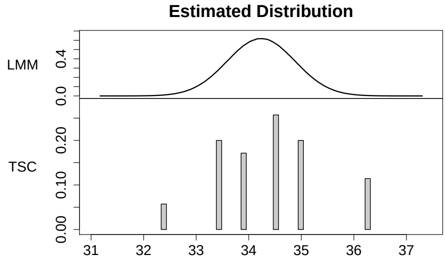


Figure 2. *Analysis of the CTB data. Comparison of the estimated distribution of a linear mixed model (LMM) and the school-specific intercepts of the tree-structured model (TSC). The distribution of the fixed effects is quite different from the normal distribution obtained for the random effects model.*

the tree-structured model are shown in Figure 1. The coefficient paths build a tree that successively partitions the schools in terms of the performance of students. The optimal number of splits that was selected by the algorithm, is marked by the dashed line. It is seen that estimates changed strongly in the first steps, but after about ten splits the estimates were very stable. A graphical comparison of the estimated normal distribution of the random effects when fitting a classical linear mixed model with **R** package `lme4` (Bates *et al.*, 2015) and the distribution of the school-specific intercepts of the tree-structured model is shown in Figure 2. It illustrates the main advantage of the tree-structured model. There is no distributional assumption on the school-specific intercepts, especially no assumption of symmetry. The number of schools in each cluster were quite different and not symmetric. The coefficient estimate for covariate gender was $\beta_{\text{gender}} = -0.088$ (95%-Bootstrap-CI: $[-0.478; 0.313]$), which showed no evidence for an effect.

References

BATES, D., MÄCHLER, M., BOLKER, B., & WALKER, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software.*, **67**, 1–48.

BERGER, M., & TUTZ, G. 2018. Tree-structured clustering in fixed effects models. *Journal of Computational and Graphical Statistics.*, **27**, 380–392.

DE BOECK, P., & WILSON, M. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* New York: Springer.

MIXTURES OF EXPERTS WITH FLEXIBLE CONCOMITANT COVARIATE EFFECTS: A BAYESIAN SOLUTION

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: Mixtures of experts models provide a framework in which concomitant variables may be included in mixture models. In this paper, we present a method to allow for flexible specification of the mixing proportions, as smoothing functions of these covariates. We propose a data augmentation algorithm for sampling the parameters from their posterior distribution within a Bayesian framework. The proposed methodology is investigated via a simulation experiment.

KEYWORDS: mixtures of experts models, data augmentation, bayesian P-splines.

1 Introduction

Mixture models are the basis of many model-based clustering methods. The use of a model-based approach to clustering allows for any uncertainty to be accounted for in a probabilistic framework. Mixtures of experts (ME) models provide a way to extend mixture models, and allow the parameters to depend on concomitant covariate information. In particular, in Jacobs et al. (1991) the components' weights are modeled as a logistic function of the covariates. Estimation of mixtures of experts models can be achieved within the Bayesian paradigm, using a Markov chain Monte Carlo (MCMC). Frühwirth-Schnatter et al. (2012) exploit data augmentation based on the differenced random utility model (dRUM) representation, thus introducing a set of auxiliary variables.

In this paper, we consider a more flexible specification of these auxiliary variables. More specifically, part of the linear predictor is substituted with a sum of smooth functions of each covariate, as in a generalized additive model. In order to achieve a parsimonious representation of these smooth functions, we use Bayesian P-splines as suggested by Lang & Brezger (2004).

2 Model specification

Consider an independent and identically distributed sample of outcome observations $\{\mathbf{y}_i\}$, with $i = 1, \ldots, n$, from a population modelled by a G components finite mixture model. Each component g (for $g = 1, \ldots, G$) is modelled by the probability density function $f(\mathbf{y}_i|\theta_g)$ with parameters denoted θ_g, and has weight p_g , such that $\sum_{g=1}^{G} p_g = 1$. Observation i has J associated covariates $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{iJ^*-1}, x_{iJ^*}, x_{iJ^*+1}, \ldots, x_{iJ})$, of which the last $J - J^*$ are metrical, with $J^* \in [1, J]$. The simple mixtures of experts model extends the finite mixture model by allowing the distribution of the latent variable to depend on the concomitant variables:

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{g=1}^{G} p_g(\mathbf{x}_i) f(\mathbf{y}_i|\theta_g). \tag{1}$$

Jacobs et al. (1991) model the components' weights using a multinomial logit regression model, which can be represented following Frühwirth-Schnatter & Frühwirth (2010) as a binary logit model conditional on knowing the regression parameters of the remaining categories.

Denote by γ_g and β_g the vectors containing the parameters respectively associated to the linear and nonlinear part of the predictor for the g-th component ($g = 1, \ldots, G - 1$):

$$\ln \frac{p_g(\mathbf{x}_i)}{p_G(\mathbf{x}_i)} = \eta_{gi} = \sum_{j=1}^{J*} \gamma_{gj} x_{ij} + \sum_{j=J^*+1}^{J} \sum_{\rho=1}^{m} \beta_{gj\rho} B_{j\rho}(x_{ij}) \tag{2}$$

where $B_{j\rho}(\cdot)$ (for $j = J^*, \ldots, J$ and $\rho = 1, \ldots, m = r + 4$) is a B-spline basis function for a cubic spline with r knots. Lang & Brezger (2004) suggest a number of knots between 20 and 40 to ensure enough flexibility, and to define the priors for the regression parameters β_{gj} in terms of a random walk:

$$\beta_{gj\rho} = \beta_{gj,\rho-1} + u_{gj\rho} \tag{3}$$

with $u_{gj\rho} \sim N(0, \tau^2_{gj})$. The amount of smoothness is controlled by the additional variance parameters τ^2_{gj}, which correspond to the inverse smoothing parameters in the classical approach. The presence of the smoothing parameter τ^2_{gj} protects against possibile overfitting if a large number of knots is chosen

Then, we can write the above-described binary logit model in the partial dRUM representation (Frühwirth-Schnatter & Frühwirth, 2010):

$$z_{gi} = \eta_{gi} - \log\left(\sum_{h \neq g} \lambda_{hi}\right) + \varepsilon_{gi}, \quad D_{gi} = \mathbf{1}(z_{gi} > 0) \tag{4}$$

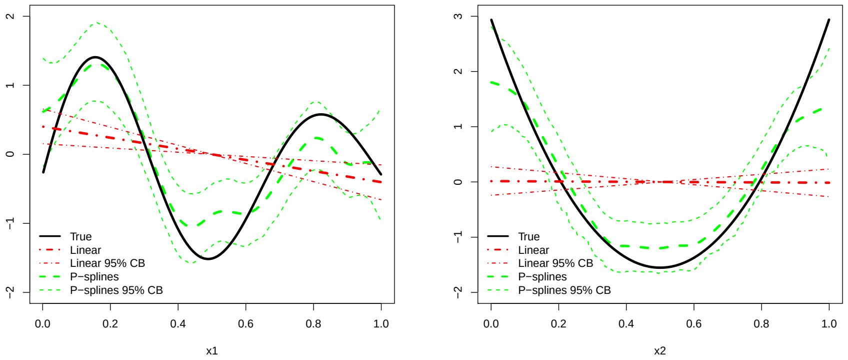


Figure 1. *True (solid line) and estimated posterior effects (with 95% posterior pointwise confidence bands) of the concomitant covariates on the log-odds of mixing proportions. Dotdashed and dashed lines are obtained using a linear predictor and an additive predictor, respectively.*

where z_{gi} is a latent variable, $\lambda_{gi} = \exp(\eta_{gi})$ and ε_{gi} are i.i.d. errors following a logistic distribution.

Given $\lambda_{1i}, \ldots, \lambda_{Gi}$ and the latent indicator variables $D_{1i}, \ldots, D_{Gi}$, the latent variables $(z_{1i}, \ldots, z_{Gi})$ follow exponential distributions and can be easily sampled in a data augmented implementation. To avoid any Metropolis-Hastings step, Frühwirth-Schnatter et al. (2012) approximate, for each ε_{gi}, the logistic distribution by a finite scale mixture of normal distributions with zero means and parameters drawn with fixed probabilities. Regarding the parameters of each component, appropriate full conditionals can be exploited in order to sample from the posterior distribution. Observations can be allocated into the G components using the maximum-a-posteriori rule.

3 Simulation study

The performances of the proposed approach are investigated in a simulated environment. Although this application concerns latent class analysis, the proposed methodology can be easily adapted to any other type of response variables, by chosing an appropriate form for $f(\mathbf{y}_i|\theta_g)$. We generate 100 independent datasets, with $n = 1000$ from a 2-components mixture distribution for 5 categorical manifest variables. The components' weights are assumed to depend on 2 uniformly distributed covariates x_1 and x_2.

Figure 1 shows the nonlinear effects of x_1 and x_2 on the additive predictor η_{1i} (solid line), along with the estimated effects obtained on one of the simulated dataset using both our method (dashed lines) and by restricting the additive predictor to be linear in x_1 and x_2 (dotdashed lines). For comparison purposes, also Bayesian latent class (BLCA) models are considered, that ignore the effects of x_1 and x_2 on the components' weights. For each dataset, we run the three algorithms setting the number of components equal to 2, 3 and 4. For all the simulated datasets, our method estimated the best model in terms of AICM (Raftery et al., 2007). In particular, the AICM suggests $G = 2$ for 95 datasets when an additive predictor with smooth effects is considered. We also fixed the right number of classes ($G = 2$) and investigated classification for each method, by comparing it to the true group membership. The average adjusted Rand index with our approach is 0.840, against a 0.797 by our competitors.

References

FRÜHWIRTH-SCHNATTER, S., & FRÜHWIRTH, R. 2010. Data augmentation and MCMC for binary and multinomial logit models. *Pages 111–132 of:* KNEIB, T., & TUTZ, G. (eds), *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*. Springer.

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.

JACOBS, R.A., JORDAN, M.I., NOWLAN, S.J., & HINTON, G.E. 1991. Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87.

LANG, S., & BREZGER, A. 2004. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.

RAFTERY, A.E., NEWTON, M.A., SATAGOPAN, J.M., & KRIVITSKY, P.N. 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Pages 1–45 of:* BERNARDO, J.M., BAYARRI, M.J., BERGER, J.O., DAWID, A.P., HECKERMAN, D., SMITH, A.F.M., & WEST, M. (eds), *Bayesian Statistics*, vol. 8. Oxford University Press.

SAMPLING PROPERTIES OF AN ORDINAL MEASURE OF INTERRATER ABSOLUTE AGREEMENT

Giuseppe Bove[1], Pier Luigi Conti[2] and Daniela Marella[1]

[1] Dipartimento di Scienze della Formazione, Università Roma TRE,
(e-mail: giuseppe.bove@uniroma3.it, daniela.marella@uniroma3.it)

[2] Dipartimento di Scienze Statistiche, Sapienza Università di Roma,
(e-mail: pierluigi.conti@uniroma1.it)

ABSTRACT: A measure of interrater absolute agreement was recently proposed capitalizing on the dispersion index for ordinal variables proposed by Giuseppe Leti. The new measure is not affected by restriction of variance problems and does not depend on the choice of a particular null distribution. In this presentation an unbiased estimator of such a measure is proposed and its variance is evaluated.

KEYWORDS: ordinal variables, interrater agreement.

1 Introduction

Ordinal rating scales are frequently developed in study designs where several raters (or judges) evaluate a group of targets. For instance, in language studies new rating scales before their routine application are tested out by a group of raters, who assess the language proficiency of a corpus of argumentative (written or oral) texts produced by a group of writers. The main interest is in analysing the extent that raters assign the same (or very similar) values on the rating scale (interrater absolute agreement), that is to establish to what extent raters evaluations are close to an equality relationship.

Bove *et al.*, 2018 proposed a new procedure to measure absolute agreement for ordinal rating scales by using the dispersion index proposed by Leti, 1983 (pp. 290-297). Such an index is given by

$$D = 2\sum_{k=1}^{K-1} F_k(1-F_k) \tag{1}$$

where K is the number of categories of the variable and F_k is the cumulative proportion associated to category k. The index D is nonnegative and it is easy to prove that $D = 0$ if and only if all the observed categories are equal (absence of dispersion). For a moderately large number of observations (N), the maximum of the index can be assumed equal to $D_{max} = (K-1)/2$ (all the observations are concentrated in the two extreme categories of the variable).Then, a measure of dispersion normalized in the interval $[0,1]$ is given by

$$d = \frac{D}{D_{max}}. \tag{2}$$

The present proposal is advantageous if compared to measures of absolute agreement in LeBreton & Senter, 2008 for two reasons. It does not depend by the formulation of a null distribution for normalization. It can never be out of the range $[0,1]$.

2 An unbiased estimator of Leti index

A sample of n_R raters and a sample of n_T targets are drawn by simple random sample without replacement. Let us denote by X_{ij} the score given by the jth rater to the ith target, for $j = 1,\dots,n_R$, $i = 1,\dots,n_T$. X_{ij}s are independent categorical random variables having K categories with $p_k^{(ij)} = P(X_{ij} = k)$, for $j = 1,\dots,n_R$, $i = 1,\dots,n_T$ and $k = 1,\dots,K$. In the sequel we assume that both the targets and the raters are homogeneus (*targets-raters homogeneity assumption*), this implies that the probability $p_k^{(ij)} = p_k$, for $j = 1,\dots,n_R$, $i = 1,\dots,n_T$ and $k = 1,\dots,K$.

As a consequence of *homogeneity assumption*, the variables X_{ij} are independent and identically distributed. As an estimator of d we consider

$$\widehat{d} = \frac{\widehat{\overline{D}}}{D_{max}} = \frac{1}{D_{max}}\left(\frac{1}{n_T}\sum_{i=1}^{n_T}\widehat{D}_i\right) \tag{3}$$

where $\widehat{D}_i$ is given by

$$\widehat{D}_i = 2\sum_{k=1}^{K-1}\widehat{F}_k^{(i)}(1-\widehat{F}_k^{(i)}) \tag{4}$$

and $\widehat{F}_k^{(i)}$ is the empirical cumulative distribution function computed on ith target.

Proposition 1 *The estimator $\widehat{d}$ has expectation*

$$E(\widehat{d}) = \left(1-\frac{1}{n_R}\right)d \tag{5}$$

and variance

$$Var(\widehat{d}) = \left(\frac{1}{D_{max}}\right)^2 \frac{V}{n_T} \tag{6}$$

where

$$\begin{aligned} V &= \left(\frac{1}{n_R^2} - \frac{1}{n_R^3}\right)(4\sigma^2 + 4(n_R - 2)J - 2(2n_R - 3)D^2) && (7) \\ \sigma^2 &= Var(X_{ij}) = \sum_{k=1}^{K} k^2 p_k - \left(\sum_{k=1}^{K} k p_k\right)^2 && (8) \\ J &= \sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{l=1}^{K} |k-h||k-l| p_k p_h p_l. && (9) \end{aligned}$$

As a consequence of Proposition 1, from (5) an unbiased estimator of d can be defined as follows

$$\widehat{d^*} = \left(\frac{n_R}{n_R - 1}\right)\widehat{d}. \tag{10}$$

3 An application on real data

In this section the ratings obtained in a research conducted at Roma Tre University are analyzed (Bove *et al.*, 2018). The aim of the study was to investigate the applicability of a six-point Likert scale for functional adequacy (an aspect of language proficiency) developed by Kuiken & Vedder, 2017 to texts produced by native and non-native writers in three different task types (narrative, instruction, and decision-making tasks). The scale comprises four subscales, corresponding to the four dimensions of functional adequacy identified by the authors of the scale: content, task requirements, comprehensibility, coherence and cohesion. In the study $n_R = 7$ raters evaluated the text produced by $n_T = 40$ targets: 20 native speakers of Italian (L1) and 20 non-native speakers of Italian (L2). For our purposes, we have selected ratings concerning only the narrative task and the subscale comprehensibility.

The results of the interrater agreement analysis for the subscale are summarized in Table 1, where the intraclass correlation $ICC(A,1)$ and the average values of r_{WG} defined as in LeBreton & Senter, 2008, the coefficient of variation CV, $\widehat{d}$ and $\widehat{d^*}$ are shown for L1, L2 and total groups. The intraclass correlation $ICC(A,1)$ provides a low-moderate level of agreement for the total group (0.67). The results for the average values of CV (12.16%), d (0.22) and d^* (0.25) seem in accord with $ICC(A,1)$, while the average value of r_{WG}

(0.87), highlights a higher level of agreement. When the analysis focuses separately on the two subgroups of L1 and L2 students, results regarding the L1 group deserve particular attention. Interrater agreement measured by intraclass correlation is very low in the L1 group ($ICC(A,1) = 0.14$). Analysing the dispersion of the ratings due to this subgroup, it comes out that most of the raters used almost exclusively levels 5 and 6 of the scale. Such a range restriction caused the very low value of the intraclass correlation, despite the substantial agreement among the raters that scored all the L1 texts in the same high levels. This problem does not regard the results for the other three indices of Table 3 ($r_{WG} = 0.90$; $CV = 8.12\%$; $\widehat{d} = 0.17$; $\widehat{d^*} = 0.19$) that show a very good level of absolute agreement.

Table 1. $ICC(A,1)$ and average of r_{WG}, CV, $\widehat{d}$ and $\widehat{d^*}$ for the comprehensibility subscale in the L1, L2 and the Total groups

Group	N	$ICC(A,1)$	r_{WG}	CV%	$\widehat{d}$	$\widehat{d^*}$
L1	20	0.14	0.90	8.12	0.17	0.19
L2	20	0.63	0.84	16.20	0.28	0.32
Total	40	0.67	0.87	12.16	0.22	0.25

References

BOVE, G., NUZZO, E., & SERAFINI, A. 2018. Measurement of interrater absolute agreement for the assessment of language proficiency. *In: S. Capecchi, Di Iorio F., Simone R. ASMOD 2018 : Proceedings of the Advanced Statistical Modelling for Ordinal Data Conference.* Università Federico II di Napoli, 24-26 October 2018. Napoli: FedOAPress, 61-68.

KUIKEN, F., & VEDDER, I. 2017. Functional adequacy in L2 writing. Towards a new rating scale. *Language Testing.*, **34**, 321–336.

LEBRETON, J.M., & SENTER, J.L. 2008. Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods.*, **11**, 815–852.

LETI, G. 1983. *Statistica descrittiva.* Bologna: Il Mulino.

TENSOR ANALYSIS CAN GIVE BETTER INSIGHT

Rasmus Bro[1]

[1] Department of Food Science, University of Copenhagen, (e-mail: rb@food.ku.dk)

ABSTRACT: Tensor analysis is also known as multi-way analysis. It allows analysing and visualizing more complex data than possible in multivariate analysis. Rather than being limited to matrices, tensor analysis allows analysing e.g. 'boxes' of data, called third-order tensors or three-way arrays. And the analysis extends easily to higher orders as well. There are several models and decompositions available for tensor data and they provide insights that are not possible to obtain with standard multivariate tools (Smilde, Bro et al. 2004).

For example, some models allow unique decomposition that completely obliviate the need for rotations towards simplicity because there is no rotational freedom whatsoever (Kruskal 1989). In particular, the PARAFAC (Harshman 1970) and PARAFAC2 (Harshman 1972) are interesting data analysis models with properties that allow solving otherwise impossible problems. For example, they allow making predictions of concentrations of chemical compounds where other methods would fail or allow to resolve completely mixed chemical signals. That is, they allow unscrambling scrambled eggs figuratively speaking.

In this presentation, we will showcase some of the interesting properties of tensor methods on a variety of data. We will mainly focus on chemical data such as fluorescence spectroscopic data for checking adulteration of food products or gas chromatography with mass spectrometry detection for untargeted chemical profiling of food products.

KEYWORDS: PARAFAC, PARAFAC2, Tucker, uniqueness.

References

HARSHMAN, A.M. 1970. Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis. *UCLA working papers in phonetics* **16**, 1-84.

HARSHMAN, A.M. 1972. PARAFAC2: Mathematical and technical notes. *UCLA working papers in phonetics*, **22**, 30-47.

KRUSKAL, J.B. 1989. Rank, decomposition, and uniqueness for 3-way and N-way arrays. *Multiway Data Analysis*. R. Coppi and S. Bolasco. Amsterdam, Elsevier: 8-18.

SMILDE, A.K., BRO, R. GELADI, P. 2004. *Multi-way analysis. Applications in the chemical sciences*, Wiley.

A BOXPLOT FOR SPHERICAL DATA

Davide Buttarazzi[1], Giuseppe Pandolfo[2], Giovanni C. Porzio[1]
and Christophe Ley[3]

[1] Department of Economics and Law, University of Cassino,
(e-mail: d.buttarazzi@unicas.it, porzio@unicas.it)

[2] Department of Industrial Engineering, University of Naples Federico II
(e-mail: giuseppe.pandolfo@unina.it)

[3] Department of Applied Mathematics, Computer Science and Statistics, Ghent University,
(e-mail: christophe.ley@UGent.be)

ABSTRACT: A boxplot for data lying on the surface of spheres is proposed. The notion of statistical depth function for directional data is adopted in order to extend the circular boxplot to spherical spaces.

KEYWORDS: Angular data depth, bagplot, graphical tool.

1 The spherical boxplot

The univariate box-and-whiskers plot (or simply the boxplot) introduced by Tukey (1977) is a well known graphical tool in exploratory data analysis. It was extended to the bivariate case by Rousseeuw *et al.* (1999), who introduced the “bagplot” by exploiting the notion of halfspace depth function (Tukey, 1975).

Here, after the boxplot for circular variables (Buttarazzi *et al.*, 2018), and in analogy with the bagplot of Rousseeuw *et al.* (1999) we propose a boxplot for spherical data which is based on the notion of angular depth function. Specifically, the angular Mahalanobis (Ley *et al.*, 2014) and angular Tukey’s depths (Liu & Singh, 1992) will be considered.

Drawing a spherical boxplot is a non-trivial task because of the peculiar features of spherical data. Spherical data arise in many scientific fields such as Earth sciences, biology, medicine and physics. They lay on the surface of a $(d-1)$-dimensional unit sphere in three dimensions, that is on $S^2 = \left\{x \in \mathbb{R}^3 : ||x|| = 1\right\}$, where $||x|| = \left(x^T x\right)^{1/2}$ is the usual L_2-norm of the vector x.

The center of the spherical boxplot will be given by the angular median corresponding to the depth function adopted (i.e., the point at which the depth is maximized). A bag containing the 50% of the data having highest depth values will be depicted. Fences will be obtained by enlarging the bag by a multiplying factor. Whiskers will be a bag including all the observed points lying within the fences area. Points outside the whiskers will be marked as far out values.

As with the univariate boxplot, the proposed spherical boxplot will allow displaying information on location, spread, and shape of a spherical distribution. Outliers may also be revealed.

For the aim of our work, we need to consider that: (*i*) the support of a spherical distribution is bounded, and hence the boxplot multiplying factor should be carefully chosen; (*ii*) a proper tool for spherical data must be rotationally invariant; (*iii*) the data spherical convex hull coincides with the whole sphere in case the data set does not lie within a hemi-sphere, and hence the extension of the ideas behind the bagplot should be carefully considered.

References

BUTTARAZZI, D., PANDOLFO, G., & PORZIO, G. C. 2018. A boxplot for circular data. *Biometrics*, **74**(4), 1492–1501.

LEY, C., SABBAH, C., & VERDEBOUT, T. 2014. A new concept of quantiles for directional data and the angular Mahalanobis depth. *Electronic Journal of Statistics*, **8**(1), 795–816.

LIU, R. Y., & SINGH, K. 1992. Ordering directional data: concepts of data depth on circles and spheres. *The Annals of Statistics*, 1468–1484.

ROUSSEEUW, P. J., RUTS, I., & TUKEY, J. W. 1999. The Bagplot: A Bivariate Boxplot. *The American Statistician*, **53**(4), 382–387.

TUKEY, J. W. 1975. Mathematics and the picturing of data. *Pages 523–531 of: Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, vol. 2.

TUKEY, J. W. 1977. *Exploratory Data Analysis*. Addison-Wesley.

MACHINE LEARNING MODELS FOR FORECASTING STOCK TRENDS

Giacomo Camba[1] and Claudio Conversano[1]

[1] Department of Economic and Business Sciences, University of Cagliari,
(e-mail: giacomo.camba@unica.it, conversa@unica.it)

ABSTRACT: This research addresses the problem of predicting the trends of two stocks and two stock indexes for the American stock market. In this study, the predictive performance of four machine learning models, are compared. The models investigated include Artificial Neural Networks (ANN), Support Vector Machine (SVM), Random Forest and Naive-Bayes. Supervised models training is performed through a 10-fold CV approach repeated 3 times, using 10 of the main indicators and oscillators of technical analysis as input. The experiments conducted show that among the 4, the Naive-Bayes model gives the worst predictive performance, the Random Forest obtains discrete results, while the SVM and the ANN are the best performing models.

KEYWORDS: machine learning, technical analysis, ann, svm, random forest.

1 Introduction

The task of predicting the evolution of stock prices and stock indexes is not easy, due to the uncertainty that characterizes this type of variables. Before buying or selling securities, analysts perform two types of analysis: fundamental analysis and technical analysis. In the fundamental analysis, the investment decision depends on the study of the variables referred to the intrinsic share value, such as the capital soundness, the ability to convert technology into value, the performance of the economic sector to which the company belongs, the political-economic climate, and so on. On the other hand, the technical analysis aims to determine the future share prices by studying the statistics generated by market activity, such as past prices and volumes. Technical analysts use stock charts and statistical tools to identify patterns, trends, cycles, which may suggest how the movement of stock price will behave in the future. The technical analysis is based on the Efficient Market Hypothesis (EMH) of Malkiel & Fama, 1970, according to which stock prices are an expression not only of fundamentals but also of all systemic variables. Therefore, if the information obtained from the prices with efficiently and appropriate algorithms is dealt, then it is possible to forecast the evolution of the stock prices and the stock indexes. For several years, in order to predict the stock performance many techniques have been developed and tested. Initially, the classic linear regression models were used, but over time more appropriate techniques such as non-linear machine learning methods were preferred (see Hastie *et al.* , 2017 for an overview). This research resumes general experimental setup that is found in the literature and it is inspired by the paper *Predicting stock and stock price index movement using Trend Deterministic Data Preparation and Machine Learning techniques* by Patel *et al.* , 2015. The goal of this experiment is to compare the forecast performances of Artificial Neural Networks, Support Vector Machine, Random Forest and Naive-Bayes Classifier on the time series of two stock indices and two stocks of the American stock market. Over ten years of data are used to compute ten technical parameters used as input for the aforementioned models. Both the securities and the indexes have a high trading volume, therefore they better express the general trading activity of the American market. The models are validated using a 10-fold Cross Validation approach repeated 3 times, which made it possible to find the best combination of parameters which minimize the forecast error. The final results showed not only the best performing models and the differences with respect to the less successful ones, but also how the predictive performance changes considerably depending on whether we consider stocks or indexes. The success or failure of a trading operation is related to the market timing and to the taken position, long or short. This work aims to help traders to move in the same direction of the market, identifying the moment in which to carry out a transaction.

2 Literature

In this section, we review some studies that focused on the application of statistical learning methods to financial time series data. Patel *et al.* , 2015 attempt to predict the direction of the trends of two stocks and two stock indices of the Indian Stock market. The study compares four prediction models, Artificial Neural Networks (ANN), Support Vector Machine (SVM), Random Forest and Naive-Bayes. Two different input approaches are presented. The first one involves the calculation of ten technical parameters using the daily trading data (opening prices, max price, min price, and closing price), while the second one consists in representing the technical parameters as deterministic

trend data. The authors evaluate the accuracy of each model with respect to the two input approaches. The assessment is carried out over 10 years of historical data, from 2003 to 2012, considering two securities, Reliance Industries and Infosys Ltd, and two stock indices, CNX Nifty and S&P Bombay Stock Exchange (BSE). The experimental results show that, when the ten technical indicators are used as continuous values, the Random Forest exceeds the other three models in terms of overall predictive performance. The research also shows that the performances of all 4 models improve significantly when the technical parameters are transformed into trend deterministic data.
Sezer *et al.* , 2017 propose a trading system in which a set of technical analysis parameters are optimized using genetic algorithms and subsequently are used as inputs of a MultiLayer Perceptron (MLP), whose outputs are buy-sell-hold signals. The model was trained on the historical series of daily stock prices belonging to the Dow 30 index for the period 1996-2016 and was subsequently tested between 2007-2016. The results suggest that the optimization of technical indicators not only improves trading performance but also provides an alternative model to other standard technical analysis approaches.
Moghaddam *et al.* , 2016 study the predictive ability of Artificial Neural Networks (ANN) on the NASDAQ stock index. Several feed-forward Neural Networks trained through the back-propagation algorithm were evaluated. The NASDAQ series was considered over a period of 100 days: the first 70 days (from 28/01/2015 to 7/03/2015) were considered as training set and the last 29 days were used to test the model forecasting ability. The authors experiment with different combinations of layers and numbers of hidden units, leading to configurations that show rather high predictive performance.

3 Results and Conclusion

We split our data into an in-sample period (*training set*, 10 yrs.) and a out-of-sample period (*test set*, 6 months), keeping the beginning *Up/Down* proposition and each model is selected using *10-fold Cross Validation* repeated 3 times. To evaluate the predictive performances, the *Accuracy, Sensitivity and Specificity* measures were used.

The experiments showed that the Naive-Bayes model performs worse than all, with an average accuracy of 68,15% on the training set and 58,93% on the test set. Support Vector Machines and Artificial Neural Networks showed the highest average performance, with an accuracy of 85.09% and 83,74% on the training set and 71,83% and 72,03% on the test set respectively. Whereas, Random Forest stood in the middle between the unlucky Naive-Bayes and the performing SVM and ANN, with an average performance of 80.33% on the training set and 69.45% on the test set. In general, all four models worked better for indexes than for securities.

Complete results (not tabulated) show that: (i) The SVM and ANN models, which on average perform better, have also a greater "horizontal" variability, i.e. the variability calculated between the estimates that each model has produced for each asset; (ii) There are so many cases of overestimation and underestimation and sometimes the differences are far from marginal. The differences in absolute value and in percentage between the average performances obtained for the in-sample and the out-of-sample periods w.r.t. Accuracy, Sensitivity, and Specificity give an idea of overestimation and underestimation errors on average. A further summary measure is represented by the average of the Accuracy percentage differences: $\overline{Err} = 0.1416$. On average, the models overestimate the accuracy, or underestimate the forecast error of 14.16%. Finally the average accuracy is computed on the training set and on the test set, considering the performances of the four models: $\overline{Accuracy_{TR}} = 79.33\%$; $\overline{Accuracy_{TE}} = 70.06\%$.

References

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. 2017. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer Texts in Statistics.

MALKIEL, B., G., & FAMA, E., F. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, **25**, 383–417.

MOGHADDAM, A., H., MOGHADDAM, M., H., & ESFANDYARI, M. 2016. Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, **21**, 89–93.

PATEL, J., SHAH, S., THAKKAR, P., & KOTECHA, K. 2015. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. *Expert Systems with Applications*, **42**, 259–268.

SEZER, O., B., OZBAYOGLU, M., & DOGDU, E. 2017. A Deep Neural-Network Based Stock Trading System Based on Evolutionary Optimized Technical Analysis Parameters. *Procedia Computer Science*, **114**, 473–480.

TREE MODELING ORDINAL RESPONSES: CUBREMOT AND ITS APPLICATIONS

Carmela Cappelli[1], Rosaria Simone[1] and Francesca Di Iorio[1]

[1] Department of Political Science, University of Naples Federico II,
(e-mail: carcappe@unina.it, rosaria.simone@unina.it,fdiiorio@unina.it)

ABSTRACT: The paper discusses a new technique for growing trees for ordinal responses in the model-based framework The class of CUB mixtures is considered which is particularly appropriate to model perceptions, judgments and evaluations, as it designs the response process as the combination of two components: a personal feeling which is related to the subject's motivations and it may be a direct measure of agreement, worry, satisfaction and an uncertainty component which expresses the inherent fuzziness of a discrete choice.

In the proposal, the partitioning process is based on the local estimation of CUB regression models to profile respondents according to feeling and uncertainty. Alternative splitting criteria which feature both inferential and fitting issues are implemented in the devoted R package which is illustrated showing how the chosen modelling framework also allows for advantageous visualization of the classification results. Various applications to real data from official surveys are presented.

KEYWORDS: Ordinal responses, Model based trees, CUBREMOT.

References

CAPPELLI, C., SIMONE, R & DI IORIO F. 2019. CUBREMOT: a tool for growing trees for ordinal response. *Expert systems with applications,***124**, 39-49.

SIMONE, R., CAPPELLI, C.,& DI IORIO F. 2019. Modelling marginal ranking distributions: the uncertainty tree. *Pattern Recognition Letters*, **25**, 278-288

ZEILEIS, A., HOTHORN, T & HORNIK,, K. 2008.Model-based recursive partitioning. *Journal of Computational and Graphical Statistics,***17***,* 492-514.

SUPERVISED LEARNING IN PRESENCE OF OUTLIERS, LABEL NOISE AND UNOBSERVED CLASSES

Andrea Cappozzo[1], Francesca Greselin[1] and Thomas Brendan Murphy[2]

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

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

ABSTRACT: Three important issues are often encountered in Supervised Classification: class-memberships are unreliable for some training units (Label Noise), a proportion of observations might depart from the bulk of the data structure (Outliers) and groups represented in the test set may have not been encountered earlier in the learning phase (Unobserved Classes). The present work introduces a Robust and Adaptive Eigenvalue-Decomposition Discriminant Analysis (RAEDDA) capable of handling situations in which one or more of the afore described problems occur. Transductive and inductive robust EM-based procedures are proposed for parameter estimation and experiments on real data, artificially adulterated, are provided to underline the benefits of the proposed method.

KEYWORDS: model-based classification, unobserved classes, label noise, outliers detection, impartial trimming, robust estimation.

1 Motivating Problem

In a standard classification framework a set of trustworthy learning data are employed to build a decision rule, with the final aim of classifying unlabelled units belonging to the test set. Therefore, unreliable learning observations can strongly undermine the classifier performance, especially if the training size is small. Additionally, the test set may include classes not previously encountered in the learning phase. For jointly overcoming these issues, we introduce a robust generalization of the AMDA methodology (Bouveyron, 2014) that accounts for outliers and label noise by detecting the observations with the lowest contributions to the overall likelihood employing impartial trimming (Gordaliza, 1991).

The rest of the paper is organized as follows: in Section 2 the notation is introduced and the main concepts about the model framework are summarized. Section 3 outlines the EM-based procedures proposed for parameter estimation. In Section 4 we employ the designed methodology in performing classification, adulteration detection and new class discovery in a food authenticity context of contaminated Irish honey samples.

2 RAEDDA Model

Consider $\{(\mathbf{x}_1,\mathbf{l}_1),\dots,(\mathbf{x}_N,\mathbf{l}_N)\}$ a complete set of learning observations, where $\mathbf{x}_n$ denotes a p-variate continuous outcome and $\mathbf{l}_n$ its associated class label, such that $l_{ng}=1$ if observation n belongs to group g and 0 otherwise, $g=1,\dots,G$. Further, denote $\mathbf{y}_m$, $m=1,\dots,M$ the set of unlabelled observations with unknown classes $\mathbf{z}_m$, where $z_{mc}=1$ if observation m belongs to group c and 0 otherwise, $c=1,\dots,C$. Note that only a subset $\mathcal{G}\subseteq\mathcal{C}$ of classes might have been encountered in the learning data, with $\mathcal{H}$ set of "hidden" classes in the test such that $\mathcal{C}=\mathcal{G}\cup\mathcal{H}$. Given a sample of N training and M test data, we construct a procedure for maximizing the *trimmed observed data log-likelihood:*

$$\begin{aligned}\ell_{trim}(\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{X},\mathbf{Y},\mathrm{l}) = &\sum_{n=1}^{N}\zeta(\boldsymbol{x}_n)\sum_{g=1}^{G}\mathrm{l}_{ng}\log\left(\tau_g\phi(\mathbf{x}_n;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\right)+\\ &+\sum_{m=1}^{M}\eta(\mathbf{y}_m)\log\left(\sum_{c=1}^{C}\tau_c\phi(\mathbf{y}_m;\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)\right)\end{aligned} \tag{1}$$

where $\phi(\cdot;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ represents the multivariate Gaussian density, τ_g denotes the probability of observing class g and $\zeta(\cdot)$, $\eta(\cdot)$ are 0-1 trimming indicator functions such that a fixed fraction α_l and α_u of observations, respectively belonging to the training and test data, is unassigned by setting $\sum_{n=1}^{N}\zeta(\boldsymbol{x}_n)=\lceil N(1-\alpha_l)\rceil$ and $\sum_{m=1}^{M}\eta(\mathbf{y}_m)=\lceil M(1-\alpha_u)\rceil$.

3 Estimation Procedure

Transductive and inductive EM-based procedures are proposed for parameter estimation and a robust model selection criteria is used for selecting the actual number of classes.

The transductive approach works on the union of learning and test sets: both samples are used to estimate model parameters. This mechanism would be equivalent to robust semi-supervised classification if $C=G$, but here we allow the procedure to also look for extra classes in the test.

The inductive approach consists of a robust learning phase and a robust discovery phase. The former performs a robust version of supervised discriminant analysis estimating model parameters for the known groups using only the training set. The latter assigns unlabelled observations to the known groups whilst searching for new classes; therefore, only the parameters for the $C-G$ extra classes need to be estimated.

In both approaches, we protect the parameter estimation from spurious solutions considering a restriction on the ratio between the maximum and the minimum eigenvalue of the group scatter matrices (Ingrassia, 2004).

4 Detect extra adulterant in samples of contaminated Irish Honey

We consider a dataset of Midinfrared spectroscopic measurements of 530 Irish honey samples recorded in the wavelength range of 3700 nm and 13600 nm (Kelly *et al.* , 2006). The experiment is carried out splitting observations in a training set composed by 145 pure honey and 60 beet sucrose adulterated samples; and a test set of 145 pure, 60 beet sucrose-adulterated and 120 dextrose syrup-adulterated honeys. In addition, 10% of beet sucrose adulterated units in the training set are wrongly labelled as pure honey. The final aim of the experiment is then three-fold: detect the wrongly labelled units in the training, discover the extra adulterant in the test and finally classify unobserved units to the correct class they belong.

The Adjusted Rand Index (Rand, 1971) is used to validate the classification accuracy in the test set for popular model-based classification methods: results for 50 random splits in training and validation are reported in Table 1. Clearly, methods that adapts to unobserved classes (i.e., AMDA and RAEDDA, estimated using either transductive or inductive approaches) display higher ARI, however the performance of AMDA is intensely affected by the presence of label noise in the learning set.

Table 1. *Adjusted Rand Index (ARI) computed on the test set for popular model-based classification methods: Eigenvalue Decomposition Discriminant Analysis (Bensmail & Celeux, 1996), Robust Mixture Discriminant Analysis (Bouveyron & Girard, 2009), Adaptive Mixture Discriminant Analysis via transductive and inductive approaches (Bouveyron, 2014), and the methods proposed in this article. Average results for 50 random splits in training and validation.*

	EDDA	RMDA	AMDAt	AMDAi	RAEDDAt	RAEDDAi
ARI	0.321	0.317	0.633	0.451	0.843	0.831

Our proposal successfully identifies the previously unseen adulterant as a hidden class and, furthermore, beet sucrose units erroneously labelled as pure honey in the training set are correctly detected by the impartial trimming 99.7% of the times in each scenario. That is, honeys that present label noise are not accounted for in the estimation procedure, enhancing the discriminating power of the classification rule.

Our methodology seems promising in effectively dealing with challenging supervised tasks, where both labelled and unlabelled units exhibit uncommon and hidden patterns. Particularly, as the application showed, practitioners involved in domains like food authenticity may benefit from the proposed approach. As a further research direction, a robust wrapper variable selection for dealing with high-dimensional problems is currently under development.

References

BENSMAIL, H. & CELEUX, G. 1996. Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, **91**(436), 1743–1748.

BOUVEYRON, C. 2014. Adaptive mixture discriminant analysis for supervised learning with unobserved classes. *Journal of Classification*, **31**(1), 49–84.

BOUVEYRON, C. & GIRARD, S. 2009. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, **42**(11), 2649–2658.

GORDALIZA, A. 1991. Best approximations to random variables based on trimming procedures. *Journal of Approximation Theory*, **64**(2), 162–180.

INGRASSIA, S. 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications*, **13**(2), 151–166.

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.

RAND, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846.

ASYMPTOTICS FOR BANDWIDTH SELECTION IN NONPARAMETRIC CLUSTERING

Alessandro Casa[1], José E. Chacón[2] and Giovanna Menardi[1]

[1] Dipartimento di Scienze Statistiche, Università degli studi di Padova,
(e-mail: casa@stat.unipd.it, menardi@stat.unipd.it)

[2] Departamento de Matemáticas, Universidad de Extremadura,
(e-mail: jechacon@unex.es)

ABSTRACT: In the framework of nonparametric clustering, clusters are defined as the domains of attraction of the modes of the density function assumed to underlie the data. To identify clusters, an estimate of the density is then needed, with kernel density estimator taking the lion's share. When resorting to these methods a fine tuning of the amount of smoothing, governing the modal structure of the density, is required. While thoroughly analyzed in the context of density estimation, this issue has been scarcely studied for clustering purposes. In this work the problem is addressed from an asymptotic perspective. A sensible distance among groupings is introduced and its asymptotic expression is derived and exploited in order to obtain a bandwidth selection procedure specifically tailored for nonparametric clustering.

KEYWORDS: modal clustering, kernel estimator, gradient bandwidth, mean shift.

1 Introduction

Density-based clustering pursues the aim of providing a statistical formalization to the widespread, yet ill-posed, problem of finding groups in a set of data. According to the nonparametric - or *modal* - formulation, clusters are seen as the domains of attraction of the modes of the density assumed to underlie the data, usually estimated by nonparametric methods. Linking the notion of cluster to the features of the underlying density frames the problem into a standard inferential context. As a consequence the concept of induced clustering, the partition implied by the characteristics of the density itself, is defined with the ideal population clustering being the one induced by the true density.

Regardless of the specific nonparametric density estimator adopted, the selection of a smoothing parameter is required. This choice represents a relevant issue since under- or over-smoothed estimates may lead to deceiving indications about the modal structure of the density, and hence about the number of groups.

The selection of the amount of smoothing is usually addressed via the minimization of some measure of distance which quantifies the discrepancy between the estimate and the target density. Standard references are the *Integrated Squared Error* and its expected value (MISE), or its asymptotic counterpart. While for the explicit task of density estimation, the distance criterion is usually selected to provide good estimates in a global sense, the same may be suboptimal in a clustering framework, where a focus on the local characteristics of the density would be more adequate to identify the modal regions.

The aim of this work is to address the problem of nonparametric density estimation for the final purpose of modal clustering. Density estimation is performed via the minimization of an appropriate metric relying on the comparison between the partitions induced by the estimated distribution and the true one, i.e. the ideal population clustering. A manageable asymptotic approximation of the considered metric is provided, which allows to define the optimal amount of smoothing for nonparametric clustering when a kernel estimator is adopted.

2 Optimal bandwidth for the asymptotic distance in measure

Let us assume that the observed data $\mathcal{X} = \{x_i\}_{i=1,\dots,n}$, are sampled from a random variable $\mathbb{X}$ with unknown density f. For mathematical tractability, in the following we restrict our attention to the univariate case, i.e. $x_i \in \mathbb{R}$.

A standard choice to estimate f is to resort to the kernel estimator

$$\hat{f}_h(x) = (1/nh)\sum_{i=1}^{n} K[(x-x_i)/h]$$

where K is a kernel function and $h > 0$ is the bandwidth which controls for the amount of smoothing and, then, the modal structure.

To tailor the choice of h for clustering purposes, we consider the *distance in measure* (Chacón, 2015) between $\hat{\mathcal{C}}_h = \{\hat{C}_1, \dots, \hat{C}_r\}$, the clustering induced by $\hat{f}_h$, and $\mathcal{C}_0 = \{C_{0,1}, \dots, C_{0,s}\}$, the ideal population one, induced by the true f:

$$d(\hat{\mathcal{C}}_h, \mathcal{C}_0) = \frac{1}{2}\min_{\sigma\in\mathcal{P}_s}\left\{\sum_{i=1}^{r}\mathbb{P}(\hat{C}_i \Delta C_{0,\sigma(i)}) + \sum_{i=r+1}^{s}\mathbb{P}(C_{0,\sigma(i)})\right\}, \qquad (1)$$

where $\mathcal{P}_s$ is the set of permutations of $\{1,\dots,s\}$, $C\Delta C_0 = (C\cap C_0^c)\cup(C^c\cap C_0)$ and with possibly $r \le s$. This distance can be seen as the minimal probability mass that needs to be moved to transform one clustering into the other. Being sample-specific, the distance in measure is subject to a random variability.

Hence, the *Expected Distance in Measure* $\mathrm{EDM(h)} = \mathbb{E}[d(\hat{\mathcal{C}}_h, \mathcal{C}_0)]$ is alternatively considered as a non-stochastic error distance. The optimal bandwidth is then defined as $h_{EDM} = \mathrm{argmin}_{\mathrm{h}>0}\mathrm{EDM(h)}$.

Under some regularity assumptions, it can be proved (Casa *et al.*, 2019) that $\mathrm{EDM(h)}$ is asymptotically equivalent to

$$\mathrm{AEDM}(h) = \sum_{j=1}^{r-1} \frac{f(m_j)}{f^{(2)}(m_j)} \psi\Big(\frac{1}{2}\mu_2(K) f^{(3)}(m_j) h^2, R(K^{(1)}) f(m_j)(nh^3)^{-1}\Big) \quad (2)$$

where $\psi(\mu,\sigma^2) = (2/\pi)^{1/2}\Big\{\sigma e^{-\mu^2/(2\sigma^2)} + |\mu| \int_0^{|\mu|/\sigma} e^{-z^2/2} dz\Big\}$, m_j is the j^{th} local minimum of f, $g^{(l)}$ denotes the l^{th} derivative of a function g, $\mu_2(K) = \int_{-\infty}^{\infty} x^2 K(x) dx$, and $R(K^{(1)}) = \int_{-\infty}^{\infty} K^{(1)}(x)^2 dx$.

Since neither the $\mathrm{EDM(h)}$ nor the $\mathrm{AEDM(h)}$ admit an explicit representation of their minima, the idea is to rely on a tight upper bound. The study of the behaviour of $\psi(\cdot,\cdot)$ allows us to introduce two different upper bounds, whose minimizers can be computed explicitly. It follows that

$$\begin{aligned} h_{AB1} &= \left(\frac{9R(K^{(1)}) \left(\sum_{j=1}^{r-1} f(m_j)^{3/2}/f^{(2)}(m_j)\right)^2}{2\pi\mu_2(K)^2 \left(\sum_{j=1}^{r-1} f(m_j)|f^{(3)}(m_j)|/f^{(2)}(m_j)\right)^2} \right)^{1/7} n^{-1/7} \\ h_{AB2} &= \left(\frac{24R(K^{(1)}) \sum_{j=1}^{r-1} f(m_j)^{3/2}/f^{(2)}(m_j)}{11\mu_2(K)^2 \sum_{j=1}^{r-1} f(m_j)^{1/2} f^{(3)}(m_j)^2/f^{(2)}(m_j)} \right)^{1/7} n^{-1/7} . \end{aligned}$$

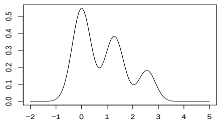
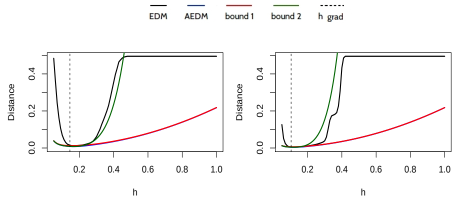
Note that, since the derived bandwidths are depending on some unknown quantities, from an operational point of view we need to resort to plug-in strategies.

3 Some results and conclusions

In this section we present an excerpt of the numerical results obtained in one-dimensional setting in order to evaluate the performances of the proposed selectors as well as the quality of the introduced asymptotic approximations.

The top panel of Table 1 shows the quality of the derived approximations to the EDM, as a function of the bandwidth, when all the involved quantities are known. The approximations improve as the sample size increases and they appear to behave satisfactorily especially around the value of h minimizing the EDM. In the bottom panel we can see the results, in terms of EDM, of the data-based bandwidth selectors over $B = 1000$ synthetic samples, along with

Table 1. *Top panel: true density (left); EDM, AEDM and the bounds vs h for* $n = 1000, 10000$ *(middle and right panels). The vertical dashed line is associated to the gradient bandwidth. Bottom panel: EDM estimates (and standard errors) at the optimum h according to the AEDM, the two bounds, and the gradient bandwidth.*

	n=1000	n=10000
Density	EDM approximation	
		
	DM estimate	
$\hat{h}_{AEDM}$	0.015 (0.018)	0.005 (0.003)
$\hat{h}_{AB1}$	0.013 (0.010)	0.005 (0.003)
$\hat{h}_{AB2}$	0.014 (0.011)	0.005 (0.003)
$\hat{h}_{GRAD}$	0.013 (0.009)	0.005 (0.003)

the performances of the gradient bandwidth, representing a sensible competitor in this framework, obtained via MISE minimization. The proposed selectors $\hat{h}_{AB1}$ and $\hat{h}_{AB2}$ led to more accurate clusterings than h_{AEDM}, with a slight preference for the former. The gradient-based bandwidth, in turn, not only produces competitive results, but its Monte Carlo average distance in measure appears lower than the one produced by the asymptotic EDM minimizers. In fact, a deeper insight into the standard errors of the obtained distances shows that $\hat{h}_{AEDM}$, as well as $\hat{h}_{AB1}$ and $\hat{h}_{AB2}$, produce more variable results, due to a higher sensitivity of the minimizers to the plugged in pilot estimates.

For a complete exposition of the results, alongside with a multivariate generalization, see Casa *et al.*, 2019.

References

CASA, A., CHACÓN, J.E., & MENARDI, G. 2019. Modal clustering asymptotics with applications to bandwidth selection. *arXiv preprint arXiv:1901.07300.*

CHACÓN, J.E. 2015. A population background for nonparametric density-based clustering. *Statistical Science*, **30**(4), 518–532.

FOREIGN IMMIGRATION AND PULL FACTORS IN ITALY: A SPATIAL APPROACH

Oliviero Casacchia[1] Luisa Natale[2] and Francesco Giovanni Truglia[2]

[1] Department of Statistical Science, Sapienza University of Rome,
(e-mail:oliviero.casacchia@uniroma1.it)

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

[3] Istat, (e-mail: truglia@istat.it)

ABSTRACT: Significant changes have affected currently internal mobility in Italy. We try to understand what are the variables that allow a place to attract population. This work focuses on the foreign population and aims to detect the factors that push immigrant population towards Italian municipalities. We want to verify whether the action is different between movements of foreigners already resident in Italy and of immigrants coming directly from abroad. Data on flows, stock of populations and socioeconomic variables on Italian municipalities from Istat, Ministry of Economy and Sole 24 were exploited. Methods used are regression analyses enriched with spatial factors with reference to the possible action of spatial variables through the building of OLS, spatial lag and spatial error models.

KEYWORDS: Foreign immigration, pull factors, internal mobility, regression, spatial analysis.

1 Background and aim

Significant changes have affected the current internal mobility in Italy. Foreign immigration, the repopulation of internal or marginal areas are important phenomena that may have played a role in the capacity of an area to attract population. We try to understand what are the variables that allow a place to attract population. Some results of a previous work (Natale, Santacroce, Truglia, 2016) show an unexpected absence of a link between the "attraction" variables identified for Italians and also those designed for foreigners. The reasons that lead natives (Italian citizens) to move within the country seem different from those of immigrants (Foreign citizens). This work focuses on the foreign population and aims to detect the capability of the foreign population already resident in the Italian municipalities to attract further flows of immigrants originated either from other municipalities or directly coming from abroad. In other words the paper tries to detect the factors underpinning the *network effect* due to foreign population resident in Italy.

2 Materials and methods

We analyze in this first phase only four Italian regions: Piedmont, The Marches, Apulia and Calabria[1]. Three sources of data are used. We consider the data concerning foreigners enrolled in the municipality population registry (demographic balance data supplied by Istat, the Italian Statistical Institute) in the about 2000 municipalities of the four regions examined. We took into account both the series recently made available by Istat, beginning with the Census data (*8mila Census*), and statistics on per capita income obtained from studies carried out by the Ministry of Economy and Finance.

We first calculated the foreign immigration rate FR_i for a generic municipality *i* observed in the years 2012-2014. The rate is defined as:

$$FR_{i,12\text{-}14}= (F_{i,2012} + F_{i,2013} + F_{i,2014})/3 \ / \ (FP_{i,1.1.2012} + FP_{i,31.12.2014})/2$$

where F_i is the sum of foreign inflows coming from other municipalities or from abroad, FP is the foreign resident population. Then we calculated two further measures: *internal* (regarding flows of foreigners resident from other Italian municipalities) and *external* (foreigners from abroad) foreign immigration rates (respectively, IFR_i and EFR_i).

In order to detect the effect of various factors and patterns of spatial association, autoregressive models are used (Anselin, 1988 and 1995). In particular, OLS, spatial lag and spatial error models are estimated. In this paper only the results concerning this strategy of analysis are showed. Anyway the results obtained with the second model are quite similar.

3 Main Results

In the four Regions the *internal* immigration rates are not so different (around 4-6%: see Table 1).

Table 1 Total, internal and external foreign immigration rates (%) in Piedmont, The Marches, Apulia and Calabria. Italy, 2012-2014.

Regions	Total Immigration Rate	Internal Immigration Rate	External Immigration Rate
Piedmont	11,5	6,3	5,2
The Marches	11,3	5,9	5,4
Apulia	14,8	5,0	9,9
Calabria	13,7	3,8	9,9
Total Regions	12,2	5,8	6,4

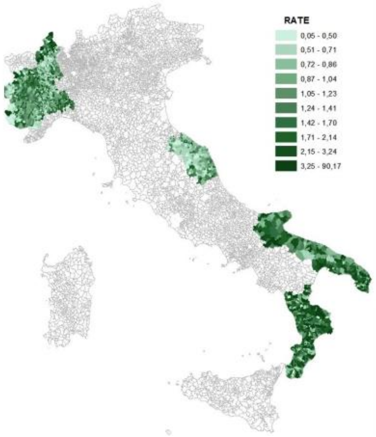
Source: own elaboration based upon Istat Resident Population Balance

A less capability of the resident foreign population to pull further flows coming from the rest of the country (the rate is equal to 3,8%) clearly emerged in Calabria.

[1] The four areas were chosen for the sake of comparison with a previous research conducted in the same Regions with reference to the attraction of Italian population: see Natale, Santacroce and Truglia (2016).

Apulia, the other Southern Region, shows low level of IFR as well. Regarding the external pull force the situation is reversed: EFR is higher in the Southern Regions (about 10% in the three years examined), well above the rates observed in the Centre-North Regions (value around 5%)[2]. Total immigration rates are clearly influenced by this different pattern: rates range from 11.3 (The Marches) to 14.8 (Apulia)[3]. Concerning the rate observed at municipal level the variability is very high, showing a surprising range of very different levels inside the same Region (see Map 1).

Map 1 Total Foreign Immigration Rate by municipalities. Italy, 2012-2014.



The autoregressive models are used prove to be useful for the analysis of the factors underlying a high or low attraction capacity of the foreign population in the municipalities chosen. The results of the preliminary analyses seem to suggest adoption both of a model with lag of the variables used and a model with autoregressive spatial disturbances (Table 2). This results are not new in the literature (see, for instance, Cracolici *et al*, 2009; Arbia, 1993; Truglia, 2013).

Some relevant variables are associated with IFR and EFR. Percentage of foreign population has a negative association with the pull force of the municipalities. It seems to be an evidence of the existence of a scarce network effect: in other words

[2] In the two Southern Regions a slight increasing presence of immigrants from African continent is observed in the 2012-2014 period: in Italy the percentage increased by 12%, in Apulia and Calabria by 25%. However to include this effect in the model didn't significantly improved the results obtained. The higher EFR in Apulia and Calabria could be also linked to the capacity of foreign population already resident in Italy to attract other components of the household from abroad. This capacity is inversely linked to the duration of stay in the country of arrival. In the Southern Regions the percentage of long sojourn residents is lower than in the Piedmont and The Marches.

[3] Regarding Italian resident population it is important to note that both the rates are below the levels observed for foreign population: nearly zero concerning the flows of Italians from abroad, more or less one third with reference to the internal migration.

municipalities with an high percentage of foreign population exert a weak force in attracting foreign flows. This could be in accordance with a theory of spatial assimilation in which foreigners tend to disperse in the territory. Unemployment has a negative effect only with reference to EFR[4]. It is interesting to note that some variables act in a different way on the two mobility measures used: in the areas with high percentage of high percentage of poor household the IER is low, the contrary happened with reference to the EFR. It is important to say that the presence of a neighbour effect emerged in the models considering spatial effects. These effects are obviously neglected by using OLS model.

Table 2: Test to determine the goodness of the model.

Test	Internal Migratory Rate		External Migratory Rate	
	Statistic	Sig.	Statistic	Sig.
Moran's I (error)	5,10	0,00	8,90	0,00
Lagrange Multiplier	17,41	0,00	55,53	0,00
Robust LM (lag)	0,43	0,51	2,91	0,09
Lagrange Multiplier (error)	22,71	0,00	72,76	0,00
Robust LM (error)	5,74	0,02	20,14	0,00
Lagrange Multiplier (SARMA)	23,14	0,00	75,67	0,00

In sum the level of the attractiveness of the foreign population in Italy is linked to the levels of the municipalities around, so that this pull force tends to be clustered in the Regions used. An extension of these results to the whole nation could lead to further interesting results.

References

ANSELIN, L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer Academic, Dordrecht.

ANSELIN L. (1995), Local Indicators of Spatial Association – LISA, *Geographical Analysis*, 27, 93-115.

ARBIA G. (1993), Recenti sviluppi nella modellistica spaziale, in S. ZANI (ed.), *Metodi statistici per le analisi territoriali*, Franco Angeli, Milano.

CRACOLICI M.F., CUFFARO M., & NIJKAMP P. (2009), A spatial analysis on Italian unemployment differences, *Statistical Methods and Applications*. 18(2), 275-291.

NATALE, L., SANTACROCE, A., & TRUGLIA, F.G. (2016), *Native immigration and pull factor evolution in Italy: a spatial approach*, communication at Parallel Specialized Session SPE-06 Spatial Analyses in Demography, SIS- Fisciano (Italy), 9 June.

TRUGLIA F.G. (2013), L'Italia incantata. Geo-statistica della diffusione dell'astensionismo, elezioni politiche 2008 e 2013, *Sociologia e Ricerca Sociale*, 101, 61-90.

[4] For sake of brevity the model is not reported here. Other information can be requested directly to the authors.

DIMENSIONALITY REDUCTION VIA HIERARCHICAL FACTORIAL STRUCTURE

Carlo Cavicchia[1], Maurizio Vichi[1] and Giorgia Zaccaria[1]

[1] Department of Statistical Sciences, University of Rome La Sapienza,
(e-mail: carlo.cavicchia@uniroma1.it,maurizio.vichi@uniroma1.it,
giorgia.zaccaria@uniroma1.it)

ABSTRACT: Manifold multidimensional concepts are explained via a tree-shape structure by taking into account the nested hierarchical partition of variables. The root of the tree is a general concept which includes more specific ones. In order to detect the different specific concepts at each level of the hierarchy, we can identify two different features regarding groups of variables: the internal consistency of a concept and the correlation between concepts. Thus, given a data positive correlation matrix, we reconstruct the latter via an ultrametric correlation matrix which detects hierarchical concepts by looking for their internal consistency and the correlation between them measured by relative indices.

KEYWORDS: ultrametric matrix, hierarchical latent concepts, correlation matrix, partition of variables.

1 Introduction

Many relevant multidimensional phenomena are represented via a tree-structure (for example well-being, sustainable development, poverty, climate change). We can hypothesize a Dimensionality Reduction model with a hierarchical structure that goes from disjoint sets of Manifest Variables (MVs) to the General Concept (GC). In other words we build a parsimonious hierarchy of classes of variables starting from a reduced number, (i.e., latent dimensions) which measure specific concepts describing the main components of the phenomenon under study up to the definition of the GC. Each cluster of MVs may be related with a factor which best represents its dimension. This is not new in many fields of research, for instance Revelle (1979) introduced a hierarchical cluster analysis method very useful to detect clusters of variables in a hierarchical approach. Our proposal can be considered into the Dimensionality Reduction framework for its ability of summarizing a big quantity of information by way of many steps of aggregation. In order to detect the hierarchy of variables, i.e., the different specific concepts at each level of the hierarchy, we identify two different features regarding clusters of variables: the internal consistency (i.e., reliability of the concept) and the correlation between concepts. Thus, given a data correlation matrix, we reconstruct the latter via an ultrametric correlation matrix which detects hierarchical concepts with the highest internal consistency and with the highest correlation between them in order to justify their fusion. The *internal consistency of concept* (i.e., variable cluster), is the global consistency of MVs based on their correlations within cluster. This is also called internal reliability and it is commonly measured by Cronbach's alpha (Cronbach, 1951). On the one hand, the reliability is connected to the concept of unidimensionality, which, on the other hand, evaluates to what extent a single latent indicator has been measured with a set of MVs. Reliability and unidimensionality are more realistic for specific dimensions, whereas, when considering a general factor, we have to hypothesize the presence of a GC (Cavicchia & Vichi, 2019). A common error is to interpret a measure of reliability as a measure of unidimensionality. Although being connected, they cannot consider as the same thing. Unidimensionality involves the homogeneity of a set of items, and internal consistency is certainly necessary for homogeneity, but it is not sufficient. We can see that, therefore, the improving of the internal consistency leads to an improvement of unidimensionality as well, but we cannot use the same index to measure both. By supposing that no variable can belong to two clusters at the same time, such that, all the clusters are disjoint at each level, we can consider another important feature which is *the correlation between clusters of variables*. This latter represents a function of the pairwise relationships between the items of the two groups and determines the bottom-up agglomerations of variable clusters. Hence, we are supposing a nested hierarchy where, starting from Q clusters of variables, all the possible combinations are taken into consideration in order to identify the aggregations which best detect reliable concepts at all levels.

2 Internal Consistency and Correlation Between

2.1 A Measure of Internal Consistency

The internal consistency of a cluster of MVs is the ability of all variables to measure the same latent concept. It is usually measured by indices based on the correlations between the MVs within the cluster. Many measures of internal consistency are reviewed by Revelle & Zinbarg (2009). In our framework, by starting from Q variable clusters at the bottom level, we have $\frac{Q(Q-1)}{2}$ clusters along the hierarchy, and as many internal consistency indexes. For each level

$q = Q, \ldots, 1$, the $(J \times q)$ membership matrix $\mathbf{V}_q$, where J is equal to the total number of MVs, tells us for each cluster which variable belongs to. Given $\mathbf{V}_q$, Cavicchia *et al.* (2019) proposed a measure of internal consistency for non-negative data correlation matrices, arranged in a $(q \times q)$ diagonal matrix as follows:

$$\widehat{\mathbf{R}}_q^W = diag\big(dg(\mathbf{V}_q'(\mathbf{R}-\mathbf{I}_J)\mathbf{V}_q)\big)[(\mathbf{V}_q'\mathbf{V}_q)^2 - \mathbf{V}_q'\mathbf{V}_q]^{-1}. \qquad (1)$$

In Eq. 1 $\mathbf{R}$ represents the $(J \times J)$ observed correlation matrix and $\mathbf{I}_J$ is the identity matrix of order J; furthermore $dg(\cdot)$ produces a vector whereas $diag(\cdot)$ builds a diagonal matrix. It is important to notice that $\widehat{\mathbf{R}}_q^W$ has q non-zero elements which are the internal consistency measures, one for each cluster. $\widehat{\mathbf{R}}_q^W$ corresponds to the Least Squares solution for reconstructing $\mathbf{R}$ via an ultrametric correlation matrix composed by a matrix which explains the internal consistency of concepts and a matrix which explains the correlation between concepts. Each value ${}_W\widehat{r}_{ll}$ $(l = 1, \ldots, q)$ of $\widehat{\mathbf{R}}_q^W$ belongs to the interval $[0,1]$, recalling that $\mathbf{R}$ has all non-negative values, thus it may be considered as a relative index. An important characteristic of the values of $\widehat{\mathbf{R}}_q^W$ is that they are not function of the number of MVs of each cluster, thus they are not affected by the size of clusters.

2.2 A Measure of Correlation Between Clusters of Variables

In order to detect all the levels of the hierarchy, it is crucial to define the correlation between clusters of MVs, each one representing a latent concept.
For each level $q = Q, \ldots, 1$ it is possible to compute the correlation between clusters of variables, and the internal consistency within clusters as well, but it is important to stress the fact that the Q-level (i.e., the level with Q variable clusters at the bottom level) is the optimal one in order to reconstruct the data correlation matrix $\mathbf{R}$. Given $\mathbf{V}_q$ and the diagonal matrix of internal consistency measures $\widehat{\mathbf{R}}_q^W$, Cavicchia *et al.* (2019) proposed a measure of correlation between clusters of MVs for non-negative data correlation matrices, arranged in a $(q \times q)$ correlation matrix as follows:

$$\widehat{\mathbf{R}}_q^B = (\mathbf{V}_q'\mathbf{V}_q)^{-1}\mathbf{V}_q'\bar{\mathbf{R}}\mathbf{V}_q(\mathbf{V}_q'\mathbf{V}_q)^{-1}. \qquad (2)$$

In Eq. 2, $\bar{\mathbf{R}} = \mathbf{R} - \mathbf{V}_q'\widehat{\mathbf{R}}_q^W\mathbf{V}_q + diag\big(dg(\mathbf{V}_q'\widehat{\mathbf{R}}_q^W\mathbf{V}_q)\big) - \mathbf{I}_J + \mathbf{V}_q'\mathbf{I}_Q\mathbf{V}_q$. The off-diagonal values within $\widehat{\mathbf{R}}_q^B$ are the between-concepts correlation whereas the diagonal elements are equal to one. $\widehat{\mathbf{R}}_q^B$ is the LS solution with respect to the matrix which explains the correlation between concepts. As for $\widehat{\mathbf{R}}_q^W$, each value ${}_B\widehat{r}_{kf}$ $(k = 1,\ldots,q;\ f = 1,\ldots,q;\ k \neq f)$ of $\widehat{\mathbf{R}}_q^B$ belongs to the interval $[0,1]$ and it turns out to be a relative index.

3 Conclusions

A correlation matrix **R** may be reconstructed via a ultrametric hierarchical structure which highlights two crucial characteristic regarding clusters of variables: the internal consistency and the correlation between clusters. In order to detect the ultrametric structure of the latent concepts, it is important to investigate in depth the reliability of each cluster of MVs and all the relations among them. For correlation matrices **R** which are composed only by non-negative elements, as common in psychometric applications, Cavicchia *et al.* (2019) presented a model that considers two main matrices, the first one which contains non-zero element only on the diagonal, that is the internal consistency measure for the related cluster, and the second one which is a correlation matrix with the off-diagonal elements that represent the correlation between clusters. The Dimensionality Reduction model with a hierarchical structure that goes from disjoint sets of Manifest Variables (MVs) to the General Concept (GC) is given by detecting consistent clusters and by following correlation between them.

References

CAVICCHIA, C., & VICHI, M. 2019. Statistical Model-based Composite Indicators for Complex Socio-Demographic and Economic Phenomena: Models, Properties and Features for tracking coherent policy conclusions. *Submitted, Social Indicators Research.*

CAVICCHIA, C., VICHI, M., & ZACCARIA, G. 2019. The Ultrametric Correlation Matrix for Modelling Hierarchical Latent Concepts. *Submitted, Advances in Data Analysis and Classification.*

CRONBACH, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika.*, **16**, 297–334.

REVELLE, L. J. 1979. Hierarchical Cluster Analysis and the Internal Structure of Tests. *Multivariate Behavioral Research.*, **14**, 57–74.

REVELLE, L. J., & ZINBARG, R. 2009. Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika.*, **74**(1), 145–154.

LIKELIHOOD-TYPE METHODS FOR COMPARING CLUSTERING SOLUTIONS

Luca Coraggio[1] and 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 solution is a longstanding problem. In model-based clustering this amounts to choose the architecture of the model mixture distribution. Decisions to be made pertain to: cluster prototype distribution; number of mixture components; (optionally) restrictions on the clusters' geometry. Classical proposals address this issue via penalized model selection criteria based on the observed likelihood function. In this study, we compare these techniques with the less explored cross-validation alternative, which is rather popular for many other data-driven optimized methods. We analyze both classical methods such as BIC, AIC, AIC3 and ICL, and several cross-validation schemes where the risk is defined in terms of minus the log-likelihood function. Selection methods are compared by using the Iris dataset.

KEYWORDS: model based clustering, model selection, penalized likelihood, cross-validation.

1 Introduction

In model-based clustering it is assumed that data are generated from a finite mixture distribution with density $f(\cdot\,;\,\theta) = \sum_{k=1}^{K} p_k f_k(\cdot\,;\,a_k)$, where $\theta = (p_1,\cdots p_K, a_1,\cdots,a_K)$, is the unknown parameter vector. Here f_k are densities representing the k-th *cluster*, $0 \leq p_k \leq 1$'s are mixing proportions, so that $\sum_{k=1}^{K} p_k = 1$, a_k is the parameter vector describing the cluster shape under f_k. Henceforth, $f_k(\cdot\,;\,a_k)$ is the Gaussian density with mean μ_k, and covariance matrix Σ_k, thus $a_k = (\mu_k, \Sigma_k)$. The definition of a member of the set of candidate models $\mathcal{M}$ requires: (i) definition of K, (ii) eventually a parameterization for the covariance matrices Σ_k. Let $a_{k,h} = (\mu_k, \Sigma_{k,h})$ be the parameters of the k-th component according to a certain parameterization h of the covariance structure. Celeux & Govaert, 1995 proposed to decompose $\Sigma_{k,h}$ into parameters describing geometrical notion of clusters' volume, orientation, and shape to reproduce different levels of model complexity.

Let $\theta(m)$ be the parameter vector representing a candidate model $m \in \mathcal{M}$. In most situations, the practice is : (i) estimate each member of $\mathcal{M}$ based on maximum likelihood (ML); (ii) choose a model m^*, and its implied clustering, based on some optimality notion. In the context of Gaussian model-based clustering the choice of m^* is typically performed by optimizing an information-theoretic statistic, based on the log-likelhood function. Section 2 introduces these and other methods. Section 3 provides a comparison of on real data.

2 Methodology

Let $X_n = \{x_1, \cdots, x_n\}$ be a sample of n data points. Let z_{ik} be the unobserved assignment, where $z_{ik} = 1$ if x_i belongs to the k-th cluster and 0 otherwise. Let $K(m)$ and $h(m)$ the values of K and h according to $m \in \mathcal{M}$. Define

$$l(\theta(m)) = \sum_{i=1}^{n} \sum_{k=1}^{K(m)} \log(p_k f_k(x_i, a_{k,h(m)})) \tag{1}$$

$$cl(\theta(m)) = \sum_{i=1}^{n} \sum_{k=1}^{K(m)} z_{i,k} \log(p_k f_k(x_i, a_{k,h(m)})) \tag{2}$$

where $l(\cdot)$ is the sampling log-likelihood function under m, and $cl(\cdot)$ is the so called complete log-likelihood function. Let $\hat{\theta}(m)$ the ML estimate of $\theta(m)$, and let $\hat{z}_{i,k}$ be the maximum a posteriori estimates of $z_{i,k}$. Replacing $\hat{\theta}(m)$ and $\hat{z}_{i,k}$ into (1) and (2) the corresponding sample estimates $\hat{l}(m)$ and $\hat{cl}(m)$ are obtained. Let ν_m be the number of free parameters in the model m, where ν_m increases with both $K(m)$, and the number of parameters required by the covariance parametrization $h(m)$. We now introduce sampling approximations of the *Bayesian Information Criterion* (BIC) of Schwarz, 1978, the *Akaike Information Critirion* (AIC) of Akaike, 1973, the modified version of the AIC (AIC3) of Bozdogan, 1983 and the *Integrated Complete Likelihood Criterion* (ICL) of Biernacki *et al.* , 2000. They are defined as:

$$AIC(m) = 2\hat{l}(m) - 2\nu_m, \qquad BIC(m) = 2\hat{l}(m) - \log(n)\nu_m,$$
$$AIC3(m) = 2\hat{l}(m) - 3\nu_m, \qquad ICL(m) = 2\hat{cl}(m) - \log(n)\nu_m.$$

A model m^* is selected in order to maximize one of the previous quantities. These criteria, although derived from different perspectives, have all the following form: "log-likehood at the MLE − penalty", where the penalty increases with model complexity, and sometimes decreases with n.

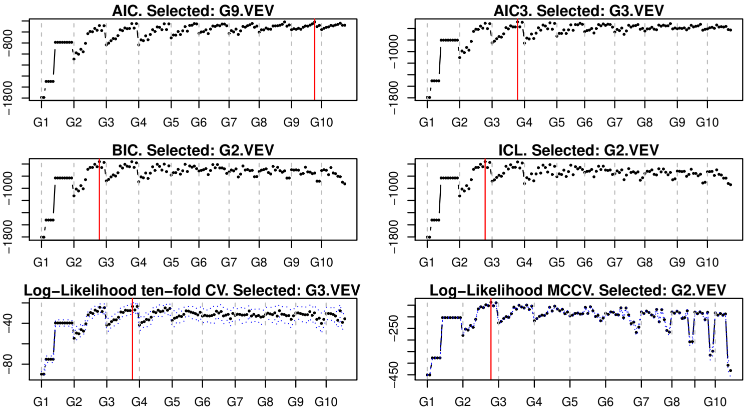


Figure 1. *x-axes show models $m \in \mathcal{M}$, ordered in terms $K(m)$ first, and then by the number of free parameters required by the covariance structure $h(m)$ (increasing complexity). E.g.* "G2" *means $K(m) = 2$. For CV plots, 95%-confidence bands for the average $CV(m)$ are shown as well.*

Another proposal, that is less explored, but still based on likelihood-type statistics, is the cross-validation (CV) method of Smyth, 2000. In CV a risk measure $CV(m)$ is computed out-of-sample by splitting the available data, and a model m^* is chosen in order to optimize $CV(m)$. For a given m the CV works as follow: (i) a partition of X_n into a training-set X_n^{tr}, and a test-set X_n^{cv} is obtained; (ii) $\hat{\theta}^{tr}(m)$ is estimated using the sample points in X^{tr}; (iii) $CV(m) = \hat{l}^{cv}(m)/n$ is computed, where $\hat{l}^{cv}$ is the estimated $\hat{l}(m)$ computed on X_n^{cv} using $\hat{\theta}^{tr}(m)$. In order to reduce the bias/variance of the CV, multiple splits are performed and the averaged value of $CV(m)$ is maximized. A model is selected in order to maximize the so computed $CV(m)$.

3 Comparing methods on real data

The comparison uses the famous *Iris* dataset (Fisher, 1936), a four dimensional dataset with $n = 150$ observations of Iris species, divided in three different classes/groups. The analysis employs the *mclust* R package (Scrucca *et al.* , 2017) for parameters estimation. $\mathcal{M}$ includes finite Gaussian mixture models with $K = 1, 2, \ldots, 10$, and the covariance parametrizations of Celeux & Govaert, 1995, for a total of 140 models. For cross-validation we compare two splitting methods: (i) *10-fold CV*: the data set is randomly partitioned into 10 non-overlapping subsets (the folds), each used once as test-set while setting the remaining 9 folds as training-set; (ii) *Monte Carlo CV* (MCCV): the dataset is partitioned $T = 100$ times into two halves, one is used as training-set, the other is used as test-set. Results for the 6 methods are summarized in Figure 1. There are two winning solutions. BIC, ICL and MCCV, select $K = 2$, ellipsoidal structures for both clusters with varying volume and orientation. This solution merges the overlapping groups corresponding to *versicolor* and *virginica* species, which might be still reasonable. AIC3 and 10-fold CV selects a solution with $K = 3$ and covariance structure as before. This solution achieves an *adjusted Rand index* = 0.9 (see Hennig *et al.* , 2015) where 3.3% of the points are misclassified in the strongly overalpping region between the *versicolor* and *virginica* species. Here we conclude that AIC3 and 10-fold CV have a superior performance. This is an interesting evidence that encourages further investigations.

References

AKAIKE, H. 1973. Information theory and an extension of the maximum likelihood principle. 267–281.

BIERNACKI, C., CELEUX, G., & GOVAERT, G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, **22**(7), 719–725.

BOZDOGAN, H. 1983. *Determining the Number of Component Clusters in the Standard Multivariate Normal Mixture Model Using Model-Selection Criteria.* Tech. rept. ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF QUANTITATIVE METHODS.

CELEUX, G., & GOVAERT, G. 1995. Gaussian parsimonious clustering models. *Pattern recognition*, **28**(5), 781–793.

FISHER, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**(2), 179–188.

HENNIG, C., MEILA, M., MURTAGH, F., & ROCCI, R. 2015. *Handbook of cluster analysis*. CRC Press.

SCHWARZ, G. 1978. Estimating the dimension of a model. *Ann. Statist.*, **6**(2), 461–464.

SCRUCCA, L., FOP, M., MURPHY, T. B., & RAFTERY, A. E. 2017. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8**(1), 205–233.

SMYTH, P. 2000. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and computing*, **10**(1), 63–72.

LABOUR MARKET ANALYSIS THROUGH TRANSFORMATIONS AND ROBUST MULTILEVEL MODELS

Aldo Corbellini[1], Marco Magnani[1] and Gianluca Morelli[1]

[1] Department of Economics and Management, University of Parma,
(e-mail: aldo.corbellini@unipr.it)

ABSTRACT: The work presents a robust approach to labor share analysis. The estimate of labour share presents various complexities related to the nature of the data sets to be analyzed. Typically, labour share is evaluated by using the discriminant analysis and linear or generalized linear models, that do not take into account the presence of missing values and possible outliers. Moreover, the variables to be considered are often characterized by a high dimensional structure. The proposed approach has the objective of improving the estimation of the model using robust multilevel regression techniques and data transformation.

KEYWORDS: labour share, robust multilevel regression, data transformation.

1 Introduction

The analysis of the labor share is a field of analysis which involves both the macro and the micro level. The relevance of this issue indeed is mostly related to the empirical analysis of the level and evolution of the aggregate labor share. A large share of the theoretical literature however has studied the dynamics of the determinants of the labor share at the micro level. This contradiction has been solved converging to a paradigm where the macro level was concealed with the micro level by assuming that a representative firm is operating in the economy. This approach characterizes most of the literature. In particular since the seminal analysis of Bentolila & Saint-Paul, 2003 where, the theoretical determinants of the labor share are summarized in the definition of the SK schedule, several studies have tried to provide an explanation for the persistent declining trend of the labor share identifying different causes for it. Most of these causes have to do with the behaviour of the representative firm, and thus concern the micro level. They include the elasticity of substitution between labor and capital (Bentolila & Saint-Paul, 2003; Lawless & Whelan, 2011) capital deepening (Piketty & Zucman, 2014; Karabarbounis & Neiman, 2013).

The present paper focuses on the micro level and analyzes the determinants of the labor share using a large set of firm-level data with the main aim to investigate the issues that in the empirical analyses at the macro level are discarded. In particular we will discuss the role of the elasticity of substitution between productive factors and its interactions with fundamental structural factors as firm size and the sectors where the firm operates. The results of our analysis contribute to the well established empirical literature studying the aggregate labor share by providing new insight on how differences between firms affect the labor share and how its determinants interact at the micro level.

2 The dataset and variables

Our main sample of firms is composed of more than thirty thousand firms in a timespan of ten years going up to year 2017, representative of the manufacturing sector and extracted from the Buerau Van Dijk's AIDA data base that contains comprehensive information on capital companies in Italy. A rich set of information is collected by this survey, including firm-specific characteristics, investment and (international) trade activities. The model variables are as follows: Y: Labour share proxied by the ratio of labour cost to value-added; This indicator is an alternative version of the ratio of wage to total company assets. Using the added value instead of total assets, this variable can assume negative or positive values. X1: The ratio of tangible fixed assets to added value. The book-value of gross investments of this year has been adjusted to account for inflation using a measure of vintage. Then the deflated value of investments in the next years has been added using sectoral deflators for gross fixed capital calculated by the ISIC/ATECO assuming year 2007 as the base reference. X2: The ratio of intangible assets to total assets: this ratio measures the percentage of investments on intellectual capital, research and development and other intangible assets over the company total assets. X3: The ratio of industrial equipment to total asset: this variable measures the theoretical productive potential of the firm and is one of the primary drivers of company value. X4: Return On Sales, (ROS), that measures firm operating profitability proxied by the ratio of operating margins to sales. We expect a negative effect on the default risk, as the higher a firm's profitability the higher the flow of internal resources available to cover debt exposure should be; X5: measures the firm's interest burden, proxied by the ratio of firm total asset to net capital; high interest burden may worsen the financial risk associated with external finance. X6: Sales, X7: Age of the firm.

3 Results

Table 1 first column shows the results of the application of the multilevel regression model on the original data. We used the method described in Nakagawa & Schielzeth, 2013 for deriving R^2 in multilevel framework. The R^2 of the multilevel model is high, but the significance of β coefficients is always near critical values, except for the intercept. The main reason of this behaviour lies in the presence of extremely high leverage points in the data that are affecting the β estimates in the linear model; this effect is well know in statistical literature and was first pointed out by Sastry and Nag, 1990 which summarized it in a theorem that states that $R^2 \to 1$ as the remoteness of the leverage units increases. Table 1 4th column shows the results of the application of the multilevel regression model after removing these outlying observations by means of the forward search, a procedure that detects multivariate outliers, Atkinson & Riani, 2000. The new pseudo R^2 is very low, near 0.01 even if the significance of the β estimates improves considerably and most of them are now significant.

	$\hat{\beta}_{ML}$		$\hat{\beta}_{MLFS}$		$\hat{\beta}_{MLFSdt}$	
Variable	Estimate	p-value	Estimate	p-value	Estimate	p-value
intercept	0.44924	6.9299e-05	0.68786	2.9993e-163	0.35525	2.5439e-239
TFA/AddVal	0.055476	0	0.018411	1.0344e-05	-0.051072	7.5262e-179
IA/TA	-0.65996	0.16129	-0.10365	0.29107	-0.66183	3.0051e-57
IE/TA	0.29917	0.78016	0.20128	0.59472	-0.49701	0.001858
ROS	0.00039737	0.95317	-0.02762	6.4095e-79	-0.094043	0
Debt ratio	0.00042643	0.70808	0.001237	0.10774	0.0019641	1.4619e-09
Sales	-1.3991e-08	0.97737	7.3252e-06	0.1368	-5.2079e-05	1.0936e-137
Age	0.0041725	0.29412	0.0020498	0.012091	0.0089754	3.1016e-148

Table 1. $\hat{\beta}$ *comparison between non robust regression, robust regression and robust regression after data transformation*

The analysis of the distribution of the regression residuals leads us to think that transformation of the response is required. To this purpose we use the non parametric conditional expectation methods (i.e. ACE and AVAS), Tibshirani, 1988, Breiman & Friedman, 1985. Applying the transformations on the cleaned dataset we were able to dramatically improve the goodness of fit, $R^2 = 0.33$. The analysis of the results still shows the presence of several regression outliers, therefore we performed again the Forward Search to remove the atypical units. Table 1 6th column shows the results of the regression model applied on the clean transformed data. The new value of the pseudo $R^2 = 0.39$, all the variables are now highly significant and the signs of the coefficients are in agreement with those suggested by the economic theory. Note that this goal was reached removing a small percentage of units that were biasing the model estimates.

4 Discussion and conclusions

The present paper studies the determinants of labor share dynamics using the approach developed by Bentolila and Saint-Paul (2003), which characterizes a one-for-one relationship between the labor share and the capital output ratio, the *SK* schedule. The sign of the relationship depends on the elasticity of substitution between labor and capital. An elasticity larger than unity implies a negative relationship, an elasticity smaller than unity implies a positive relationship, and unit elasticity implies that the labor share is constant. In the present context, the coefficient multiplying the capital output ratio, measured as the book-value of tangible assets on value added, highlighting that, as in most of the literature using micro-data, the productive factors capital and labor are largely substitute.

References

ATKINSON, A. C., & RIANI, M. 2000. *Robust Diagnostic Regression Analysis*. New York: Springer–Verlag.

BENTOLILA, S., & SAINT-PAUL, G. 2003. Explaining movements in the labor share. *Contributions in Macroeconomics*, **3**(1).

BREIMAN, L., & FRIEDMAN, J. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, **80**, 580–597.

KARABARBOUNIS, L., & NEIMAN, B. 2013. The global decline of the labor share. *The Quarterly journal of economics*, **129**(1), 61–103.

LAWLESS, M., & WHELAN, K. T. 2011. Understanding the dynamics of labor shares and inflation. *Journal of Macroeconomics*, **33**(2), 121–136.

NAKAGAWA, S., & SCHIELZETH, H. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**(2), 133–142.

PIKETTY, T., & ZUCMAN, G. 2014. Capital is back: Wealth-income ratios in rich countries 1700–2010. *The Quarterly Journal of Economics*, **129**(3), 1255–1310.

TIBSHIRANI, R. 1988. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, **83**(402), 394–405.

MODELLING CONSUMERS' QUALITATIVE PERCEPTIONS OF INFLATION

Marcella Corduas[1], Rosaria Simone[1] and Domenico Piccolo[1]

[1] Department of Political Sciences, University of Naples Federico II,
(e-mail: marcella.corduas@unina.it, rosaria.simone@unina.it,
domenico.piccolo@unina.it)

ABSTRACT: This article proposes an innovative model, based on a mixture distribution, for ordinal time series data. The method is illustrated by its application to the qualitative perceptions of inflation in Italy.

KEYWORDS: CUB model, ordinal data time series, minimum distance.

1 Introduction

Repeated surveys about opinions, perceptions or attitudes of the interviewees are regularly carried out by national statistical offices. This is the case of the surveys concerning the qualitative assessment or anticipations on price level that ISTAT carries out every month. Earlier studies of perceived and expected inflation focussed either on quantifying the observed ordinal data in order to derive indices of perceived (or expected) inflation or on searching explicative models that could describe data in terms of economic explanatory variables (Simmons & Weiserbs, 1992). In this article, we discuss an innovative model for time series ordinal data, that extends the well established CUB model (Piccolo *et al.*, 2018) to allow for time varying parameters. For illustrative purpose the method is applied to consumers' perceptions of inflation in Italy.

2 The methodology

Let $\{Y_t, t = 1, ..., T\}$ be a collection of random variables describing ordinal data observed at different time points. We assume that at time t, the variable Y_t is characterized by the following GeCUB distribution:

$$P(Y_t = y|I_{zwv}) = \delta_t D_t + (1-\delta_t)\left[\pi_t \begin{pmatrix} m-1 \\ y-1 \end{pmatrix} (1-\xi_t)^{y-1}\xi_t^{m-y} + (1-\pi_t)\frac{1}{m}\right]), \quad y = 1, 2, ..., m.$$

with:

$$\pi_t = \frac{1}{1+e^{-\beta_0-\beta_1 z_{t-1}\dots\beta_p z_{t-p}}};\ \xi_t = \frac{1}{1+e^{-\gamma_0-\gamma_1 w_{t-1}\dots\gamma_s w_{t-s}}};$$
$$\delta_t = \frac{1}{1+e^{-\alpha_0-\alpha_1 v_{t-1}\dots\alpha_k v_{t-k}}}; \qquad (1)$$

where z_t, w_t and v_t are explanatory variables, I_{zwv}, is the set of information concerning each of these variables until time $(t-1)$. Moreover, D_t is a degenerate distribution such that: $D_t = 1$ for the shelter category and $D_t = 0$ for the remaining categories; $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ and $\gamma = (\gamma_0, \gamma_1, ..., \gamma_s)'$, and $\alpha = (\alpha_0, \alpha_1, ..., \alpha_k)'$ are the parameter vectors. The model can be easily generalized to the case when each GeCUB parameter is affected by several explanatory variables. When the shelter effect is not present the model (1) collapses to the CUB formulation. Let us denote with $[f_{1t}, f_{2t}, ..., f_{mt}]$ the relative frequencies from a random sample of n observations drawn from Y_t. We propose to estimate the model by minimizing the sum of the Pearson's chi-square distances (see, Harris & Kanji, 1983 and references therein) between the observed relative frequencies and the probabilities implied by the model:

$$G(\alpha,\beta,\gamma) = \sum_{t=1}^{T}\sum_{i=1}^{m}[f_{it} - p_{it}]^2/p_{it} \qquad (2)$$

where $p_{it} = P(Y_t = i)$. The goodness of fit of the model is assessed by comparing $G_{min} = G(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma})$ with the distance G_U, evaluated using the probability: $p_{it} = m^{-1}$, $\forall(i,t)$. The uniform distribution, in fact, reflects pure ignorance about the ordinal data distribution at time t.

3 The empirical study

As an illustration, we have considered data from the survey on consumer qualitative perception and expectation of inflation, carried out by ISTAT every month among about 2000 individuals. The link between inflation perceptions and actual inflation had been quite strong before 2002, but it collapsed following the euro cash changeover in 2002 in all EU countries. In Italy, this gap was exceptionally large and persistent, and a similar gap also affected perceived and expected inflation. Only towards the end of 2009, after the global economic crises, the distance between those measures disappeared as shown by the pattern of the balance statistic for the expectations and perceptions in Figure 1. We have applied model (1) to ordinal data originated by the question

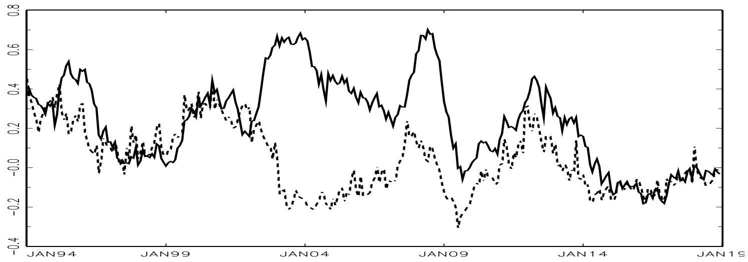


Figure 1. *Balance statistic of perceived (solid line) and expected (dashes) inflation*

concerning the perception of past price development in the above mentioned survey: How do you think that consumer prices have developed over the last 12 months? They have: risen a lot; risen moderately; risen slightly; stayed about the same; fallen. The analysis refers to observations from 1994.01 to 2018.12. The categories have been recoded from 1 (fallen) to 5 (risen a lot). The initial points for explanatory variables have been derived from previous surveys. In particular, we have specified the dynamics of the GeCUB coefficients as follows:

$$\xi_t = \frac{1}{1+e^{-\gamma_0-\gamma_1 w_{t-1}}}; \; \pi_t = \frac{1}{1+e^{-\beta_0-\beta_1 z_{t-1}}}; \; \delta_t = \frac{1}{1+e^{-\alpha_0-\alpha_1 v_{t-1}}}; \qquad (3)$$

where, for any t:

- the parameter ξ_t depends on w_{t-1}, the mean of the price past trend perceptions (this is simply the mean of the observed ratings) at time $t-1$;
- the parameter π_t depends on z_{t-1}, the mean of the expectations about future price level at time $t-1$;
- $D_t = 1$ for the category: stayed about the same, and 0 otherwise. The corresponding coefficient δ_t depends on $v_{t-1} = w_{t-1} - z_{t-1}$, the gap between price trend perceptions and future trend expectations. When this gap is small, the perception that prices stayed about the same becomes stronger (see Greitemeyer *et al.*, 2005 for a discussion of the influence of expectations on price level judgements).

Table 1 illustrates the estimated coefficients of the model with their standard errors in parenthesis. Figure 2 shows the pattern of the time varying estimates $(\widehat{\delta_t}, \widehat{\pi_t}, \widehat{\xi_t})$. When the perceptions and expectations start having a divergent pattern (from 2002 onwards) the weights π_t and ξ_t show a rapid change. They both increase, whereas the weight of the shelter category rapidly falls to zero. As matter of facts the GeCUB distribution is left skewed because a great

Table 1. *Estimation results (standard errors in parenthesis)*

$\hat{\gamma}_0 = \ \ 3.581(0.199)$	$\hat{\gamma}_1 = -1.229(0.054)$	Fitting measures
$\hat{\beta}_0 = -0.819(0.883)$	$\hat{\beta}_1 = \ \ 0.658(0.282)$	$G_{min} = 46.78$
$\hat{\alpha}_0 = \ \ 2.213(0.116)$	$\hat{\alpha}_1 = -1.050(0.250)$	$G_U = 157.32$

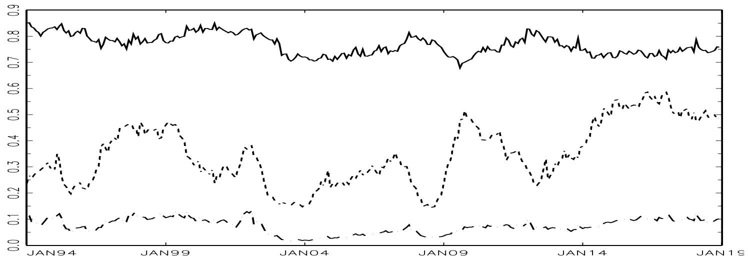


Figure 2. *Time varying coefficients:* π_t *(solid line),* ξ_t *(short dashes),* δ_t *(long dashes)*

part of respondents believe that inflation has increased. For sake of space, it is not possible to comment further these results, but it is worth pointing out that the proposed model provides a very parsimonious formulation that well describes the perceptions of inflation in Italy in the considered years.

References

GREITEMEYER, T., SCHULZ-HARDT, S., TRAUT-MATTAUSCH, E., & FREY, D. 2005. The influence of price trend expectations on price trend perceptions: Why the Euro seems to make life more expensive? *Journal of Economic Psychology*, **26**, 541–548.

HARRIS, R. R., & KANJI, G. K. 1983. On the use of minimum chi-square estimation. *Journal of the Royal Statistical Society (D)*, **32**, 379–394.

PICCOLO, D., SIMONE, R., & IANNARIO, M. 2018. Cumulative and CUB models for rating data: a comparative analysis. *Intl. Statistical Review*, 1–30.

SIMMONS, P., & WEISERBS, D. 1992. Consumer Price Perceptions and Expectations. *Oxford Economic Papers*, **44**, 35–50.

NOISE RESISTANT CLUSTERING OF HIGH-DIMENSIONAL GENE EXPRESSION DATA

Pietro Coretto[1], Angela Serra[2] and Roberto Tagliaferri[3]

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

[2] Faculty of Medicine and Health Technology, Tampere University,
(e-mail: angela.serra@tuni.fi)

[3] NeuRoNeLab, Department of Management and Innovation Systems,
University of Salerno, (e-mail: robtag@unisa.it)

ABSTRACT: Discovery of disease sub-types is one of the fundamental problem in clinical applications. This is usually accomplished by grouping patients based on gene expression data. However, microarray data sampling is terribly noisy, and this undermines the possibility to reach scientific consensus on the empirical evidence. In this work we discuss the need of robust data analysis methods for gene expression data. We introduce and discuss recent proposals of clustering methods and algorithms that can handle noise effectively, and that can scale scale with the typical dimension of microarray data. The methods and algorithms are tested on a selection of data sets obtained from the well known "*The Cancer Genome Atlas*" repository.

KEYWORDS: clustering, high-dimensional data, gene expression, otrimle, snf.

1 Introduction

Sub-typing is the precision medicine task of identifying sub-populations of similar patients that can lead to more accurate diagnostic and treatment strategies (see Saria & Goldenberg, 2015 and references therein) . Sub-typing has an enormous practical impact in clinical practice because it allows to refine prognosis for similar individuals, and this reduces the uncertainty in the expected outcome of a medical treatment.

The main technique to sub-type patients is to use statistics and machine learning methods to identify clusters of individuals with similar genetic patterns. The problem is particularly difficult for several reasons. First, subtyping is to find clusters which is an unsupervised task, therefore the underlying group structure is totally unobservable. Second, there are various sources of genetic information from different omics data types (miRNA, methylation, etc), all these data types have huge dimensionality while few sample units are usually available. Moreover there is no guarantee that each data types carries the same information about the same groups, so there is even a difficulty to choose which data types to look at. Third, although high-throughput omics-technologies have progressed substantially, these type of data remains terribly noisy (Marshall, 2004).

In Section 2 we review recent noise-free clustering methods for patient sub-typing. In Section 3 we discuss applications to cancer data, and we outline the main conclusions.

2 Clustering methods

There is an abundance of clustering methods used in genomics. Some of these methods are specifically designed for gene expression data, other methods consists in tuned versions of classical methods (e.g. k-means, hierarchical methods, etc.). A recent systematic review is given in Kiselev *et al.* , 2019. However, none of the classical tools used in this field is noise-resistant. It is well known that genomic data is terribly noisy, and research have made terrible efforts to cure data acquisition technologies. Despite the huge progresses, this type of data remain dramatically subject to noise, contamination, and heavy-tailedness (see Serra *et al.* , 2018). In this paper we introduce and discuss two recent additions that, although built from completely different perspectives, both are designed to be noise-resistant, and both established remarkable performances in benchmark cancer data sets.

SNF ALGORITHM. Wang *et al.* , 2014 introduced the *Similarity Network Fusion* algorithm (SNF). The SNF integrates many different types of measurements (e.g. mRNA expression data, DNA methylation, miRNA expression, etc.). A similarity network is built for each data input, and a final single data set is built performing network fusion. Working in the sample network space allows SNF to overcome the twists caused by different scales, data acquisition bias, and noise that strongly varies across data input types. The fused network sample is clustered based on Spectral Clustering. The SNF algorithm has several input tunings, however it is shown (experimentally) that the method is not too sensitive to them. The SNF algorithm gained a wide popularity and it is considered the state-of-the-art method for genomic data integration and patient sub-typing.

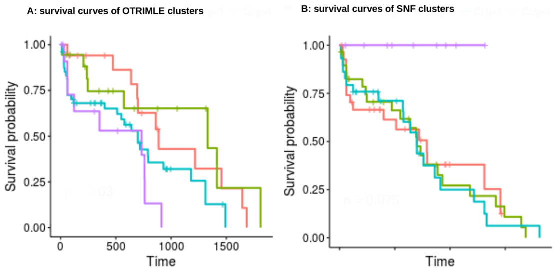


Figure 1. *Survival curves for the "Lung Cancer Data" for the OTRIMLE-based method (left), and the sNF method (right).*

OTRIMLE-BASED ALGORITHM. This was introduced by Coretto *et al.* , 2018, and integrates several ideas from robust data analysis and clustering. Differently form the SNF, this methods does not integrates different data types. It uses only two data inputs: (i) gene expressions (e.g. mRNA expressions), (ii) patients survival data. A typical situation may be that observe about $p = 3,500$ genes on $n = 100$ patients, but much higher concentration ratios p/n are not so unusual. The correlation structure is captured based on the *Robust and Sparse Correlation* matrix estimator (RSC) of Serra *et al.* , 2018. As for PCA, the original high-dimensional data gene expression data matrix is projected over the direction of $m << p$ eigenvectors of the RSC matrix. The OTRIMLE algorithm of Coretto & Hennig, 2016 and Coretto & Hennig, 2017 recovers the Gaussian-shaped clusters over the projected subspace. The OTRIMLE adapts to the noise level, but it needs an input parameter, that is the eigen-ratio constraint γ, which restricts the relative discrepancy between clusters' elliptical shapes. The method looks for several clustering solutions based on different values of m and γ. The final solution is chosen in order to minimize a criterion (called RLEDMIN) which measures the overall separation of the cluster-wise survival curves.

3 Results and conclusions

The two algorithms are extensively compared in Coretto *et al.* , 2018 on five distinct experimental data sets from the TCGA database*. For 4 cancer data sets out of 5, the OTRIMLE-based method outperformed the state-of-the-art SNF in terms of survival patters separation. As an example in Figure 1 we

*Available at https://portal.gdc.cancer.gov/

report the results for the LUNG cancer data set. The figure shows how the OTRIMLE-based method can lead to well distinct survival patterns across the recovered clusters. Of course this is due to the fact that the OTRIMLE-based method is optimized in order to achieve the best separation in terms of survival curves. However, in comparative studies survival separation is always used as the ultimate validation criterion. And in fact the main advantage the OTRIMLE-based algorithm is to optimize the procedure on a data space, the survival data, different from that where the clusters are assumed to belong.

The method introduced in Coretto *et al.* , 2018 deserves further investigations. For example one may change the dimensional reduction technique (projection using the RSC matrix), or the clustering technique (OTRIMLE) in the clustering step. However, this is for future researches.

References

CORETTO, P., & HENNIG, C.. 2016. Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering. *Journal of the American Statistical Association*, **111**(516).

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.

CORETTO, P., SERRA, A., & TAGLIAFERRI, R.. 2018. Robust clustering of noisy high-dimensional gene expression data for patients subtyping. *Bioinformatics*, **34**(23), 40644072.

KISELEV, V. Y., ANDREWS, T. S., & HEMBERG, M.. 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, **20**(5), 273–282.

MARSHALL, E. 2004. Getting the Noise Out of Gene Arrays. *Science*, **306**(5696), 630–631.

SARIA, S., & GOLDENBERG, A.. 2015. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems*, **30**(4), 70–75.

SERRA, A., CORETTO, P., FRATELLO, M., & TAGLIAFERRI, R.. 2018. Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data. *Bioinformatics*, **34**(4), 625–634.

WANG, B., MEZLINI, A. M., DEMIR, F., FIUME, M., TU, Z., BRUDNO, M., HAIBE-KAINS, B., & GOLDENBERG, A.. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, **11**(3), 333–337.

CLASSIFY X-RAY IMAGES USING CONVOLUTIONAL NEURAL NETWORKS

Federica Crobu[1] and Agostino Di Ciaccio[2]

[1] Sapienza Università di Roma, (e-mail: federicacrobu@gmail.com)

[2] Department of Statistics, Sapienza Università di Roma,
(e-mail: agostino.diciaccio@uniroma1.it)

ABSTRACT: In recent years, computer-assisted diagnostic systems have gained increasing interest through the use of deep learning techniques. In this work we show how it is possible to classify X-ray images through a multi-input convolutional neural network. The use of clinical information together with the images allowed to obtain better results than those present in the literature on the same data.

KEYWORDS: deep learning, convolutional neural networks, chest x-ray, multi-input neural networks.

1 Introduction

In recent years, the deep neural networks (DNN) and in particular the deep convolutional neural networks (DCNN) have attracted the attention of the researchers for their great ability to analyse images. One of the most fascinating and advantageous branches for the application of these models is medicine. Thanks to them we can now imagine a future in which doctors are helped by computers to recognize diseases and make diagnoses. Furthermore, it could be a drastic improvement in the underdeveloped countries where the availability of doctors is problematic and pathologies such as pneumonia are still one of the main causes of death.

In the classical context of image recognition, the goal is to classify what is contained in an image, however, in the analysis of medical images, the challenge is quite different. In fact, to emulate the role of the doctor, the model needs much more information that cannot be deduced from the analysis of radiographic images only. Therefore, it is also necessary to consider many other information collected on the patients such as clinical and demographic details.

The correlation of certain pathologies with age or smoking is well known, for example. Other diseases may have genetic predispositions and many diseases can be related to each other. Usually, doctors can obtain and use all this information and it is advantageous to provide them also to the predictive model.

From the technical point of view, the goal of including more inputs of different nature, images and numerical values, has been achieved using a multi-input neural network architecture. Through this model we were able to obtain a very accurate classification, as it is shown in the following sections.

2 The X-ray data and the previous works

The availability of large medical databases containing both images and clinical information is scarce. Currently, the largest database is the ChestX-ray14, chosen for this application. It was released by the United States National Institutes of Health (NIH) and contains over 112,000 radiographic frontal chest images of 30,805 patients. Each of them can be healthy or sick, with one or more of the following 14 diseases: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural thickening, Pneumonia, Pneumothorax. Furthermore, a "no finding" category represents the images in which none of the previously mentioned diseases have been detected.

As can be understood by analysing the database, for many patients are available multiple results of the tests, which can be useful for capturing the progress of diseases over time.

The labels, corresponding to the pathologies identified in each image, were extracted from radiological reports using natural language processing techniques with an accuracy that is declared by authors over 90% (Rajpurkar et al. 2017). Therefore, we cannot fully trust the labelling process and, furthermore, some researchers have raised many doubts about the correctness of the labels. The criticism of the radiologist Luke Oakden-Rayner (2017) that, after observing the images, states that there are incorrect labels. Finally, it should be noted that many diagnoses present more concomitant diseases.

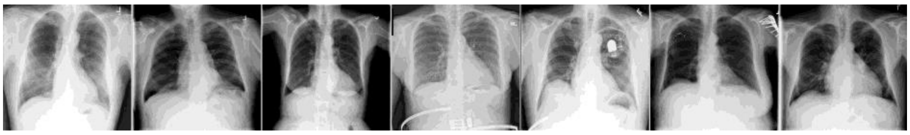


Fig.1 – Some images of the database ChestX-ray14

This dataset has been already used by many other researchers. Surely, the best-known work was made by a Stanford’s team (Rajpurkar et al. 2017). They proposed an architecture called *CheXNet* based on the usage of the DCNN architecture called *DenseNet121* (Huang et al. 2017). This work represents, at this moment, the state-of-the-art results in terms of AUC scores.

Other important works are the one of Yao et al. (2017) and the one of Wang et al. (2017). The first is mainly based on an architecture consisting of a DenseNet as encoder and on a recurrent neural network as decoder. Wang tries to apply some of the most famous CNN architectures (excluding DenseNet), achieving the best results with ResNet-50. Other interesting and more recent works are the ones of Baltrushat (2018) and Guendel (2018). Baltrushat based his work on a ResNet-50 to analyze the images, supported by the use of age, gender and view position.

3 The proposal and the results

Inspired by the work of the Stanford team, we decided to improve the model by exploiting the few clinical and demographic information available with these images. We have considered age, sex, sight position and 14 new variables containing the patient's information obtained from previous pathological history present in the same data.

The goal was therefore to improve the DenseNet121 model with another parallel neural network, with two small dense layers (32 and 16 neurons), which processes the non-image characteristics. The two independent networks are then concatenated and connected to the output layer based on 14 neurons with sigmoid activation function, whose task is to estimate the probability of the presence of each disease in the X-ray image. The final network has a complex structure with 123 'main' layers and 7,053,182 parameters. We used the pretrained weights of DenseNet121 on Imagenet as initialization of the network.

To solve this multi-input multi-class problem, we have employed a weighted binary cross-entropy loss function with data augmentation.

Our results provide an interesting improvement of the state-of-the-art, confirming our intuition of the architecture's power. Following the literature, we have adopted the AUC index as main tool to evaluate the quality of the predictions. In the table 1 we can see the comparison of the performances of our model with the best results obtained by other researchers in terms of the mean AUC scores.

	Wang et al. (2017)	***Yao et al. (2018)***	***CheXNet (2017)***	***Our Multi-input***
Official split	Yes	No	No	Yes
Atelectasis	0.716	0.772	0.809	**0.816**
Cardiomegaly	0.807	0.904	0.925	**0.925**
Effusion	0.784	0.859	0.864	**0.867**
Infiltration	0.609	0.695	0.735	0.731
Mass	0.706	0.792	0.868	**0.897**
Nodule	0.671	0.717	0.780	**0.827**
Pneumonia	0.633	0.713	0.768	**0.776**
Pneumothorax	0.806	0.841	0.889	**0.927**
Consolidation	0.708	0.788	0.790	**0.801**
Edema	0.835	0.882	0.888	**0.893**
Emphysema	0.815	0.829	0.937	**0.946**
Fibrosis	0.769	0.767	0.805	**0.881**
Pleural Thickening	0.708	0.765	0.806	**0.827**
Hernia	0.767	0.914	0.916	**0.963**
Average	0.738	0.803	0.841	**0.863**

Table 1. Comparison of the AUC on test data.

We have chosen the subdivision suggested by the data authors, and we have also verified that the previously proposed approaches with different splits still obtain the same results with this subdivision. The size of the test-set on which the AUC was measured has dimensions greater than 25,000 and therefore guarantees a great stability of the results with respect to the possible subdivisions. It is evident in the table that the average AUC has been significantly improved by our approach and, for most classes, we have clearly outperformed previous jobs.

4 Conclusions

The results of this application have confirmed the validity of our approach: a multi-input neural network architecture can significantly improve predictions. Clearly, the idea of combining different sources of heterogeneous information can be applied in other fields of medicine, as in the analysis of MRI scans. Whenever the patient's clinical and/or demographic information is available, it is possible and fruitful to apply this approach. Similarly, this technique can be used in other application areas.

References

BALTRUSCHAT, I.M., NICKISCH, H., GRASS, M., KNOPP, T., SAALBACH, A. 2018. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *arXiv:1803.02315*.

GUENDEL, S., GRBIC, S., GEORGESCU, B., ZHOU, K., RITSCHL, L., MEIER, A., COMANICIU, D. 2018. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. *arXiv preprintarXiv:1803.04565*.

HUANG, G., LIU, Z., VAN DER MAATEN, L., WEINBERGER, K. Q. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 2261-2269.

OAKDEN-RAYNER, L. 2017. *Exploring the ChestXray14 dataset: problems*, https:// lukeoakdenrayner.wordpress.com.

RAJPURKAR, P.,IRVIN, J., ZHU, K., YANG, B., MEHTA, H., DUAN, T., DING, D., BAGUL, A., BALL, R. L., LANGLOTZ, C., SHPANSKAYA, K., LUNGREN, M. P., NG, A. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, *arXiv preprint arXiv:1711.05225*.

WANG, X., PENG, Y., LU, L., LU, Z., BAGHERI, M., SUMMERS, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Honolulu, HI, 2017, pp. 3462-3471.

YAO, L., POBLENZ, E., DAGUNTS, D., COVINGTON, B., BERNARD, D., LYMAN, K. 2017. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*.

A COMPOSITIONAL ANALYSIS APPROACH ASSESSING THE SPATIAL DISTRIBUTION OF TREES IN GUADALAJARA, MEXICO

Marco Antonio Cruz[1], Maribel Ortego[1] and Elisabet Roca[1]

[1] Department of Civil and Environmental Engineering, Polytechnic University of Catalonia BarcelonaTech, (e-mail: marco.antonio.cruz@upc.edu, ma.isabel.ortego@upc.edu, elisabet.roca@upc.edu)

ABSTRACT: Urban green infrastructure such as parks, gardens and trees, provide several ecosystem services and benefits. Particularly trees provide a broad amount of services in urban areas, such as improving air quality, mitigating carbon pollution and heat-island effect, attenuating storm-water floods, reducing noise and serving as habitat for different species among others. Likewise, urban trees provide different social (i.e., social cohesion), economic (i.e., increase in property value), psychological (i.e., stress reduction) and medical (i.e., increase in longevity of life) benefits (Landry, 2009; Roy et al., 2012; Battisti et al., 2019). Although it is well documented that trees are essential for the well-being and health of urban areas and their inhabitants, trees are not evenly distributed in urban areas. Previous studies have found that urban residents with a deprived socioeconomic status are associated with low coverage of urban trees in their communities (Hernández and Villaseñor, 2017; Park and Kwan, 2017; Wang and Qiu, 2018). Therefore, environmental justice seeks to ensure that green infrastructure and its benefits are distributed equally throughout the territory (Anguelovski, 2013; Gould and Lewis, 2017). The objective of this study is to determine whether the distribution of urban trees in the city of Guadalajara, Mexico is distributed equally or not among its colonies and urban districts. The information is obtained from the first and only tree census conducted in the city on June 2018 and treated with geographic information systems (GIS). The attributes of the tree dataset include their location (urban blocks, streets, parks and gardens), heights and diameters of their canopy (Government of Guadalajara, 2019). For the analysis and due to the compositional nature of the data, compositional analysis techniques are applied (see Aitchison, 1986; Pawlowsky-Glahn, et al., 2015; Filzmoser et al., 2018). With this novel approach, we contribute to the existing literature. Additionally, Principal Component Analysis (PCA) and cluster analysis are performed to identify the distribution of trees in the city. Likewise, to observe the relationship between trees and socio-economic variables, a multivariable linear regression is carried out respecting the compositional nature of the data. The results from PCA and cluster analysis show a clear differentiation in the distribution of trees between the East-West of the city, mainly in the compositions with respect to their height and diameter. Likewise, from the multivariate linear regression, considerable significance ($p<0.05$) is found in socio-economic variables.

KEYWORDS: compositional data analysis, environmental justice, trees, Guadalajara.

References

AITCHISON, J. 1986. *The Statistical Analysis of Compositional Data. Monographs on Satistics and Applied Probability.* Edited by Chapman & Hall Ltd. London.

ANGUELOVSKI, I. 2013. New Directions in Urban Environmental Justice. *Journal of Planning Education and Research*, 33(2), pp. 160–175. doi: 10.1177/0739456x13478019.

BATTISTI, L. *et al.* 2019. Residential Greenery: State of the Art and Health-Related Ecosystem Services and Disservices in the City of Berlin. *Sustainability*, 11(6), p. 1815. doi: 10.3390/su11061815.

FILZMOSER, P., HRON, K. & TEMPL, M. 2018. *Applied Compositional Data Analysis*. 1st edn. Edited by Z. S. Diggle Peter, Gather Ursula. Cham, Switzerland: Springer Series in Statistics. doi: 10.1007/978-3-319-96422-5.

GOULD, K. A. & LEWIS, T. L. 2017. *Green gentrification.Urban Sustainability and the struggle for environmental justice*. 1st edn. Edited by J. Agyeman and Z. Patel. New York: Routledge equity, justice and the sustainable city series.

GOVERNMENT OF GUADALAJARA. 2019. *GeoGDL, Inventario Arbolado LIDAR*. Available at: https://mapa.guadalajara.gob.mx/geomap.

HERNÁNDEZ, H. J. & VILLASEÑOR, N. R. 2017. Twelve-year change in tree diversity and spatial segregation in the Mediterranean city of Santiago, Chile. *Urban Forestry & Urban Greening*. Elsevier GmbH. doi: 10.1016/j.ufug.2017.10.017.

LANDRY, S. M. 2009. Street trees and equity : evaluating the spatial distribution of an urban amenity, 41, pp. 2651–2671. doi: 10.1068/a41236.

PARK, Y. M. & Kwan, M. P. 2017. Multi-contextual segregation and environmental justice research: Toward fine-scale spatiotemporal approaches. *International Journal of Environmental Research and Public Health*, 14(10). doi: 10.3390/ijerph14101205.

PAWLOWSKY-GLAHN, V., EGOZCUE, J. J. & TOLOSANA-DELGADO, R. 2015. *Modeling and Analysis of Compositional Data*. First. London: John Wiley & Sons.

ROY, S., BYRNE, J. & PICKERING, C. 2012. Urban Forestry & Urban Greening A systematic quantitative review of urban tree benefits , costs , and assessment methods across cities in different climatic zones’. Elsevier GmbH., 11, pp. 351–363.

WANG, H. & QIU, F. 2018. Urban Forestry & Urban Greening Spatial disparities in neighborhood public tree coverage : Do modes of transportation matter ?. *Urban Forestry & Urban Greening*. Elsevier, 29(November 2017), pp. 58–67. doi: 10.1016/j.ufug.2017.11.001.

JOINING FACTORIAL METHODS AND BLOCKMODELING FOR THE ANALYSIS OF AFFILIATION NETWORKS

Daniela D'Ambrosio[1], Marco Serino[2] and Giancarlo Ragozini[2]

[1] Interdepartmental Research Center Urban/Eco, University of Naples Federico II,
(e-mail: daniela.dambrosio2@unina.it)

[2] Department of Political Science, University of Naples Federico II,
(e-mail: marco.serino@unina.it, giragoz@unina.it)

ABSTRACT: In this paper we explore a new strategy to jointly use factorial methods and blockmodeling to analyse affiliation (two-mode) networks. Among the methods that group simultaneously and directly individuals and variables for binary matrices, we propose using cluster correspondence analysis, in order to (*i*) look at the way network positions can be incorporated in the cluster CA; (*ii*) verify if cluster CA is apt to represent specific network structures. Finally, an empirical application on an affiliation network of stage co-productions will be provided.

KEYWORDS: affiliation networks, blockmodeling, cluster ca, data classification.

1 Introduction

Affiliation networks are a special case of two-mode networks which consist of two disjoint sets: a set of actors and a set of events in which those actors are involved. One of the main concerns in studying such networks is to establish equivalent classes of actors that are similarly embedded in the whole network, following some criterion of equivalence, such as structural equivalence. Blockmodeling, with its recent extensions (Doreian *et al.*, 2005), allow to perform a clustering of the affiliation network units.

However, other methods proved equally apt to find relational patterns within affiliation networks. Factorial methods, such as Multiple Correspondence Analysis (MCA) (Greenacre & Blasius, 2006), permit to synthesize, analyse and graphically represent the relational structure in a metric space. Thanks to the relationships between MCA and blockmodeling, as for the measures that capture structural similarities in the network (D'Esposito *et al.*, 2014a; D'Esposito *et al.*, 2014b), a joint approach has been proposed to apply a clustering method, i.e. blockmodeling, along with a given factorial method - but not simultaneously (Serino *et al.*, 2017; Ragozini *et al.*, 2018).

Hence, in this paper we propose using another method, namely cluster correspondence analysis (cluster CA) (van de Velden *et al.*, 2017), that groups simultaneously individuals and variables for binary matrices and also permits to evaluate the relations among groups in terms of proper distances. We present an application of this approach by analysing the affiliation network of the stage co-productions released in Campania (Italian region) during the 2012/2013 season.

2 Factorial methods and blockmodeling for analysing affiliation networks

Recently, a *joint approach* has been proposed that uses MCA and blockmodeling for affiliation networks, relying upon the relationships that exist between factorial methods and blockmodeling. The network positions (i.e. the clusters), as derived from the blockmodeling, are incorporated in the analysis made by MCA as supplementary variables and represented in the metric space (Serino *et al.*, 2017; Ragozini *et al.*, 2018). In this approach, clustering and factorial methods, albeit jointly used to analyse the network structure, are kept separated in the analytic process. In this paper, as an advancement of such research line, we propose using a factorial method that performs simultaneously a clustering of individuals and variables for binary matrices, the latter being no less than the type of variable concerned with event affiliations (participation or non-participation to a given event).

The method we propose using in this work, namely cluster CA, combines cluster analysis and CA and allows to obtain both a low-dimensional representation of clusters and attributes and a clustering of individuals relying on the profiles related to the categorical variables(van de Velden *et al.*, 2017). Therefore, it permits to obtain dimension reduction and clustering of categorical data simultaneously (van de Velden *et al.*, 2017).

3 Applying cluster CA and blockmodeling to affiliation networks

An affiliation network $\mathcal{G}$ can be represented by a triple $\mathcal{G}(V_1, V_2, \mathcal{R})$ composed of two disjoint sets of nodes, V_1 and V_2 of cardinality n and m, and a set of edges or arcs, $\mathcal{R} \subseteq V_1 \times V_2$. By definition $V_1 \cap V_2 = \emptyset$, the two disjoint sets V_1 and V_2 refer to different entities i.e. the set $V_1 = \{a_1, a_2, \ldots, a_n\}$ represents the actor set whereas the other, $V_2 = \{e_1, e_2, \ldots, e_m\}$, represents the set of m relational events. The edge $r_{ij} = (a_i, e_j)$, $r_{ij} \in \mathcal{R}$, is an ordered couple, and indicates if an actor a_i attends an event e_j. The set $V_1 \times V_2$ can be fully represented by a binary matrix, the affiliation matrix, $\mathbf{F}(n \times m)$, with element $f_{ij} = 1$ if $(a_i, e_j) \in \mathcal{R}$ and 0 otherwise.

In affiliation networks the structural equivalence principle states that two actors are equivalent if they participate exactly to the same events (Pizarro, 2007). Formally, given two actors a_i and $a_{i'}$, structural equivalence property $\equiv$ states that: $a_i \equiv a_{i'}$ if and only if $r_{ij} = r_{i'j}$ $\forall j$. If two actors a_i and $a_{i'}$ are structurally equivalent they are indistinguishable, and one equivalent actor can substitute for the other one because the two relational patterns are the same.

In order to discover the relational structure embedded in $\mathbf{F}$, it is possible to consider it as an usual *case-by-variable* matrix and, than, apply a factorial method like the MCA. In the latter application the indicator matrix $\mathbf{Z}$ is derived from the matrix $\mathbf{F}$ through the full disjunctive coding. Given that each relational event e_j is a dichotomous variable, the indicator matrix $\mathbf{Z}$ contains two columns for each e_j, namely e_j^+ and e_j^-, where e_j^+ is the value of a dummy variable coding the participation to the event, and e_j^- is the value of a dummy variable coding the non participation. As all the variables in $\mathbf{F}$ are dichotomous, the corresponding indicator matrix $\mathbf{Z}$ turn to be a *doubled matrix*.

Given our affiliation matrix $\mathbf{F}$ and the (doubled) indicator matrix $\mathbf{Z}$ derived from the former, and following the approach proposed by van de Velden *et al.* (2017), we aim to find $\mathbf{Z}_K$, i.e. the indicator matrix of dimensionality $n \times K$ which includes the cluster membership considered as a categorical variable such that $\mathbf{F}^c = \mathbf{Z}'_K\mathbf{Z}$ is the table cross-tabulation that includes the associations between the cluster membership and the binary variables coding the participation (and non-participation) in events.

Following the iterative procedure described by van de Velden *et al.* (2017), skipping its technical details, we propose to apply the algorithm for cluster CA as follows: 1) generate an initial cluster allocation $\mathbf{Z}_K$; 2) find category quantifications by using the usual CA algorithm; 3) construct an initial configuration of the relational patterns for the actors $\mathbf{Y}$ (as defined by van de Velden *et al.* (2017)); 4) update the membership matrix $\mathbf{Z}_K$ by applying a clustering methods to $\mathbf{Y}$; 5) repeat the procedure (i.e. go back to step 2) until convergence. In the original paper the first solution has been proposed to be randomly assigning while the clustering algorithm is the k-means. In this paper we compare the performance of such method with the use of blockmodeling to provide both the initial cluster allocation $\mathbf{Z}_K$ and their updating. In this way, the network positions should be optimally separated with respect to the distributions over the events and, simultaneously, events with different participation patterns should be optimally separated (van de Velden *et al.*, 2017).

Hence, our main goals are i) to look at the way network positions, as they result from blockmodeling analysis, can be incorporated in the cluster CA method, and to assess the advantages of this strategy with respect to the one provided by Ragozini *et al.* (2018) (see also Serino *et al.*, 2017); ii) to analyse specific network structures (e.g. core-periphery and/or segmentation) and to verify if cluster CA is able to reveal and clearly represent such structures. The proposed approach will be shown by analysing an affiliation network made of 45 co-productions that 44 theatre companies located in the Campania Region (Italy) jointly released during the 2012/2013 season. In this data structure, where the rows represent the companies and the columns represent the stage co-productions, thanks to this approach we expect to find groups of theatre companies that share similar participation patterns and that are involved in co-productions with similar characteristics (i.e. belonging to the same genres). At the same time, we attempt to evaluate the structural similarities between the groups of companies on the basis of their projections in the metric space.

References

D'ESPOSITO, M. R., DE STEFANO, D., & RAGOZINI, G. 2014a. A Comparison of χ^2 Metrics for the Assessment of Relational Similarities in Affiliation Networks. *Pages 113–122 of: Analysis and Modeling of Complex Data in Behavioral and Social Sciences*. Springer.

D'ESPOSITO, M. R., DE STEFANO, D., & RAGOZINI, G. 2014b. On the use of multiple correspondence analysis to visually explore affiliation networks. *Social Networks*, **38**, 28–40.

DOREIAN, P., BATAGELJ, V., & FERLIGOJ, A. 2005. *Generalized blockmodeling*. Vol. 25. Cambridge university press.

GREENACRE, M., & BLASIUS, J. 2006. *Multiple correspondence analysis and related methods*. CRC Press.

PIZARRO, N. 2007. Structural identity and equivalence of individuals in Social Networks: beyond duality. *International Sociology*, **22**(6), 767–792.

RAGOZINI, G., SERINO, M., & D'AMBROSIO, D. 2018. On the Analysis of Time-Varying Affiliation Networks: The Case of Stage Co-productions. *Pages 119–129 of: Convegno della Società Italiana di Statistica*. Springer.

SERINO, M., D'AMBROSIO, D., & RAGOZINI, G. 2017. Bridging social network analysis and field theory through multidimensional data analysis: the case of the theatrical field. *Poetics*, **62**, 66–80.

VAN DE VELDEN, M., IODICE D'ENZA, A., & PALUMBO, F. 2017. Cluster correspondence analysis. *psychometrika*, **82**(1), 158–185.

A LATENT SPACE MODEL FOR CLUSTERING IN MULTIPLEX DATA

Silvia D'Angelo[1, 2, 3] and Michael Fop[1]

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

[2] Institute of Food and Health, School of Agriculture and Food Science,
University College Dublin,

[3] Insight Centre for Data Analytics, University College Dublin.

ABSTRACT: Network data are relational data recorded among a group of individuals, the nodes. Multiple relations observed among the same set of nodes may be represented by means of different networks, using a so-called multidimensional network, or multiplex. We propose a latent space model for network data that enables clustering of the nodes in a latent space, with clusters in this space corresponding to communities of nodes. The clustering structure is modelled using an infinite mixture distribution framework, which allows to perform joint inference on the number of clusters and the cluster parameters. An application to terrorist network data will be discussed.

KEYWORDS: multidimensional network, mixture model, latent space model.

1 Introduction

A network is defined by a set of nodes, among which a relation can be established. Binary networks record relations that are either present or absent between nodes, with presences corresponding to edges linking pairs of nodes. When multiple relations are recorded for a constant set of nodes, a multidimensional network, or multiplex, arises, where different relations coincide with different networks. Observed connections in network data are hard to interpret, due to the complexity and potential high dimensionality of networks themselves. Latent space models (Hoff *et al.*, 2002) are a class of models which aims at explaining the connections observed in network data in terms of unobserved similarities among the nodes. In distance latent space models (Hoff *et al.*, 2002), such similarities are modelled as distances between unobserved nodes coordinates in a latent space. A sub-class of latent space models (Handcock *et al.*, 2007) addresses the issue of clustering of the nodes, by clustering nodes latent coordinates.

We propose an extension to the Latent position cluster model (Handcock *et al.*, 2007), which allows clustering of the nodes both for single and multidimensional network data. An infinite mixture distribution framework is adopted, so that joint inference on both the number of clusters and the cluster parameters can be performed.

2 The model

2.1 Latent position cluster model

The model introduced by Handcock *et al.*, 2007 postulates that nodes have latent coordinates in a p-dimensional Euclidean latent space, z_i, $i = 1, \dots, n$, drawn from a mixture of G spherical Gaussian distributions,

$$z_i \sim \sum_{g=1}^{G} \pi_g MVN_p\left(\mu_g, \sigma_g^2 \mathbf{I}\right), \quad g = 1, \dots, G,$$

where π_g, $g = 1, \dots, G$, denote the mixture weights and μ_g, σ_g^2 the component-specific means and variances.

2.2 Infinite latent position cluster model

We propose to extend the model by Handcock *et al.*, 2007 assuming that the latent coordinates are distributed according to an infinite mixture of p-variate Gaussian components:

$$z_i \sim \sum_{g=1}^{\infty} \pi_g \mathcal{MVN}_p\left(\mu_g, \Sigma_g\right),$$

where Σ_g is the covariance matrix of the g^{th} component and component parameters are taken to be realizations of a Dirichlet process.

In general, for K-dimensional network data, the probability of observing an edge between any two nodes i and j in the k^{th} network ($k = 1, \dots, K$) is modelled as a function of their distance, $d(\cdot)$, and some other parameters (D'Angelo *et al.*, 2019):

$$P\left(y_{ij}^{(k)} \mid \alpha^{(k)}, \beta^{(k)}, z_i, z_j\right) = \frac{\exp\left(\alpha^{(k)} - \beta^{(k)} d(z_i, z_j)\right)}{1 + \exp\left(\alpha^{(k)} - \beta^{(k)} d(z_i, z_j)\right)}$$

The above equation simplifies to that of edge probabilities for single networks when $K = 1$. Inference for this model is performed within a hierarchical

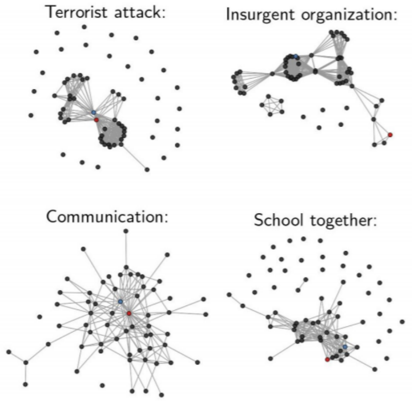


Figure 1. *Noordin Top data. The networks.*

Bayesian framework, where estimates of model parameters and latent coordinates are obtained using an MCMC algorithm.

3 Noordin Top multiplex data

To illustrate the proposed model, we have used it to analyse the Noordin Top multiplex data. The data concern four different relationships recorded among members of the Noordin Top terrorist organization, active in Indonesia in the early 2000s (see Figure 1).
An Infinite latent position cluster model with diagonal covariance matrices was estimated, and four different components were found in the latent space for the Noordin Top data, see Figure 2. Also, Noordin Top was positioned close to Azahari Husin, who was believed to be Noordin Top right-hand man ("star" coordinates in Figure 2). Both terrorists are assigned to the same component, which is also the largest.

4 Discussion

We have introduced an Infinite latent position cluster model to perform clustering of the nodes in network and multidimensional network data, by means of clustering of their latent coordinates in a latent space representation of the data. Thanks to the infinite mixture framework, and differently from previous methods (Handcock *et al.*, 2007), the proposed model is able to perform

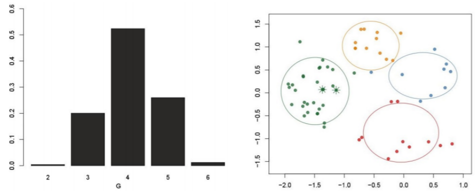


Figure 2. *Noordin Top data. Estimated posterior distribution of the number of components and estimated nodes latent coordinates and mixture components.*

joint inference on the latent coordinates, the component parameters, and the number of mixture components. Applying this model to the Noordin Top multiplex data we were able to recover four different components, among which a larger one (the green component in Figure 2) included the organization most influential members. From Figure 2, we may also notice that few nodes latent coordinates are located quite distant from the center of the components they were assigned to. Such issue may be addressed using a different specification of the components covariance matrices Σ_g. Another possible solution for latent coordinates that would still be located far away from components centres could be to investigate whether such coordinates should be clustered at all. Indeed, some nodes may not exhibit a clustering behaviour, either because they connect to only few others or because they connect randomly across different networks in a multiplex. Such nodes should not be forced to belong to one of the Gaussian components, as these correspond to social groups in the data. An extra component, arising from a different distribution (as for example a Uniform distribution), could be added to the infinite mixture framework, with the purpose of grouping together "poorly interacting" nodes.

References

D'Angelo, S., Murphy, T.B., & Alfò, M. 2019. Latent space modelling of multidimensional networks with application to the exchange of votes in Eurovision Song Contest. *The Annals of Applied Statistics*, **13**.

Handcock, M., Raftery, A., & Tantrum, J. 2007. Model-based clustering for social networks (with discussion). *Journal of the Royal Statistical Society: Series A*, **170**.

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

POST PROCESSING OF TWO DIMENSIONAL ROAD PROFILES: VARIOGRAM SCHEME APPLICATION AND SECTIONING PROCEDURE

Mauro D'Apuzzo[1], Rose-Line Spacagna[1], Azzurra Evangelisti[1], Daniela Santilli[1]
and Vittorio Nicolosi [2]

[1] Department of Civil and Mechanical Engineering (DICeM), University of Cassino,
(e-mail: dapuzzo@unicas.it, rlspacagna@unicas.it,
aevangelisti.ing@gmail.com, daniela.santilli@unicas.it)

[2] Department of Enterprise Engineering "Mario Lucertini", University of Rome,
"Tor Vergata" Rome, (e-mail: nicolosi@uniroma2.it)

ABSTRACT: Road sectioning plays a crucial role in Road Asset Management Systems and High Speed laser-based devices are able to collect a huge amount of data on pavement surface characteristics. However, this implies a high computational effort in identifying road homogeneous sections following a long and meticulous post processing analysis. The Geostatistic methodology, in terms of Variogram scheme has been applied for characterizing road surface: "Range" and "Sill" values, deriving from the Variogram application, have been proposed as macrotexture synthetic indices to characterized different road surfaces. Then a dynamic sectioning procedure has been employed to detect homogeneous road pavement sections. Preliminary results seem to highlight that the Variogram variables can be promising in identifying homogeneous sections in terms of pavement surface macrotexture.

KEYWORDS: pavement management, road surface macrotexture, dynamic sectioning, geostatistics variogram scheme, spatial data analysis.

1 Introduction

The quality and the quantity of the data collected by high speed laser-based (HSL) texture measuring devices for pavement road monitoring and programming of maintenance interventions, open new challenge to Pavement Managers in fact, in this context, new skills for filtering, analysing and interpreting of data are requested.

In order to apply the Pavement Management Systems (PMS) principles, an identification of homogeneous sections for subdividing road network is needed. These homogeneous sections can be defined as road sections in which the parameters, that generally affect the maintenance strategies, can be considered as almost constant. Usually the road profile texture data, collected by HSL (here from now on called HSL data), can be described as "time series" characterized by information on position and height with a fixed sampling frequency on a straight alignment. HSL data usually undergo to a pre-processing (filtering) procedure in order to remove noise and invalid readings (as spikes or drop-outs) according to several approaches [Losa & Leandri 2011; D'Apuzzo et al. 2015].

Relevant macrotexture descriptive indexes, such as *Estimated Texture Depth* (*ETD*) evaluated according to [ASTM E1845], can be derived from road surface filtered profiles, although more reliable macrotexture synthetic indexes have been recently proposed [D'Apuzzo et al. 2015].

In this paper an innovative approach to describe the macrotexture of road surface employing the Geostatistical method applied to characterized 2D road profiles by means of the Variogram scheme, is proposed. Transformed data so obtained undergo to a sectioning procedure, in order to identify the homogeneous pavement sections.

2 Methodology

Geostatistics is a field of the Statistics focused on the study of spatial or regionalized phenomena, which are characterized by a spatial correlation. Thanks to this peculiarity, several applications within environmental aspects have been performed [Chilès & Delfinet 1999; Spacagna et al. 2019] and encouraging results have been achieved from preliminary attempts for the road profiles analysis [M. Ech et al. 2007]. The spatial law can be defined by means of the Variogram, which describes the relation between two point at “h” distance and it presents the following structure:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \big(Z(x_i + h) - Z(x_i)\big)^2$$

Where:
γ(h) = Variogram;
h= distance between couple of points;
N(h) = number of couple of points at h distance;
Z(x_i) = value at x point;
Z(x_{i+h}) = value at x+h point.

In the literature, different Variogram models are presented [Chilès & Delfinet, 1999] and, in this study, the Spherical model has been used.

In general, the Variogram is characterized by two values the *Sill* and the *Range*. Within the *Range*, Z(x) and Z(x+h) values are related, outside are independent. For these reasons it is possible to define the *Sill* and the *Range* as the measure of the maximum variability and the distance where the variables are correlated, respectively.

Applying the Variogram to the filtered pavement profile, two new "time series", the *Sill* and the *Range* profiles, are produced and, to identify the homogeneous pavement sections, a dynamic sectioning process must be performed. Several methods, such as Bayesian methods, Cumulative Sum or Difference (CUMSUM) methods, Dichotomic method, minimum standard deviation based methods (MINRMS) and Linear models with Multiple Structural Change (LMSC), are available to identify homogeneous pavement sections from a series of measured data, and an interesting benchmarking has been previously proposed [D'Apuzzo et al., 2012]. In this paper the Dichotomic Method has been employed.

3 Case Study and Data Analysis

Pavement profiles measurements have been collected at the Virginia Smart Road, (Blacksburg, Montgomery County, Virginia) where more than 15 different pavement types and mixes have been laid. HSL device, which performs dynamic measurements on a straight alignment, with a laser spot of 0.2 mm and a sampling frequency of 64 kHz, has been used for the profile measurements. An example of pavement profile collected by HSL device along the entire Smart Road track (about 2300 m) has been reported in the Figure 1a.

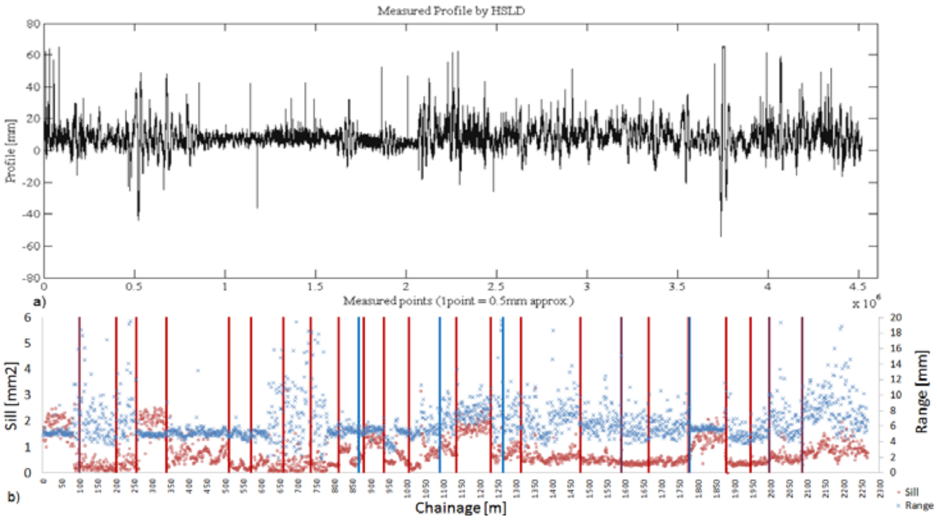


Figure 1: a) Measured profile; b) Sill and Range representation and Dichotomic sectioning restitution.

Following the profile cleaning phase, then the Variogram, with lag = 0.5mm and nlag = 40 (20mm), has been calculated, the Spherical Model has been applied and, on a window of 1m, Sill and Range have been evaluated with an autofitting process. In particular, 1 m long window has been identify as an optimal trade-off between precision sectioning needs in pavement asset management and computational effort required.

Graphical result has been summarized in the Figure 1b. As it is possible to see, the "time series" describe two different features of the same measured profile thus providing additional information on structural changes that can be used by sectioning methods. The Dichotomic method, with significance level (α) = 5% and sample size of 50, has been used for the identification of the homogeneous pavement road sections, and the results has been represented in the Figure 2b.

4 Conclusion

A Variogram scheme has been applied to the filtered road profile, measured by means of the HSL Device. Preliminary results show that Sill and Range can be considered as effective macrotexture indices since they can better highlight changes in pavement type and mixes. Dynamic sectioning by means of Dichotomic Method has been applied, yielding an identification rate of about 90% of real break points. Further studies are needed, nevertheless the developed methodology seems promising.

References

LOSA, M., & LEANDRI, P. 2011. The reliability of tests and data processing procedures for pavement macrotexture evaluation. *International Journal of Pavement Engineering*. Vol. **12**, No. 1, 59–73, Taylor and Francis. DOI: 10.1080/10298436.2010.501866.

D'APUZZO, M.; EVANGELISTI, A., FLINTSCH; G. W., DE L. IZEPPI, E., MOGROVEJO, D. E. & NICOLOSI, V. 2015. Evaluation of Variability of Macrotexture Measurement with Different Laser-Based Devices. *Airfield and Highway Pavements: Innovative and Cost-Effective Pavements for a Sustainable Future.* 294-305. TRIS, ASCE. DOI: 10.1061/9780784479216.027.

ASTM E1845, 2009. Standard Practice for Calculating Pavement Macrotexture Mean Profile Depth. *American Society for Testing and Materials*.

CHILÈS, J.P. & DELFINER, P. 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York.

SPACAGNA, R.L., MODONI, G., SAROLI, M, 2019. An integrated model for the assessment of subsidence risk in the area of Bologna (Italy). *Geotechnical Research for Land Protection and Development- Proceedings of CNRIG 2019 -* © Springer Nature Switzerland AG 2020 F. Calvetti et al. (Eds.): CNRIG 2019, LNCE 40, pp. 358–368, 2020. https://doi.org/10.1007/978-3-030-21359-6_38.

ECH M., S. MOREL, B. POUTEAU, YOTTE S., BREYSSE D. 2007. Laboratory evaluation of pavement macrotexture durability, *Revue Européenne de Génie Civil*, 11:5, 643-662. http://dx.doi.org/10.1080/17747120.2007.9692949.

D'APUZZO, M. & NICOLOSI, V. 2012 Detecting Homogeneous Pavement Section Using Econometric Test for Structural Changes in Linear Model. *Transportation Research Board 91st Annual Meeting* Paper n. 12-2125 , 0–18, Transportation Research Board, Washington DC, United States.

A NEW APPROACH TO PREFERENCE MAPPING THROUGH QUANTILE REGRESSION

Cristina Davino[1], Tormod Næs[2], Rosaria Romano[1] and Domenico Vistocco[3]

[1] Department of Economics and Statistics, University Federico II of Naples,
(e-mail: cristina.davino@unina.it, rosaroma@unina.it)

[2] Nofima, Norway, (e-mail: tormod.naes@nofima.no)

[3] Department of Political Science, University Federico II of Naples,
(e-mail: domenico.vistocco@unina.it)

ABSTRACT: The aim of the paper is to propose a new approach to preference mapping by exploiting quantile regression. The proposal consists into a multi-steps procedure combining principal component analysis, least squares and quantile regression. Results of the procedure on a case study show how the classical preference map can be enriched by information on the variability along the direction of the most preferred products. Such an additional information is obtained by the use of quantile regression.

KEYWORDS: preference mapping , least squares regression, quantile regression.

1 Introduction

Preference mapping (PREFMAP) exploits multivariate statistical techniques to analyze consumer acceptance of products. It consists of a two step procedure combining principal component analysis (PCA) and least squares regression (LSR) (Næs *et al.* , 2011). In the first step, a perceptual map of the products is obtained through a PCA of the product-by-attribute sensory matrix, and the obtained principal components are called key sensory dimensions. In the second step, a regression model is used to fit the liking of each consumer in the perceptual space. The main assumption is that the preference of each consumer depends linearly on the sensory attributes. Furthermore, as the method is grounded on LSR, it focuses on the average effects of sensory dimensions. Evaluating the effect of the sensory dimensions on the whole distribution of the liking can be a relevant challenge. At this aim, quantile regression (QR) (Koenker, 2005) has been recently introduced in consumer study for relating liking to consumer factors (Davino *et al.* , 2015), and for handling consumer heterogeneity (Davino *et al.* , 2018). Note that QR estimates as many models as the number of selected quantiles (Davino *et al.* , 2013, Furno & Vistocco, 2018). The aim of this study is to extend the use of QR to the PREFMAP in order to provide additional information on the variability along the direction of the most liked samples for each consumer. This is the most interesting direction in the perceptual space from a marketing perspective. The proposed approach will be discussed through a case study from consumer analysis based on the liking of yogurts. Specifically, 8 samples were profiled by a sensory panel according to 21 attributes: six odour attributes, three taste attributes, six flavour attributes and six texture attributes. The same samples were evaluated by a consumer panel consisting of 101 consumers on a scale from 0=dislike extremely to 100=like extremely. The details of the experiment can be found in (Nguyen *et al.* , 2018).

2 Quantile regression in preference mapping

The proposal consists into a multi-step procedure. In the first step a PCA of the product-by-attribute sensory matrix is used to obtain a perceptual map of the products. The score and the loading plots on yogurt data are shown in Figure 1. Along the first component, one can notice a clear distinction between the samples on the right side (P3, P4, P7, P8) and the ones on the left side (P1, P2, P5, P6). The second component is mostly related to distinguishing product 7, characterized by sickening odour and flavour, from product 2 characterized by fullness and thickness.

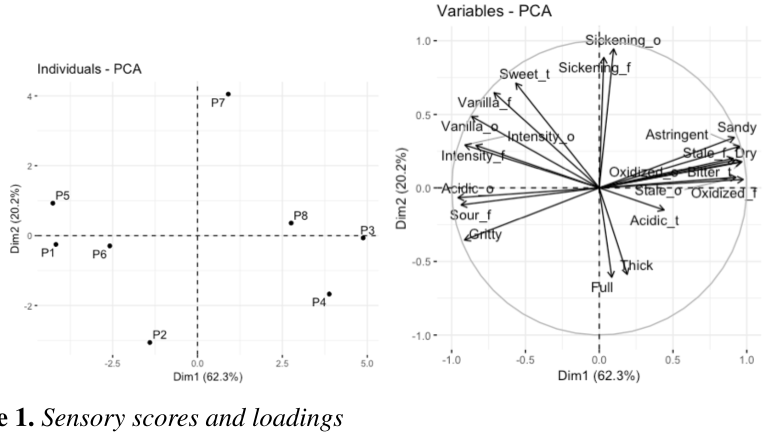


Figure 1. *Sensory scores and loadings*

In the second step, a regression model is used to fit each consumer in the perceptual space. Let **Y** be the matrix of liking values ($I \times J$), where the entry y_{ik} is the measured liking value of product i and consumer j ($j = 1, \ldots, J$). The liking values for each consumer are regressed onto the first *sensory dimensions*, most often the first two PC's:

$$y_{ij} = \beta_{j1} t_{i1} + \beta_{j2} t_{i2} + f_{ij} \tag{1}$$

where t_{i1} and t_{i2} comes from the PCA model, the β's represent the regression coefficients (also called consumer loadings) and f_{ij} represents the residuals. The intercept can be avoided here since the variables are centered.

In the third step, the direction of the most liked samples is identified by the $\beta_{\mathbf{j}}$ regression coefficients. Here, each sample i is projected onto the direction identified in the previous step (in the t_1, t_2 space):

$$s_{ij} = \left(\hat{\beta}_j^T \hat{\beta}_j\right)^{-1} \hat{\beta}_j^T \mathbf{t}_i \quad \text{where} \quad \mathbf{t}_i = (\mathbf{t}_{i1}, \mathbf{t}_{i2}) \tag{2}$$

Finally, a QR is exploited to evaluate if the distribution of liking is wider or narrower in the direction of increased liking. Specifically, two quantile regression models are estimated for $\theta = 0.25$ and $\theta = 0.75$:

$$\hat{y}_{ij}(\theta) = \hat{\beta}_{j0}(\theta) + \hat{\beta}_j(\theta) s_{ij} \tag{3}$$

For each consumer the two QR lines can diverge or converge as a function of s thus providing information on the variability of the liking along this direction. Consumers can then be classified according to whether the variability is larger for the most liked area in the sensory space than for the least liked samples. In order to measure the degree of such variability, the distance between fitted values at $\theta = 0.25$ and $\theta = 0.75$ has been computed at two fixed values of the **s** regressor corresponding to the first and third quartiles. In case the two distances between the fitted values differ not more than a fixed threshold the lines are considered parallel (the choice of the threshold is data driven). Based on this, we decided to consider 3 consumer categories, parallel, diverging and converging. Figure 2 depicts the consumer loadings plot from standard PREFMAP, but now the size of the points is proportional to the variability measure previously computed (based on the two values of **s**) and the shape is related to the distribution around the regression line in the direction of preference (converging, diverging, parallel). Three consumers C57, C75 and C87 are highlighted as they show different tendencies (C75 is represented by a very small star above a close diverging consumer). As can be seen from Figure 2, there is a relatively clear tendency of more convergence to the left and divergence to the right. In other words, for the sensory region represented by samples P3, P4 and P8 in Figure 1, the liking is more ‘flexible’ than in the opposite direction. With the exception of a few, the parallel consumers seem to be quite centrally positioned, i.e. most of them are consumers with low or moderately strong preference pattern (coefficients moderately large).

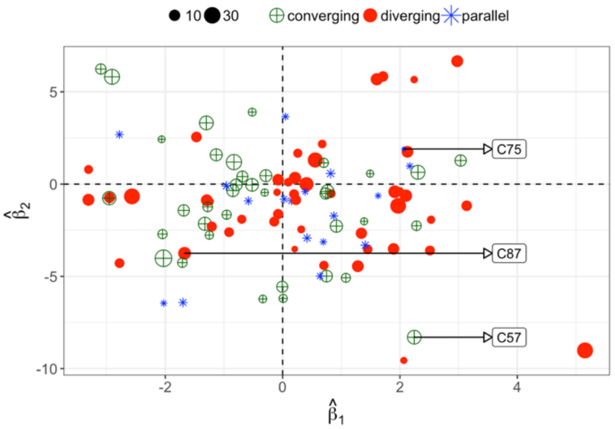


Figure 2. *Consumer loadings plot where the size of the points is proportional to the QR variability measure.*

References

Davino, C, Furno, M, & Vistocco, D. 2013. *Quantile regression: theory and applications*. John Wiley & Sons.

Davino, C, Romano, R, & Næs, T. 2015. The use of quantile regression in consumer studies. *Food quality and preference*, **40**, 230–239.

Davino, C, Romano, R, & Domenico, V. 2018. Modelling drivers of consumer liking handling consumer and product effects. *Italian Journal of Applied Statistics*, **30**, 359–372.

Furno, M, & Vistocco, D. 2018. *Quantile regression: estimation and simulation*. John Wiley & Sons.

Koenker, R. 2005. *Quantile Regression (Econometric Society monographs; no. 38)*. Cambridge university press.

Næs, T, Brockhoff, P, & Tomic, O. 2011. *Statistics for sensory and consumer science*. John Wiley & Sons.

Nguyen, QC, Næs, T, & Varela, P. 2018. When the choice of the temporal method does make a difference: TCATA, TDS and TDS by modality for characterizing semi-solid foods. *Food quality and preference*, **66**, 95–106.

ON THE ROBUSTNESS OF THE COSINE DISTRIBUTION DEPTH CLASSIFIER

Houyem Demni[1, 2], Amor Messaoud[2] and Giovanni C. Porzio[1]

[1] Department of Economics and Law, University of Cassino and Southern Lazio, (e-mail: houyem66gmail.com, porzio@unicas.it)

[2] University of Tunis, (e-mail: amor.messaoud@gmail.com)

ABSTRACT: Investigating how classifiers perform under some data contaminations is an important issue in robustness studies. While some research is available on the robustness of classifiers, a little is known about directional classifiers. This work thus investigates the robustness of the cosine depth distribution classifier, a classification technique recently introduced for directional data. This latter is a non-parametric method and it is based on the distribution function of the cosine depth.

KEYWORDS: directional data, supervised classification, unit vectors.

1 Introduction

Directional data occur when observations are recorded as directions. They can be described as unit vectors on the surface of the $(d-1)$ dimensional hypersphere $S^{(d-1)} := \{x : x^T x = 1\}$. This kind of data can be found in many scientific areas such as medicine, astronomy, biology and geology, to cite a few. Applications include cases with $d = 2$ (circular data), $d = 3$ (Mardia & Jupp, 2000) and in higher dimensions (Buchta *et al.*, 2012).

In this work, we consider the problem of classifying directional data according to some supervised classification technique, and in particular on a technique which relies on data depth.

Data depth functions provide basis for nonparametric inference given that they aim at ordering data in a d-dimensional space according to some centrality measures. The particular properties of directional data and the complexity of the sample space imply the need of specific methods to analyze them.

Within the framework of classification, the use of data depth has been extensively investigated and successfully applied. The max depth classifier has been firstly developed (Ghosh & Chaudhuri, 2005, after Liu *et al.*, 1990). Later, the idea has been extended and the DD-classifier has been introduced (Li *et al.*, 2012).

A recent interest arises on the use of depth based classifier for directional data: the use of the directional max-depth classifier based on some new depth functions has been investigated (Pandolfo *et al.*, 2018a), and the DD-plot classifier for circular data has been discussed (Pandolfo *et al.*, 2018b).

Even more recently, a depth based distribution classifier was introduced in the framework of supervised classification to assign points lying on the surface of hyper-spheres (spherical data) to groups (Demni *et al.*, 2019). It was based on the cosine depth, and called the cosine distribution depth classifier. Simulation results showed that the cosine depth distribution classifier outperforms the max depth classifier in term of average misclassification rate also in many settings.

In supervised classification, the presence of anomalous observations in the training set can greatly reduce the effectiveness of the classification method adopted (Vencalek & Pokotylo, 2018). For this reason, it is always of interest to investigate the robustness of these kind of techniques. Several works dealt with robust based classifiers (see Dutta & Ghosh, 2012; Li *et al.*, 2012). Pandolfo investigated some robustness aspects of the DD-classifier for directional distributions (Pandolfo, 2017).

Here, the focus will be on the cosine depth distribution classifier. By means of a simulation study, it will be investigated to what extent this classifier is able to deal with contaminated training sets. The rest of the work is organized as follows. Section 2 introduces the directional cosine depth distribution classifier, while in Section 3 the simulation scheme that will be used to assess its robustness is provided.

2 The cosine depth distribution classifier

Directions in d-dimensional spaces can be represented as unit vectors x on the sphere $S^{(d-1)} := \{x : x^T x = 1\}$ with unit radius and center at the origin. A distribution H with support $\Omega \subseteq S^{(d-1)}$ is called a directional distribution. By definition, the cosine depth of a point $x \in S^{(d-1)}$ with respect to H is given by:

$$D_{cos}(x,H) := 2 - E_H[(1 - x'W)],$$

where $E[.]$ is the expected value, and W is a random variable from H.

The cumulative distribution function of the cosine depth function $F_D^H(x)$ is given by:

$$F_D^H(x) := P(D_{cos}(X,H) \leq D_{cos}(x,H))$$

Suppose now observations come from either the distribution (group) H_1 or H_2. Then, the directional depth distribution classification rule (Demni *et al.*, 2019) is given by:

$$\begin{cases} F_D^{\hat{H}_1}(x) > F_D^{\hat{H}_2}(x) \Longrightarrow \text{assign } x \text{ to population 1} \\ F_D^{\hat{H}_1}(x) < F_D^{\hat{H}_2}(x) \Longrightarrow \text{assign } x \text{ to population 2,} \end{cases}$$

where $\hat{H}$ refers to the empirical distribution.

If $F_D^{\hat{H}_1}(x) = F_D^{\hat{H}_2}(x)$, the classification rule will randomly assign the observation to one of the two groups with equal probability.

3 A simulation scheme to study the robustness of the cosine depth distribution classifier

To investigate the robustness properties of the cosine depth distribution classifier for directional data, the following simulation setting will be used.

Let H_1 and H_2 be two von Mises-Fisher distributions (vMF). That is, their corresponding density functions $h()$ are given by

$$h(x;\mu,c) := \left(\frac{c}{n}\right)^{d/2-1} \frac{1}{\Gamma^{(d/2)} I_{d/2-1}(c)} \exp\{c\mu^T x\},$$

where $c \geq 0$, $||\mu|| = 1$, and I_v denotes the modified Bessel function of the first kind and order v. The parameters μ and c are the mean direction and the concentration parameter, respectively.

The training set size will be 1000 (500 from each group), while the size of the testing set will be 500. The number of replications will be set equal to 150 times. For the concentration parameters c_1 and c_2 of H_1 and H_2, we consider two cases: equal concentration ($c_1 = c_2 = 5$), and different concentration ($c_1 = 2$ and $c_2 = 6$).

The location parameters for μ_1 and μ_2 are set to be equal to $(0,0,1)$, $(1,0,0)$ in dimension $d = 3$, respectively. The training observations from H_1 are contaminated with observations generated from VmF with location parameter equal to $\mu = (0,0,-1)$ and concentration parameter $c = 8$.

The location parameters are set to be equal to $\mu_1 = (0,0,0,0,0,0,0,0,0,1)$, and $\mu_2 = (1,0,0,0,0,0,0,0,0,0)$ in dimension $d = 10$.
Contaminated observations are generated from VmF with location parameter $\mu = (0,0,0,0,0,0,0,0,0,-1)$ and concentration parameter $c = 8$.

Finally, the contamination levels will be set equal to 0%, 10%, 20%.

References

BUCHTA, C, KOBER, M, FEINERER, I, & HORNIK, K. 2012. Spherical k-means clustering. *Journal of Statistical Software*, **50**(10), 1-22.

DEMNI, H, MESSAOUD, A, & PORZIO, G.C. 2019. The Cosine depth distribution classifier for directional data. *In:* ICKSTADT K, TRAUTMANN H, SZEPANNEK G LÜBKE K, & N, BAUER (eds), *Applications in Statistical Computing, Chapter 4*. Springer Nature Switzerland AG, in press. `https://doi.org/10.1007/978-3-030-25147-5_4`.

DUTTA, S, & GHOSH, A. K. 2012. On robust classification using projection depth. *Annals of the Institute of Statistical Mathematics*, **64**(3), 657-676.

GHOSH, A. K, & CHAUDHURI, P. 2005. On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, **32**(2), 327-350.

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, *et al.* 1990. On a notion of data depth based on random simplices. *The Annals of Statistics*, **18**(1), 405-414.

MARDIA, K.V, & JUPP, P.E. 2000. Directional Statistics, John Wiley and Sons, London.

PANDOLFO, G. 2017. Robustness aspects of DD-classifiers for directional data. *In:* GRESELIN, F, MOLA F, & ZENGA, M (eds), *CLADAG 2017 Book of Short Papers*. Universitas Studiorum S.r.l. Casa Editrice, Mantova.

PANDOLFO, G, PAINDAVEINE, D, & PORZIO, G. C. 2018a. Distance-based depths for directional data. *Canadian Journal of Statistics*, **46**(4), 593-609.

PANDOLFO, G, D'AMBROSIO, A, & PORZIO, G.C. 2018b. A note on depth-based classification of circular data. *Electronic Journal of Applied Statistical Analysis*, **11**(2), 447-462.

VENCALEK, O., & POKOTYLO, O. 2018. Depth-weighted Bayes classification. *Computational Statistics and Data Analysis*, **123**, 1-12.

NETWORK EFFECT ON INDIVIDUAL SCIENTIFIC PERFORMANCE: A LONGITUDINAL STUDY ON AN ITALIAN SCIENTIFIC COMMUNITY

Domenico De Stefano[1], Giuseppe Giordano[2] and Susanna Zaccarin[3]

[1] Department of Political and Social Sciences, University of Trieste, (e-mail: ddestefano@units.it)

[2] Department of Political and Social Studies, University of Salerno, (e-mail: ggiordano@unisa.it)

[3] Department of Economics, Business, Mathematics and Statistics University of Trieste, (e-mail: susanna.zaccarin@deams.units.it)

ABSTRACT: One of the most debated questions in scientific network analysis is the impact of collaboration on scientific performance, that is the effect of actors' embeddedness in co-authorship networks on their individual research outputs. Recent literature showed that specific centrality measures (e.g., closeness, betweenness) are correlated with indicators of scientific performance. This contribution intends to explore the influence of actors' embeddedness in co-authorship networks in a longitudinal framework. By adopting a Stochastic Actor-Oriented Model, we will model scientific performance (and its measurement) and authors' collaborative behaviour as a particular mechanism of 'social influence' over time.

KEYWORDS: scientific collaboration, co-authorship networks, SAO models, social influence.

1 Introduction

Several studies have shown that scientific productivity depends, among other factors, on scientists' attitudes towards collaboration in research (see Lee & Bozeman, 2005 and Wuchty *et al.*, 2007). In their collaborative interactions, scientists can benefit by both methodological and technological complementarity and synergy, improving the quality and quantity of their research output. In this stream of research, Social Network Analysis (SNA) has become the privileged theoretical and statistical approach to study the typical collaboration patterns within disciplines (for instance, see De Stefano & Zaccarin, 2016, Ferligoj *et al.*, 2015). Collaboration among scientists can be represented as a network, in which the actors are scholars and ties may be represented by various forms of scientific collaboration among them. The most frequent way of specifying such networks is to take into account formal research activities, especially co-authorship (i.e., co-production of scientific publications, Bellotti *et al.*, 2016). One of the most debated questions in collaboration network analysis is the impact of collaboration on scientific performance, that is the effect of actors' embeddedness in co-authorship networks on their individual research outputs (Abbasi *et al.*, 2011).

In the light of these findings, our contribution intends to add new empirical evidence on the topic of the impact of collaboration on scientific performance, exploring the influence of actors' embeddedness in co-authorship networks in a longitudinal framework. We will model scientific performance (and its measurement, e.g. *h-index*) as a particular mechanism of 'social influence' over time. To this end we will use performance and co-authorship data on Italian statisticians in convenient time periods before and after the two Italian research evaluation exercises (VQR1 and VQR2 respectively on products published in the periods 2004-2010 and 2011-2014).

2 Theoretical framework

Several studies recognized research collaboration as a key element in knowledge advancement because it facilitates interactions, exchanges, sharing methods and techniques – even from different fields – allowing a fertile ground for the development of new ideas. A further aspect of research collaboration, investigated in empirical studies, is the association with scientific performance, especially at individual level.

Melin, 2000 underlined the increase on knowledge and quality deriving from collaboration. Baker, 2015 documented its crucial role in individuals' job mobility and academic success. Other authors found that collaboration is a strong predictor of publishing productivity (Lee & Bozeman, 2005) although with controversial results depending on the choice of the productivity measure, while other authors (Abbasi *et al.*, 2011) found evidence of a positive correlation between performance and several network measures.

Combining co-authorship data from different sources (ISI-WoS, Current Index to Statistics, and publications in nationally funded projects), De Stefano *et al.*, 2013 and De Stefano & Zaccarin, 2016 analysed the impact of collaboration on scientific performance of the Italian academic statisticians. Their findings show that specific centrality measures (e.g., closeness, betweenness) are correlated with indicators of scientific performance, even if this impact is affected by heterogeneity depending on the discipline and on the data source used to construct the co-authorship networks.

3 Data and modelling

The research hypothesis of the present contribution relies on the idea that across the two evaluation exercises the community under study – Italian academic statisticians – tends to change their collaboration behavior. Our aim is to analyze the relation between co-authorship network indicators across the two VQR exercises in the period 2004-2010 and 2011-2014 and how authors' position in the co-authorship network affects their scientific performance in a longitudinal perspective.

In particular, we analyze the co-authorship networks across these two periods as retrieved from the scientific production of the Italian academic statisticians. That is, those scientists classified as belonging to one of the five subfields established by the governmental official classification: Statistics (Stat), Statistics for Experimental and Technological Research (Stat for E&T), Economic Statistics (Economic Stat), Demography (Demo), and Social Statistics (Social Stat).We recover the scientific production and the bibliographic metadata of the Italian academic statisticians from the novel IRIS platform for publications data storage (`https://www.cineca.it`). From the retrieved metadata we will compute *ad hoc* indicator for measuring individual scientific performance. Then, we will treat authors' performance as a behavioral variable in a Stochastic Actor-Oriented Model (SAOM) in order to disentangle how co-authorship affects performance ('behavior') and *viceversa*.

SAOM approach allows to model if and what type of local network configuration is associated to the increase or decrease of the individual scientific performance. The comparison between different periods will consider explicitly the temporal dimension. The SAOM describes the development of a network through time as a result of the relational choices of a set of individual actors in order to maximize their utility Snijders *et al.*, 2010. It is a combination of random utility model, continuous time Markov model and simulation. When actors change their personal network they may face several options, for example, to create a new collaboration tie, dropping an existing one or leaving ties unchanged. Under certain condition, the probability of these choices can be specified as a multinomial logit model with the utility functions being the linear cores. The utility function expresses the characteristics of actors' personal networks toward which the actors seem to be attracted.

Adding to longitudinal networks the so-called behavior consists in considering one or more changing nodal variables – performance measures in our case – that are also treated as dependent variables. The network will influence the dynamics of the behavior, as well as the behavior will influence the dynamics of the network. Roughly speaking, this means to consider the co-evolution of networks and behavior. In particular, by means of this approach we will model the change in authors' performance indicator (for instance the propensity to publish in high impact journals) depending on the embeddedness in the co-authorship networks between the two VQR exercises periods.

References

ABBASI, A., ALTMANN, J., & HOSSAIN, L. 2011. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *J. Informetrics*, **5**.

BAKER, A. 2015. Non-tenured post-doctoral researchers job mobility and research output: An analysis of the role of research discipline, department size, and coauthors. *Research Policy*, **44**(3), 634 – 650.

BELLOTTI, E., KRONEGGER, L., & GUADALUPI, L. 2016. The evolution of research collaboration within and across disciplines in Italian Academia. *Scientometrics*, **109**(2), 783–811.

DE STEFANO, D., & ZACCARIN, S. 2016. Co-authorship networks and scientific performance: an empirical analysis using the generalized extreme value distribution. *Journal of Applied Statistics*, **43**(1), 262–279.

DE STEFANO, D., FUCCELLA, V., VITALE, M.P., & ZACCARIN, S. 2013. The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks*, **35**(3), 370 – 381.

FERLIGOJ, A., KRONEGGER, L., MALI, F., SNIJDERS, T. A. B., & DOREIAN, P. 2015. Scientific collaboration dynamics in a national scientific system. *Scientometrics*, **104**(3), 985–1012.

LEE, S., & BOZEMAN, B. 2005. The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science*, **35**(5), 673–702.

MELIN, G. 2000. Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, **29**(1), 31–40.

SNIJDERS, T.A.B., VAN DE BUNT, G. G., & STEGLICH, C.E.G. 2010. Introduction to stochastic actor-based models for network dynamics. *Social Networks*, **32**(1), 44 – 60. Dynamics of Social Networks.

WUCHTY, S., JONES, B.F., & UZZI, B. 2007. The Increasing Dominance of Teams in Production of Knowledge. *Science*, **316**(5827), 1036–1039.

PENALIZED VS CONSTRAINED MAXIMUM LIKELIHOOD APPROACHES FOR CLUSTERWISE LINEAR REGRESSION MODELING

Roberto Di Mari[1], Stefano Antonio Gattone[2] and Roberto Rocci[3, 4]

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

[2] Department of Philosophical and Social Sciences, Economics and Quantitative Methods,
University G. d'Annunzio, Chieti-Pescara, (e-mail: gattone@unich.it)

[3] Department of Statistical Sciences, University of Rome La Sapienza,

[4] Department of Economics and Finance, University of Rome Tor Vergata,
(e-mail: roberto.rocci@uniroma2.it)

ABSTRACT: Several approaches exist to avoid singular and spurious solutions in maximum likelihood (ML) estimation of clusterwise linear regression models. We propose to solve the degeneracy problem by using a penalized approach: this is done by adding a penalty term to the log-likelihood function which increasingly penalizes smaller values of the scale parameters and the tuning of the penalty term is done based on the data. Another traditional solution to degeneracy consists in imposing constraints on the variances of the regression error terms (constrained approach). We will compare the penalized approach to the constrained approach in a broad simulation study and an empirical application, providing practical guidelines on which approach to use under different circumstances.

KEYWORDS: clusterwise linear regression, penalized likelihood, scale constraints.

1 Introduction

Let $\mathrm{y}_1, \dots, \mathrm{y}_n$ be a sample of independent observations drawn from the response random variable Y_i, each observed alongside with a vector of J explanatory variables $\mathbf{x}_1, \dots, \mathbf{x}_n$. Let us assume $\mathrm{Y}_i|\mathbf{x}_i$ to be distributed as a finite mixture of linear regression models, that is

$$f(\mathrm{y}_i|\mathbf{x}_i;\boldsymbol{\psi}) = \sum_{g=1}^{G} p_g \phi_g(\mathrm{y}_i|\mathbf{x}_i, \sigma_g^2, \boldsymbol{\beta}_g) = \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left[-\frac{(\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\beta}_g)^2}{2\sigma_g^2}\right], \tag{1}$$

where G is the number of clusters and p_g, $\boldsymbol{\beta}_g$, and σ_g^2 are the mixing proportion, the vector of $J+1$ regression coefficients that includes an intercept, and the variance term for the g-th cluster. The set of all model parameters is given by $\boldsymbol{\psi} = \{(p_1, \ldots, p_G; \boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_G; \sigma_1^2, \ldots, \sigma_G^2) \in \mathbb{R}^{(G-1)+(J+1)G+G} : p_1 + \cdots + p_G = 1, p_g > 0, \sigma_g^2 > 0$, for $g = 1, \ldots, G\}$.

The likelihood function can be specified as

$$\mathcal{L}(\boldsymbol{\psi}) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{G} p_g \frac{1}{\sqrt{2\pi\sigma_g{}^2}} \exp\left[-\frac{(\mathrm{y}_i - \mathbf{x}_i'\boldsymbol{\beta}_g)^2}{2\sigma_g^2} \right] \right\}, \tag{2}$$

which we maximize to estimate $\boldsymbol{\psi}$ either by means of direct maximization or with the perhaps more popular EM algorithm (Dempster *et al.*, 1977). However, there is a well-known complication in ML estimation of this class of models: the likelihood function of mixtures of (conditional) normals with cluster-specific variances is unbounded (Kiefer & Wolfowitz, 1956; Day, 1969).

A traditional solution to the problem of unboundedness is based on the seminal work of Hathaway (1985) which, in order to have the likelihood function of univariate mixtures of normals bounded, suggested to impose a lower bound to the ratios of the scale parameters in the maximization step. The method is equivariant under linear affine transformations of the data. That is, if the data are linearly transformed, the estimated posterior probabilities do not change and the clustering remains unaltered. Recently, in the multivariate case, Rocci *et al.* (2018) incorporated constraints on the eigenvalues of the component covariances of Gaussian mixtures that are tuned on the data based on a cross–validation strategy. These constraints are built upon Ingrassia (2004)'s reformulation and are an equivariant sufficient condition for Hathaway's constraints. Estimation is done in a familiar ML environment Ingrassia & Rocci (2007), with data–driven selection of the scale balance. Di Mari *et al.* (2017) adapted Rocci *et al.* (2018)'s method to clusterwise linear regression, further investigating its properties.

Another possible approach for handling unboundedness is to modify the log-likelihood function by adding a penalty term, in which smaller values of the scale parameters are increasingly penalized. Representative examples can be found in Chen & Tan (2009); Chen *et al.* (2008); Ciuperca *et al.* (2003).

In this work we review the constrained approach of Di Mari *et al.* (2017) and develop a data-driven equivariant penalized approach for ML estimation. Next, we sketch the bulk of the methodologies.

2 The methodology

2.1 The constrained approach

Di Mari *et al.* (2017) proposed relative constraints on the group conditional variances σ_g^2 of the kind

$$\sqrt{c} \leq \frac{\sigma_g^2}{\bar{\sigma}^2} \leq \frac{1}{\sqrt{c}}, \tag{3}$$

or equivalently

$$\bar{\sigma}^2\sqrt{c} \leq \sigma_g^2 \leq \bar{\sigma}^2\frac{1}{\sqrt{c}}. \tag{4}$$

The above constraints are equivariant and have the effect of shrinking the variances to a suitably chosen $\bar{\sigma}^2$, the *target* variance term, and the level of shrinkage is given by the value of c. This constraints are easily implementable within the EM algorithm (Ingrassia, 2004; Ingrassia & Rocci, 2007), which is fully available in closed-form, and the selection of c is based on the data.

2.2 The penalized approach

An alternative to the constrained estimator is the penalized approach, in which a penalty $s_n(\sigma_1^2,\ldots,\sigma_G^2)$ is put on the component variances and it is added to the log-likelihood. Under certain conditions on the penalty function, the penalized estimator is know to be consistent (Chen & Tan, 2009). A function s_n that satisfies these conditions is

$$s_n(\sigma_1^2,\ldots,\sigma_G^2) = -\lambda\sum_{g=1}^{G}\left(\frac{\bar{\sigma}^2}{\sigma_g^2}+\log(\sigma_g^2)\right), \tag{5}$$

where $\bar{\sigma}^2$, the *target* variance, can be seen as our *prior* information on the scale structure and λ is the penalizing constant that is selected based on the data. Thus, the penalized log-likelihood can be written as

$$p\ell(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) + s_n(\sigma_1^2,\ldots,\sigma_G^2) \tag{6}$$

and the set of unknown parameters is found by ML with computation done by means of an EM algorithm that is available in closed-form. As well as with the constrained approach, the penalized approach is equivariant with respect to linear transformation in the response.

References

CHEN, J., & TAN, X. 2009. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, **100**(7), 1367–1383.

CHEN, J., TAN, X., & ZHANG, R. 2008. Inference for normal mixtures in mean and variance. *Statistica Sinica*, 443–465.

CIUPERCA, G., RIDOLFI, A., & IDIER, J. 2003. Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, **30**(1), 45–59.

DAY, N.E. 1969. Estimating the components of a mixture of normal distributions. *Biometrika*, **56**(3), 463–474.

DEMPSTER, A.P., LAIRD, N.M., & RUBIN, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

DI MARI, R., ROCCI, R., & GATTONE, S.A. 2017. Clusterwise linear regression modeling with soft scale constraints. *International Journal of Approximate Reasoning*, **91**, 160–178.

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

INGRASSIA, S. 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications*, **13**(2), 151–166.

INGRASSIA, S., & ROCCI, R. 2007. Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics & Data Analysis*, **51**(11), 5339–5351.

KIEFER, J., & WOLFOWITZ, J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.

ROCCI, R., GATTONE, S.A., & DI MARI, R. 2018. A data driven equivariant approach to constrained Gaussian mixture modeling. *Advances in Data Analysis and Classification*, **12**(2), 235–260.

LOCAL FITTING OF ANGULAR VARIABLES OBSERVED WITH ERROR

Marco Di Marzio[1], Stefania Fensore[1], Agnese Panzera[2]
and Charles C. Taylor[3]

[1] DSFPEQ, University of Chieti-Pescara, (e-mail: marco.dimarzio@unich.it, stefania.fensore@unich.it)

[2] DISIA, University of Florence, (e-mail: a.panzera@disia.unifi.it)

[3] Department of Statistics, University of Leeds,
(e-mail: charles@maths.leeds. ac.uk)

ABSTRACT: The problem of estimating a circular regression when the predictor is contaminated by errors is studied. Other than some estimators, we also present a novel smoothing degree selection rule.

KEYWORDS: deconvolution, measurement error, Simex.

1 Introduction

Statistical regression models are generally based on the assumption that the independent variables have been measured exactly. However, sometimes the regressors are, for some reason, not directly observable or measured with errors. When this is the case specific models, known as *errors-in-variables* or *measurement error models*, have to be taken into account.

Formally, suppose that we are interested in estimating the regression of Y on X^*, denoted as m, and that our data are realizations from variables $X = X^* + \eta$ and Y, say $(x_1, y_1), \ldots, (x_n, y_n)$. A general model for this case could be

$$\begin{aligned} y_i &= m(x_i^*) + \zeta_i \\ x_i &= x_i^* + \eta_i \end{aligned} \tag{1}$$

for $i = 1, \ldots, n$, where X^* and Y respectively refer to the predictor and response variable, ζ_is are observations of the random error term ζ, η_is are realizations of η. The unobserved variable X^* is always referred as latent or true variable. Usual assumptions include that ζ is independent from both X^* and η, the distribution of ζ is unknown but has mean 0 and constant variance, while the distribution of η is known.

Let f_X, f_{X^*} and f_η respectively denote the probability density function of X, X^* and η. Basic theoretical considerations suggest that f_X is the convolution between f_{X^*} and f_η:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X^*}(x-\nu)dF_\eta(\nu), \tag{2}$$

where F_η denotes the distribution function of η. As the consequence, the estimators of the free-error model are clearly not consistent. In such a context there are two approaches to obtain accurate estimates: deconvolution methods and explicit bias estimation and correction.

In this paper we address the measurement error case when data can be represented as points on a circumference. Specifically, we present a nonparametric deconvolution estimator along with a rule for smoothness selection.

2 Circular data

Angular or circular data are collected whenever observations are measured by means of a periodic scale. They are usually represented as points on the circumference of a circle with unit radius. Classical examples of such data are wind directions, animal movements, any phenomenon measured by the 24 h clock, etc. Once a zero direction and a sense of rotation have been arbitrarily chosen, these observations can be expressed as angles. Due to their periodic nature, circular data cannot be analysed by standard real-line methods, therefore in the last decades great attention has been devoted to circular statistics. For a comprehensive account, see the survey paper by Lee, 2010, and the references therein.

3 The estimator

Consider a pair of random angles (Θ, Δ), i.e. variables taking values on $[0, 2\pi)$. Given the random sample $(\Phi_1, \Delta_1), \ldots, (\Phi_n, \Delta_n)$, we can write model (1) as

$$\begin{aligned} \Delta_i &= (m(\Theta_i) + \varepsilon_i)\mathsf{mod}(2\pi), \\ \Phi_i &= \Theta_i + u_i, \end{aligned} \tag{3}$$

where Θ_is are independent copies of the circular latent variable Θ, the ε_is are i.i.d. random angles independent of the Θ_is, with zero mean direction and finite concentration, and the u_is are realizations of the random angle U independent of the Θ_is.

A local estimator for m at $\theta \in [0,2\pi)$ can be defined as

$$\hat{m}(\theta;\kappa) = \mathsf{atan2}(\hat{m}_s(\theta;\kappa), \hat{m}_c(\theta;\kappa)), \tag{4}$$

with

$$\hat{m}_s(\theta;\kappa) = \sum_{i=1}^{n} \sin(\Delta_i) L_\kappa(\Theta_i - \theta),$$
$$\hat{m}_c(\theta;\kappa) = \sum_{i=1}^{n} \cos(\Delta_i) L_\kappa(\Theta_i - \theta),$$

where the function $\mathsf{atan2}(y,x)$ returns the angle between the x-axis and the vector from the origin to (x,y), and L_κ is a circular *deconvolution kernel* function depending on $\gamma_\ell(\kappa)$ and $\lambda_\ell(\kappa_U)$ which are, for $\ell \in \mathbb{Z}$, respectively, the ℓth Fourier coefficient of the periodic weight function K_κ and the error density f_U whose concentration parameter is κ_U:

$$L_\kappa(\theta) = \frac{1}{2\pi}\left\{1 + 2\sum_{\ell=1}^{\infty} \frac{\gamma_\ell(\kappa)}{\lambda_\ell(\kappa_U)} \cos(\ell\theta)\right\}. \tag{5}$$

4 Smoothing degree selection

In the context of measurement error the standard cross-validation criterion for the selection of the smoothing degree κ is not suitable. Indeed, if we knew the values $\Theta_1, \ldots, \Theta_n$ in addition to $(\Phi_1, \Delta_1), \ldots, (\Phi_n, \Delta_n)$ then we could compute the conventional cross-validation smoothing degree $\hat{\kappa}_0 = argmin CV_0(\kappa)$, with

$$CV_0(\kappa) = \frac{1}{n}\sum_{i=1}^{n}(1 - \cos(\Delta_i - \hat{m}_{-i}(\Theta_i))), \tag{6}$$

where $\hat{m}_{-i}$ denotes the version of $\hat{m}$ computed by omitting the ith pair of the sample. However, since Θ_is are unknown above criterion is not attainable.

However, a cross-validation idea could still be employed through a SIMEX (simulation-extrapolation) approach proposed by Delaigle and Hall, 2008 by following the steps listed below:

1. Generate two i.i.d. samples from U denoted as $u_1^*, \ldots, u_n^*$ and $u_1^{**}, \ldots, u_n^{**}$. Then, for $i = 1, \ldots, n$, define $\Phi_i^* = \Phi_i + u_i^*$ and $\Phi_i^{**} = \Phi_i + u_i^* + u_i^{**}$ and consider the problem of estimating two regression functions, m_1 and m_2, respectively from the contaminated data (Φ_i^*, Δ_i) and (Φ_i^{**}, Δ_i).

2. Define the objective functions $CV^*(\kappa)$ and $CV^{**}(\kappa)$

$$CV^*(\kappa) = \frac{1}{n}\sum_{i=1}^{n}(1-\cos(\Delta_i - \hat{m}_{1,-i}(\Phi_i)))$$
$$CV^{**}(\kappa) = \frac{1}{n}\sum_{i=1}^{n}(1-\cos(\Delta_i - \hat{m}_{2,-i}(\Phi_i^*)))$$

   in order to obtain $\hat{\kappa}_1^* = argmin CV^*(\kappa)$ and $\hat{\kappa}_2^{**} = argmin CV^{**}(\kappa)$.

3. The dependence of $\hat{\kappa}_1^*$ on Φ_i^* and $\hat{\kappa}_2^{**}$ on Φ_i^{**} can be removed by averaging over a large number, say B, of CV^* and CV^{**} for different simulated sequences of $u_1^*,\ldots,u_n^*$ and $u_1^{**},\ldots,u_n^{**}$:

$$CV_1 = \frac{1}{B}\sum_{b=1}^{B} CV_b^*$$
$$CV_2 = \frac{1}{B}\sum_{b=1}^{B} CV_b^{**}$$

4. Then, we define, for $j = 0,1,2$,

$$\hat{\kappa}_j = argmin CV_j(\kappa). \qquad (7)$$

   Now, Φ^{**} approximates Φ^* in the same way that Φ^* approximates Φ and Φ approximates Θ. Therefore we expect that the relationship between $\hat{\kappa}_0$ and $\hat{\kappa}_1$ is similar to that between $\hat{\kappa}_1$ and $\hat{\kappa}_2$. As the final result, we get

$$\hat{\kappa}_0 = \hat{\kappa}_1^2/\hat{\kappa}_2. \qquad (8)$$

References

LEE, A. 2010. Circular data. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 477–486.

DELAIGLE, A., HALL, P. 2008. Using SIMEX for Smoothing-Parameter Choice in Errors-in-Variables Problems. *Journal of the American Statistical Association*, **103**, 280–287.

QUANTILE COMPOSITE-BASED PATH MODELING TO ESTIMATE THE CONDITIONAL QUANTILES OF HEALTH INDICATORS

Pasquale Dolce[1], Cristina Davino[2], Stefania Taralli[3] and Domenico Vistocco[4]

[1] Department of Public Health, University of Naples Federico II,
(e-mail: pasquale.dolce@unina.it)

[2] Department of Economics and Statistics, University of Naples Federico II,
(e-mail: cristina.davino@unina.it)

[3] Istat, Italian National Institute of Statistics, (e-mail: taralli@istat.it)

[4] Department of Political Science, University of Naples Federico II,
(e-mail: domenico.vistocco@unina.it)

ABSTRACT: Quantile Composed-based Path Modeling complements the classical PLS Path Modeling. The latter is widely used to model relationships among latent variables and between the manifest variables and their corresponding latent variables. Since it essentially exploits classical least square regressions, PLS Path Modeling focuses on the effect the predictors exert on the conditional means of the different outcome variables involved in models. Quantile Composed-based Path Modeling extends the analysis to the whole conditional distributions of the outcomes. This paper proposes a procedure to estimate the conditional quantiles for the manifest variables of the outcome blocks. Starting from the information related to a grid of conditional quantiles, it is possible to define the most accurate model for each health indicator and the best predictive model for each Italian province. The proposed method is shown in action both on artificial and real data. The real data concerns the prediction of health indicators.

KEYWORDS: PLS Path Modeling, Quantile Composite-based Path Modeling, Conditional Quantile Model-based Prediction.

1 Introduction

Partial Least Squares Path Modeling (PLS-PM) is a multivariate statistical method for studying relationships among latent variables (LVs), each one represented through a set of observed variables usually called manifest variables (MVs). The general model consists of two sub-models: the structural model and the measurement model. The measurement model relates each MV to its own LV, assuming that the conditional mean of each MV is a linear function of the corresponding LV. The structural model specifies the linear relationships between LVs. The estimation of the model parameters proceeds through an iterative algorithm essentially based on a sequence of simple and multiple Ordinary Least Squares (OLS) regressions. The obtained coefficients measure the rates of change in the conditional mean of the dependent

MVs and LVs as a function of changes in the correspondent set of predictors. The same holds both in the measurement and structural model.

Theories behind the application of PLS-PM focus on the estimation of conditional expected values, regardless of the distribution of response variables. Although PLS-PM does not require any assumption about the distribution of LVs, MVs and error terms, and it seems to be robust with respect to departure from normality, it is known that heavy-tailed or highly skewed distribution may inflate standard errors obtained from bootstrapping and that influential outliers affect the OLS regression estimates (Hair et al., 2017). Moreover, modeling only the conditional mean may be inadequate when the effects of the investigated relationships are expected to vary across the different locations of the responses.

To this end, Davino and Esposito Vinzi (2016) introduced Quantile Composite-based Path Modelling (QC-PM). This method exploits Quantile regression (QR) (Koenker and Basset, 1978) and Quantile correlation (QC) (Li et al., 2014) in the classical PLS-PM algorithm for estimating the model parameters. Conditional quantile modeling provides a complete description of the relationship among LVs, considering the whole distribution of the outcome variables (and not only their conditional means). QC-PM is a complementary method to PLS-PM, used to investigate if the relationships among LVs change across different parts of the dependent LV distributions, when there are outliers in the data and when MVs distributions are heavy-tailed or highly skewed. Furthermore, QC-PM can be used to provide conditional quantile predictions of the MVs of the dependent blocks given the explanatory blocks, which is the main objective of this paper.

2 Conditional Quantile Model-based Prediction for Health Indicators.

The focus of the present work is on the performance of the QC-PM for predicting the θ-th ($0 < \theta < 1$) conditional quantile for the dependent MVs (i.e., the MVs related to the endogenous LVs), given the values of all explanatory MVs.

The proposed approach is mainly based on the idea that using a dense grid of quantiles, conditional quantiles offer more flexibility than the conditional mean in capturing the unobserved heterogeneity among the statistical units. The use of statistical models, tailored to discover, incorporate and exploit such an unobserved heterogeneity, is an old and wide explored issue in the regression model framework (Spath, 1979). Following the procedure proposed by Davino and Vistocco (2018) to handle heterogeneity in quantile regression, the focus here is on the heterogeneity of the dependent block variables and predictor block variables and its use to predict or influence the different parts of the conditional distribution of dependent variables.

The merit of the proposed method will be illustrated through a study concerning the relationships among three well-being domains (Education, Economic Well-being and Health) measured on Italian provinces. The interest in such an application concerns both advances in knowledge about the dynamics that determine the well-being outcomes at local level (multiplier effects or trade-offs) and a more complete measurement of regional inequalities of well-being. At the province level, inequalities can strengthen each other affecting multiple disadvantages or advantages. Therefore, in assessing well-being outcomes the conditions within those outcomes are determined should be properly considered. In the path model in figure 1, Health variables are placed as response variables. The underlying hypothesis, supported by literature and empirical studies, is that Economic well-being and Education affect Health. We consider both the direct effects on Health and the interaction between Economic well-being and Education (wealthy territories offer better job opportunities and therefore attract higher skilled people; human capital is a factor of economic growth).

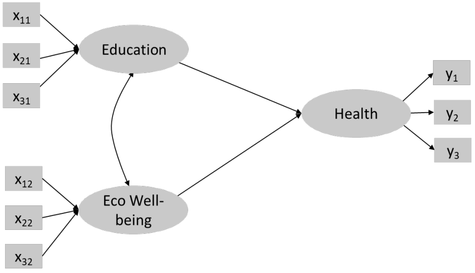


Fig. 1 A path model to predict Health outcomes from Education and Economic well-being at local level.

To provide a more in-depth assessment of Health inequalities, the specified path model is estimated for a dense grid of equally spaced quantiles through QC-PM, producing *m* estimates for each parameter of the model, where *m* is the number of chosen quantiles. The conditional quantile prediction for each health indicator (i.e., the MVs belonging to the health block) can be estimated in correspondence of each quantile θ, $(0 < \theta < 1)$. The accuracy of prediction is evaluated through quantile scoring based on the so-called pinball loss function, the loss function used as the objective function in quantile regression (Grushka-Cockayne et al., 2017), the lower the pinball loss, the more accurate the conditional quantile prediction. Moreover, a modified version of the Conditional Quantile Plot (Wilks, 2005), a graphical approach to evaluate the model performance for continuous measurements, will be used as a diagnostic verification technique. This plot will show, for each MV, the joint distribution between the estimated conditional quantiles (for the median, 25/75th and 10/90th quantiles) and the corresponding observed values. The estimated conditional quantile distributions are compared to the 1:1 diagonal line representing perfect prediction, to visualize which predicted conditional quantiles most agree with observations across the full MV unconditional distribution.

Finally, the best predictive model for each Italian province and for each Health indicator is defined as follows. Let y_{ip}, $(i = 1, . . . ,110)$, $(p = 1, . . . , 3)$, denote the observed value for province *i* on the health indicator *p,* the best predictive model is identified by the quantile that best predict the observed value, namely through the quantile which minimizes the absolute difference between the observed value and the estimated value:

$$\theta_{ip}^{best} = \underset{\theta=1,\ldots,m}{\arg\min} \left| y_{ip} - \hat{y}_{ip}(\theta) \right|,$$

where $\theta_{ip}{}^{best}$ represents the quantile associated to the best predictive model for each unit and each indicator, while $\hat{y}_{ip}\left(\theta^{best}\right)$ is the correspondent best prediction for y_{ip}.

The proposed procedure is used also to compare the best (i.e., optimal) conditional quantile predictions for the Health outcomes, given the Education and Economic Well-being levels, with the observed unconditional quantiles. A comparison for each province between the conditional θ^{best} value and the unconditional quantile be very informative. Best predictive model is obviously subject to overfitting, but this does not actually matter here and in general when dataset contains all the population units and the objective is not to generalize on different data.

Finally, the optimal conditional quantile predictions deliver a better prediction accuracy than using a single quantile approach or estimating only the conditional mean. This is obvious for a single regression model, but it is not for composite-based path modeling, where a number of regression models is analysed simultaneously, and the predictive model is a combination of two separate models, i.e., the measurement and the structural model. First results are very promising and show that the predictive model and the proposed procedure drastically improve the in-sample predictive capability of models.

Disclaimer*: The paper is the result of collaboration among the authors. Istat is not responsible for the contents.*

References

DAVINO, C., & VINZI, V. E. (2016). Quantile composite-based path modeling. *Advances in Data Analysis and Classification. Theory, Methods, and Applications in Data Science,* **10**(4), 491-520.

DAVINO, C. AND VISTOCCO, D. (2018). Handling heterogeneity among units in quantile regression. Investigating the impact of students' features on university outcome. *Statistics and its Interface*. **11**, 541-556.

HAIR, J.F., HULT, G.T.M., RINGLE, C.M. AND SARSTEDT, M. (2017), *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM),* 2nd ed., Sage, Thousand Oaks, CA.

KOENKER, R., & BASSET, G. (1978). Regression quantiles. *Econometrica,* **46**, 33–50.

LI, G., LI, Y., & TSAI, C. 2014. Quantile correlations and quantile autoregressive modeling. *Journal of the American Statistical Association,* **110(509)**, 233–245.

SPATH, H. (1979). Algorithm 39: Clusterwise linear-regression. *Computing*, **22**, 367-373.

TENENHAUS M., ESPOSITO V. V., CHATELIN Y. M., AND LAURO C. (2005). PLS path modeling. *Computational Statistics & Data Analysis* **48(1),** 159–205.

WILKS, D. S. (2005). *Statistical Methods in the Atmospheric Sciences*, Volume 91, Second Edition (International Geophysics), 2nd Edition. Academic Press.

Y. GRUSHKA-COCKAYNE, K.C. LICHTENDAHL JR, V.R.R. JOSE, R.L. WINKLER (2017). Quantile evaluation, sensitivity to bracketing, and sharing business payoffs. *Operations Research* **65 (3)**, 712-728.

AUC-BASED GRADIENT BOOSTING FOR IMBALANCED CLASSIFICATION

Martina Dossi[1] and Giovanna Menardi[2]

[1] European Central Bank[†]
(e-mail: martina.dossi@ecb.europa.eu)

[2] Dipartimento di Scienze Statistiche, Universita` di Padova,
(e-mail: menardi@stat.unipd.it)

ABSTRACT: Classification problems with imbalanced class distributions are pervasive in a plurality of real-world applications, such as network intrusion detection, fraud detection and rare disease diagnosis. In this context, most of standard classification models are heavily compromised, as they tend to focus on the majority class, yet the minority class is often the one of greatest importance. To tackle the problem, we combine *XGBoost*, a powerful and recent formulation of the *gradient boosting*, with a loss function specifically derived to optimise the Area Under the ROC curve, an evaluation metric more robust towards class imbalance.

KEYWORDS: AUC, boosting, classification, class imbalance.

1 Introduction

Class imbalance refers to all supervised classification tasks which suffer of uneven class distributions. The issue has gained ground with some further implicit assumptions, such that imbalanced data are expected to have rare instances belonging to the class of greatest interest and a (relatively) large number of units from the other classes. An imbalanced class distribution may severely affect the performance of classification algorithms, by interfering with both model estimation and accuracy evaluation phases. Disregarding each model own specificities, model estimation is typically driven by the optimisation of a global loss function, which favours classification rules ignoring the rare units as overwhelmed by the prevalent class. A number of techniques have been developed to cope with imbalanced classes: data level approaches attempt to re-balance the class distribution before building learning models, whereas classifier level approaches aim to adapt existing algorithms to focus on the minority class. The latter group includes cost-sensitive techniques, methods that replace the loss function with more meaningful measures and combinations of classifiers, that follow the logic of *boosting*, *bagging* and *random forests*.

[†]Disclaimer: this document reflects authors' views, not necessarily shared by ECB.

Under imbalanced scenarios, assessing the performance of a classifier plays a role that is at least as crucial as its estimation. Accuracy, which is the most commonly used metric for classification tasks, is not sufficient, as it is governed by the majority class. Other performance metrics which account for the class distribution are preferred in this context, as the G-mean, the F-measure, and especially the Area Under the ROC Curve (AUC). See Menardi & Torelli (2014) for a more comprehensive discussion about the imbalance problem.

Within the logic of the approaches at a classifier level, in this work we derive a differentiable loss function that optimises the AUC to train a gradient-based model within the *boosting* family, in order to extend the benefits of the AUC as evaluation metric to the phase of model estimation. After presenting the building blocks relevant for a full comprehension of the proposed method, we discuss our contribution and show some numerical results.

2 Gradient boosting optimisation based on the AUC

Given a training set $\mathcal{T}_n$ containing n *i.i.d.* pairs $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of attributes and $y_i \in \{\mathcal{Y}_0, \mathcal{Y}_1\}$ is a response variable whose classes are conventionally labeled as negative and positive respectively, a classifier $\mathcal{H} : \mathcal{X} \mapsto \mathbb{R}$ is a function that allows to predict the response variable y, based on the observed $\mathbf{x}$. The output $\mathcal{H}(\mathbf{x})$ measures the confidence of $\mathbf{x}$ belonging to the positive class, whereas the predicted label $\hat{y}$ is defined on the basis of a threshold $k \in \mathbb{R}$ such that $\hat{y} = \mathcal{Y}_0$ if $\mathcal{H}(\mathbf{x}) < k$ and $\hat{y} = \mathcal{Y}_1$ otherwise. A non-negative loss function $\mathcal{L}(y, \hat{y})$, that measures the discrepancy between observed and fitted values, is used either to optimize the classifier during the learning process and to assess the performance of the model.

Even if not specifically developed to tackle the class imbalance problem, the *gradient boosting* (Friedman, 2001) has showed to achieve competitive results in this domain. In broad terms, it exploits the connection between *AdaBoost*, the first applicable approach of *boosting*, that relies on the idea of increasing the weight of the hardest to classify units, and a forward-stagewise additive modeling approach. At each iteration of the algorithm, a functional gradient descent optimisation is applied to a loss function, in the n-dimensional space of the fitted values, and it is then approximated by some simple model. The final rule is a linear combination of all the previous estimated functions. A specific formulation of the *gradient boosting* is *XGBoost* (Chen & Guestrin, 2016), which, at each iteration, approximates the objective loss function by a second order Taylor's series expansion, and estimates a classification model via its minimisation. This implementation easily supports different loss functions, as it is sufficient to provide the algorithm with its first two derivatives.

The rationale behind the proposed approach is to integrate into the *XGBoost* a loss function independent on the class distribution. In this perspective, the AUC - its ones' complement, in fact - represents a sensible candidate.

Let n_+ and n_- be the sample size of positive and negative observations respectively, and assume that $\mathcal{H}(\mathbf{x}_i^+)$ and $\mathcal{H}(\mathbf{x}_j^-)$ are the fitted scores respectively for the i-th positive and the j-th negative instances. The AUC is equivalent to the normalized Wilcoxon Mann-Whitney statistic, in the form:

$$AUC = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \mathbb{I}_{0.5}(\mathcal{H}(\mathbf{x}_i^+) - \mathcal{H}(\mathbf{x}_j^-)), \tag{1}$$

where $\mathbb{I}_{0.5}(t)$ is 0 if $t < 0$, 0.5 if $t = 0$, 1 otherwise. The AUC estimates the probability that a positive unit receives a higher score than a negative one by means of comparisons between instances belonging to different classes. While the global accuracy of a classifier depends on the choice of a classification threshold, the AUC evaluates its discriminating ability as the threshold varies over all its range. This allows to cater for the presence of rare units as, by construction, it does not place more emphasis on one class over the other.

Unfortunately, two issues prevent the expression (1) from being directly used as a loss function: first and foremost, the function is non differentiable, secondly, its argument is not the single observation but rather refers to pairs of instances. To overcome the first limitation, we consider the following differentiable approximation (Yan *et al.*, 2003):

$$\mathcal{U}_s = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \mathcal{S}(\mathcal{H}(\mathbf{x}_i^+), \mathcal{H}(\mathbf{x}_j^-)), \text{ where:} \tag{2}$$

$$\mathcal{S}(\mathcal{H}(\mathbf{x}_i^+), \mathcal{H}(\mathbf{x}_j^-)) = \begin{cases} (-(\mathcal{H}(\mathbf{x}_i^+) - \mathcal{H}(\mathbf{x}_j^-) - \tau))^p & \text{if } \mathcal{H}(\mathbf{x}_i^+) - \mathcal{H}(\mathbf{x}_j^-) < \tau, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

for a given $\tau \in (0,1]$ and $p > 1$ selected by the user. A pair of observations contributes to the loss function when the score of a positive unit exceeds the one of a negative unit by τ. The authors suggest to choose $\tau \in [0.1, 0.7]$ and $p \in \{2,3\}$. The quantity $\mathcal{U}_s$ is then reformulated to refer to unique instances:

$$\mathcal{U}_s = \frac{1}{n_+ n_-} \sum_{i=1}^{n} \left[\mathbb{I}_{(y_i=1)} \sum_{i'=1}^{i-1} \mathcal{S}_{i'}^+ + \mathbb{I}_{(y_i=-1)} \sum_{i'=1}^{i-1} \mathcal{S}_{i'}^- \right], \text{ where:} \tag{4}$$

$\mathcal{S}_{i'}^+ = \mathbb{I}_{(y_{i'}=-1)} \mathcal{S}(\mathcal{H}(\mathbf{x}_i), \mathcal{H}(\mathbf{x}_{i'}))$ and $\mathcal{S}_{i'}^- = \mathbb{I}_{(y_{i'}=1)} \mathcal{S}(\mathcal{H}(\mathbf{x}_{i'}), \mathcal{H}(\mathbf{x}_i))$. Once the parameters are defined, the computation of the first two derivatives is straightforward and the method can be implemented.

Empirical results reveal that the proposed approach outperforms many other competitive classifiers, especially in scenarios of extreme rarity and nontrivial data patterns. In the bidimensional setting illustrated in Figure 1, as well as in its generalisation in 5 dimensions, rare units lie in small disjunct sets, overlapping with the majority class at the margins of each box. The results of the analysis are outlined in Table 1. As expected, standard models as the logistic regression and the classification tree fail in this domain. The algorithm *SMOTEBoost* (Chawla *et al.*, 2003), specifically developed to address the imbalance, performs even worse than the original *AdaBoost*. Conversely, the modified *XGBoost* achieves better results in the majority of the cases, including the hardest.

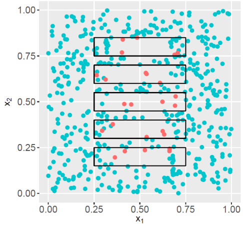


Figure 1: Simulated data in the bidimensional space. Red dots represent the rare instances.

%	dim.	*Logistic Reg.*	*Tree (Gini)*	*Ada-Boost*	*SMOTE-Boost*	*Gradient boosting*	*Modified XGBoost*
0.6	2	0.500 (0.004)	0.500 (0.000)	0.772 (0.047)	0.602 (0.059)	0.782 (0.043)	**0.790** (0.041)
	5	0.500 (0.012)	0.500 (0.000)	0.721 (0.041)	0.563 (0.059)	0.712 (0.040)	**0.736** (0.042)
1	2	0.500 (0.003)	0.501 (0.008)	0.830 (0.034)	0.632 (0.059)	**0.838** (0.030)	0.833 (0.030)
	5	0.499 (0.008)	0.501 (0.014)	0.786 (0.032)	0.609 (0.067)	0.777 (0.030)	**0.790** (0.031)

Table 1: Average AUC (and standard deviation) over 300 Monte Carlo samples of size 1000, with dimension 2 and 5, rare class frequency of 0.6% and 1%. For *boosting* algorithms 200 iterations were considered; for the AUC-based loss function $\tau = 0.7$ and $p = 2$.

References

CHAWLA, N. V., *et al.* 2003. SMOTEBoost: Improving prediction of the minority class in boosting. *European conference on principles of data mining and knowledge discovery*, Springer, 107-119.

CHEN, T. & GUESTRIN, C. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, 785-794.

FRIEDMAN, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

MENARDI, G., & TORELLI, N. 2014. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, **28(1)**, 92-122.

YAN, L., *et al.* 2003. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. *Proceedings of the 20th International Conference on Machine Learning*, 848-855.

How to measure material deprivation? A Latent Markov Model based approach

Francesco Dotto[1]

[1] Department of Economics, University Roma Tre,
(e-mail: francesco.dotto@uniroma3.it)

Abstract: Material deprivation can be used to assess poverty in a society. The status of poverty is not directly observable, but can be measured with error for instance through a list of deprivation items. Given two unobservable classes, corresponding to the poor and not poor, we develop a time-inhomogeneous latent Markov model which allows us to classify households according to their current and inter-temporal poverty status, and to identify transitions between classes that may occur year-by-year. Households are grouped by estimating their posterior probability of belonging to the latent status of poverty.

Keywords: latent markov, material deprivation, EU SILC.

1 Introduction

Measurement of material deprivation has generally followed the "counting approach", that is a parsimonious way of classifying a society according to the number of zero-one deprivation indicators, that lead to a deprivation score. An individual score of deprivation results from the (possibly weighted) sum of the dichotomous indicators listed in Table 1. Two individuals with the same deprivation score are treated equally, even though they do not necessarily lack the same items. The cut-off, the list of items, and their associated weights have been a matter of concern and dispute, since they can affect the results and the consequent policy. To overcome such issues we develop a *dynamic* latent state model able to classify individuals (or households) according to their unobserved poverty status from their observed current and inter-temporal deprivations and to estimate movements into and out of poverty during the whole observation period (Dotto *et al.*, 2019). In this dynamic perspective, the probability of being *persistently* poor is estimated as the joint probability of being poor over the whole period and transitions between classes (poor and not poor) that may occur year-by-year can be estimated.

2 A sketch of the model

Individuals belong to the latent state of poverty with a probability that depends on the presence/absence of a specific combination of deprivation items. More formally, let $Y_{it} = (Y_{it1}, \ldots, Y_{itR})$ be the outcome for the i-th individual at time t is the R-dimensional configuration. Given the R-dimensional outcome measures, with error, let U_{it} be a binary latent variable which represents for each individual i an indicator of being in the poverty status in the simplest case of $k = 2$, and an indicator of being in the j-th latent group at time t, $j = 1, \ldots, k$ in the more general case. Subjects are allowed to move from one latent state to another between each measurement occasion, hence U_{it} is not necessarily constant over time. In what follows we assume the i-th subject has been measured at times $1, \ldots, T_i$, with $T = \max_i T_i$, and that missing measurements are not informative. In our analysis $T_i = T = 4$. The resulting likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} \left(\sum_{U_{i1}=1}^{k} \sum_{U_{i2}=1}^{k} \cdots \sum_{U_{iT_i}=1}^{k} \Pr(U_{i1}) \prod_{t=2}^{T_i} \Pr(U_{it}|U_{i,t-1}) \prod_{t=1}^{T_i} \prod_{r=1}^{R} \Pr(Y_{itr}|U_{it}) \right)^{s_i}, \tag{1}$$

where, in (1), θ is a short-hand notation for all parameters involved and s_i, $i = 1, \cdots, n$, the longitudinal sampling weights. To maximize (1) we use an EM-type procedure whose details are outlined in Bartolucci *et al.*, 2012. At convergence of the algorithm, the obtained MLE for the parameters of interest can be used for inference, prediction, and their interpretation is explained within the next section.

3 Results

To validate the proposed methodology we use the 2013 longitudinal component of EU-SILC (UDB SILC 2013 rev.2), released in August 2016. The unit of analysis is the household. Table 1 reports the association between each item r and the latent categorical variable. For each country, the first column indicates the estimated probability of being poor ($j = 2$) in a specific item given that the latent variable assumes the status of poverty, $\hat{p}_{2r}$, and it is a measure of how *sensitive* the item is. The second column, instead, indicates the *specificity* of each item r, $1 - \hat{p}_{1r}$, that is the probability of not lacking item r given that the household is not poor. Ideally, item r should have *sensitivity* and *specificity* equal to 100%: whoever is poor lacks that item and whoever is not poor does not lack that item. It can be seen that generally durable goods (telephone,

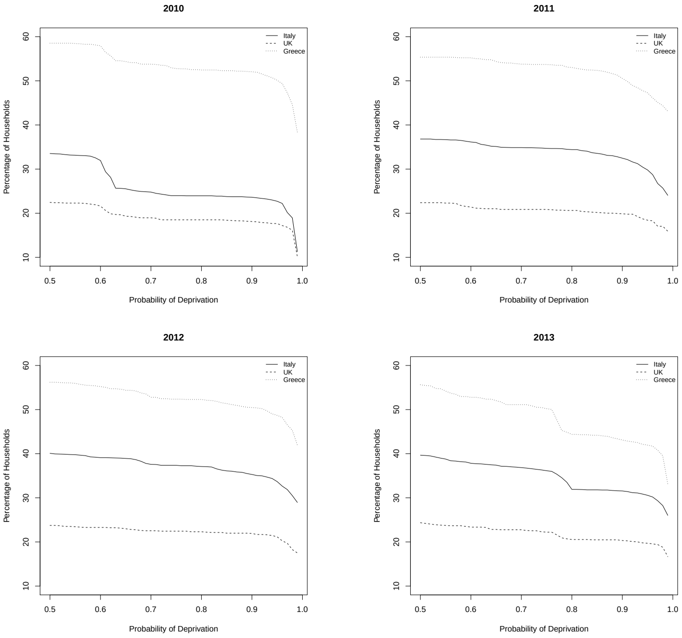


Figure 1. *Percentage of deprived households according to varying thresholds*

TV, washing machine) are very specific, but not sensitive, attributes. The incapacity to afford a meal and to keep the house adequately warm are also very specific but also quite sensitive. More balanced items, and on the whole more discriminating, are the incapacity of having one week annual holiday away from home and of facing unexpected expenses.

Table 1. *Estimated probability (percentage) of lacking item r given that the latent state is poverty ($\hat{p}_{2r}$=sensitivity) and probability of not lacking item r given that the latent state is non-poverty ($1-\hat{p}_{1r}$=specificity). Greece, Italy, and UK separately and as a whole: 2010–2013.*

		Greece		Italy		UK		Pooled	
Item	description	$\hat{p}_{2r}$	$1-\hat{p}_{1r}$	$\hat{p}_{2r}$	$1-\hat{p}_{1r}$	$\hat{p}_{2r}$	$1-\hat{p}_{1r}$	$\hat{p}_{2r}$	$1-\hat{p}_{1r}$
1	keep the house warm	49.6	92.9	43.4	98.0	21.8	98.1	34.5	98.0
2	one week holiday	88.9	76.0	92.4	82.4	81.0	95.7	87.4	87.5
3	afford a meal	31.7	99.0	30.8	98.9	20.9	99.8	25.8	99.5
4	unexpected expenses	87.3	88.8	83.4	90.3	85.3	91.5	83.5	90.9
5	telephone	1.2	100.0	0.8	100.0	0.2	100.0	0.7	100.0
6	color TV	0.1	100.0	0.8	100.0	0.3	100.0	0.5	100.0
7	washing machine	2.5	99.7	0.9	100.0	1.6	100.0	1.3	100.0
8	car	15.5	97.6	7.9	99.8	17.9	99.2	12.3	99.5
9	arrears	58.5	82.9	26.8	98.3	28.7	99.5	29.8	98.5

Additionally, for each household i, we estimated the probability of being in state of current deprivation based on the configuration of the vector of the nine outcomes: $p_{it} = Pr(\mathbf{Y_{it}}|U_{it} = 1)$. These estimated conditional probabilities allow to classify households into the state of deprivation or non-deprivation according to a given threshold, τ. Each household is not classified according to an established cut-off, but with uncertainty. This does not prevent calculation of the usual deprivation statistics such as the deprivation rate but leads to a *continuum* of solutions represented by curves of the estimates, permitting an evaluation of their robustness (comapre Figure 1).

4 Conclusions

Measurement of material deprivation which is a relative concept, is still challenging since involves both methodological and substantive issues. Herein we proposed a latent Markov model for categorical longitudinal data able to solve some of the issues raised in measuring deprivation. Our model is able to study the evolution of individual characteristics that are not directly observable.

References

BARTOLUCCI, F., FARCOMENI, A., & PENNONI, F. 2012. *Latent Markov models for longitudinal data*. Chapman and Hall/CRC.

DOTTO, F., FARCOMENI, A., PITTAU, M. G., & ZELLI, R. 2019. A dynamic inhomogeneous latent state model for measuring material deprivation. *JRSS: Series A*, **182**(2), 495–516.

DECOMPOSITION OF THE INTERVAL BASED COMPOSITE INDICATORS BY MEANS OF BICLUSTERING

Carlo Drago[1, 2]

[1] University “Niccolo Cusano” in Rome, (e-mail: carlo.drago@unicusano.it)

[2] NCI University in London, Faculty of Business and Management

ABSTRACT: Interval-based composite indicators are useful as subjectivities exist in the choices leading to the construction of a composite indicator. The interval shows the level of variation which is able to be determined by the different factors considered on the construction of the composite indicator. We will explore and analyze different results of the underlying composite indicators computed on the Monte Carlo simulations using biclustering. The results offer an understanding and explanation of the sensitivity of the composite indicator outcomes to the inputs under consideration.

KEYWORDS: composite indicators, interval data, biclustering.

1 Interval-Based Composite Indicators

A relevant problem* in the construction of the composite indicators is the existence of subjectivity in some relevant decisions (JRC European Commission and OECD, 2008). For instance the choice of the different weights could be subjective. A sensitivity analysis can be performed in order to evaluate the robustness of the different results due to changes in the stated assumptions (for instance the choice of the weights). The construction of a composite indicator should take into consideration both sensitivity and robustness analysis in order to validate the results. A possible solution is the use of interval data (Drago, 2017). These interval data take into account the different characteristics of the different results by their features. A typical composite indicator can be considered (Aiello & Attanasio, 2006)

$$Y = f[T_1(y_1), T_2(y_2), \dots, T_s(y_s)] \tag{1}$$

*Thanks to the anonymous referees for the useful suggestions and professor Filomena Maggino and dr. Leonardo Alaimo for the productive discussions. Eventual mistakes are mine where $y_1, y_2 \dots y_s$ are some indicators and $T_1, T_2 \dots T_s$ are transformation functions of the data, and finally f is a specific aggregation function considered. From the different combinations of possible inputs we can construct the composite indicator. If we consider the entire set of the output of the composite indicator considering all the changes on the underlying c assumptions (for instance weighting) we can have the interval of the different composite indicators (outputs):

$$I[Y] = [\underline{Y}^c, \overline{Y}^c] = \{Y \in \mathbb{R} : \underline{Y}^c \leq Y^c \leq \overline{Y}^c\} \tag{2}$$

where $\underline{Y}^c$ and $\overline{Y}^c$ are respectively the lower and the upper bounds for the assumptions considered c for $c = 1, \dots, C$. It is important to note that it is possible to take into account the center

$$Y_{center} = \frac{1}{2}(\underline{Y}^c + \overline{Y}^c) \tag{3}$$

and also the radius

$$Y_{radius} = \frac{1}{2}(\overline{Y}^c - \underline{Y}^c) \tag{4}$$

At this point it is of great importance to interpret the different intervals constructed by considering the different blocks obtained.

2 Decomposition of the Interval Composite Indicators

It is possible to decompose the interval composite indicator into different intervals which can be contained in the original one.

$$[\underline{Y_j^1}, \overline{Y_j^1}], [\underline{Y_j^2}, \overline{Y_j^2}], \dots, [\underline{Y_j^C}, \overline{Y_j^C}] \tag{5}$$

where c for $c = 1, \dots, C$ are the different assumptions considered , $\underline{Y_j^1}$ and $\overline{Y_j^1}$ are the lower and the upper bounds for each statistical unit j with $j = 1, \dots, J$. We can start from the general matrix A of the simulations obtained. Each simulation returns a single composite indicator for the each statistical unit and a rank. At this point we can consider an approach of biclustering in order to analyze in more depth the different results obtained from the different simulations (the different simulations returning different outputs or composite indicators are in columns whereas the statistical units are in the rows). In particular we are interested in comparing the different simulations used assuming different weightings or changing the structure of the composite indicator. We can validate the biclusters obtained by examining the Jaccard index (Kaiser & Leisch, 2008):

$$JI(B1,B2) = \frac{|B1 \cap B2|}{|B1| + |B2| - |B1 \cap B2|}$$

where $B1$ and $B2$ are two bicluster results. In this way, from the consideration of the biclustering results it is possible to decompose the interval of the composite indicator. In fact we are able to analyze and explore the different results of the factor which cause the variation of the composite indicator. This result is very important for the operational use of the composite indicator (we are interested in both in the score of the composite indicator and also in its variability).

3 Analysis of the Criminal Rates in the United States

We consider a composite indicator useful to measure the level of crime in the US by combining the information of the different criminal indicators. The data are in McNeil, 1977 and consist of statistics relating to the arrests for assault, rape and murder in the US (year 1973). The data are related to the 50 US states. All the different data are per 100,000 habitants. So in this sense we firstly standardize the different indicators, then we consider the construction of the interval based composite indicator. The analysis of the matrix of the simulations is visualized and analyzed by a heatmap (figure 1). Then we apply the biclustering approach in order to analyze the simulation matrix. Overall the variables tend to vary much more in the composite indicators than in other factors such as the weights. We used the biclustering algorithms developed in Kaiser *et al.*, 2018.

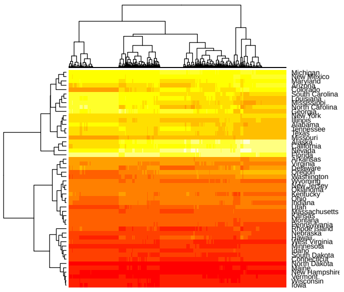


Figure 1. *Simulations for the Interval Based Composite Indicator (in columns the different inputs considered; in rows the statistical units).*

4 Conclusions

Composite indicators are often characterized by assumptions which call for some subjective choices. In this sense it is usually relevant to perform a sensitivity analysis in order to evaluate the robustness of the composite indicator constructed. Interval composite indicators allow to take into account all possible variation sources on a single interval outcome based on all the possible outputs obtained by the different inputs. In this context biclustering can be usefully applied so as to detect groups of statistical units which do not vary on the same simulations. The interval decomposition allows the evaluation and the exploration of the variability patterns relating to the inputs and more specifically where inputs lead to higher variation outcomes.

References

AIELLO, F., & ATTANASIO, M. 2006. Some issues in constructing composite indicators. *Pages 11–13 of: VIII international meeting on quantitative methods for applied sciences, Certosa di Pontignano.*

BILLARD, L. 2008. Some analyses of interval data. *Journal of computing and information technology*, **16**(4), 225–233.

DRAGO, C. 2017. *Interval Based Composite Indicators.* FEEM Working Paper.

JRC EUROPEAN COMMISSION AND OECD. 2008. *Handbook on constructing composite indicators: methodology and user guide.* OECD publishing.

KAISER, S., & LEISCH, F. 2008. *biclust- A Toolbox for Bicluster Analysis in R.* Presentation at UseR 2008 13.8.2008 Dortmund.

KAISER, S., SANTAMARIA, R., KHAMIAKOVA, T., SILL, M., THERON, R., QUINTALES, L., LEISCH, F., & DE TROYER, F. 2018. *biclust: BiCluster Algorithms.* R package version 2.0.1.

MCNEIL, D.R. 1977. *Interactive Data Analysis.* New York: Wiley.

CONSENSUS CLUSTERING VIA PIVOTAL METHODS

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

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

ABSTRACT: We propose an approach to the cluster ensemble problem based on pivotal units extracted from a co-association matrix. It can be seen as a modified version of *K*-means method, which utilizes pivots for careful seeding. Different criteria for identifying the pivots are discussed, as well as preliminary results concerning the comparison with alternative ensemble methods.

KEYWORDS: cluster ensemble, pivot, K-means.

1 Introduction

Ensembles methods have recently emerged as a valid alternative to conventional clustering techniques and have shown to effectively improve the quality of clustering results and achieve robustness (see, e.g., Strehl & Ghosh, 2002, Jain, 2010). Such methods require a strategy to generate multiple clusterings of the same data set (the ensemble) and then combine them into a *consensus* partition (presumably superior), by following the idea of evidence accumulation, i.e., by viewing each clustering result as an independent evidence of data structure. A common way to do this is to obtain a new pairwise similarity matrix, or co-association matrix, by taking the co-occurrences of pairs of points in the same group across all partitions (Fred & Jain, 2005). Then, a similarity-based clustering algorithm can be applied to this matrix to yield the final partition.

We propose to use the co-association matrix to find some specific units (hereafter, pivots) which are representative of the group they belong to (because they never or very rarely co-occur with members of other groups). Various criteria for detecting the pivots are proposed in Section 2. Section 3 illustrates the use of pivotal methods for data clustering, and compare the proposed approach with classical *K*-means and other common ensemble methods.

Pivotal methods and related clustering procedures are implemented via the R package `pivmet`, available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=pivmet.

2 Pivotal methods based on co-association

Let $\mathbf{Y} = (y_1,\ldots,y_n)$ be a set of n observations, where $y_i \in \mathbb{R}^d$. Consider a set $\mathcal{P} = \{P^1, P^2, \ldots, P^H\}$ of H partitions of the data points into K disjoint clusters, derived from an arbitrary clustering algorithm. Note that the number of groups is pre-specified and equal for all P^h. $\mathcal{P}$ can be summarized via the $n \times n$ co-association matrix C with generic element

$$c_{i,j} = \frac{1}{H} \sum_{h=1}^{H} |P^h(y_i) = P^h(y_j)|, \tag{1}$$

where $|\cdot|$ denotes the indicator function, and $P^h(y_i)$, $P^h(y_j)$, represent the clusters of the objects y_i and y_j in P^h, respectively. Clearly, units which are very dissimilar from each other are likely to have zero co-occurrences; as a consequence, C is expected to contain a non-negligible number of zeros. Given a large and sparse 0-1 matrix, the Maxima Units Search (MUS) algorithm seeks those elements, among a pre-specified number of candidate pivots, whose corresponding rows contain more zeros compared to all other units (Egidi *et al.*, 2018c). Define a reference partition, $G_1,\ldots,G_K$ of $y_1,\ldots,y_n$ obtained by applying, for instance, an agglomerative hierarchical algorithm into K groups. The MUS procedure takes C as input and outputs a set of K units–one for each group of the reference partition–that exhibit the highest degree of separation (Egidi *et al.*, 2018b). As an alternative approach, the pivot y_{i_k} for group G_k can be chosen so that it is as far as possible from units that might belong to other groups and/or as close as possible to units that belong to the same group, according to one of the following objective functions

$$(a)\ \max_{i_k} \sum_{j \in G_k} c_{i_k,j} \quad (b)\ \min_{i_k} \sum_{j \notin G_k} c_{i_k,j} \quad (c)\ \max_{i_k} \sum_{j \in G_k} c_{i_k,j} - \sum_{j \notin G_k} c_{i_k,j}, \tag{2}$$

where $c_{i,j}$ is defined as in (1). Ideally, the $K \times K$ submatrix of C with only the rows and columns corresponding to $i_1,\ldots,i_K$ will be the identity matrix. In practice, it may contain few nonzero elements off the diagonal.

3 A simulation experiment

In order to illustrate the proposed algorithm, we simulate bivariate data from a mixture of three Gaussian distributions with mean vectors $\boldsymbol{\mu}_1 = (1,5)$, $\boldsymbol{\mu}_2 =$

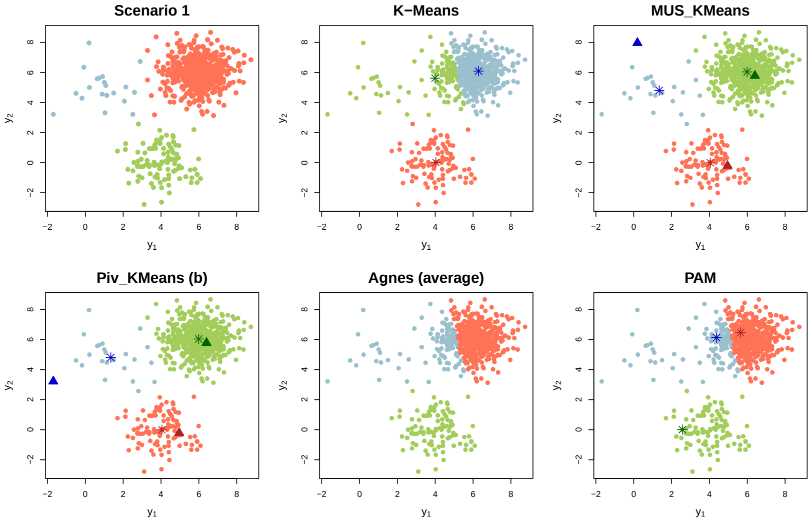


Figure 1. *Mixture of three Gaussian distributions (sample size n=620). Cluster centers and/or pivots for each method are marked via asterisks and triangles, respectively.*

$(4,0)$, $\boldsymbol{\mu}_3 = (6,6)$, and the 2×2 identity matrix as covariance matrix. The components have sample size 20, 100 and 500, respectively (see Figure 1, top-left panel). The K-means algorithm with random seeds is used to generate a cluster ensemble of $H = 1000$ partitions, and obtain the co-association matrix C. For each simulated dataset, we proceed as follows:

1. For a given number of clusters K, obtain a partition of the data $G_1, \dots, G_K$ (reference partition);
2. Apply the MUS algorithm or one alternative criterion in (2) to the matrix C to find K (distinct) pivots $y_{i_1}, \dots, y_{i_K}$;
3. Run the K-means algorithm using the pivots as initial cluster centers.

The proposed modification of the standard K-means technique introduces a pivot-based initialization step with the aim of reducing the effect of random seeding (see also Egidi *et al.*, 2018a). An alternative approach to careful seeding can be found in Arthur & Vassilvitskii, 2007. Figure 1 shows the solution from K-means, using $K = 3$, and by pivotal methods MUS and criterion (b) in Eq. (2), where Average-Linkage (AL) agglomerative clustering is used to obtain the reference partition. The results of consensus clustering using PAM (Partitioning Around Medoids) method and AL-agglomerative hierarchical clustering (agnes) are also shown (Single Linkage (SL) and Complete Linkage (CL) give similar results). Table 1 reports the comparison between the different methods in terms of Adjusted Rand Index (ARI), used to quantify the agreement between two partitions. The mean value is considered for 1000 simulations. Preliminary results suggest that the pivot-based approach outperforms the competing similarity-based ensemble methods and the standard K-means, which gives a mean ARI of 0.659.

Table 1. *2D Gaussian data: mean ARI (*1000 *simulations) between the final clustering and the true data partition. Ensemble methods use dissimilarities* $1 - c_{i,j}$.

Pivotal methods	MUS 0.857	(a) 0.865	(b) 0.883	(c) 0.779
Ensemble methods	agnes (AL) 0.512	agnes (SL) 0.535	agnes (CL) 0.514	PAM 0.506

References

ARTHUR, D., & VASSILVITSKII, S. 2007. k-means++: The advantages of careful seeding. *Pages 1027–1035 of: Proceedings of the eighteenth annual ACM-SIAM Symposium on Discrete algorithms.*

EGIDI, L., PAPPADÀ, R., PAULI, F., & TORELLI, N. 2018a. K-means seeding via MUS algorithm. *Pages 256–262 of: Book of Short Papers SIS 2018.*

EGIDI, L., PAPPADÀ, R., PAULI, F., & TORELLI, N. 2018b. Maxima Units Search (MUS) algorithm: methodology and applications. *Pages 71–81 of: Studies in Theoretical and Applied Statistics.*

EGIDI, L., PAPPADÀ, R., PAULI, F., & TORELLI, N. 2018c. Relabelling in Bayesian mixture models by pivotal units. *Statistics and Computing*, **28**, 957–969.

FRED, A. L. N., & JAIN, A. K. 2005. Combining Multiple Clusterings Using Evidence Accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 835–850.

JAIN, A. K. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 651 – 666.

STREHL, A., & GHOSH, J. 2002. Cluster ensembles - A knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research*, **3**, 583–617.

ROBUST MODEL-BASED CLUSTERING WITH MILD AND GROSS OUTLIERS

Alessio Farcomeni[1] and Antonio Punzo[2]

[1] Department of Public Health and Infectious Diseases, Sapienza - University of Rome,
(e-mail: alessio.farcomeni@uniroma1.it)

[2] Department of Economics and Business, University of Catania,
(e-mail: antonio.punzo@unict.it)

ABSTRACT: We propose a model-based clustering procedure for mild and gross outliers. Our mixture model is based on heavy-tailed components (e.g., the contaminated normal distribution), but it is assumed to apply only to a subset of the data. Consequently, a proportion of observations is trimmed. We propose a penalized likelihood approach for estimation and selection of the proportions of mild and gross outliers, where the penalty parameter is fixed by formal optimality arguments. We conclude with an original real data example on the identification of the source from illicit drug shipments seized in Italy and Spain.

KEYWORDS: tclust, contaminated normal, penalized likelihood.

1 Introduction

In clustering based on the normal mixture model there are two main approaches to deal with contamination. One is based on the use of heavy-tailed or skewed component distributions. A recent example in this direction, preserving elliptical contours of clusters, are mixtures of contaminated normal (CN) distributions (Punzo & McNicholas, 2016). Component-wise methods are well suited to work with mild outliers (Ritter, 2015), and are sometimes labeled as weakly robust. A separate body of literature has instead worked with outliers in more general position, including gross outliers, and has usually proceeded by discarding or at least downweighting a proportion of the observations (Farcomeni & Greco, 2015). A good example is `tclust` (García-Escudero *et al.* , 2008), where a fixed proportion of observations is trimmed and the rest is assumed to follow a normal mixture model. These procedures have often formal robustness properties, e.g., positive breakdown point asymptotically.

In this work we merge the two approaches above by estimating a CN mixture after trimming a fixed proportion of gross outliers. Our model can be seen from two different perspectives. On the one hand, clusters having a distribution with slightly heavy tails might be desired in order to assign as many observations to clusters as possible. In this case, it is indeed assumed that clean observations arise from, for example, a CN model. On the other hand, the trade off between mild and gross outliers is exploited in order to increase efficiency: some (mild) outliers are assigned to a cluster and contribute to centroid estimation, therefore decreasing the final mean squared error (MSE).

In this work we tackle also an additional open problem with trimming procedures, that of selecting the trimming proportion. Our proposal is based on a penalized likelihood approach, where the trimming proportion is in practice substituted by a penalty parameter. The advantage is that we can identify a heuristic but theoretically justified way of choosing an optimal penalty level, and therefore an optimal trimming proportion. Our fixed-penalty approach in some sense solves the issue of selecting the trimming proportion both for our model and the special case of trimmed normal mixture models (`tclust`). The methodology proposed in this paper has been implemented in `R` functions which can be downloaded from `https://github.com/afarcome/cntclust`.

2 Methodology

Let $x_1,\dots,x_i,\dots,x_n$ be a sample of n observations in d dimensions. Moreover, let $\alpha_0 \geq 0$ denote a trimming proportion of outliers which shall not be used to estimate model parameters. We assume data arise from the contaminated spurious outlier model

$$\prod_{i\in R}\sum_{j=1}^{k}\pi_j f_{\text{CN}}(x_i;\mu_j,\Sigma_j,\alpha_j,\eta_j)\prod_{i\notin R} g_i(x_i), \tag{1}$$

where R denotes a set of non-trimmed observations of cardinality $\lfloor (1-\alpha_0)n \rfloor$ and g_i are pdfs generating the outliers in general position. Let $f_{\text{N}}(\cdot;\mu,\Sigma)$ denote the probability density function (pdf) of a d-variate normal (N) distribution with mean vector μ and covariance matrix Σ. In (1), $f_{\text{CN}}(x;\mu,\Sigma,\alpha,\eta) = (1-\alpha) f_{\text{N}}(x;\mu,\Sigma) + \alpha f_{\text{N}}(x;\mu,\eta\Sigma)$ denotes the pdf of a d-variate CN distribution with mean vector μ, scale matrix Σ, proportion of mild outliers $\alpha \in (0,1)$, and degree of contamination $\eta > 1$.

To estimate the parameters, we optimize the profile likelihood

$$\ell(\vartheta) = \sum_{j=1}^{k}\sum_{i\in R_j}\ell_i(\vartheta) = \sum_{j=1}^{k}\sum_{i\in R_j}\left[\ln \pi_j + \ln f_{\text{CN}}\left(x_i;\mu_j,\Sigma_j,\alpha_j,\eta_j\right)\right], \tag{2}$$

where R_j denotes the set of observations assigned to the j-th cluster. To make maximization of (2) a well defined problem, we adopt the classical eigenvalue ratio constraint proposed by García-Escudero *et al.* , 2008.

Model (1) involves the difficult choice of $\alpha_0, \alpha_1, \ldots, \alpha_k$, where α_0 controls the proportion of gross outliers and α_j the proportion of mild outliers in the j-th cluster. We propose a LASSO-type penalized likelihood approach enforcing a sparse model selection in which some values in the set $(\alpha_0, \alpha_1, \ldots, \alpha_k)$ might be set to zero. A general form of penalized log-likelihood is given by

$$\ell(\vartheta) + P(\alpha_0, \alpha_1, \ldots, \alpha_k), \tag{3}$$

and we propose using $P(\alpha_0, \ldots, \alpha_k) = -\log(n)\sum_{j=0}^{k} \nu_j \alpha_j$. In order to reduce the number of penalty parameters, we set $\nu_0 = n\nu$ and $\nu_j = \nu$ for $j > 0$.

The choice of the penalty parameter ν has got direct consequences on the estimated trimming proportion α_0. If also $\alpha_1, \ldots, \alpha_k$ are included in the penalty, it also affects their estimates. Surprisingly enough, mapping the problem of selecting contaminating proportions to the scale of the likelihood gives an asymptotically "optimal" fixed value, $\nu = \sqrt{2d}$, which under certain assumptions guarantees that observations outside a chi-square type ellipse from a bulk of the data are trimmed.

Maximization of (2), and for fixed ν of (3), is carried out using a classification expectation-conditional maximization (CECM) algorithm, where eigenvalue ratio constraints are activated at the conditional maximization step is needed,

3 Example about clustering illicit drug shipments

We analyze data about $n = 151$ seizures of shipments of cocaine and heroin in Italy and Spain. They were sent to the forensic laboratories for checking the nature of the substance and quantifying the absolute and relative contents of each of several chemical compounds. In modern forensics it is believed that the contents of certain solvents might be useful for identifying the source, that is, clustering packages with respect to the illicit laboratory where the drug was processed. We verify this assumption by focusing on $d = 3$ compounds: hexane, acetone, and 2-propanol. We fix $k = 2$ and estimate a classical normal mixture model and a contaminated normal mixture model without trimming first. Then we use robust clustering methods: `tclust` and the contaminated-normal mixture model with trimming. In Table 1 we report, for values of the trimming level chosen using our penalized likelihood approach, the adjusted

Rand-index (ARI) showing the agreement between the class labels and the true underlying Italy/Spain location of seizure. With no or insufficient trimming one might conclude that there is no relationship between solvent contents and seizure location. On the other hand, after trimming the agreement becomes fairly high. As expected we note that the optimal trimming level using `tclust` is slightly larger than those using CNTCLUST0. While in our low sample size example this might not have strong consequences in terms of MSE, $\lceil 151(0.066-0.053)\rceil = 2$ seizures will not be attributed to a location using `tclust`, which can have forensic consequences.

Table 1. *Adjusted Rand-index (ARI) for location of drug seizure and clustering. In parentheses the trimming level. NM: normal mixture, CNM: contaminated normal mixture,* `tclust`*: trimmed NM, CNTCLUST0: trimmed CNM, CNTCLUST: penalized trimmed CNM with* $\nu = \sqrt{2d}$ *and fixed trimming level. The trimming level selected with our fixed-penalty approach is indicated with* $\hat{\alpha}_0$*.*

Method	ARI
NM	-0.076
CNM	-0.069
`tclust` ($\hat{\alpha}_0 = 0.066$)	0.657

Method	ARI
CNTCLUST0($\hat{\alpha}_0 = 0.053$)	0.660
CNTCLUST($\hat{\alpha}_0 = 0.053$)	0.658

References

FARCOMENI, A., & GRECO, L. 2015. *Robust Methods for Data Reduction*. Boca Raton, FL: CRC Press.

GARCÍA-ESCUDERO, L. A., GORDALIZA, A., MATRAN, C., & MAYO-ISCAR, A. 2008. A general trimming approach to robust cluster analysis. *Annals of Statistics*, **36**, 1324–1345.

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

RITTER, G. 2015. *Robust Cluster Analysis and Variable Selection*. CRC Press.

GAUSSIAN PROCESSES FOR CURVE PREDICTION AND CLASSIFICATION

Sara Fontanella[1], Lara Fontanella[2], Rosalba Ignaccolo[1], Luigi Ippoliti[2]
and Pasquale Valentini[2]

[1] University of Torino, (e-mail: sara.fontanella@unito.it, ignaccolo@econ.unito.it)

[2] University "G. d'Annunzio" of Chieti-Pescara,
-(e-mail: lara.fontanella@unich.it, luigi.ippoliti@unich.it, pasquale.valentini@unich.it)

ABSTRACT: Effective statistical modelling under complex designs for functional data is still under development and requires innovative theories. In this work, we discuss an approach for modelling multivariate dependent functional data, where the dependence can arise via multiple responses, temporal or spatial effects. Specifically, we consider bivariate functional data and illustrate the proposed methodology in the frameworks of spatial patterns detection and curve prediction. To account for dominant structural features, we rely on the theory of Gaussian Processes (GPs) and extend hierarchical dynamic linear models for multivariate time series to functional data setting. An interesting feature of the proposed framework is that it allows to leverage knowledge from one process when solving an inferential task for another and to use derivative data for curve prediction. This framework also leads to the notion of derivative principal component analysis, which complements functional principal component analysis, one of the most popular tools of functional data analysis and facilitates the use of multivariate statistical techniques..

KEYWORDS: Gaussian processes, functional data, derivative process.

A NEW PROPOSAL FOR BUILDING IMMIGRANT INTEGRATION COMPOSITE INDICATOR

Mario Fordellone[1], Venera Tomaselli[2] and Maurizio Vichi[1]

[1] Department of Statistical Sciences, Sapienza University of Rome,
(e-mail: mario.fordellone@uniroma1.it, maurizio.vichi@uniroma1.it)

[2] Department of Political and Social Sciences, University of Catania,
(e-mail: tomavene@unict.it)

ABSTRACT: Integration consists in a multidimensional process, which can take place in different ways and in different times in relation to each single economic, social, cultural, and political dimension. In this paper, we aim at providing a methodological proposal based on PLS-SEM to build a composite immigrant integration indicator.

KEYWORDS: partial least squares, immigrant integration, composite indicator, structural equation modelling.

1 Measuring immigrant integration

Integration consists in a multidimensional process, which can take place in different ways and in different times in relation to each single economic, social, cultural, and political dimension. It aims at pursuing mutual respect of ethno-cultural differences and peaceful coexistence among populations within a historical and social reality. Its goal cannot be reached once for all but must be continuously pursued distinguishing different integration processes at economic, cultural, social, and political level. A high economic integration level may be quickly achieved, indeed, along with scarce or no social or political integration. Each single dimension, diachronically positioned over time, generates different integration levels. Hence, examining each single dimension is important as well as building composite indexes simultaneously comprehensive of all dimensions in order to obtain a full description of a complex phenomenon and to convey a suitable set of information.

According to the literature (Entzinger, 2000), the concept of integration can be broken down into different dimensions. Firstly, the socio-economic dimension refers to housing conditions, work conditions and income. Including mostly the theme of citizenship, also the legal-political dimension takes into account two sub-dimensions. The other sub-dimension concerns the rights of political participation - from the freedom of association to the voting right - which in some countries can be used at local government elections even without having achieved the citizenship status of the host country. Finally, the cultural and social dimension considers several elements, among which knowledge of the Italian language, free times activities and access to information.

In this paper, we aim at providing a methodological proposal to build a composite immigrant integration indicator, able to measure the different aspects related to integration, such as employment, education, social inclusion, and active citizenship. With this in mind, we analyse the data collected from European Social Survey (ESS), Round 8, on immigration by the Partial Least Squares Path Modelling (PLSPM) approach (Tenenhaus et al., 2005). The PLSPM models are Structural Equation Modelling suitable to estimate interaction and main effects among multiple sets of latent variables. In the present study we use a simultaneous non-hierarchical clustering and Partial Least Squares Modelling, named Partial Least Squares K-Means (PLS-KM), recently proposed by Fordellone and Vichi (2017). In this model, centroids are laying the reduced space of the latent variables, ensuring the optimal partition of the statistical units on the best latent hyperplane. Estimating the measurement relations by the SEM pre-specified model, the latent structure is defined.

2 ESS data

The data from the eighth iteration of the survey for ESS are until now available from 18 of the 24 countries, which undertook fieldwork in 2016. The 18 countries included in this initial release are: Austria, Belgium, Czech Republic, Estonia, Finland, France, Germany, Iceland, Ireland, Israel, Norway, Netherlands, Poland, Russia, Slovenia, Sweden, Switzerland and United Kingdom. The included questions asked in every round since 2002 on topics including crime, democracy and politics, human values, immigration, media consumption, national and ethnic identity, perceived discrimination, religion, social exclusion, social trust/trust in institutions, subjective wellbeing and socio-demographics and public attitudinal data towards welfare, climate change and energy security, personal norms, efficacy and trust and energy preferences. The data must be weighted to adjust for different selection probabilities, for sampling error and non-response bias as well as different selection probabilities. The table 1 shows the topics covered by the survey in the collection of questions, classified into two main parts: a core section and a rotating section. The core module contains items measuring a range of topics of enduring interest to the social sciences as well as the most comprehensive set of socio-structural variables of any cross-national survey. The rotating modules are carried out by multi-national teams of researchers selected to contribute to the design of survey.

Table 1 - *Topics and items of ESS.*

Items	**Topics**
Core A1-A6	Media use; internet use; social trust
Core B1-B43	Politics, including: political interest, trust, electoral and other forms of participation, party allegiance, socio-political orientations, immigration
Core C1-C44	Subjective wellbeing, social exclusion, crime, religion, perceived discrimination, national and ethnic identity, test questions (sect. I), refugees
Rotating D1-D32	Climate change and energy, including: attitudes, perceptions module and policy preferences
Rotating E1-E42	Welfare, including attitudes towards welfare provision, size of module claimant groups, attitudes towards service delivery and likely future dependence on welfare, vote intention in EU referendum
Core F1-F61	Socio-demographic profile, including household composition, sex, age, marital status, type of area, education and occupation, partner, parents, union membership, income and ancestry
Core Section H	Human values scale
Core Section l	Test questions

Source: www.europeansocialsurvey.org.

The ESS sampling strategy is based on the design and implementation of workable and equivalent sampling plans in all participating countries, following key principles:

- samples must be representative of all persons aged 15 and over (no upper age limit) resident within private households in each country, regardless of their nationality, citizenship or language
- individuals are selected by strict random probability methods at every stage
- sampling frames of individuals, households and addresses may be used
- all countries must aim for a minimum 'effective achieved sample size' of 1,500 or 800 in countries with ESS populations of less than 2 million after discounting for design effects
- quota sampling is not permitted at any stage
- substitution of non-responding households or individuals (whether 'refusals', 'non-contacts' or 'ineligibles') is not permitted at any stage.

In the present paper, we use ESS Multilevel Data resource in order to analyse the ESS-respondents with reference to the context they live in. The resource contains data about:

- individuals (the ESS respondents)
- regions (mainly data collected from EUROSTAT)
- countries (data collected from different sources)

3 Methodology

Given the $n{\times}J$ data matrix $\mathbf{X}$, the $n{\times}K$ membership matrix $\mathbf{U}$, the $K{\times}J$ centroids matrix $\mathbf{C}$, the $J{\times}P$ loadings matrix $\mathbf{\Lambda} = [\mathbf{\Lambda}_{\mathrm{H}}, \mathbf{\Lambda}_{\mathrm{L}}]$, and the errors matrices $\mathbf{E}$, $\mathbf{Z}$, $\mathbf{D}$, the Partial Least Squares Structural Equation Modelling *K*-Means approach can be written as follows (Fordellone and Vichi, 2017; Fordellone et al., 2018):

$$\begin{aligned}
&\mathbf{H} = \mathbf{H}\mathbf{B}^{\mathrm{T}} + \mathbf{\Xi}\mathbf{\Gamma}^{\mathrm{T}} + \mathbf{Z},\\
&\mathbf{X} = \mathbf{\Xi}\mathbf{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{H}\mathbf{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{E},\\
&\mathbf{X} = \mathbf{U}\mathbf{C}\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} = \mathbf{U}\mathbf{C}\mathbf{\Lambda}_{\mathrm{H}}\mathbf{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{U}\mathbf{C}\mathbf{\Lambda}_{\mathrm{L}}\mathbf{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{D},
\end{aligned} \tag{1}$$

under constraints: *(i)* $\mathbf{\Lambda}^{\mathrm{T}}\mathbf{\Lambda} = \mathbf{I}$; and *(ii)* $\mathbf{U} \in \{0,1\}$, $\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$. Where, *H* is the n×L matrix of the endogenous LVs with generic element $\eta_{i.l}$, $\mathbf{\Xi}$ be the $n{\times}H$ matrix of the exogenous LVs with generic element $\xi_{i.h}$, $\mathbf{B}$ is the $L{\times}L$ matrix of the path coefficients $\beta_{l.l}$ associated to the endogenous latent variables, $\mathbf{\Gamma}$ is the $L{\times}H$ matrix of the path coefficients $\gamma_{l.h}$ associated to the exogenous latent variables, $\mathbf{\Lambda}_{\mathrm{H}}$ is the $J{\times}H$ loadings matrix of the exogenous latent constructs with generic element $\lambda_{i.h}$, and $\mathbf{\Lambda}_{\mathrm{L}}$is the $J{\times}L$ loadings matrix of the endogenous latent constructs with generic element $\lambda_{i.l}$. Thus, the PLS-SEM-KM model includes the SEM estimated via Partial Least Squares (PLS) and the clustering equations. The simultaneous estimation of the three sets of equations will produce the estimation of the pre-specified SEM describing relations among variables and the corresponding best partitioning of units

There is a relevant aspect to considerate in the application of PLS-SEM-KM procedure: when we applying PLS-SEM-KM, the number of groups is unknown and the identification of an appropriate number of *K* clusters is not straightforward. Then, often you need to rely on some statistical criterion. In particular, the PLS-SEM-KM algorithm includes the choice of the number of clusters *K* classes according the *gap method* criterion (Fordellone and Vichi, 2017).

References

ENTZINGER, H. 2000. The Dynamics of Integration Policies: A Multidimensional Model. Challenging Immigration and Ethnic Relations Politics. *R. Koopmans and P. Statham, Oxford: University Press*.

FORDELLONE, M., VICHI, M. 2017. Partial Least Squares Modelling and simultaneous clustering. *Cladag 2017 Book of Short Papers. Universitas Studiorum*.

FORDELLONE, M., TOMASELLI V., AND VICHI M. 2018. From Tandem to Simultaneous Dimensionality Reduction And Clustering Of Tourism Data. *Rivista Italiana di Economia Demografia e Statistica 72.1*.

TENENHAUS, M., VINZI, E.V., CHATELIN, Y.M., LAURO, N.C. 2005. PLS path modeling. *Computational Statistics & Data Analysis. Vol. 48 No. 1, pp. 159–205*.

BIODIVERSITY SPATIAL CLUSTERING

F. Fortuna[1], F. Maturo[2] and T. Di Battista[1]

[1] DISFPEQ Department, G.dAnnunzio University of Chieti-Pescara,
(e-mail: francesca.fortuna@unich.it, dibattis@unich.it)

[2] ”Luigi Vanvitelli” University of Caserta,
(e-mail: fabrizio.maturo@unicampania.it)

ABSTRACT: The recognition of spatial heterogeneity through spatial techniques is essential to guide decision-making regarding biodiversity conservation. Many ecological studies concerning a spatial approach for biodiversity have focused only on species richness or evenness, leading to a partial overview of this complex concept. For this reason, we propose a spatial functional approach to diversity profiles for assessing spatial biodiversity and identifying groups of curves which are similar in spatial patterns. Specifically, the distance-based LISA algorithm has been extended to the case of functional diversity profiles in lattice, after smoothing the discretized curves and specifying a suitable distance measure.

KEYWORDS: spatial FDA, spatial lattice data, LISA map, diversity profile.

1 Introduction

The identification of spatial patterns in species diversity represents an essential issue to establish conservation strategies and monitoring programs (Hernndez-Stefanoni *et al.*, 2011). Specifically, mapping biodiversity is crucial to investigate spatial variations in natural communities. Although spatial patterns of richness and diversity indices are among the most-studied patterns in ecology, they do not provide a reliable biodiversity representation as they neglect the multivariate nature of this complex concept. The use of diversity profiles has been recommended in the literature to solve this issue (Patil & Taillie, 1982). Indeed, they provide a graphical representation of a collection of indices belonging to the same parametric family. Since diversity profiles are presented as curves, they have been analyzed in a functional framework (Di Battista *et al.*, 2016, Di Battista & Fortuna, 2017, Maturo & Di Battista, 2018). However, these studies have focused on independent curves, which is not a reasonable assumption in the environmental fields.

For this reason, we propose a spatial functional approach to diversity profiles for identifying groups of curves which are similar in spatial patterns and evaluating the possibility of improving the accuracy of biodiversity maps. Specifically, a distance-based LISA map (Delicado & Broner, 2008) has been applied to functional diversity profiles in a spatial finite lattice. The main advantage of our approach is that it allows to identify spatial patterns by jointly considering the two fundamental aspects of biodiversity, that is the richness and the evenness. Moreover, regarding the data as functions has the advantage of overcoming some of the problems that are associated with irregularly spaced or sparse data (Haggarty *et al.*, 2015).

2 Unsupervised spatial classification of functions in lattice

Following Delicado et al. (2010), a spatial functional process can be defined as follows:

$$\left\{\mathcal{F}_s(x) : s \in \mathcal{D} \subset \mathbb{R}^d, x \in \mathcal{X} \subset \mathbb{R}\right\} \tag{1}$$

where s is a generic data location in the d-dimensional Euclidean space, $\mathcal{F}_s(x)$ are functional random variables, which are defined as random elements taking values in an infinite dimensional space, $x \in \mathcal{X}$ is the domain of the functions, and the set $\mathcal{D} \subseteq \mathbb{R}^d$ can be fixed or random. The realization of a spatial process, $f_{s_1}(x), f_{s_2}(x), ..., f_{s_n}(x)$, $s_i \in \mathcal{D}$, $i = 1, 2, ..., n$, constitutes a set of functional spatial data. The nature of the set $\mathcal{D}$ allows to classify spatial functional data (Cressie, 1993) in geostatistical functional data, functional marked point pattern and functional areal data. We focus on the latter case, that is on functions observed on a regular grid containing a finite number of sites whose whole constitutes the entire study region. To detect for the existence of spatial dependence and identify spatial clusters among curves, the distance-based LISA maps algorithm (Delicado & Broner, 2008) has been applied. It is a generalization of the well-known LISA maps for univariate data in lattice (Anselin, 1995) and can be applied to a wide range of data types, provided that a dissimilarity measure can be defined between any pair of observations.

In the functional context, for each location, a number of noise-corrupted raw-data, say $\{y_i(x_j)\}_{j=1}^{J}$, are sampled form a random trajectory $\mathcal{F}_i(x)$ at J equispaced points of the functional domain. Indeed, functional data are recorded only for discrete values of $x \in \mathcal{X}$; thus, for each i-th site, a linear approximation of the observed discretized trajectory can be computed using spline functions

(Ramsay & Silverman, 2005) as follows:

$$f_{s_i}(x) = \sum_{b=1}^{B} c_{ib}\phi_b(x) = \boldsymbol{c}_i^T \boldsymbol{\Phi}(x), \ \ i = 1,2,...,n \tag{2}$$

where $\boldsymbol{c}_i = (c_{i1}, c_{i2}, ..., c_{iB})$, is the coefficient vector, which defines the linear combination and $\boldsymbol{\Phi}(x) = \big(\phi_1(x), \phi_2(x), ..., \phi_B(x)\big)$ is the vector of basis functions (Ramsay & Silverman, 2005). In our case, $f_{s_i}(x)$ is a diversity profile, which represents a summary function of the biodiversity of the s_i area. A spatial cluster is defined as a set of areas that are close to each other having similar observed values for the variable of interest. This kind of clusters would exist when the functional variable $\mathcal{F}(x)$ presents spatial dependence at local level. To summarize the spatial relationship among n spatial units, a $n \times n$ spatial weight matrix $\boldsymbol{W}$ is specified. It is often defined by neighboring information, thus its elements w_{ij} are equal to one if s_i and s_j are neighbors and zero otherwise. Once the neighborhood matrix $\boldsymbol{W}$ has been defined, the distance-based LISA algorithm can be applied, after introducing a distance measure among the n observed functions. In the $L^2(\mathcal{X})$ space, a suitable distance is the L^2 norm:

$$d\Big(f_{s_i}(x), f_{s_j}(x)\Big) = \sqrt{\int_X \Big(f_{s_i}(x) - f_{s_j}(x)\Big)^2 dt} \tag{3}$$

which can be written as follows:

$$d_{ij} = \sqrt{\int_X (\boldsymbol{c}_i - \boldsymbol{c}_j)^T \boldsymbol{M} (\boldsymbol{c}_i - \boldsymbol{c}_j)} \tag{4}$$

where $\boldsymbol{M} = \int_X \boldsymbol{\Phi}(x)\boldsymbol{\Phi}^T(x)\,dt$ is a symmetric square matrix of order equal to the number of basis functions, and $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ are the coefficients of the basis expansion for $f_{s_i}(x)$ and $f_{s_j}(x)$, respectively. Then, the distance-based LISA maps algorithm consists into five steps (Delicado & Broner, 2008):

- **Step 1**: Detect global outliers.
- **Step 2**: Mark tracts significantly similar to (and significantly different from) their neighbors.
- **Step 3**: Mark non-marked tracts that are similar to a neighbor marked tract.
- **Step 4**: Identify spatial clusters by applying any standard clustering algorithm to the ares marked at Steps 2 and 3.
- **Step 5**: Draw the map.

Regarding the clustering step, standard unsupervised classification algorithms can be applied to the coefficients of basis functions.

References

ANSELIN, L. 1995. Local indicators of spatial association-LISA. *Geographical Analysis*, **2**, 93–115.

CRESSIE, N. 1993. *Statistics for Spatial Data*. New York: John Wiley & Sons.

DELICADO, P., & BRONER, S. 2008. *Distance-based LISA maps for multivariate lattice data*. Tech. rept. Universitat Politecnica de Catalunya.

DI BATTISTA, T., & FORTUNA, F. 2017. Functional confidence bands for lichen biodiversity profiles: A case study in Tuscany region (central Italy). *Statistical Analysis and Data Mining: The ASA Data Sci Journal*, **10**, 21–28.

DI BATTISTA, T., FORTUNA, F., & MATURO, F. 2016. Environmental monitoring through functional biodiversity tools. *Ecological Indicators*, **60**, 237–247.

HAGGARTY, R.A., MILLER, C.A., & SCOTT, E.M. 2015. Spatially weighted functional clustering of river network data. *Rojal Statistical Society - Series C*, **64**, 491–506.

HERNNDEZ-STEFANONI, J.L., GALLARDO-CRUZ, J.A., MEAVE, J.A., & DUPUY, J.M. 2011. Combining geostatistical models and remotely sensed data to improve tropical tree richness mapping. *Ecological Indicators*, **11**, 1046–1056.

MATURO, F., & DI BATTISTA, T. 2018. A functional approach to Hill's numbers for assessing changes in species variety of ecological communities over time. *Ecological Indicators*, **84**, 70–81.

PATIL, G. P., & TAILLIE, C. 1982. Diversity as a concept and its measurement. *Journal of the American Statistical Association*, **77**, 548–567.

RAMSAY, J.O., & SILVERMAN, B. W. 2005. *Functional Data Analysis, 2nd edn*. New York: Springer.

SKEWED DISTRIBUTIONS OR TRANSFORMATIONS? INCORPORATING SKEWNESS IN A CLUSTER ANALYSIS

Michael P.B. Gallaugher[1], Paul D. McNicholas[1], Volodymyr Melnykov[2] and Xuwen Zhu[3]

[1] Department of Mathematics and Statistics, McMaster University, (e-mail: gallaump@mcmaster.ca, paulmc@mcmaster.ca)

[2] Culverhouse College of Business, University of Alabama, (e-mail: vmel-nykov@cba.ua.edu)

[3] Department of Mathematics, University of Louisville, (e-mail: xuwen.zhu@louisville.edu)

ABSTRACT: Due to its mathematical tractability, the Gaussian mixture model holds a special place in the literature. However, in a clustering scenario, using a Gaussian mixture model when skewness or outliers are present can be problematic. As a result, in recent years, many different methods have been proposed to account for skewed clusters. The two most prevalent methods in the literature are modelling skewness directly by using skewed distributions, and performing clustering alongside a suitable transformation. Although both these methods have been studied extensively in the literature and compared for select datasets in terms of relative performance, no extensive study has been performed to motivate in which situation to use one method over another. Using many different real datasets, and looking at their underlying properties, such as measures of overlap between clusters, skewness, and kurtosis, we aim to provide more insight as to when one method - i.e., transformation or a skewed distribution - might be preferable to another. Simulated data and a large number of multivariate datasets will be considered.

KEYWORDS: skewed distributions, transformations, mixture models, clustering.

ROBUST PARSIMONIOUS CLUSTERING MODELS

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

[1] Departamento de Estadìstica e Investigación Operativa and IMUVA, Universidad de Valladolid, (e-mail: lagarcia@eio.uva.es, agustin@med.uva.es)

[2] Dipartimento di Scienze Economiche e Aziendali and Ro.S.A., Università di Parma, (e-mail: mriani@unipr.it)

ABSTRACT: In model based clustering, there are two main distinct approaches depending on whether the mixture or the classification likelihood function is used. It is well known that both likelihoods are unbounded without any constraint on the cluster scatter matrices. Constraints also prevent traditional EM and CEM algorithms from being trapped in (spurious) local maxima. Controlling the maximal ratio between the eigenvalues of the scatter matrices to be smaller than a fixed constant $c \geq 1$ is the traditional way for setting such constraints. In this paper we discuss other types of constraints and extend them to the family of the parsimonious Gaussian clustering models.

KEYWORDS: clustering, mixtures, EM algorith, CEM algoritm.

1 Introduction and notation

The traditional approach of unsupervised learning assumes multivariate normal components and adopts a maximum likelihood approach for clustering purposes. With this idea in mind, well-known classification and mixture likelihood approaches can be used.

In this work, we denote with symbol $\phi(\cdot;\mu,\Sigma)$ the probability density function of a p-variate normal distribution with mean μ and covariance matrix Σ.

In the *classification likelihood* approach, given a sample of observations $\{x_1,\cdots,x_n\}$ in $\mathbb{R}^p$, we search for a partition $\{H_1,...,H_k\}$ of the indices $\{1,\cdots,n\}$, centres $\mu_1,\cdots,\mu_k$ in $\mathbb{R}^p$, symmetric positive semidefinite $p\times p$ scatter matrices $\Sigma_1,\cdots,\Sigma_k$ and positive weights $\pi_1,\cdots,\pi_k$ with $\sum_{j=1}^k \pi_j = 1$, which maximize

$$\sum_{j=1}^{k}\sum_{i\in H_j}\log\left(\pi_j\phi(x_i;\mu_j,\Sigma_j)\right). \tag{1}$$

On the other hand, in the *mixture likelihood* approach, the idea is to maximize the expression below

$$\sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j \phi(x_i; \mu_j, \Sigma_j) \right), \tag{2}$$

In this latter approach, a partition into k groups can be also obtained, from the fitted mixture model, by assigning each observation to the cluster-component with the highest posterior probability.

It is well-known in the statistical literature that the maximization of "log-likelihoods" like (1) and (2) without constraints on the Σ_j matrices is a mathematically ill-posed problem, e.g. Day, 1969. It is possible to appreciate this problem taking $\mu_k = x_1$, $\pi_k = 1$ and $|\Sigma_k| \to 0$ making (2) to diverge to infinity or (1) also diverge with $H_1 = \{1\}$.

A simple way of tackling the lack of boundedness is to consider local maxima of the likelihood target functions. However, a lot of local solutions are often found and it is difficult to know which are the most interesting ones. See McLachlan & Peel, 2000 for a detailed discussion of this issue. In the literature, non-interesting local maxima are named "spurious" solutions. They usually are formed by some, almost collinear, observations and are often detected by the Classification EM algorithm (CEM), traditionally applied when maximizing (1), and by the EM algorithm, traditionally applied when maximizing (2). A paper which tackles this problem together with suggestions for reducing spurious solutions can be found in García-Escudero *et al.*, 2018.

The use of constraints on the relative sizes of the determinant of the Σ_j matrices may be seen as a simple and useful way to overcome these degeneracy issues and to apply affine equivariant constraints. This approach has been proposed by McLachlan & Peel, 2000 and lies behind the EVV (equal volume, variable shape and orientation) parametrization within the well-known Gaussian parsimonious clustering models Celeux & Govaert, 1992; Banfield & Raftery, 1993. In the following section we show this approach does not fully avoid the detection of degenerate (spurious) solutions. Moreover, different clustering approaches can be defined depending on the strength of these two, determinant and shape, types of constraints.

2 An approach based on determinant-and-shape constraints

We have seen that $|\Sigma_j| \to 0$, for any j, may be problematic. We could therefore consider the maximization of (1) and (2) but under

Determinant constraints: we force

$$\frac{\max_{j=1,...,k} |\Sigma_j|}{\min_{j=1,...,k} |\Sigma_j|} \leq c_1, \tag{3}$$

for a given fixed constant $c_1 \geq 1$.

The particular case $c_1 = 1$ forces all the determinants of the scatter matrices to be equal, i.e. $|\Sigma_1| = ... = |\Sigma_k|$. This case corresponds to the approach in McLachlan & Peel, 2000 and to the EVV (equal volume, variable shape and orientation) parametrization within the Gaussian parsimonious family. When considering $1 < c_1 < \infty$, we relax the exact "equal determinant" assumption without leaving determinants completely free.

Notice that (3) implies that if any of the determinants $|\Sigma_j|$ goes to 0 then all the other determinants also have to go to 0 and this solution is not interesting. It is also trivial to see that this type of constraints is affine equivariant.

However, even when all the $|\Sigma_j|$ determinants are kept away from 0, degeneracy troubles still may take place, because some eigenvalues of the Σ_j matrices may still go to 0. More in detail, let us consider the well-known decomposition for the covariance matrices Σ_j

$$\Sigma_j = \lambda_j^{1/p} \Omega_j \Gamma_j \Omega_j',$$

where Ω_j is an orthogonal matrix of eigenvectors, Γ_j is a diagonal matrix with $|\Gamma_j| = 1$ and with elements $\{\gamma_{j1}, ..., \gamma_{jp}\}$ in its diagonal (proportional to the eigenvalues of the Σ_j matrix) and $|\Sigma_j| = \lambda_j$. These Γ_j matrices are commonly known as "shape" matrices, because they determine the shape of the fitted cluster components. Notice that (3) can be rewritten as

$$\frac{\max_{j=1,...,k} \lambda_j}{\min_{j=1,...,k} \lambda_j} \leq c_1.$$

To see that degeneracy problems may still happen, even with controlled determinant sizes, it is enough to set $p = 2$ and take $\mu_k = x_1$, $\pi_k > 0$, $\lambda_k = 1$, $\gamma_{k1} = C$ and $\gamma_{k2} = 1/C$. The remaining Σ_j matrices, $j = 2, ..., k$, are arbitrarily chosen but satisfying $|\Sigma_2| = ... = |\Sigma_k| = 1$. Note that the smallest eigenvalue of Σ_1 converges to 0 when $C \uparrow \infty$ and, then, one of the fitted components can be made arbitrarily close to a degenerate normal component.

In order to tackle the above explained source for degeneracy, we may consider, besides (3), an additional type of constraint which controls the elements of the "shape" matrices as:

Shape constraints: consider the following k constraints:

$$\frac{\max_{l=1,\ldots,p} \gamma_{jl}}{\min_{l=1,\ldots,p} \gamma_{jl}} \leq c_2, \qquad \text{for} \qquad j = 1, \ldots, k, \tag{4}$$

where $c_2 \geq 1$.

Notice that (4) imposes k independent set of constraints, one for each shape matrix, and nothing relates the shape matrix elements of one component to the other components.

The combination of different combinations of c_1 and c_2 values, with the constraints $1 \leq c_1 < \infty$ and $1 \leq c_2 < \infty$, enables us to consider different clustering approaches throughout their associated constrained maximizations.

Note that with a very large c_2 value (e.g., $c_2 = 10^{10}$) we are virtually affine equivariant. That choice would constitute just mild constraints on the scatter matrices "condition numbers" (ratios between the largest and smallest eigenvalues). This type of constraint has to be considered as a sort of convenient "computational precision" protection especially when dimension increases. In the extended version of the paper the above concept are applied to each member (when necessary) of the family of Gaussian parsimonious clustering models.

References

BANFIELD, J. D., & RAFTERY, ADRIAN E. 1993. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, **49**, 803–821.

CELEUX, G., & GOVAERT, A. 1992. A Classification EM algorithm for clustering and two stochastic versions. *Comput Stat Data Anal*, **14**, 315–332.

DAY, N.E. 1969. Estimating the Components of a Mixture of Normal Distributions. *Biometrika*, **56**, 463–474.

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

MCLACHLAN, G., & PEEL, D.A. 2000. *Finite mixture models*. New York: Wiley Series in Probability and Statistics.

PROJECTION-BASED UNIFORMITY TESTS FOR DIRECTIONAL DATA

Eduardo García-Portugués[1], Paula Navarro-Esteban[2]
and Juan Antonio Cuesta-Albertos[2]

[1] Department of Statistics, University Carlos III of Madrid,
(e-mail: edgarcia@estecon.uc3m.es)

[2] Department of Mathematics, Statistics and Computation, University of Cantabria,
(e-mail: paula.navarro@unican.es, juan.cuesta@unican.es)

ABSTRACT: Testing uniformity of a sample supported on the hypersphere is one of the first steps when analyzing multivariate data for which only the directions (and not the magnitudes) are of interest. In this work, a projection-based class of uniformity tests on the hypersphere is introduced. The new class allows for extensions of circular-only uniformity tests and introduces the first instance of an Anderson–Darling test in the context of directional data. A simulation study corroborates the theoretical findings. Finally, a real data example illustrates the usage of the new tests.

KEYWORDS: circular data, directional data, hypersphere, Sobolev tests, uniformity.

1 Setting

Testing uniformity of a sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$ of a random vector $\mathbf{X}$ supported on the hypersphere $\Omega_q := \{\mathbf{x} \in \mathbb{R}^{q+1} : \mathbf{x}'\mathbf{x} = 1\}$ of $\mathbb{R}^{q+1}$, with $q \geq 1$ is one of the first steps when analysing multivariate data for which only the directions (and not the magnitudes) are of interest – the so-called *directional data*. This kind of data arise in many applied disciplines, such as astronomy, biology, etc.

The inspiration for this contribution comes from the projection-based test of Cuesta-Albertos *et al.* , 2009, which is based on the fact that the distribution of $\mathbf{X}$ is determined by that of a one-dimensional *random* projection, $\gamma'\mathbf{X}$. For each γ (uniformly distributed on Ω_q and independent of the sample), Cuesta-Albertos *et al.* , 2009 considered a Kolmogorov–Smirnov test statistic on the projected sample $\gamma'\mathbf{x}_1,\ldots,\gamma'\mathbf{x}_n$. This test clearly depends on γ, which Cuesta-Albertos *et al.* , 2009 mitigates by taking k random directions $\gamma_1,\ldots,\gamma_k$ and combining the p-values associated to each of the k tests.

2 Results

Differently from Cuesta-Albertos *et al.* , 2009, we consider for each γ the well-known weighted quadratic norm by Anderson & Darling, 1954:

$$Q^w_{n,q,\gamma} := n \int_{-1}^{1} \left(F_{n,\gamma}(x) - F_q(x)\right)^2 w(F_q(x))\, \mathrm{d}F_q(x), \tag{1}$$

where w is a weight function, $F_{n,\gamma}$ and F_q are the empirical cumulative distribution function and the cumulative distribution function of the projected sample, respectively. In addition, instead of drawing several random directions and aggregating afterwards the outcomes of the associated tests, our statistic itself gathers information from all the directions on Ω_q: it is defined as the *expectation* of (1) with respect to γ. The new class of uniformity tests is thus the one indexed by the weights w.

Using this formulation, simple expressions for several test statistics are obtained for the circle and sphere, and relatively tractable forms for higher dimensions. Despite their different origins, the proposed class and the well-studied Sobolev class of uniformity tests (see Prentice, 1978) are shown to be related. Our new parametrization proves itself advantageous by allowing to derive new tests for hyperspherical data that neatly extend the circular tests by Watson, Ajne, and Rothman, and by introducing the first instance of an Anderson–Darling-like test in such context. The asymptotic distributions and the local optimality against certain alternatives of the new tests are obtained.

References

ANDERSON, T. W., & DARLING, D. A. 1954. A test of goodness of fit. *Journal of the American Statistical Association*, **49**, 765–769.

CUESTA-ALBERTOS, J. A., CUEVAS, A., & FRAIMAN, R. 2009. On projection-based tests for directional and compositional data. *Statistics and Computing*, **19**(4), 367–380.

PRENTICE, M. J. 1978. On invariant tests of uniformity for directions and orientations. *The Annals of Statistics*, **6**(1), 169–176.

GRAPH-BASED CLUSTERING OF VISITORS' ARAJECTORIES AT AXHIBITIONS *

Martina Gentilin[1], Pietro Lovato[2], Gloria Menegaz[2], Marco Cristani[2]
and Marco Minozzo[1]

[1] Department of Economics, University of Verona,
(e-mail: (martina.gentilin_01,marco.minozzo)@univr.it)

[2] Department of Computer Science, University of Verona,
(e-mail: (pietro.lovato, marco.cristani, gloria.menegaz)@univr.it)

ABSTRACT: In this paper we apply graph theory techniques on real data visitors' paths recorded during an exhibition to detect clusters of stands. We consider in particular the dominant set clustering technique, which finds complete heavy subgraphs in weighted undirected graphs. The resulting overlapping clusters could be used to set a travel recommendation system, identify market segments and assess stand assignment effectiveness.

KEYWORDS: trajectory clustering, dominant set, graph theory, fuzzy method.

1 Trajectory unsupervised classification

The spreading of new location referencing systems in smartphones and other personal devices is favouring the collection of huge amounts of trajectory data. As showed by Zheng (2015), nowadays many algorithms can be used to extract interesting insights from these path information. Different techniques have been used to: preprocess and manage raw data, mine patterns, detect outliers, classificate trajectories and transform them into graphs, matrices and tensors. Given the quantity of data collectable and the strict privacy policies spreading worldwide (that often do not permit to analyze jointly trajectory data and other information about users), unsupervised classification methods (clustering techniques) are particularly interesting. The main objectives of these techniques in trajectory data mining are to: identify representative paths or subpaths, find the most popular route, find the most likely route, detect underlying problems in a network, calculate similarity between users, discover cluster of locations with denser connections, calculate user's interest in unvisited location and set a travel recommendation system.

*This research has been carried out in collaboration with Veronafiere S.p.A. and funded by Regione Veneto with the European Social Fund (1695-29-11-2018, DGR 11/2018).

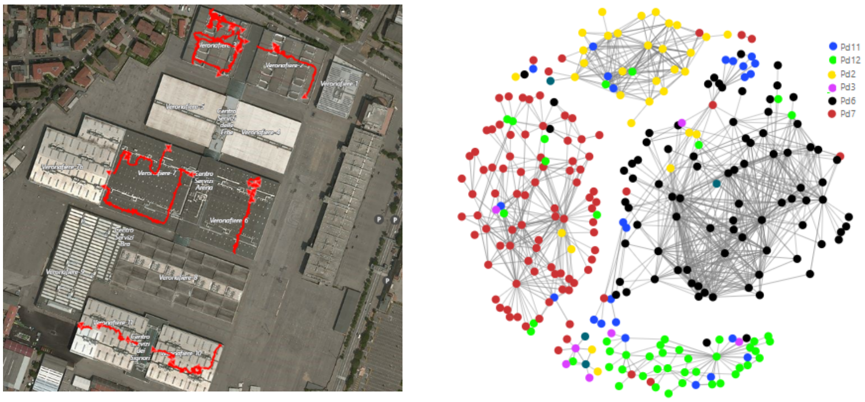


Figure 1: Raw trajectory (left) and undirected graphs (right).

In this paper we apply the dominant set clustering technique (Pavan & Pelillo, 2003), a state-of-the-art unsupervised classification algorithm, to analyze trajectories recorded during an exhibition. Based only on the information embedded in the trajectories, we propose a method that can be used to detect if the visits of certain stands can give information about logistics, provide a next visit recommendation system for visitors and identify market segments of stand exhibitors.

The raw information we started working on is a pilot collection of trajectories recorded during a four days marble exhibition through an accurate real-time positioning technology using Bluetooth Low energy signal from smartphones and HAIP Locators mounted on the ceiling of six exhibition halls, along an area of approximately 58 thousands sqm. The raw data consisted of 1,192 trajectories defined by a sequence of datetime, latitude and longitude information, with a theorical capture of one registration per second per device (see an example of a trajectory in Figure 1 (left)). In total we counted about 1,946 thousands points. However, not all this information has beed used since permanence at stands showed to be very poor, mainly due to smartphone sleep settings and open air areas presence. To analyze these data, we transformed the raw trajectories into undirected edge-weighted graphs with no-self loops: $G = (V, E, w)$, where $V = (1, \ldots, n)$ is the set of *nodes* representing stands visited for more than 2 seconds in a given trajectory (*stay points*), $\mathrm{E} \subseteq \mathrm{V} \times \mathrm{V}$ is the *edge* set, each edge representing that at least one visitor passed by both stands in the connected nodes and stayed there for more than 2 seconds, and

$w: E \to \mathbb{R}_+^*$ is the positive weight function, counting the number of trajectories including both stands. A representation of the resulting graphs is shown in Figure 1 (right), where different colours identify nodes (stay points) belonging to the six exhibition halls. The graphical representation gives an immediate idea of some possible clusters. However, this representation does not distinguish between edges with different weight and does not identify complete subgraphs (in which each node is connected to each other node in the subgraph). These problems call for the use of a trajectory clustering algorithm.

2 Dominant set algorithm for clustering

Graphical theoretic algorithms basically consist of searching for certain structures in the graph, such as a spanning tree, minimum cut or maximal complete subgraph. Pavan & Pelillo (2003) proposed an optimization function and an easy algorithm to find maximal complete subgraphs (*dominant set*) in weighted undirected graphs. Basically, the idea is to find a cluster defined by a set of vertices with higher edge-weights on average. To detect this dominant set of vertices, they first calculate each node importance, compared to a given set of vertices, in terms of average edge-weight. If this value is positive (connection higher than average) the node becomes part of the maximal complete subgraph, otherwise it is kept out. The initial specified set of vertices has to be changed and calculations must be repeated till convergence. Given a symmetric nonnegative $n \times n$ matrix A (called *weighted adjacency matrix*), with elements equal to $w(i, j)$, if $i, j \in E$ and 0, otherwise, they demonstrate that finding the dominant set is equivalent to find the local maximum x of

$$f(x) = x^T Ax, \text{ with } x \text{ in the standard simplex } \Delta \text{ of } \mathbb{R}^n, \tag{1}$$

where x is an n-dimensional positive vector representing the participation of each node to the cluster, the function $f(x)$ represents the cohesiveness of the cluster and the standard simplex constraint serves to normalize x. They solved this optimization problem using replicator dynamics taken from evolutionary game theory (more details about the optimization algorithm can be found in Pavan & Pelillo, 2003). In order to detect more than one dominant set they also proposed an iterative procedure which alternates the search of the dominant set in the graph and the deletion of its edge-weights.

The application of the dominant set algorithm to the marble exhibition trajectory data allowed to identify overlapping clusters of stands. An example of three dominant sets located in the sixth exhibition hall is shown in Figure 2

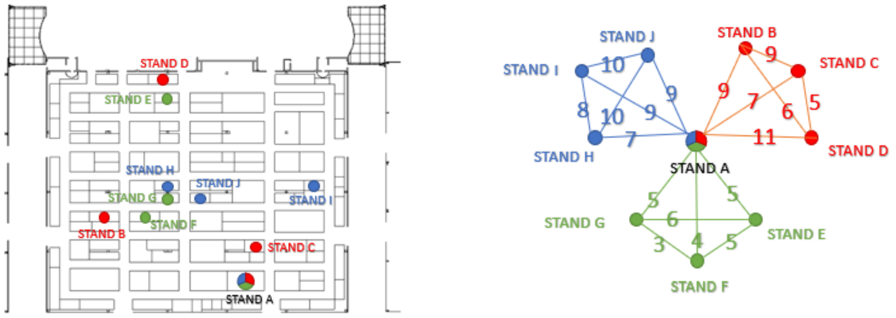


Figure 2: Example of overlapping clusters of visits inside an exhibition hall.

(left and right): one cluster of stands is represented by the blue dots (mainly cave owners), another by the red dots (mainly natural stones traders) and a third one by the green dots (mainly design products sellers). The stand represented by a three colours dot in the bottom right of the exhibition hall (Figure 2, left) belongs to the three different clusters.

This clustering can be used, for example, to suggest visitors of stands A and C to visit stands B and D (ordered by edge weights), setting a travel recommendation system. Moreover, it can be used to enhance stand assignment by detecting if visitors stop by stands belonging to the same dominant set, but located in different buildings. Further, the belonging of a stand to different clusters might be exploited to characterize segments of visitors.

We chose this approach for our clustering problem as it does not supply a full dendrogram (which is burdensome in the case of huge amounts of data), and does provide a flexible number of clusters. Moreover it offers a natural measure of within cluster's cohesiveness (average edge-weight) and an evaluation of nodes participation to each cluster (corresponding node value in the x vector), which are desirable features in an exhibition context.

References

PAVAN, M., & PELILLO, M. 2003. A new graph-theoretic approach to clustering and segmentation. *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

ZHENG, Y. 2015. Trajectory data mining: an overview. *ACM Trans. Intell. Syst. Technol. 6*, **3**, 29:1–29:41.

Symmetry in Graph Clustering

Andreas Geyer-Schulz[1] and Fabian Ball[1]

[1] Informationsdienste und elektronische Märkte, Karlsruhe Institute of Technology (KIT), Karlsruhe, (e-mail: andreas.geyer-schulz@kit.edu, fabian.ball@kit.edu)

Abstract: In this contribution we give a survey of our results on analyzing graph clustering results of graphs with more or less symmetry. These results fall into two different classes. The first class is purely mathematical: What is the impact of symmetry on the uniqueness and the stability of optimal partitions? And, how do we compare optimal partitions of symmetric graphs? The second class is empirical: Are these results relevant for applications of graph clustering in real life or are they just *l'art pour l'art*?

Keywords: graph clustering, automorphisms, pseudo-metric spaces, invariant partition comparison measures.

1 Introduction

In this contribution we give a survey of our results on analyzing graph clustering results of graphs with more or less symmetry. These results fall into two different classes. The first class is purely mathematical: What is the impact of symmetry on the uniqueness and the stability of optimal partitions? And, how do we compare optimal partitions of symmetric graphs? The second class is empirical: Are these results relevant for applications of graph clustering in real life or are they just *l'art pour l'art*? To answer these questions, we investigated the presence of symmetries in large sample of graphs from an Internet repository and the effect of symmetries on the uniqueness and stability of optimal graph partitions computed by the randomized greedy algorithm.

2 The Automorphism Group a Graph

Graphs with symmetry have non-trivial automorphism groups which are finite permutation groups (Wielandt, 1964). Recent advances in the implementation of algorithms for the analysis of the automorphism group of a graph (e.g. Darga *et al.* , 2008 and McKay & Piperno, 2014) allow the extraction of the set of generating permutations of the automorphism group of a graph.

The existence of such a non-trivial automorphism group of a graph implies that isomorphisms between at least some graph partitions exist. Given a partition, the set of all partitions that is generated by the automorphism group of the graph forms an equivalence class of graph partitions for this partition. When analyzing graphs with symmetry, we consider the pseudometric space of the equivalence classes generated by the automorphism group of the graph.

Whenever the equivalence class of the optimal partition of a graph cluster algorithm contains more than one element, the clustering solution is unstable and not unique. This solves the analysis of multiple optimal graph partitions which result from symmetry (see Geyer-Schulz & Ball, 2013). While this is a progress, this still leaves open the automatic analysis of multiple optimal graph partitions which are structurally different.

The problem of comparing graph partitions of symmetric graphs has also been introduced at the CLADAG 2013 conference. We now present its solution: We start with a minimal example which demonstrates problems of the Rand Index. Then we prove that this problem affects all existing graph partition comparison measures: They do not work for partitions of graphs with non-trivial automorphism groups.

As a remedy, we present three ways of building invariant graph comparison measures based on Hausdorff's and von Neumann's construction of invariant measures on a pseudo-metric space. By a combination of a pseudo-metric and a metric space we provide a measure decomposition which separates an invariant part which captures the structural difference and a part which is attributed to the action of the graph automorphism group on the partitions compared. See Ball & Geyer-Schulz, 2017, and, especially, Ball & Geyer-Schulz, 2018c.

3 Toy Examples: The Karate and the Petersen Graph

We finish the mathematical part with two examples: We show that for Zachary's Karate graph the optimal solution is not affected by symmetry, before we turn to the Petersen graph (Holton & Sheehan, 1993) which is a fully transitive graph. As far as we are aware, this is the first full analysis of clustering a fully transitive graph. For doing this, we use an extended version of the randomized greedy clustering algorithm (see e.g. Stein & Geyer-Schulz, 2013) and its ensemble variant (see Ovelgönne & Geyer-Schulz, 2013) and invariant measures for partition comparison (see Ball & Geyer-Schulz, 2018d and Ball & Geyer-Schulz, 2020).

4 Investigations of Graph Symmetry in Real-World Graphs

However, there remains the question of the relevance of the analysis of symmetry for applications in practice. Or as one reviewer has put it: *This research is completely irrelevant for practical applications and it will never be published in this journal.* For network sciences, for example, in social sciences, computer science and data science, only a few small-scale and restricted studies of the symmetry of complex real-world graphs exist. These studies show the existence of symmetry, but not the effects of symmetry e.g. on the stability of optimal partitions.

In the following, we report on our research on the existence of symmetries in real-world graphs, and the effects of symmetries on modularity-optimal solutions of real-world graphs.

The answer to the question of existence of symmetries is published in Ball & Geyer-Schulz, 2018a. In this study an analysis of over 1500 graph datasets from the meta-repository `networkrepository.com` has been carried out and a normalized version of the *network redundancy* measure has been presented. It quantifies graph symmetry in terms of the number of orbits of the symmetry group from zero (no symmetries) to one (completely symmetric), and improves the recognition of asymmetric graphs. Over 70% of the analyzed graphs contain symmetries (i.e., graph automorphisms), independent of size and modularity. Therefore, we conclude that real-world graphs are likely to contain symmetries. This contribution is the first larger-scale study of symmetry in graphs and it shows the necessity of handling symmetry in data analysis e.g. by the mathematical tools presented in the previous section.

The second study (Ball & Geyer-Schulz, 2018b) investigates the effect of graph symmetry on modularity optimal graph clustering partitions and it gives an insight to the effects of symmetry on optimal graph partitions. The key finding is that there actually exists an impact of graph symmetry, as more than 22% of the analyzed graphs have an unstable partition. The results are based on an empirical analysis of 1254 symmetric graphs, which are a subset of the 1699 graphs that were analyzed by Ball & Geyer-Schulz, 2018a. For each graph a modularity optimal partition is computed by one of the leading graph clustering algorithms. Additionally, generators for the automorphism group of each graph are obtained. All computed partitions are tested for stability (see Ball & Geyer-Schulz, 2018d), which means that the symmetry that is captured by the automorphism group does not change this partition. Furthermore, definitions that allow to distinguish local and global symmetry of graphs are presented.

References

BALL, F. 2019. *Impact of Symmetries in Graph Clustering*. Ph.D. thesis, Karlsruher Institut für Technologie, Karlsruhe.

BALL, F., & GEYER-SCHULZ, A. 2017. Weak Invariants of Actions of the Automorphism Group of a Graph. *Archives of Data Science, Series A*, **2**(1), 123 – 144.

BALL, F., & GEYER-SCHULZ, A. 2018a. How Symmetric Are Real-World Graphs? A Large-Scale Study. *Symmetry*, **10**(1(29)), 1 – 17.

BALL, F., & GEYER-SCHULZ, A. 2018b. The Impact of Graph Symmetry on Clustering. *Archives of Data Science, Series A*, **5**(1), 1 – 19.

BALL, F., & GEYER-SCHULZ, A. 2018c. Invariant Graph Partition Comparison Measures. *Symmetry*, **10**(10), 1 – 24.

BALL, F., & GEYER-SCHULZ, A. 2018d. Symmetry-based graph clustering partition stability. *Archives of Data Science, Series A*, **4**(1), 1 – 21.

BALL, F., & GEYER-SCHULZ, A. 2020. *Comparing Partitions of the Petersen Graph*. To appear.

DARGA, P. T., SAKALLAH, K. A., & MARKOV, I. L. 2008. Faster symmetry discovery using sparsity of symmetries. *Pages 149 – 154 of: 45th ACM/IEEE Design Automation Conference 2008*.

GEYER-SCHULZ, A., & BALL, F. 2013. Formal Diagnostics for Graph Clustering: The Role of Graph Automorphisms. *Pages 211 – 214 of:* MINERVA, T. ET. AL. (ed), *CLADAG 2013 – Book of Abstracts*. Modena: CLEUP.

HOLTON, D. A., & SHEEHAN, J. 1993. *The Petersen Graph: Australian Mathematical Society*. Cambridge: Cambridge University Press.

MCKAY, B. D., & PIPERNO, A. 2014. Practical graph isomorphism, II. *Journal of Symbolic Computation*, **60**(0), 94 – 112.

OVELGÖNNE, M., & GEYER-SCHULZ, A. 2013. An Ensemble Learning Strategy for Graph Clustering. *Pages 187–205 of:* BADER, D. A. ET AL. (ed), *Graph Partitioning and Graph Clustering*. Contemporary Mathematics, vol. 588. Providence: American Mathematical Society.

STEIN, M., & GEYER-SCHULZ, A. 2013. A Comparison of Five Programming Languages in a Graph Clustering Scenario. *Journal of Universal Computer Science*, **19**(3), 428 – 456.

WIELANDT, H. 1964. *Finite Permutation Groups*. New York: Academic Press.

BAYESIAN NETWORKS FOR THE ANALYSIS OF ENTREPRENEURIAL MICROCREDIT: EVIDENCE FROM ITALY

Lorenzo Giammei[1] and Paola Vicard[2]

[1] Department of Economics, Sapienza Università di Roma,
(e-mail: lorenzo.giammei@uniroma1.it)

[2] Department of Economics, Università Roma Tre, (e-mail: paola.vicard@uniroma3.it)

ABSTRACT: The aim of this study is to understand whether Bayesian networks are an appropriate tool to analyse and improve the performance of a microcredit initiative. This technique is employed to study simultaneously all the interactions between variables and perform *what-if* analyses. The analysed dataset originates from an important microcredit initiative aimed to help damaged firms, after the earthquake struck Italy in 2009. The model appears to provide a clear picture of the subject matter and seems to be appropriate to both assess risk connected to microcredit and support its development.

KEYWORDS: Bayesian network, microcredit, NPC algorithm.

1 Introduction

Modern microfinance was born in the 1970s as a financial instrument of social and economic integration, but rapidly spread all around the world. The first Italian law concerning microcredit was introduced in 2010, following Directive 2008/48/EC. The laws on the subject evolved during the years. The resulting regulation defines a financial instrument tailored on the needs of small firms and individuals, facing social and economic vulnerability. The amount of the loan is low but, together with every microcredit granted, the lender must provide some supplementary services, such as a business plan for firms and a help on how to manage the family budget for individuals. This is a very important component of Italian microcredit, since it addresses specific needs of the beneficiary and can consistently diminish credit risk.

A study from Borgomeo&co (2016) highlights that in Italy, from 2005 to 2014, individuals and firms are increasingly resorting to microcredit to access financial resources. However, performance studies related to recent microcredit programs, are barely available to date. It is crucial to start to analyse data, to understand strengths and weaknesses of entrepreneurial microcredit, as regulated by the new legislation.

Bayesian networks (BN) are proposed as a tool to perform the mentioned analysis. BNs are causal networks in which the strength of the relation is defined by probabilities and they are an effective instrument when reasoning under uncertainty. Through BN analysis, beneficiary and loan characteristics will be studied, in order to understand their role in determining the performance of a microcredit provision. The results can be useful to promote a healthy development of microcredit regulation and help microcredit firms evaluating credit risk.

2 Motivating data

The dataset used to fit the model is connected with an Italian microcredit initiative which took place between 2011 and 2017 and was promoted by an Italian firm called MXIT. The initiative targeted firms affected by the earthquake which struck Italy in 2009. The dataset consists of 21 variables and around 1000 units. Variables are divided into groups and put in a causal/logical order. Every box, shown in Figure 1, contains a different group of variables whereas arrows represent the direction of causal/logical relations between groups. The proposed configuration originates from interviews with MXIT and schematises previous knowledge on the subject. The structure summed up in Figure 1, will help defining the skeleton of the model, since it will be assumed that each variable can potentially affect only variables contained in its group, or in groups situated to the right of the considered variable group.

Demographic variables	Firm variables	Financial variables	Rejection variables	Repayment variables	Performance variables
Age Gender Continent Artisan City dimension Firm age Family firm Italian region Earthquake Firm peculiarity	Economic sector Legal status	Guarantee Lending bank Loan term Loan instalment Interest rate	Rejected?	Loan age Paid in full?	Performance

Figure 1. *Logical groups of variables*

The first group of variables contains demographics and some basic characteristics of financed firms. The second group consists of specific firm-related variables, whereas the third group contains the main financial characteristics of the credit. Fourth and fifth contain variables which operate a distinction between rejected and granted credit applications, or between partially and fully repaid credits. The variable performance, contained in the last box, indicates if all the instalments were paid on time and is used as a measure of the performance of the loan.

3 Bayesian network

A BN consists of a directed acyclic graph, where nodes and directed edges respectively identify variables and relations between variables. Each variable must have a finite set of mutually exclusive states. When a directed edge points from variable A to variable B, A is called parent of B. The strength of these relations is described by conditional probability tables assigned to each variable given its parents. Nodes without parents, are instead associated with marginal probability tables (Kjærulff and Madsen, 2013). Probability tables are a crucial element of a BN and can be either computed through a dataset or derived from experts of the analysed phenomenon. The structure of the BN is learnt through the necessary path condition algorithm (Kjærulff and Madsen, 2013). This algorithm first discovers the undirected graph performing conditional independence tests between variables, then assigns the direction to the unoriented edges. It also provides the user with the possibility of introducing further subject matter knowledge into the model, allowing to determine the presence and the direction of some arcs of the graph, if more than one solution is available.

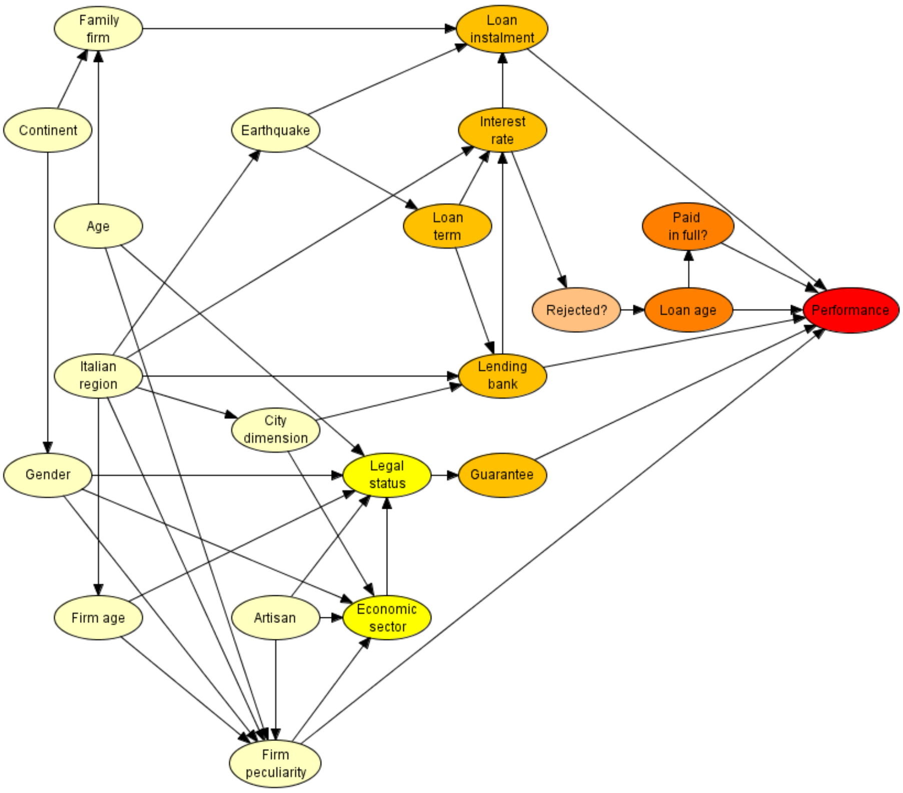


Figure 2. *Bayesian network built on the dataset*

The software Hugin has been used. The obtained BN (Figure 2) shows several connections between variables, which are coherent with the mechanics of the provision of microcredit. Every group of variables listed in the previous section is identified by nodes of a different colour. Nodes of the graph will be indicated with teletype font. Demographics variables are connected between them and with financial and firm-related variables, coherently with what expected. For example, `City dimension` affects `Economic sector` and `Italian region` is connected with `Lending bank`. Each variable of the network shows a direct or indirect connection with `Performance`. This could suggest that every variable plays a role in determining the probability of paying all the loan instalments on time.

Once the network has been estimated, the associated inference engine enables the users to efficiently make inference on the probability tables. The network can thus be interrogated, and *what-if* analysis can be carried out: different scenarios are simulated and their impact on the target variable is evaluated in terms of change in probability tables. The algorithms used for propagating evidence and updating the marginal probability tables are based on the junction tree (Kjærulff and Madsen, 2013). The efficiency of the propagation algorithms with the easy-to-read graphical representation of the relations among the variables, are the main reasons why BNs are increasingly used as a tool to support decision under uncertainty. In the obtained model for example, if we enter evidence about a firm hit by the earthquake and run by an adult male, after propagating the evidence throughout the network, we find a probability of 57.65% associated to the scenario where all the instalments are paid on time. If we consider the same situation, with a female entrepreneur, the same probability drops to 52.91%. The observed decrease could be due to the gender inequality that female entrepreneurs still face when running a business. *What-if* analysis can also be performed backwards, for instance by entering evidence about a specific state of `performance`, to find out the most probable profile of entrepreneur associated to that state.

4 Conclusion

BNs seem to provide a clear picture of how all the selected variables interact in the provision of microcredit. *What-if* analysis allows to study the strength and the effects of these interactions, in order to assess the risk connected to a specific microcredit provision. On the other hand, it allows to analyse which kind of microcredit provision is more suitable to a particular beneficiary, to promote a healthy development of the instrument. Existing BN could be also enlarged with additional modules accounting for new variables that are important in the light of the fast evolution of microcredit. For example, it will be very interesting to study how supplementary services affect performance results and which kind of service works best in a certain situation. The flexibility of the model, its clear graphical interface, the possibility to take into account subject matter knowledge and its efficient inference engine could make BN an appropriate model to analyse and improve the performance of entrepreneurial microcredit in the future.

References

BORGOMEO, C. & CO. 2016. *10 anni di microcredito, Le principali esperienze in Italia dal 2005 al 2014*. Rubbettino Editore.

KJÆRULFF U. B., & MADSEN A. L. 2013. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. New York: Springer.

THE PARAFAC MODEL IN THE MAXIMUM LIKELIHOOD APPROACH

Paolo Giordani[1], Roberto Rocci[2] and Giuseppe Bove[3]

[1] Dipartimento di Scienze Statistiche, Sapienza Università di Roma,
(e-mail: paolo.giordani@uniroma1.it)

[2] Dipartimento di Economia e Finanza, Università degli Studi di Roma "Tor Vergata",
(e-mail: roberto.rocci@uniroma2.it)

[3] Dipartimento di Scienze della Formazione, Università degli Studi Roma Tre,
(e-mail: giuseppe.bove@uniroma3.it)

ABSTRACT: Factor analysis is a well-known model for describing the covariance structure among a set of manifest variables through a limited number of unobserved factors. When the observed variables are collected at various occasions on the same statistical units, the data have a three-way structure and standard factor analysis may fail to discover the interrelations among the variables. To overcome these limitations, three-way models can be adopted. Among them, the so-called Parallel Factor (Parafac) model can be applied. In this article, the structural version of such a model, i.e. as a reparameterization of the covariance matrix, is studied by discussing under what conditions factor uniqueness is preserved.

KEYWORDS: three-way factor analysis, maximum likelihood, factor uniqueness property.

1 Introduction

Factor analysis (FA) (Bartholomew et al., 2011) is a well-known method explaining the relationships among a set of manifest variables, observed on a sample of statistical units, in terms of a limited number of latent variables. In FA data are stored in a matrix, say X, of order ($I \times J$) being I and J the number of statistical units and variables, respectively. Thus, FA deals with two-way two-mode data, where the modes are the entities of the data matrix, i.e., statistical units and manifest variables, and the ways are the indexes of the elements of **X**, i.e., $i = 1, \ldots, I$ and $j = 1, \ldots, J$. In many practical situations, it may occur that the scores on the same manifest variables with respect to a sample of statistical units are replicated across K different occasions, e.g. time, locations, conditions etc. In this case, there are three sets of entities (statistical units, manifest variables and occasions), hence three modes and the available information is stored in the so-called array, or tensor, usually denoted by $\underline{\mathbf{X}}$ of order ($I \times J \times K$). Its generic element is x_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$, expressing the score of statistical unit i on manifest variable j at occasion k.

Therefore, the elements have three indexes and the array three ways. For all of these reasons data are three-way three-mode (see, e.g., Kroonenberg, 2008).

The basic FA model is not adequate to handle three-way three-mode data. It has been extended in order to take into account and exploit the increasing complexity of three-way three-mode data. The most famous three-way three-mode extensions of FA are the Tucker3 (Tucker, 1966) and Parafac (Harshman, 1970) models, where the latter can be seen as a particular case of the former with a useful property of parameter uniqueness (Kruskal, 1977). Such extensions were born as suitable generalizations of Principal Component Analysis (PCA) and are mainly devoted to fit the model to the data according to a certain criterion. Some authors revised these proposals as structural models for the covariance structure of the manifest variables (e.g., Bentler et al., 1988). In this paper, after recalling the main features of the Parafac model following the above-mentioned two approaches, a structural extension of Parafac is considered and its uniqueness property is analysed when some specific factors are correlated across occasions, or variables.

2 The Parafac model

The Parafac model (Harshman, 1970) summarizes the three-way three-mode tensor $\underline{\mathbf{X}}$ by looking for a limited number of components for the modes. Let $\mathbf{X}_A$ be the matrix of order ($I \times JK$) obtained by juxtaposing next to each other the frontal slabs of $\underline{\mathbf{X}}$, i.e. the standard two-way two-mode matrices $\mathbf{X}_k$ ($k = 1, \ldots, K$) of order ($I \times J$) collected at the different occasions. The Parafac model can be formulated as

$$\mathbf{X}_A = \mathbf{A}(\mathbf{C}\bullet\mathbf{B})' + \mathbf{E}_A, \tag{1}$$

where the symbol '•' denotes the Khatri-Rao product of matrices, i.e., it is $\mathbf{C}\bullet\mathbf{B} = [\mathbf{c}_1\otimes\mathbf{b}_1, \ldots, \mathbf{c}_S\otimes\mathbf{b}_S]$, where $\mathbf{b}_s$ and $\mathbf{c}_s$ are the s-th columns of **B** and **C**, respectively ($s = 1, \ldots, S$), being S the number of components for the modes, and the symbol '⊗' denotes the Kronecker product of matrices. The matrices **A**, **B**, **C** have order ($I \times S$), ($J \times S$), ($K \times S$), respectively, and give the scores of the entities of the various modes on the components. Like Principal Component Analysis, the parameter estimates are found in the ordinary least squares (OLS) sense by minimizing the sum of squares of the error term $\mathbf{E}_A$. For this purpose, alternating least squares (ALS) algorithms can be applied.

The most interesting feature of Parafac is that under mild conditions the factors are essentially unique. This point has been deeply investigated by Kruskal (1977), who has found the following result. Let us denote by k-rank(**Z**) the so-called k-rank of a matrix **Z**. It is defined as the largest number k such that every subset of k columns of **Z** is linearly independent. Moreover, let (**A**, **B**, **C**) and ($\mathbf{A}_T$, $\mathbf{B}_T$, $\mathbf{C}_T$) be two optimal Parafac solutions. Kruskal (1977) has shown that if

$$k\text{-rank}(\mathbf{A}) + k\text{-rank}(\mathbf{B}) + k\text{-rank}(\mathbf{C}) \geq 2S + 2 \tag{2}$$

then, by considering (1),

$$\mathbf{A}(\mathbf{C}\bullet\mathbf{B})' = \mathbf{A}_{\mathrm{T}}(\mathbf{C}_{\mathrm{T}}\bullet\mathbf{B}_{\mathrm{T}})' \tag{3}$$

implies that there exists a permutation matrix $\mathbf{P}$ and three diagonal matrices $\mathbf{D}_{\mathrm{A}}$, $\mathbf{D}_{\mathrm{B}}$ and $\mathbf{D}_{\mathrm{C}}$, for which $\mathbf{D}_{\mathrm{A}}\mathbf{D}_{\mathrm{B}}\mathbf{D}_{\mathrm{C}} = \mathbf{I}$, such that

$$\mathbf{A}_{\mathrm{T}} = \mathbf{APD}_{\mathrm{A}},\ \mathbf{B}_{\mathrm{T}} = \mathbf{BPD}_{\mathrm{B}},\ \mathbf{C}_{\mathrm{T}} = \mathbf{CPD}_{\mathrm{C}}. \tag{4}$$

Starting from the original formulation in (1) we can derive what is the corresponding covariance structure. We limit our attention to the i-th row of $\mathbf{X}_{\mathrm{A}}$, say $\mathbf{x}_{Ai}'$, pertaining to the i-th statistical unit. $\mathbf{x}_{Ai}'$ is the vector of length JK containing the scores of statistical unit i on the J manifest variables during the K occasions. By explicitly considering a vector of intercepts and rewriting the model in terms of column vectors, we get

$$\mathbf{x}_{Ai} = \boldsymbol{\mu} + (\mathbf{C}\bullet\mathbf{B})\mathbf{a}_i + \mathbf{e}_{Ai}. \tag{5}$$

As usual in standard FA, we assume that the common factors $\mathbf{a}_i$ and the specific factors $\mathbf{e}_{Ai}$ are random with $\mathrm{E}(\mathbf{a}_i) = \mathbf{0}$ and $\mathrm{E}(\mathbf{e}_{Ai}) = \mathbf{0}$, without loss of generality because of $\boldsymbol{\mu}$, and $\mathrm{E}(\mathbf{a}_i\mathbf{e}_{Ai}') = \mathbf{0}$. If $\mathrm{E}(\mathbf{a}_i\mathbf{a}_i') = \boldsymbol{\Phi}$ and $\mathrm{E}(\mathbf{e}_{Ai}\mathbf{e}_{Ai}') = \boldsymbol{\Psi}$ are positive definite, then the covariance matrix of $\mathbf{x}_{Ai}$ is given by

$$\boldsymbol{\Sigma} = \mathrm{E}[(\mathbf{x}_{Ai} - \boldsymbol{\mu})(\mathbf{x}_{Ai} - \boldsymbol{\mu})'] = (\mathbf{C}\bullet\mathbf{B})\boldsymbol{\Phi}(\mathbf{C}\bullet\mathbf{B})' + \boldsymbol{\Psi}. \tag{6}$$

The generic element of the matrix $\boldsymbol{\Sigma}$ (of order $JK \times JK$), $\sigma_{jk,j'k'}$, holds the covariance between manifest variable j at occasion k and manifest variable j' at occasion k' ($j, j' = 1, \ldots, J$; $k, k' = 1, \ldots, K$). Bearing in mind the standard FA model, it should be clear that the Parafac model is a constrained version of standard FA. If we set $\boldsymbol{\Lambda} = (\mathbf{C}\bullet\mathbf{B})$, then (6) coincides with the oblique FA model where $\boldsymbol{\Lambda}$ is the matrix of factor loadings having a particular form depending on the three-way three-mode structure of the data. Maximum likelihood theory is used for estimating the parameters of the structural Parafac model assuming that the vectors $\mathbf{x}_{Ai}$, $i = 1, \ldots, I$, are independent and identically distributed as a multivariate normal.

3 Results

In this work we analyzed whether the constraints $\boldsymbol{\Lambda} = (\mathbf{C}\bullet\mathbf{B})$ affect the parameter identifiability under different covariance structure of the specific factors. In particular, we proved that the Parafac model in the structural formulation maintains the uniqueness property when, as in the standard FA model, the specific factors are assumed to be uncorrelated, i.e. the matrix $\boldsymbol{\Psi}$ is diagonal, and when the specific factors of the different variables are correlated within the same occasion, i.e. the matrix $\boldsymbol{\Psi}$ is block-diagonal, i.e.,

$$\mathbf{\Psi} = \mathrm{diag}(\mathbf{\Psi}_{11}, \ldots, \mathbf{\Psi}_{kk}, \ldots, \mathbf{\Psi}_{KK}), \tag{7}$$

where $\mathbf{\Psi}_{kk}$ denotes the covariance matrix of order ($J \times J$) for the specific factors at occasion k, $k = 1, \ldots, K$. The Parafac covariance model in (6) with the correlation structure of the specific factors given in (7) represents a more realistic model able to fit reasonably well in many practical three-way three-mode studies. It is important to note that what follows can be extended to the case where the specific factors of the same variable are correlated across the different occasions. Such an extension can be easily obtained by exploiting the symmetry of the model with respect to variables and occasions. When $\mathbf{\Psi}$ is diagonal, the proof is based on Theorem 5.1 of Anderson & Rubin (1956). When $\mathbf{\Psi}$ is block-diagonal, the conditions of Anderson & Rubin (1956) cannot be longer applied. To prove the uniqueness, the results of Browne (1980), formulated in the context of the FA model for multiple batteries of tests, is considered. For further details, see Giordani et al. (2019). During the meeting, we show how the factor uniqueness property hold in the above described cases. Moreover, we illustrate the effectiveness of the proposal by means of a real-life example in the multitrait-multimethod analysis framework.

References

ANDERSON, T.W., & RUBIN, H. 1956. Statistical inference in factor analysis, in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry, University of California, California, 111-150.

BARTHOLOMEW, D.J., KNOTT, M., & MOUSTAKI, I. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*, 3rd Edition, Chichester: Wiley.

BENTLER, P.M., POON, W.-Y., & LEE, S.-Y. 1988. Generalized multimode latent variable models: implementation by standard programs. *Computational Statistics & Data Analysis*., **6**, 107-118.

BROWNE, M.W. 1980. Factor analysis of multiple batteries by maximum likelihood. *British Journal of Mathematical and Statistical Psychology*., **33**, 184-199.

HARSHMAN, R.A. 1970. Foundations of the PARAFAC procedure: models and conditions for an ‘explanatory’ multi-model factor analysis. *UCLA Working papers in Phonetics*., **16**, 1-84.

GIORDANI, P., ROCCI, R., & BOVE, G. 2019. Uniqueness of the structural Parafac model. Submitted.

KROONENBERG, P.M. 2008. *Applied Multiway Data Analysis*. Hoboken: Wiley.

KRUSKAL, J.B. 1977. Three-way arrays: rank and uniqueness of trilinear decomposition, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*., **18**, 95-138.

TUCKER, L.R 1966, Some mathematical notes on three-mode factor analysis, *Psychometrika*, 31, 279-311.

STRUCTURE DISCOVERING IN NONPARAMETRIC REGRESSION BY THE GRID PROCEDURE

Francesco Giordano[1], Soumendra Nath Lahiri[2] and Maria Lucia Parrella[1]

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

[2] Department of Statistics, Noth Carolina State University, (e-mail: snlahiri@ncsu.edu)

ABSTRACT: A method for variable selection and structure discovery in the context of nonparametric regression in high dimensions is proposed in a forthcoming paper, where a small subset of variables are relevant and may have nonlinear effects on the response. The proposed method, called the GRID, is an extension of the RODEO method of Lafferty & Wasserman, 2008 (which only makes variable selection). In this paper we briefly describe the method and present the main theoretical foundations of the two stages of the procedure: (i) variable selection with linear/nonlinear classification of the covariates and (ii) identification of interactions.

KEYWORDS: Variable selection, nonparametric regression, high dimension.

1 The GRID method

In this paper we describe a new method, called the *GRID* method, for simultaneous variable selection, classification of the relevant covariates between linear and nonlinear, and estimation of the low-dimensional structure of the regression function. This method is an extension of the RODEO method proposed by Lafferty & Wasserman, 2008 and it is proposed and deeply investigated in a forthcoming paper by Giordano *et al.*, 2019. To briefly describe the methodology, consider the nonparametric regression model

$$Y_t = m(X_t) + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{1}$$

where the X_t represents the $\mathbb{R}^d$-valued covariates and the errors ε_t are *iid* with zero mean and variance σ^2. The errors ε_t are independent of X_t, and are assumed to be Gaussian, as in Lafferty & Wasserman, 2008. Here $m(X_t) = E(Y_t|X_t) : \mathbb{R}^d \to \mathbb{R}$ is the multivariate conditional mean function. We use the notation $X_t = (X_{t1}, \ldots, X_{td})$ to refer to the covariates. We assume that the number of covariates $d \to \infty$ but only r of these covariates are relevant for model (1), where $r \ll d$ is considered bounded or unbounded.

The acronym GRID derives from *Gradient Relevant Identification of Derivatives*, meaning that the procedure is based on testing the significance of partial derivative estimators (derived by the Local Linear Estimation methodology). We now illustrate the idea behind the GRID procedure with the following example: let $d = 10$ and let the true model be given by

$$Y_t = 2X_{t1} + X_{t2}^2 X_{t3} + 10X_{t4}X_{t5}X_{t6} + \exp(X_{t7})X_{t2} + \varepsilon_t, \quad t = 1, \ldots, n. \qquad (2)$$

The first stage of the GRID procedure identifies (the indices of) the following sets of covariates (variable selection and classification).

$$C = \{2,7\}, \quad A = \{1,3,4,5,6\}, \quad U = \{8,9,10\}.$$

The selected variables are automatically classified by the procedure as linear (denoted by the set A) and nonlinear (denoted by the set C). The other ones constitute the set U of irrelevant variables.

The second stage of the GRID procedure derives (the indices of) the following sets of interactions

$$I^1 = \{1\},\ I^2 = \{2,3,7\},\ I^3 = \{3,2\},\ I^4 = \{4,5,6\},\ I^5 = \{5,4,6\},$$
$$I^6 = \{6,4,5\},\ I^7 = \{7,2\},$$

where I^j includes the interactions of variable j with other covariates. By default, each set I^j automatically includes the index j (self-interaction). Therefore, if the set I^j has the only component j, then X_j appears in the model as an isolated additive covariate, like X_1 in model (2).

2 Theoretical basis for the two stages of the GRID algorithm

Local linear estimation (LLE) is a nonparametric method for estimating the regression function $m(\cdot)$ in (1) (cf. Ruppert & Wand, 1994). To estimate $m(\cdot)$ at $x = (x_1, \ldots, x_d)$, the LLE performs a locally weighted least squares fit of a linear function. Let

$$\hat{\beta}(x;H) \equiv \arg\min_{\beta_0,\beta_1} \sum_{t=1}^{n} \left\{Y_t - \beta_0 - \beta_1^T (X_t - x)\right\}^2 K_H(X_t - x), \qquad (1)$$

where the function $K_H(u) = |H|^{-1} K(H^{-1}u)$ gives the local weights with a d-variate product Kernel function $K(u) = \prod_{j=1}^{d} K_1(u_j)$. The bandwidth matrix H controls the bias and the variance of the resulting LLE of $m(x)$. For simplicity, we shall suppose that $H = diag(h_1,\dots,h_d)$ is a diagonal matrix with strictly positive entries. The estimator $\hat{\beta}(x)$ can be written in a closed form as:

$$\hat{\beta}(x;H) = (\Gamma^T W\Gamma)^{-1}\Gamma^T W\Upsilon, \tag{2}$$

where $\Upsilon = (Y_1,\dots,Y_n)^T$ and

$$\Gamma = \begin{pmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{pmatrix}, \quad W = \begin{pmatrix} K_H(X_1 - x) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & K_H(X_n - x) \end{pmatrix}.$$

Note that $\hat{\beta}(x;H)$ gives estimators of the function $m(x)$ and its gradient:

$$\hat{\beta}(x;H) = \begin{pmatrix} \hat{\beta}_0(x;H) \\ \hat{\beta}_1(x;H) \end{pmatrix} \equiv \begin{pmatrix} \hat{m}(x;H) \\ \hat{\mathbb{D}}(x;H) \end{pmatrix}. \tag{3}$$

The theoretical foundations of the GRID procedure are based on the following assumptions and the theorem below.

A1) The bandwidth H is a diagonal matrix with strictly positive diagonal entries: $H = diag(h_1,\dots,h_d)$, with $c_1 \leq h_j$ for $j = 1,\dots,d$ for some $c_1 \in (0,\infty)$.
A2) The d-variate kernel function K is a product kernel, based on a nonnegative and symmetric univariate kernel density function $K_1 \in \mathcal{C}^1[-c_2,c_2]$ for some $c_2 > 0$ such that $0 < x_j - c_2 h_j < x_j + c_2 h_j < 1$ for all $j = 1,\dots,d$.
A3) All the partial derivatives of the function $m(x)$ up to and including order five are bounded.
A4) X_1 is uniformly distributed on the unit cube $(0,1)^d$.

Theorem 1 *Under model (1) and assumptions A1-A4, we have:*

$$E\left\{ \frac{\partial \hat{m}(x;H)}{\partial h_j} \middle| \chi_n \right\} = \begin{cases} \theta_{0j} \neq 0 & \textit{if } j \in C \\ 0 & \textit{otherwise} \end{cases} + o_p(1) \tag{4}$$

$$E\left\{ \frac{\partial \hat{\mathbb{D}}^{(i)}(x;H)}{\partial h_j}, i \neq j \middle| \chi_n \right\} = \begin{cases} \theta_{ij} \neq 0 & \textit{if } i \in I^j, j \in C \\ 0 & \textit{otherwise} \end{cases} + o_p(1) \tag{5}$$

for $i, j = 1,\dots,d$ and $i \neq j$, where $\chi_n = \{X_t : t = 1\dots,n\}$ and the exact expressions for θ_{0j} and θ_{ij} can be derived following Giordano et al.*, 2019.*

Note that Theorem 1 can be used to identify the relevant (nonlinear!) covariates and interactions, by using a proper threshold technique on θ_{0j} and θ_{ij},

		$d = 20$			$d = n/2$			$d = 2n$		
	n	R	C	$I(6,7)$	R	C	$I(6,7)$	R	C	$I(6,7)$
X_6	300	0.975	0.330	0.900	0.855	0.335	0.720	0.630	0.330	0.365
	500	1.000	0.610	1.000	0.990	0.595	0.985	0.810	0.480	0.620
	1000	1.000	0.910	1.000	1.000	0.915	1.000	0.910	0.835	0.815
X_7	300	0.940	0.325	0.900	0.875	0.335	0.720	0.580	0.250	0.365
	500	1.000	0.370	1.000	0.995	0.635	0.985	0.765	0.515	0.620
	1000	1.000	0.935	1.000	1.000	0.890	1.000	0.835	0.815	0.815
X_{10}	300	1.000	*	-	1.000	*	-	0.995	0.035	-
	500	1.000	*	-	1.000	*	-	1.000	*	-
	1000	1.000	*	-	1.000	*	-	1.000	*	-

Table 1. *Simulation results for different dimensions d and sample sizes n. The values show the proportion of times that a given covariate X_i is classified as a relevant covariate (R), as a nonlinear covariate (C), and as part of an interaction term (I). The symbol $(*)$ denotes a value ≤ 0.025 while the symbol $(-)$ means zero.*

as suggested in Lafferty & Wasserman, 2008 and Giordano *et al.*, 2019. However, this theorem cannot be used to identify the linear covariates. To overcome this, we consider an auxiliary regression where all those covariates that have not been selected in the first pass, are to be transformed, so that the *linear covariates* of the original model become *nonlinear* in the auxiliary model.

3 Some simulation results

The Monte Carlo simulation is based on 200 iterations. The covariates are uniformly distributed. We consider the model $Y_t = m(X_t) + \varepsilon_t$ with $m(x) = x_6^3 x_7^3 + x_{10}$ and $\varepsilon_t \sim N(0,1)$ for all t. The additive components of the model are standardized so that they all have variance equal to one, to make them comparable each other. The Kernel function is $K_1(u) = 1/C_1 \left(5 - u^2\right) \mathbb{I}_{\{|u| \leq \sqrt{5}\}}$, as in Lafferty & Wasserman, 2008, where C_1 is a scale factor to make the integral equal one. The simulation results are shown in Table 1.

References

GIORDANO, F., LAHIRI, S.N., & PARRELLA, M.L. 2019. GRID: A variable selection and structure discovery method for high dimensional nonparametric regression. *To appear on The Annals of Statistics.*

LAFFERTY, J., & WASSERMAN, L. 2008. RODEO: sparse, greedy nonparametric regression. *The Annals of Statistics.*, **36**, 28–63.

RUPPERT, D., & WAND, P. 1994. Multivariate locally weighted least squares regression. *The Annals of Statistics.*, **22**, 1346–1370.

A MICROBLOG AUXILIARY PART-OF-SPEECH TAGGER BASED ON BAYESIAN NETWORKS

Silvia Golia[1] and Paola Zola[1]

[1] Department of Economics and Management, University of Brescia,
(e-mail: silvia.golia@unibs.it, paola.zola@unibs.it)

ABSTRACT: Part-of-speech (POS) tagging is the basis of many Natural Language Processing tasks and, nowadays, there exist several algorithms able to determine the POS tag for a specific word. However, the increasing usage of Internet and the explosion of blogs and microblogs changed the way people communicate, and POS taggers trained on structured corpora lost the ability to catch this new tendency. The proposed algorithm is an auxiliary POS tagger which aims at predicting unknown POS tags. It is based on the Bayesian Networks and it uses information regarding POS tags that precede and follow the unknown POS tag. The well-known Brown Corpus and the more recent Ark dataset are the datasets over which the proposed methodology is tested.

KEYWORDS: microblogs, part-of-speech tagger, bayesian networks.

1 Introduction

Part-of-speech (POS) tagging is the basis of many Natural Language Processing (NLP) tasks and there are several algorithms able to determine the POS tag for a specific word. However, the increasing usage of Internet and the explosion of blogs and microblogs changed the way people communicate, involving an increasing usage of slang, abbreviations, symbols and emoticons, creating the so called cyber-slang. To extract and analyze such novel information from Web2.0 is a pretty new challenge for many NLP tasks, as POS taggers. In fact, traditional POS taggers, trained on structured corpora, lost their ability when applied to blog and microblog data. POS tagging's earlier works were mainly based on grammar rules and morphemes. Thanks to the progress in computational technology and the growing interest in machine learning models, new research has been done focusing, for example, on Markov Models and their variants (Cutting *et al.*, 1992) and deep learning algorithms (Plank *et al.*, 2016). Some research tried to extend traditional POS tagger to blogs and microblogs data, obtaining poor performances, as shown for example in Nand and Perera (2015). Following the limitations of existing POS taggers for blogs and microblogs, this paper wants to propose a novel approach to assign a POS tag to unknown words based on the information deriving from the POS tag sequence. The proposed method can be interpreted as an auxiliary POS tagger because it intervenes after the initial POS tagging step of the corpus, predicting the remaining unknown POS tags that, for any reasons, do not match any vocabulary. It uses a Bayesian Network as predictor of the probability distribution of the unknown POS tag.

2 The proposed approach and preliminary results

In order to predict the unknown POS tag, the proposed approach needs to identify a suitable Bayesian Network (BN). The BN is a model that explicits, through a Directed Acyclic Graph (DAG), a set of (conditional) dependence and independence properties among the variables under study (Kjaerulff & Madsen, 2008). A DAG G is composed by a set of nodes V, which corresponds to a set of random variables X_V indexed by V, and a set E of directed links between pairs of nodes in V. A BN is composed by the pair $(G, \mathcal{P})$, where G is a DAG and $\mathcal{P}$ is the set of conditional probabilities involved in the factorization, according to G, of the joint probability distribution $P(X_V) = \prod_{\mathrm{v} \in V} P(X_\mathrm{v} | X_{pa(\mathrm{v})})$, where $X_{pa(\mathrm{v})}$ denotes the set of parent variables of the variable X_v for each node $\mathrm{v} \in V$. A BN can be used, for example, to compute the effect of a new piece of information on one or more target variables, computing the corresponding posterior distribution (Koller & Friedman, 2009). In order to construct a BN, firstly the DAG is identified, then the joint probability distribution is computed, estimating the set of conditional probability distributions $P(X_\mathrm{v} | X_{pa(\mathrm{v})})$. Several algorithms have been proposed to automatically find the structure of a BN. In this paper two score-based algorithms, including Hill Climbing (HC) and Tabu search, were evaluated, considering the Bayesian Dirichlet equivalent uniform score (BDE) and the Bayesian Information criterion (BIC) as possible scores.

In order to predict the unknown POS tag, denoted by tag_t, it is necessary to identify the most suitable length of the tag sequence and to extract the predicted attribute from the probability distribution given by the estimated BN. Most of the previous works only rely on the information linked to the two preceding POS tags, however, in the subsequent analysis three possible sets of information were investigated:
$Tag_{t-/+1} = \{tag_{t-1}, tag_t, tag_{t+1}\}$,
$Tag_{t-/+2} = \{tag_{t-2}, tag_{t-1}, tag_t, tag_{t+1}, tag_{t+2}\}$,
$Tag_{t-/+3} = \{tag_{t-3}, tag_{t-2}, tag_{t-1}, tag_t, tag_{t+1}, tag_{t+2}, tag_{t+3}\}$.

$t-i$ $(t+i)$ indicates the position of the tag that precedes (follows) tag_t.

Regarding the way in which the predicted probability distribution of the unknown POS tag can be summarized in a predicted POS tag, three criteria were evaluated, including the *Mode criterion*, the *Max. Dist. criterion*, which consists in computing the difference between the predicted POS tag probabilities and the corresponding sample frequencies and choosing the attribute corresponding to the maximum difference, and the *Hybrid criterion*, which uses the Max. Dist. criterion when each frequency associated with the modal attribute is less than 0.5, and the Modal criterion otherwise.

The identification and estimation of a suitable BN was performed with *bnlearn* R package (Scutari, 2010), making use of the Brown Corpus, which is a classical and widely used POS tagged dataset. Its predictive performance was evaluated through the following metrics: Area Under the Curve (AUC) of the Receiver Operating Characteristic curve, average accuracy (Av Acc), macro precision (M Prec), macro-averaging F1-score (MAF1) and overall accuracy (Acc) (Sokolova & Lapalme, 2009, Witten *et al.* 2016). A 10-fold cross-validation procedure was applied to select the best information set combination, predictive criterion and BN, resulting in the following choice: $Tag_{t-/+3}$ set of variables, Max. Dist. criterion, and BN obtained applying the HC algorithm with BDE score (iss=5000). Figure 1 shows the chosen BN.

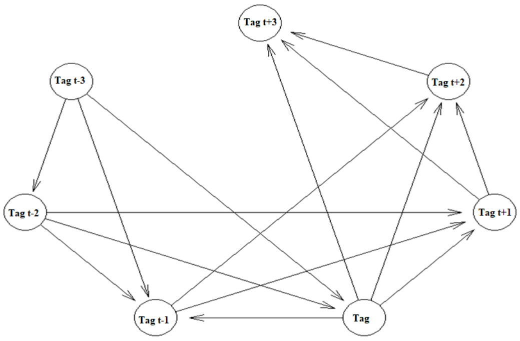


Figure 1. *The chosen BN*

The predictive performances of the best BN in estimating the unknown POS tags were evaluated on the Brown Corpus and the Ark Dataset, which comprises POS tagged Twitter messages. Moreover, a domain adaptation analysis was performed, consisting in using the Brown Corpus as a training domain, and the ARK dataset as a target domain. The choice to perform the domain adaptation analysis is due to the fact that few and relatively small labeled datasets for Twitter and Web 2.0 data are available. Table 1 reports the results in terms of evaluation metrics for the Brown Corpus, the ARK Dataset and the domain adaptation case obtained by computing the metrics on each fold of a 10-fold cross validation procedure, and then averaging the 10 outcomes. The results obtained on the Brown Corpus are overall better than the ones on the ARK dataset. Comparing the performances of the models that used only the ARK dataset (ARK → ARK) with respect to the cross domain setting (Brown → ARK), one notices a slight decrease of the evaluation metrics.

Table 1. *Evaluation metrics*

	AUC	**Av Acc**	**MAF1**	**M Prec**	**Acc**
Brown → Brown	0.731	0.444	0.533	0.344	0.624
ARK → ARK	0.629	0.355	0.364	0.259	0.474
Brown→ ARK	0.613	0.337	0.318	0.242	0.399

References

CUTTING D. & KUPIEC J. & PEDERSEN J. & SIBUN P. 1992. A practical part-of-speech tagger. In: *ANLC '92 Proceedings of the third conference on Applied natural language processing*, 133–140.

KJÆRULFF U.B. & MADSEN A.L. 2013. *Bayesian networks and influence diagrams: a guide to construction and analysis*. Springer, New York.

KOLLER D. & FRIEDMAN N. 2009. *Probabilistic graphical models: principles and techniques*. MIT Press.

NAND P. & PERERA R. 2015. An Evaluation of POS tagging for Tweets Using HMM Modeling. In: *Proceedings of the 38th Australasian Computer Science Conference (ACSC 2015)*, 83–89.

PLANK B. & SØGAARD A. & GOLDBERG Y. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 412–418.

SCUTARI M. 2010. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, **35(3)**, 1–22.

SOKOLOVA M. & LAPALME G. 2009. A systematic analysis of performance measures for classification tasks, *Information Processing & Management*, **45 (4)**, 427–437.

WITTEN I.H. & FRANK E & HALL M.A. & PAL C.J. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

RECENT ADVANCES IN MODEL-BASED CLUSTERING OF HIGH DIMENSIONAL DATA

Isobel Claire Gormley[1]

[1] School of Mathematics and Statistics, Insight Centre for Data Analytics,
University College Dublin, (e-mail: claire.gormley@ucd.ie)

ABSTRACT:

The model-based clustering framework provides well established methods that uncover sub-groups of observations in data. Such methods bestow several desirable benefits: reproducibility due to their statistical modelling basis, objectivity through the availability of principled model selection tools and interpretability through the provision of parameter estimates and their associated uncertainties.

However, model-based clustering approaches begin to lose traction as data dimension increases, whether in terms of number of observations, variables, timepoints etc. This loss of applicability is often due to stability issues associated with high dimensional covariance matrices, optimisation difficulties and/or the expensive nature of computing the likelihood function.

Here we consider recent advances in model-based methods to clustering data where the number of variables p is large. In particular, we explore developments in factor analytic approaches, which are well known models for big p data, and recent work utilising composite likelihood methods that facilitate computation of intractable likelihood functions. The utility of such methods is illustrated through benchmark and real data sets.

KEYWORDS: high dimensional data, factor analytic models, composite likelihood.

References

MURPHY, K., VIROLI, C., & GORMLEY, I.C. 2019. Infinite Mixtures of Infinite Factor Analysers. *Bayesian Analysis*, To appear.

TREE EMBEDDED LINEAR MIXED MODELS

Anna Gottard[1], Leonardo Grilli[1], Carla Rampichini[1] and Giulia Vannucci[1]

[1] Department of Statistics, Computer Science, Applications "G. Parenti",
University of Florence (e-mail: anna.gottard@unifi.it,
leonardo.grilli@unifi.it, carla.rampichini@unifi.it,
giulia.vannucci@unifi.it)

ABSTRACT: This work gives a contribution to the emerging literature on the use of regression trees for hierarchical data to increase the flexibility and the predictive ability of random effects models. The proposed procedure extends random effect regression trees considering a random effect model with both a tree component and a linear component. Moreover, it is suggested to decompose the effects of predictors within and between clusters. The performance of the proposed procedure is evaluated through a simulation study and an application to INVALSI data on students achievement.

KEYWORDS: CART, hierarchical data, random effects.

1 Introduction

Mixed or multilevel models (Snijders & Bosker, 2012) are useful tools to deal with hierarchical data. In general, hierarchical data are composed by level 1 units nested into level 2 units (clusters), such students within schools (individual cross-sectional data) or children growth evaluated at several time points (repeated measures). Model specification is a challenging task in mixed models. A worthwhile approach exploits regression trees (Breiman *et al.*, 1984) to capture nonlinear fixed effects. This technique has been extended to clustered data by modelling fixed effects with a decision tree, while accounting for random effects with a linear mixed model in a separate step (Hajjem & Larocque, 2011; Sela & Simonoff, 2012). It is shown that random effect regression trees are less sensitive to parametric assumptions and provide improved predictive power compared to linear models with random effects and regression trees without random effects. The literature has grown with variants and extensions (e.g. Hajjem & Larocque, 2014; Miller & Lubke, 2017).

Our proposal extends random effect regression trees in two directions: (i) incorporating a linear component in the final random effect model, and (ii) allowing a decomposition of the effect of a given predictor within and between clusters.

2 A tree embedded linear mixed model

To take into account both non-linear and interaction effects and cluster mean dependencies, we are proposing here a random effect model, called *Tree embedded linear mixed model*, where the regression function is piecewise-linear, consisting in the sum of a tree component and a mixed effect linear component. The proposal is the mixed effect version of the semi-linear regression trees (Vannucci, 2019; Vannucci & Gottard, 2019).

The prosed model can be ideally divided into three parts: a fixed effect linear part, a fixed effect non-linear part based on a tree and a random effect part. In this work, we limit our attention to the case of random intercept mixed models, but the extension to random slopes is straightforward. The resulting model can be formulated as

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\gamma} + T(\mathbf{X}_{ij}, \mathbf{Z}_j) + U_j + \varepsilon_{ij} \tag{1}$$

where Y_{ij} is the response variable for level 1 unit i belonging to level 2 unit j, $\mathbf{X}_{ij}$ is the vector of the level 1 predictors, $\boldsymbol{\beta}$ the associated fixed effect coefficients, $\mathbf{Z}_j$ is the vector of the level 2 predictors, $\boldsymbol{\gamma}$ the associated fixed effect coefficients. Then, $T(\mathbf{X}_{ij}, \mathbf{Z}_j)$ is the tree based predictor depending on some or all the level 1 and the level 2 explanatory variables. Finally, $U_j \sim N(0, \sigma_u^2)$ is the random intercept for level 2 unit j and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$.

The model is additive in its components where the tree-component acts as a region-specific intercept. As an alternative, the model can be written as

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\gamma} + \sum_{m=1}^{M} \mu_m \mathbb{I}\{(\mathbf{X}_{ij}, \mathbf{Z}_j) \in R_m\} + U_j + \varepsilon_{ij}, \tag{2}$$

where $R_1, \ldots, R_M$ is the partition of the predictor space corresponding to the tree-component. When the unknown regression function can be assumed to be quasi-linear (Wermuth & Cox, 1998), the number of leaf nodes M can be kept small to avoid overfitting.

To disentangle the within and between effects of an level 1 predictor, say W_{ij}, we decompose W_{ij} into the cluster mean $\overline{W}_j = (1/n_j)\sum_{i=1}^{n_j} W_{ij}$ and the deviation $\widetilde{W}_{ij} = W_{ij} - \overline{W}_j$. Then, we include $\overline{W}_j$ in $\mathbf{Z}_j$ and the deviation $\widetilde{W}_{ij}$ in $\mathbf{X}_{ij}$ (Snijders & Bosker, 2012).

Model fitting is obtained by the iterative procedure described in Algorithm 1. This procedure is based on the backfitting algorithm (Breiman & Friedman, 1985), and recently applied in semilinear regression trees (Vannucci, 2019; Vannucci & Gottard, 2019). The convergence of the algorithm is evaluated comparing the mean square error in two successive iterations. At the final step, model (2) is fitted using the partition associated to the tree selected at convergence. The leaf node parameters μ_m are estimated jointly with the other model parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ_u^2, σ_ε^2. Algorithm 1 is implemented in a user written R code.

Algorithm 1: Backfitting algorithm for tree embedded linear mixed models

```
Data: $(Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_j)$, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$
Result: Fitting of the tree embedded linear mixed model (2)
Initialization step: The tree is initialized at depth 0: $\widehat{T}(\mathbf{X}_{ij}, \mathbf{Z}_j) = \overline{Y}_{ij}$;
Iteration step: repeat
    Compute the tree-based residuals $Y_{ij}^{*} = Y_{ij} - \widehat{T}(\mathbf{X}_{ij}, \mathbf{Z}_j)$;
    Fit a linear random intercept model of $Y_{ij}^{*}$ on $\mathbf{X}_{ij}$ and $\mathbf{Z}_j$ and
      compute the predicted $\widehat{Y}_{ij}^{\mathrm{re}}$ (fixed part + $\widehat{U}_j$);
    Compute the model-based residuals: $Y_{ij}^{**} = Y_{ij} - \widehat{Y}_{ij}^{\mathrm{re}}$;
    Fit the regression tree of $Y_{ij}^{**}$ on $\mathbf{X}_{ij}$ and $\mathbf{Z}_j$ and compute the
      predicted values $\widehat{T}(\mathbf{X}_{ij}, \mathbf{Z}_j)$;
until convergence criteria is met;
Estimation step: Estimate the parameters of model (2) using the
  partition selected by the tree at convergence.
```

The main difference of our procedure with respect to previous proposals (Hajjem & Larocque, 2011; Sela & Simonoff, 2012), is the inclusion of the linear component $\mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_j\boldsymbol{\gamma}$ in the random effect model (2). This inclusion allows to avoid overfitting and helps interpretation. Moreover, since the μ_m are jointly estimated in the final step, standard hypothesis tests and confidence intervals can be used for model selection and evaluation, together with the mean squared error computed on a test data set for prediction accuracy evaluation.

We will show via a simulation study and an application to INVALSI data on students achievement that our proposal improves the predictive performance of the model in presence of quasi-linear relationships (Wermuth & Cox, 1998), avoiding overfitting and facilitating interpretation.

References

BREIMAN, L, & FRIEDMAN, JH. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, **80**, 580–598.

BREIMAN, L., FRIEDMAN, J., STONE, C.J., & OLSHEN, R.A. 1984. *Classification and regression trees*. CRC press.

HAJJEM, A., BELLAVANCE F., & LAROCQUE, D. 2011. Mixed effects regression trees for clustered data. *Statistics and Probability Letters*, 451459.

HAJJEM, A., BELLAVANCE F., & LAROCQUE, D. 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 1–18.

MILLER, P.J., MCARTOR D.B., & LUBKE, G.H. 2017. metboost: Exploratory regression analysis with hierarchically clustered data. *arXiv:1702.03994v1 [stat.ML]*.

SELA, R.J., & SIMONOFF, J.S. 2012. RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, 1–37.

SNIJDERS, T.A.B., & BOSKER, R.J. 2012. *Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.)*. SAGE Publications Inc.

VANNUCCI, G. 2019. Interpretable semilinear regression trees. FLORE, FLOrence REsearch repository.

VANNUCCI, G., & GOTTARD, A. 2019. Semilinear regression trees. submitted.

WERMUTH, N., & COX, D.R. 1998. On association models defined over independence graphs. *Bernoulli*, **4**(4), 477–495.

WEIGHTED LIKELIHOOD ESTIMATION OF MIXTURES

Luca Greco[1] and Claudio Agostinelli[2]

[1] DEMM Department, University of Sannio,(e-mail: luca.greco@unisannio.it)

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

ABSTRACT: This contribution deals with robust estimation of mixtures by developing a weighted likelihood methodology, which relies on a suitable modification of the EM (or Classification EM) algorithm. In the proposed algorithm, the likelihood equations in the M-step are replaced by weighted likelihood estimating equations, which are characterized by the presence of data dependent weights aimed at downweighting outliers. The weights are based on the Pearson residuals and the residual adjustment function. Formal rules for robust clustering and outlier detection can be defined based on the fitted mixture model. Mixtures of multivariate Gaussian components and regression models will be considered.

KEYWORDS: classification, EM, mixture, outliers, Pearson residuals.

1 Introduction

It is well known that maximum likelihood estimation (MLE) is likely to lead to unreliable results when the sample data are contaminated by the occurrence of outliers. In mixture modeling, in the presence of such data inadequacies, the bias of at least one of the component parameters estimate can be arbitrarily large and model based clustering strategies become unfeasible in recovering the true underlying grouping structure in the data at hand. Actually, the occurrence of outliers could lead to find spurious clusters and/or merge together genuine separate groups. The reader is pointed to the book by Farcomeni & Greco, 2015 for a gentle introduction to robustness issues with a particular emphasis on multivariate problems and cluster analysis.

Here, in order to take into account the possible presence of outliers, it is suggested to replace maximum likelihood by weighted likelihood estimation. Maximum likelihood estimation of mixture models is commonly obtained by resorting to the EM algorithm. An alternative strategy is given by the (penalized) Classification EM (CEM) algorithm. Weighted likelihood estimation of mixture models can be achieved by developing a modified version of the EM (or CEM) algorithm. Actually, in the M-step, the likelihood equations are replaced by a different set of estimating equations whose single term contributions are attached a weight aimed at downweighting outliers. In particular, weighted likelihood estimation is achieved by evaluating weights stemming from Pearson residuals (Markatou *et al.* , 1998). The Pearson residual gives a measure of the agreement between the assumed model $m(y;\tau)$ and the data, that are summarized by a non-parametric density estimate $\hat{m}_n(y) = n^{-1}\sum_{i=1}^{n} k(y;y_i,h)$, based on a kernel $k(y;t,h)$ indexed by a bandwidth h, that is

$$\delta(y) = \frac{\hat{m}_n(y)}{m(y;\tau)} - 1 \ , \tag{1}$$

with $\delta \in [-1,\infty)$. In regression and multivarate problems, the Pearson residuals can be evaluated as

$$\delta(y) = \frac{\hat{m}_n(g(y;\tau))}{m(y)} - 1 \ , \tag{2}$$

where $g(y;\tau)$ is an appropriate pivotal transformation: (standardized) residuals in regression (Agostinelli & Markatou, 1998, Alqallaf & Agostinelli, 2016) and Mahalanobis distances in multivariate estimation (Agostinelli & Greco, 2018). The weight function is defined as

$$w(\delta(y)) = \frac{[A(\delta(y))+1]^{+}}{\delta(y)+1} \ , \tag{3}$$

where $[\cdot]^{+}$ denotes the positive part and $A(\delta)$ is the Residual Adjustment Function (RAF, Basu & Lindsay, 1994). When the model is correctly specified, the Pearson residual function (1, 2) evaluated at the true parameter value converges almost surely to zero, whereas, otherwise, for each value of the parameters, large Pearson residuals detect regions where the observation is unlikely to occur under the assumed model. The RAF plays the role to bound the effect of large residuals on the fitting procedure, as well as the Huber and Tukey-bisquare function bound large distances in M-estimation and we assume is such that $|A(\delta)| < |\delta|$. One can consider the families of RAF stemming from the Symmetric Chi-Squared divergence, the family of Power divergence or Generalized Kullback-Leibler divergence measures. The resulting weight function (3) is unimodal and decline smoothly to zero as $\delta(y) \rightarrow -1$ or $\delta(y) \rightarrow \infty$. Hence, those observations lying in such regions are attached a weight that decreases with increasing Pearson residual. Large Pearson residuals and small weights will correspond to data points that are likely to be outliers.

2 Weighted EM and CEM

Let $y = (y_1, y_2, \ldots, y_n)^{\mathrm{T}}$ be a random sample of size n from the mixture model

$$m(y;\tau) = \sum_{k=1}^{K} \pi_k p(y_i;\theta_k) \ ,$$

where $\tau = (\pi_1, \ldots, pi_K, \theta_1, \ldots, \theta_K)$, θ_k denotes the vector of component specific parameters and K is the number of groups, that is assumed to be fixed in advance. The weighted EM (WEM) algorithm iteratively alternates between the standard E-step, in which posterior probabilities $u_{ik} \propto \pi_k p(y_i;\theta_k)$ are obtained, and a weighted M-step in which one solves the estimating equations

$$\sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij} \frac{\partial}{\partial \tau} \left[\log \pi_j + \log \phi_p(y_i; \mu_j, \Sigma_j) \right] w_{ij} = 0 \ , \tag{4}$$

where w_{ij} denotes the weight for the i-th unit with respect to the j-th component. In the weighted CEM algorithm (WCEM), after the E step, let $k_i = \text{argmax}_k u_{ik}$, then $u_{ik_i} = 1$ and $u_{ik} = 0$ for $k \neq k_i$. Therefore, in the modified M-step one is allowed to compute one single weight per unit, conditionally on the current cluster assignments, in equation (4), i.e $w_{ij} = w_{ik_i}$.

The WCEM automatically provides a classification of the sample units, whereas a Maximum-A-Posteriori criterion can be used for cluster assignment after running the WEM algorithm. Such criteria lead to classify all the observations, both genuine and contaminated data, meaning that also outliers are assigned to a cluster. Actually, we are not interested in classifying outliers and for purely clustering purposes outliers have to be discarded. Outlier detection should be based on the robust fitted model and performed separately by using formal rules. Outlyingness of each data point is measured conditionally on the final assignment. For instance, for a mixture of Gaussian components, a common rule is to flag outliers when $d^2_{ik_i} > \chi^2_{p;1-\alpha}$, where d_{ik_i} is a robust distance and $\chi^2_{p;1-\alpha}$ is the $(1-\alpha)$-quantile of a χ^2_p variate. In the case of mixtures of linear regressions, in a similar fashion, the outlier detection rule can be based on standardized residuals and their reference standard normal distribution.

Let us consider a couple of illustrative examples on synthetic data. Figure 1 displays the results stemming from the WEM algorithm for a mixture of bivariate normal distributions (left) and a mixture of linear regressions (right) in the presence of outliers. The cluster assignments and the detected outliers are also given. The classical procedures fail, whereas the proposed methods lead to robust solutions with a satisfactory accuracy in fiitting, clustering and outlier detection.

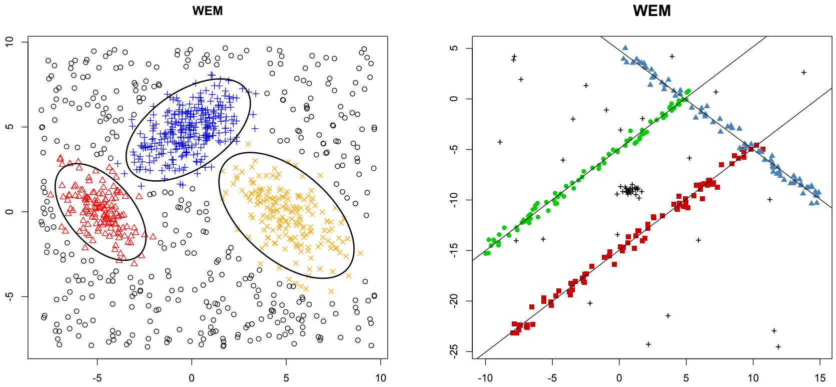


Figure 1. *Fitted mixture of bivariate normal distributions with outliers (left). Fitted mixture of linear regressions with outliers (right). Clusters are denoted by using different colors and symbols. Outliers are represented as circles (left) or crosses (right).*

References

AGOSTINELLI, C., & GRECO, L. 2018. Weighted likelihood estimation of multivariate location and scatter. *Test*.

AGOSTINELLI, C., & MARKATOU, M. 1998. A one-step robust estimator for regression based on the weighted likelihood reweighting scheme. *Statistics & probability letters*, **37**(4), 341–350.

ALQALLAF, F., & AGOSTINELLI, C. 2016. Robust inference in generalized linear models. *Communications in Statistics-Simulation and Computation*, **45**(9), 3053–3073.

BASU, A., & LINDSAY, B.G. 1994. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, **46**(4), 683–705.

FARCOMENI, A., & GRECO, L. 2015. *Robust methods for data reduction*. CRC press.

GRECO, L., & AGOSTINELLI, C. 2018. Weighted likelihood mixture modeling and model based clustering. *arXiv preprint arXiv:1811.06899*.

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

A CANONICAL REPRESENTATION FOR MULTIBLOCK METHODS

Mohamed Hanafi[1]

[1] StatSC, FRANCE, (e-mail: mohamed.hanafi@oniris-nantes.fr)

ABSTRACT: We introduce a representation called canonical representation of multiblock methods from a factorization lemma for partitioned matrices. We show that this canonical representation highlights the strategy adopted by these methods for analyzing multiblock data. This strategy involves two analyzes: (i) a global analysis described by a factorization of the whole data matrix. (ii) a block analysis described by the factorization of each block. The link between parameters of these two analyses is simple and will be presented in detail. The interpretation and visualization of parameters are based on the same principle as the usual Principal Component Analysis.

KEYWORDS: multiblock data analysis, matrix factorization, principal component analysis.

1 Introduction and motivation

Extracting relevant information from multiblock data by reducing dimensionality, summarizing the information in an understandable way or visualizing multiblock data for interpretation purposes, are challenges often raised in chemometrics. When K data blocks denoted $\mathbf{X}_k \left(1 \leq k \leq K\right)$ are available and each data block $\mathbf{X}_k$ reflects the measurements of p_k quantitative variables on n individuals, several multiblock methods are proposed in the literature. We limit ourselves to five methods widely used in chemometrics as listed in Table 1.

Table 1. *List of widely exploratory multiblock methods used in chemometrics*

HPCA	Hierarchical Principal Component Analysis [1,2]
CPCA	Consensus Principal Component Analysis [1,2]
MCOA	Multiple Co-inertia Analysis [5,6]
CCSWA	Common Components and Specific Weights Analysis [3,4,5]
STATIS	Structuration de Tableaux A Trois Indices de la Statistique [7,8]

Linking these methods to each other aims to have a comprehensible picture. This issue was modestly studied in the literature. A monotony property of HPCA was disclosed and an optimization criterion was exhibited [10] pinpointing the equivalence between HPCA and CCSWA. In the same manner [9], new properties of CPCA were disclosed and pinpoint its connection to MCOA and PCA of the whole blocks. Indeed, CPCA and PCA of the whole blocks are equivalent and the main difference between CPCA and MCOA being in the deflation step. In addition, a new formulation of CCSWA was introduced [11] by means of a new criterion which brought it closer to MCOA and CPCA.

Despite these important clarifications, the access to these methods by users continues to be difficult. One of the reasons resides in the heterogeneity of the outputs which makes evaluation of methods difficult. From one method to another, the outputs have neither the same aspect nor the same form. The user is often lost.

The present paper introduces a "canonical representation" of multiblock methods listed in table 1 in order to harmonize their outputs. Several data sets will be used to show how canonical representation of methods listed in table 1 makes easy the evaluation and comparison of these methods.

2 Main contribution

The main idea of canonical representation of multiblock methods takes its origin in the following factorization lemma of a partitioned matrix.

Let $\mathbf{X}_\bullet$ be a matrix of dimension (n, p) partitioned by columns in K blocks $\mathbf{X}_k$ with dimension (n, p_k), there exist a matrix $\mathbf{V}_\bullet$, K matrices $\mathbf{U}_k$ and $(K+1)$ diagonal matrices $\mathbf{D}_\bullet$, $\mathbf{D}_k$ $(1 \le k \le K)$ such that :

$$\mathbf{X}_k = \mathbf{V}_\bullet \mathbf{D}_\bullet \mathbf{U}_k^T \quad (1 \le k \le K),$$

$$\mathbf{X}_\bullet = \mathbf{V}_\bullet \mathbf{D}_\bullet \mathbf{U}_\bullet^T$$

with $\mathbf{V}_\bullet \mathbf{V}_\bullet^T = \mathbf{I}_r$, $diagonal(\mathbf{V}_\bullet \mathbf{V}_\bullet^T) = diagonal(\mathbf{U}_k \mathbf{U}_k^T) = \mathbf{I}_r$ $(1 \le k \le K)$ and r is the rank of $\mathbf{X}_\bullet$.

The decomposition of data blocks $\mathbf{X}_k$ $(1 \le k \le K)$ as described by the lemma is called: canonical representation. It will be shown that multiblock methods listed in table 1 looking for a factorization as presented in the above lemma. In other words, the matrices $\mathbf{V}_\bullet, \mathbf{U}_k$ $(1 \le k \le K)$, $\mathbf{D}_k$ $(1 \le k \le K)$ and $\mathbf{D}_\bullet$ are the main parameters of these methods. Although the parameters differ from one method to another the representation of the parameters remains the same through the factorization.

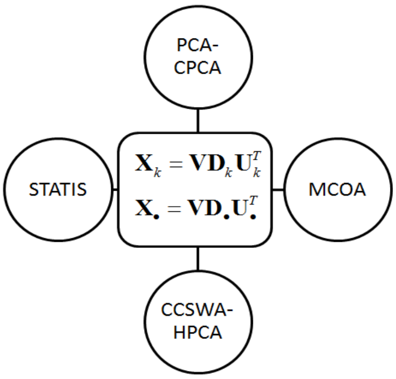

PCA-
CPCA
STATIS
$\mathbf{X}_k = \mathbf{V}\mathbf{D}_k\mathbf{U}_k^T$
$\mathbf{X}_\bullet = \mathbf{V}\mathbf{D}_\bullet\mathbf{U}_\bullet^T$
MCOA
CCSWA-
HPCA

Figure 1. *Highlighting the main parameters of multiblock methods through their canonical representation*

Also, canonical representation highlighted the strategy adopted by these methods. This strategy involves two analyses: (i) a global analysis described by the factorization of $\mathbf{X}_\bullet$, (ii) a block analysis described by the factorization of $\mathbf{X}_k$. The link between parameters of these two analyses is simple and will be presented in detail. The interpretation and visualization of parameters are based on the same principle as the usual Principal Component Analysis.

References

Chessel, D., Hanafi, M. 1996. Analyse de la co-inertie de K nuages de points. *Revue de Statistique Appliquée,* XLVI, (**2**), 35-60

Gourvenec, S., Stanimirova, I., Saby, C-A., Airiau C.Y., Massart. D.L. 2005. Monitoring batch processes with the STATIS approach. *Journal of Chemometrics,* **19**: 288–300

Hanafi, M., Kohler, A., Qannari E. M. 2011. Connections between Multiple Co-inertia Analysis and Consensus Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, **106**(1) : 37-40.

Hanafi, M., Kohler, A., Qannari E. M. 2010. Shedding new light on Hierarchical Principal Component Analysis. *Journal of Chemometrics*, **24**(1): 703-709.

Hanafi, M., Qannari E.M. 2008. Nouvelles propriétés de l'Analyse en Composantes Communes et Poids Spécifiques. *Journal de la Société Française de Statistique*, **149**(2): 75-97.

Wold, S., Kettaneh, N. and Tjessem, K. 1996. Hierarchical multi-block PLS and PC models for easier interpretation and as an alternative to variable selection. *Journal of Chemometrics*, (**10**), 463-482.

Hanafi, M., Mazerolles, G., Dufour, E., Qannari, E. M. 2006. Common components and specific weight analysis and multiple Co-inertia analysis applied to the coupling of several measurement techniques. *Journal of Chemometrics,* (**20**)5, 172-183.

LAVIT, C., ESCOUFIER,Y., SABATIER, R., TRAISSAC. P. 1994. The ACT (Statis method). *Computational Statistics & Data Analysis*, **18**(1):97 119.

MAZEROLLES, G., HANAFI, M., DUFOUR, E., QANNARI, E. M., BERTRAND, D. 2006. Common Components and specific weights analysis: a chemometric method for dealing with complexity of food products. *Chemometrics and Intelligent Laboratory Systems*, (**81**), 41- 49.

QANNARI E. M,. WAKELING I., COURCOUX PH., MACFIE M.F. 2000. Common Components and specific weights analysis performed on preference data. *Food Quality and Preference*, **11**, 151-154.

WESTERHUIS, J. A, KOURTI, T., MACGREGOR J. F. 1998. Analysis of Multiblock and Hierarchical PCA and PLS Models. *Journal of Chemometrics*, (**12**), 301-321.

AN ADEQUACY APPROACH TO ESTIMATING THE NUMBER OF CLUSTERS

Christian Hennig[1]

[1] Department of Statistical Sciences "Paolo Fortunati", University of Bologna,
(e-mail: christian.hennig@unibo.it)

ABSTRACT: I propose a general approach for estimating the number of clusters in a model-based setting. The idea is to choose the smallest number of clusters that provides an "adequate" model, where "adequacy" means that according to one or more suitable criteria, the dataset to be analysed looks like a (more or less) typical dataset generated from the model. Parametric bootstrap can be used to generate datasets from the model, and adequacy can then be assessed by bootstrap tests. For finding meaningful clusters, it may often not be required that the model fits perfectly, so adequacy could be assessed based on smaller than the actually available sample sizes to allow for imprecise fits. Adequacy criteria and application to some model-based clustering methods are discussed.

KEYWORDS: Model-based clustering, parametric bootstrap, OTRMLE, k-quantiles clustering.

1 Introduction

A problem with standard methods to estimate the number of clusters in model-based clustering such as the BIC is that they are critically dependent on the model assumptions. For example, for Gaussian mixtures, if clusters are not exactly Gaussian and datasets are big enough, the BIC will often fit more than one Gaussian distribution to every cluster.

The proposal introduced here is to choose the smallest number of clusters that provides an "adequate" model. The principle of adequacy goes back to Davies, 1995. The idea is that a model is "adequate" for a dataset if the dataset cannot be distinguished from (more or less) typical datasets generated from the model by some criteria that are relevant to the application. Davies & Kovac, 2004 applied the approach to density estimation; they tried to find the density with the smallest number of modes that is "adequate" in terms of the Kolmogorov distance between the fitted and the observed distribution.

The assessment of adequacy relies on criteria measuring the quality of approximation. In cluster analysis, it is often of interest to find meaningful clusters that do deviate slightly from the model assumptions. When fitting a Gaussian mixture, some unimodal and fairly symmetric data subsets should be counted as a single cluster even if they could be fitted slightly better by two or more very close Gaussian mixture components. Therefore, very high precision of approximating the dataset by the fitted model is not required. Rather the modelled clusters should correspond to meaningful clusters in the data. This can be reflected by appropriate quality criteria. Some imprecision in the fit that makes it possible to tolerate clusters for which the model assumption is only roughly appropriate can be achieved by basing the adequacy assessment on a number of observations that is lower than the number of actual observations. Given an adequacy criterion, adequacy can be assessed using parametric bootstrap testing.

2 The adequacy algorithm

Here is an outline of the general adequacy approach.

1. Apply a model-based clustering method to dataset D for a range R of numbers of clusters $G \in R$.
2. For all $G \in R$, generate B datasets $D_{G,b},\ b = 1,\dots,B$ from the fitted models (parametric bootstrap).
3. Apply the clustering method to $D_{G,b},\ b = 1,\dots,B$, fixing G.
4. Compute statistics S that measure the quality of the clustering for all fitted clusterings on the real data and on the bootstrapped data.
5. G is adequate if $S(D,G)$ is not significantly worse than the distribution of $S(D_{G,b},G)$.
6. Choose the smallest adequate G.

3 Clustering methods, outliers

The adequacy principle can be applied to all model-based clustering methods, and even to clustering methods that are not model based, as long as a model can be specified and fitted to generate the parametric bootstrap datasets (see Hennig & Lin, 2015 for the use of parametric bootstrap with non-model based clustering methods). In case of model-based clustering, the model to be used is obviously the fitted model.

Two methods for which up to now no other methods have been proposed to estimate the number of clusters are k-quantiles clustering (Hennig *et al.* , 2018), based on a fixed partition model with asymmetric Laplace distributions, and the Optimally Tuned Robust Improper Maximum Likelihood Estimator (OTRIMLE, Coretto & Hennig, 2016; Coretto & Hennig, 2017), which fits a Gaussian mixture allowing for noise and outliers. In the latter case, some observations are not assigned to any cluster, and better clustering of the non-outliers can be achieved if more observations are classified as outliers. This can be taken into account by looking for the adequate clustering with smallest $G + \frac{\hat{\pi}_0}{p_0}$, where $\hat{\pi}_0$ is the estimated proportion of outliers, and p_0 is the borderline proportion of outliers that the user is willing to trade in for a model with one cluster more.

4 Clustering quality measures

Many clustering quality measures can be used as statistic S. Using multiple testing corrections (Davies, 1995), even more than one statistic can be used (see Hennig, 2017 for some proposals).

Here is one proposal that measures to what extent the found clusters are unimodal. The measure is first defined for one-dimensional data; for more dimensions variable-wise (or principal component-wise) measures can be aggregated.

Apply the following to every cluster with $j \in \{1, \ldots, G\}$ being the current number of clusters:

(a) Compute a kernel density estimator at q equidistant points $y_1 < y_2 < \ldots < y_q$ covering a large probability range (say 99%) under the fitted model, yielding $\hat{f}(y_1), \ldots, \hat{f}(y_q)$.

(b) Separately sort those on the left and those on the right side of the mode of the fitted model: $\hat{f}_l^{(1)} \leq \ldots \leq \hat{f}_l^{(q_l)}$, $\hat{f}_r^{(1)} \geq \ldots \geq \hat{f}_r^{(q_r)}$, $q = q_l + q_r$.

(d) Compare with kernel density left and right of the mode of the fitted model:

$$s_l = \sum_{i=1}^{q_l} (\hat{f}(y_i) - \hat{f}_l^{(i)})^2,$$

$$s_r = \sum_{i=q_l+1}^{q} (\hat{f}(y_i) - \hat{f}_r^{(i)})^2,$$

$$T_j(y_1, \ldots, y_q) = \sqrt{\frac{1}{q}(s_l + s_r)}.$$

In case of unimodality where the mode is as close as possible to the fitted model mode, this yields $T_j = 0$. It may be advisable to standardise T_j by its mean and variance under the assumed model in order to make clusters of different sizes comparable.

The resulting measures T_j need to be aggregated over clusters:

$$S(D,G) = \sqrt{\sum_{j=1}^{G} (T_j)^2}.$$

A version that averages pairs of density values and aggregates the same density values on the left and right side of the mode can be defined if clusters are meant to be symmetric.

5 Conclusion

The algorthm introduced in Section 2 with S as defined in Section 4 can be used to find the smallest G so that the empirical within-cluster distribution does not deviate significantly from the quality S as expected if the fitted model is in fact true; i.e., clusters look as unimodal (and symmetric, if required) as generated by the fitted model.

References

CORETTO, P., & HENNIG, C. 2016. Robust Improper Maximum Likelihood: Tuning, Computation, and a Comparison With Other Methods for Robust Gaussian Clustering. *Journal of the American Statistical Association*, **111**, 1648–1659.

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

DAVIES, P. L. 1995. Data features. *Statistica Neerlandica*, **49**(2), 185–245.

DAVIES, P. LAURIE, & KOVAC, ARNE. 2004. Densities, spectral densities and modality. *Ann. Statist.*, **32**(3), 1093–1136.

HENNIG, C. 2017. Cluster validation by measurement of clustering characteristics relevant to the user. *arXiv:1703.09282*.

HENNIG, C., & LIN, C.-J. 2015. Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. *Statistics and Computing*, **25**, 821–833.

HENNIG, C., VIROLI, C., & ANDERLUCCI, L. 2018. Quantile-based clustering. *arXiv:1806.10403*.

CLASSIFICATION WITH WEIGHTED COMPOSITIONS

Karel Hron[1], Julie Rendlová[1] and Peter Filzmoser[2]

[1] Department of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, (e-mail: karel.hron@upol.cz, julie.rendlova@gmail.com)

[2] Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, (e-mail: peter.filzmoser@tuwien.ac.at)

ABSTRACT: In classification tasks with geochemical of chemometric data, it frequently happens that observations are of relative (compositional) nature. It means that the relevant information is contained in ratios rather than in the absolute values of components due to the possible influence of the size effect. The logratio approach to compositional data analysis offers a concise methodology replacing the original scale invariant positive data by reasonable real variables, which are formed by ratios of the components or their amalgamation, prior to further statistical processing. The preferred type of such logratio variables corresponds to orthonormal coordinates with respect to the Aitchison geometry of compositional data, and particularly to such a coordinate system, where the first coordinate aggregates all logratios with the specific part of interest and can be thus linked to that component - we refer to so-called pivot coordinates. However, including all respective logratios into the first pivot coordinate, specifically those logratios reflecting differences between groups to a coordinate corresponding to a non-biomarker, may lead to an artificial occurrence of false positive biomarker detection. Accordingly, biased picture concerning sources for classification of groups can be expected. Therefore, in the contribution, we propose a method excluding aberrant logratios so that the coordinate which is afterward considered to be the pivot one in the resulting coordinate system contains already just the cleaned information about the relative dominance of the specific component. Importantly, the alternative choice of pivot coordinates, which we suggest to call selective pivot coordinates, does not influence the quality of classification itself since both coordinate systems are just rotations of each other. The effect of such a choice of coordinates will be presented with the partial least squares regression - discriminant analysis of metabolomic data.

KEYWORDS: compositional data, logratio coordinates, partial least squares regression - discriminant analysis, biomarker detection.

MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers

Mia Hubert[1], Peter J. Rousseeuw[1, 2] and Wannes Van den Bossche[1]

[1] Department of Mathematics, KU Leuven, Belgium,

[2] Presenting author, (e-mail: peter@rousseeuw.net)

Abstract: Multivariate data are typically represented by a rectangular matrix (table) in which the rows are the objects (cases) and the columns are the variables (measurements). When there are many variables one often reduces the dimension by principal component analysis (PCA), which in its basic form is not robust to outliers. Much research has focused on handling rowwise outliers, i.e. rows that deviate from the majority of the rows in the data (for instance, they might belong to a different population). In recent years also cellwise outliers are receiving attention. These are suspicious cells (entries) that can occur anywhere in the table. Even a relatively small proportion of outlying cells can contaminate over half the rows, which causes rowwise robust methods to break down.

In this paper a new PCA method is constructed which combines the strengths of two existing robust methods, DetectDeviatingCells and ROBPCA, in order to be robust against both cellwise and rowwise outliers. At the same time, the algorithm can cope with missing values. As of yet it is the only PCA method that can deal with all three problems simultaneously. Its name MacroPCA stands for **PCA** allowing for **M**issings **A**nd **C**ellwise & **R**owwise **O**utliers. Several simulations and real data sets illustrate its robustness. New residual maps are introduced, which help to determine which variables are responsible for the outlying behavior. The method is well-suited for online process control. The function MacroPCA has been incorporated in the R package *cellWise* on CRAN, which also contains a vignette with real data examples.

Keywords: detecting deviating cells, outlier map, residual map.

References

Hubert, M., Rousseeuw, P.J., & Van den Bossche, W. 2019. MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics*, in press.

RAYMAEKERS, J., ROUSSEEUW, P.J., VAN DEN BOSSCHE, W., & HUBERT, M. 2019. *cellWise: Analyzing Data with Cellwise Outliers. R package, CRAN.*

ROUSSEEUW, P.J., & VAN DEN BOSSCHE, W. 2019. Detecting Deviating Data Cells. *Technometrics*, **60**, 123–145.

Marginal effects for comparing groups in regression models for ordinal outcome when uncertainty is present

Maria Iannario[1] and Claudia Tarantola[2]

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

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

Abstract: This contribution deals with effect measures for covariates in ordinal data models to address the interpretation of the results on the extreme categories of the scales. It provides a simpler interpretation than model parameters both in standard cumulative models with proportional odds assumption and in the recent extension of the CUP models, the mixture models to account for uncertainty in the process of selection of the score. Visualization tools for the effect of covariates are proposed and the measure of relative size and marginal effects based on rates of change are evaluated by use of a case study.

Keywords: cumulative link models, CUP models, extreme categories, marginal effects, uncertainty.

1 Background and Preliminaries

Traditional models for rating data analysis are *G*eneralized *L*inear *M*odels (GLM) that employ nonlinear link functions to cumulative probabilities (McCullagh, 1980). Recent attention on the uncertainty detected when a subject selects a score on a rating question led to the alternative CUP models (Tutz *et al.*, 2017). They represent a special-case in the framework of the *Ge*neralized *M*ixture with Uncertainty (GEM) models (Iannario & Piccolo, 2016) in which CUB models are the starting point. Indeed, they are a *C*ombination of two components referred to the individual indecision (*U*ncertainty) expressed on the selection or motivated by the context and to a deliberate choice of a response category determinated by the *P*reference of the respondent.

As a consequence of the nonlinearity model parameters are not as simple to interpret as slopes and correlations for ordinary linear regression. The model effect parameters relate to measures, such as odds ratios and probits, may not be easily understood or can even be misinterpreted. Furthermore, the interest in some specific fields to the correct interpretation of the effect of categories in the extreme of the scale (the worst/best of the selection) motivates the present contribution. Indeed it surveys simpler ways to interpret the effects of the explanatory variables simplifying the interpretation of the models, describing and visualizing average and global marginal effects. Section 2 is devoted to the introduction of the model and marginal effects whereas Section 3 concludes with a case study and some remarks.

2 Marginal effect measures for covariates in CUP models

In a CUP model (Tutz *et al.*, 2017), the probability distribution of the ordinal response variable R_i, for $i = 1,2,\ldots,n$, describing the rating assigned by respondent i, is given by

$$P(R_i = r|\boldsymbol{x}_i) = \pi_i P_M(Y_i = r|\boldsymbol{x}_i) + (1-\pi_i)P(U_i = r),\ r = 1,2,\ldots,m,$$

where $P_M(Y_i = r|\boldsymbol{x}_i)$ (*P*refence part) is obtained via a cumulative link model on an appropriate set of covariates, and a logit link is usually assigned on the uncertainty parameter π_i. Here, the second component of the mixture $P(U_i = r)$ follows a discrete Uniform distribution.

One natural way to interpret the effect of one explanatory variable is to consider the corresponding marginal effects (MEs). A ME shows how a variation in one variable affects the outcome distribution, holding all the other variables constant. We refer to Greene & Hensher, 2010 for a discussion of the interpretation of marginal effects in ordered response models.

As an exemplification, the marginal effect of a continuous explanatory variable on $P(R = i)$ will be reported. The rate of change in $P(R = 1)$ with respect to a continuous variable x_{ik} involved in the preference part of the model is the partial derivative of $P(R = 1)$ with respect to x_{ik}

$$\frac{\partial P(Y = 1|\boldsymbol{x}_i^*)}{\partial x_{ik}} = -\gamma_k f(\alpha_1 - \boldsymbol{x}_i\gamma),$$

where $f()$ is the density function corresponding to the examined cumulative model, and the other explanatory variables having fixed values $\boldsymbol{x}_i^*$. In a similar way we obtain the rate of change in $P(Y = 1)$ with respect to y_{ik} involved in the uncertainty part of the model

$$\frac{\partial P(R = 1)}{\partial y_{ik}} = \beta_1 f(\beta_0 + y_{ik}\beta_1)(F(\alpha_1 - \boldsymbol{x}\gamma) - 1/k).$$

Here, $F()$ is the cumulative distribution funnction. Further details are in Iannario & Tarantola, 2019.

3 Example

Data was provided by the Survey of Health, Ageing and Retirement in Europe (SHARE) from wave 1, 2004. In this contribution, a rating concerning the perceived *Pain* collected on a 4 points Likert scale (Never=1, Rarely=2, Every Ones in a While=3, Almost Always=4) has been analyzed. Covariates introduced for the analysis are *Gender* ($0 =$ Male,$1 =$ Female) and Body Mass Index (*BMI* from 2.563 to 76.950, with average=26.592 and s.d.=4.310). The sample of $n = 3458$ elderly people (average *age*=62) is overall overweight (average *BMI*=26.590)

Table 1. CUP *models fitted to perceived pain assessment.*

$\hat{\beta}_1$ (*Gen*)	$\hat{\beta}_2$ (*BMI*)	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$
0.770 (*0.089*)	0.085 (*0.011*)	1.992 (*0.334*)	3.646 (*0.374*)	5.468 (*0.476*)

Estimated results of CUP models are reported in Table 1. It lists estimated parameters $\hat{\beta}_j$, $j = 1,2$ and cutpoints $\hat{\alpha}_j$, $j = 1,2,3$ with asymptotic standard errors (in parentheses). Here the AIC index is 8603.609 compared with a standard CUB model with AIC=8610.932 whereas the π parameter is $0.916(\mathit{0.051})$ with respect to $\pi_{CUB} = 0.789(\mathit{0.022})$ highlighting the different role of uncertainty in the selected model. Average ME are in Table 2. Given the sign convention, as expected, it is possible to observe a positive effect on the Female (*Gen*) (they perceived more pain) and on increasing level of *BMI* on perceived pain. Group comparisons with relative marginal effects are in Figure 1. There is evidence that the first category effectively thresholds those having absolutely

Table 2. *Average Marginal Effect for CUP models - SHARE data.*

$ME.1$				
	effect	std.error	z.value	p.value
Gen	-0.152	0.017	-9.002	0.001
BMI	-0.017	0.002	-7.713	0.001
$ME.4$				
	effect	std.error	z.value	p.value
Gen	0.041	0.006	7.470	0.002
BMI	0.005	0.001	7.537	0.005

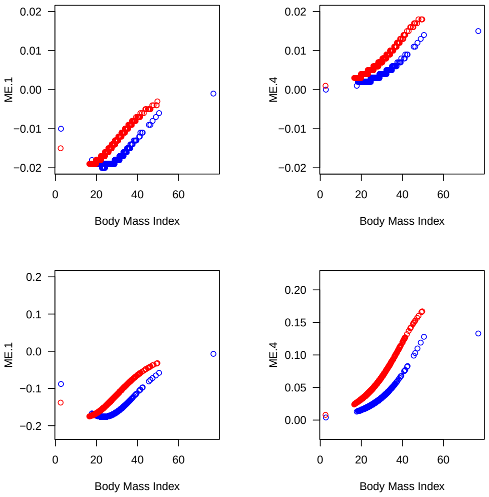


Figure 1. *Group comparisons, Male (blue) and Female (red) versus BMI, for marginal effects (First marginal effect on left panel, last in right panel). Top panel is about BMI marginal effect, bottom panel on gender marginal effect.*

no pain. The last one highlights the difference in gender groups.

An extended study has been planned to validate the efficacy of the proposal and the impact of the results.

References

GREENE, W.H. 2008. *Econometric Analysis*. 6th edn. Upper Saddle River, NJ: Pearson Prentice Hall.

GREENE, W.H., & HENSHER, D.A. 2010. *Modeling Ordered Choices: A Primer*. Cambridge University Press.

IANNARIO, M., & PICCOLO, D. 2016. A comprehensive framework of regression models for ordinal data. *METRON*, **74**, 233–252.

IANNARIO, M., & TARANTOLA, C. 2019. How to interpret the effect of covariates on the extreme categories in rating data models. *Manuscript*.

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

TUTZ, G., SCHNEIDER, M., IANNARIO, M., & PICCOLO, D. 2017. Mixture models for ordinal responses to account for uncertainty of choice. *Advances in Data Analysis and Classification*, **11**, 281–305.

A MULTI-CRITERIA APPROACH IN A FINANCIAL PORTFOLIO SELECTION FRAMEWORK

Carmela Iorio[1], Giuseppe Pandolfo[1] and Roberta Siciliano[1]

[1] Department of Industrial Engineering, University of Naples Federico II,
(e-mail: carmela.iorio@unina, giuseppe.pandolfo@unina.it,
roberta@unina.it)

ABSTRACT: The key role of a portfolio manager is to establish a suitable strategy of asset allocation. The composition of a portfolio does not only depend on both return and risk of each asset, but it is also influenced by various factors. The final decision making belongs to a multiple criteria problem. Our aim is to apply a multi-criteria approach to select the attractive securities for a portfolio according to the resulting clustering of time-varying beta of the stocks. To reach this aim, we propose a two-step approach that consists in applying before a k-means algorithm on the time-varying beta computed on a suitable Capital Asset Pricing Model. Then, we rank these stocks by a Multi Criteria Decision Making model.

KEYWORDS: CAPM, time-varying beta coefficient, P-spline, cluster analysis, MCDM.

1 Introduction

From the milestone work of Markowitz, 1952, Capital Asset Pricing Model (CAPM) proposed by Sharpe, 1964 was the most famous model of financial market equilibrium. The CAPM states a linear relationship between a stock return and its risk, measured by a coefficient known as beta. It explains the systemic risk that is related to market itself (thus not decreases by the diversification step). Under the (unrealistic) CAPM hypothesis, the beta coefficients do not vary over time. This characteristic is very restrictive and not readily found in the reality. In facts, the beta of the assets can vary at any point in time depending on (among the other) the information available at the given time about the financial markets, the overall economic conditions, the specific firms. Taking into account all the above mentioned characteristics, the portfolio selection problem belongs to a Multi Criteria Decision Making (MCDM) framework (Xidonas *et al.*, 2010). It consists of a set of different methodologies taking into consideration conflicting several criteria to support decision makers in solving a decision problem. MCDM are useful tools in portfolio selection and management (see, e.g. Zopounidis *et al.*, 2015). Within the MCDM problems, the aggregation of all the evaluation criteria can be carried out by using different models (outranking relations, utility function or decision rules). One of them is the Elimination and Choice Translating Reality (ELECTRE) method proposed by Roy, 1968. The ELECTRE family in MCDM problems consists of two steps: *(i)* the outranking relations are constructed then *(ii)* the procedures of choosing, selection, sorting or ranking among the alternatives are applied. During the years, the ELECTRE method evolved into a number of other variants that are based on the same foundation, but they differ slightly. Among these methods, ELECTRE III (Roy, 1991) was designed for ranking problems, also providing different advantages in a decision making process.

2 The key idea

The static (with constant βs) CAPM formulation is given by:

$$r_i(t) = r_f + \beta_i(r_m(t) - r_f), \tag{1}$$

where r_i is the return for asset i, r_f is the risk-free rate (which is known), β_i is the sensitivity of the expected asset returns to the the market returns r_m (as measured by a stock market index for example). To allow the risk factors to vary over time, we follow the varying-coefficient model proposed by Hastie & Tibshirani, 1993. The following relationship holds:

$$r_i(t) = r_f + \beta_i(t)(r_m(t) - r_f). \tag{2}$$

In this paper, we propose to model $\beta_i(t)$ using P-spline (Eilers & Marx, 2002). Equation (2) can then be formulated as

$$y_i(t) = a_{0,i} + \text{diag}\{x(t)\}Ba_{i,i} + \varepsilon_i(t) = (\mathbf{1}|U)\alpha_i + \varepsilon_i(t) = Q\alpha_i + \varepsilon_i(t), \tag{3}$$

where $\varepsilon_i(t)$ is a zero mean error term with constant variance, $y_i(t) = r_i(t) - r_f$, $x(t) = rm(t) - r_f$, $\alpha_i = (a_{0,i}^\top, a_{1,i}^\top)^\top$, $a_{0,i}$ is an asset-specific intercept term, $a_{1,i}$ is the vector of spline coefficients for the time-varying risk factor for asset i, B is a B-spline matrix and $U = \text{diag}\{x(t)\}B$ with $\text{diag}\{x(t)\}$ aligning the predictors with the appropriate smooth slope values. If $Q = (\mathbf{1}|U)$, then the penalized estimation problem for asset i becomes:

$$S_i = \|y_i(t) - Q\alpha_i\|^2 + \lambda_i \left\|\breve{D}_d\alpha_i\right\|^2, \tag{4}$$

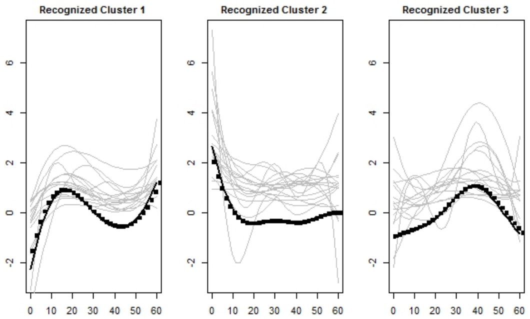


Figure 1. *Beta varying coefficients clustered by k-means procedure with Pearson's correlation coefficient based distance. For each subplot the horizontal axis represents the time and the vertical axis the systematic risk level. The gray lines reproduce the beta series assigned to each cluster. The dots indicate the estimated optimal beta coefficients for cluster center. The black solid lines indicate the center functionals.*

where $\breve{D}_d$ shrinks only the $a_{1,i}$ coefficients in α_i and λ_i is a smoothing parameter. The solution of (4) is then

$$\hat{\alpha}_i = (Q^\top Q + \lambda_i \breve{D}_d^\top \breve{D}_d)^{-1} Q^\top y_i(t), \tag{5}$$

from which it follows that $\hat{\beta}_i(t) = B\hat{a}_{1,i}$.

We propose a two-step procedure combining hard clustering of risk factors and ELECTRE III ranking procedure, for the selection of asset to compose an investment portfolio by evaluating the associated systematic risk. In analogy with Iorio *et al.*, 2016, we propose to cluster the beta coefficients for a set of assets. We model the risk indicators by means of P-spline whose coefficients are clustered so that each group is characterized by stocks with similar systemic risk profiles. Figure 1 shows the results of our proposal on a data set of 48 stocks constituent the S&P500 Index collected monthly from january 2006 to december 2010 (source *yahoo.finance.com*). In a second step, we compute for each stock a series of risk-adjusted performance measures that are used as criteria of ELECTRE III method to obtain a stock ranking useful for the asset selection step. Then a portfolio manager can select the $n < N$ top stocks according to ranking, given the previous screening clustering based on different profiles of systemic risk, ensuring a better diversification of portfolio.

3 Conclusion

In this paper, we proposed a portfolio composition method within a multi-criteria framework. The procedure is based on a cluster analysis of the time-varying betas, estimated by using P-spline, so that each resulting group contains different level of systemic risk. Then we computed some risk adjusted performance measures for the stocks of the recognized clusters. Finally, we use these indexes as input of the ELECTRE III method to obtain a stock ranking useful for the asset selection phase.

References

EILERS, P. H. C ., & MARX, B. D. 2002. Generalized Linear Additive Smooth Structures. *Journal of Computational and Graphical Statistics*, **11**(4), 758–783.

HASTIE, T., & TIBSHIRANI, R. 1993. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, **55**(4), 757–779.

IORIO, C., FRASSO, G., DAMBROSIO, A., & SICILIANO, R. 2016. Parsimonious time series clustering using P-splines. *Expert Systems with Applications*, **52**, 26–38.

MARKOWITZ, H. 1952. Portfolio selection. *The journal of finance*, **7**(1), 77–91.

ROY, B. 1968. Classement et choix en présence de points de vue multiples. *Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle*, **2**(1), 57–75.

ROY, B. 1991. The outranking approach and the foundations of ELECTRE methods. *Theory and decision*, **31**(1), 49–73.

SHARPE, W. F. 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, **19**(3), 425–442.

XIDONAS, P., MAVROTAS, G., & PSARRAS, J. 2010. A multiple criteria decision-making approach for the selection of stocks. *Journal of the Operational Research Society*, **61**(8), 1273–1287.

ZOPOUNIDIS, C., GALARIOTIS, E., DOUMPOS, M., SARRI, S., & ANDRIOSOPOULOS, K. 2015. Multiple criteria decision aiding for finance: An updated bibliographic survey. *European Journal of Operational Research*, **247**(2), 339–348.

CLUSTERING OF TRAJECTORIES USING ADAPTIVE DISTANCES AND WARPING

Antonio Irpino[1] and Antonio Balzanella[1]

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

ABSTRACT: The paper deals with the clustering trajectories of moving objects. A prototype-based clustering using Euclidean distance between piece-wise linear curves is used. The main novelty of the paper is the opportunity of considering in the clustering procedure two steps: a step that automatically weights the importance of sub-trajectories of the original ones and an alignment step for expressing the prototypal trajectory which uses the Dynamic Time Warping algorithm. The algorithm uses an adaptive distances approach and a cluster-wise weighting. The algorithm is tested against some workbench trajectory datasets.

KEYWORDS: trajectory clustering, adaptive distances, time warping.

1 Introduction

Nowadays, surveillance systems or the global positioning system (GPS) sensors integrated into devices produce a huge amount of data about moving objects expressed as trajectories. The extraction of patterns from trajectories is increasingly challenging and demanding. Clustering is a very useful tool for extracting patterns and trajectory clustering has some peculiarities involving spatial and time information.

Thus, the problem of clustering trajectories depends on how trajectories are compared, or if a trajectory is considered as a set of sub-trajectories or not. Depending on time, a trajectory can be considered as a two or three dimensional time-series. Trajectories clustering looks for groups of trajectories, or of sub-trajectories, such that they represent a movement pattern in the data. The subject is surveyed in Yuan *et al.*, 2017. In the literature, two main algorithms are considered: the Lee *et al.*, 2007 and the Nivan *et al.*, 2013. In Lee *et al.*, 2007, a distance between sub-trajectories is defined and the algorithm implements an extension of a density dased algorithm for grouping set of sub-trajectories. In Nivan *et al.*, 2013, the idea is to estimate *k* predefined vector fields that represent group of trajectories observed in a 2D space. This application, is inspired by the problem of monitoring and predicting storm or hurricane paths. In a functional data analysis approach, a trajectory is considered as a curve in a 2D or 3D space. In order to be analyzed a smoothing, interpolation, or alignment step is performed and then the trajectory are analyzed Sangalli *et al.*, 2010.

In this paper, we consider a prototype based approach for grouping trajectories. We show how to decompose the Euclidean distance between two trajectories and use such a decomposition for explaining some aspects of the compared trajectories. We enrich the algorithm with a step that automatically assign a relevance weights to the aspects. Further, considering that trajectories may be misaligned in time, we introduce an alignment step for defining a prototype of the cluster using Dynamic Time Warping (DTW). We remark that, considering that DTW bassed distances do not allow convex optimization problems, the proposed algorithm is only inspired to the classical k-means one and we show its convergence to a stable result only empirically.

Finally, we show some preliminary results on some benchmark datasets.

2 Data and distances

A trajectory is a sequence of ordered space-time points (namely, a point has two or three spatial coordinates and a time-stamp), where the order follows time. A trajectory P_i is a collection of ordered pairs of data $(\mathbf{s}^i_j, t^i_j)$, $j = 1, \ldots, T$, sampled in T time-points where $\mathbf{s}^i_j$ is a spatial location (namely. a 2D or a 3D vector of spatial coordinates) and t^i_j is a time-stamp. A set of N trajectories is a collection of trajectories denoted as P_i. We assume that each trajectory may have a different number of sampled time-points T_i. Clustering is based on a distance/dissimilarity measure between objects. In our case, the computation of a distance between two trajectories may require a normalization step for comparing them. Such a step, depending on the application domain and on the aim of analysis, may be questionable.

The hypothesis that a trajectory is piece-wise linear curve is computationally useful for computing a continuous version of the Euclidean distance between two trajectories.

Under this assumption, the Euclidean distance between two 2D trajectories[*] having the same k time stamps normalized in $[0,1]$ as follows. Given two normalized trajectories $P_1 = \left\{\{(x^1_0, y^1_0), 0\}, \ldots, \{(x^1_j, y^1_j), \tau^1_j\}, \ldots, (x^1_{T_1}, y^1_{T_1}), 1\}\right\}$

[*] The trajectory is on a plane, but the extension to 3D spaces is straightforward.

and $P_2 = \left\{ \{(x_0^2, y_0^2), 0\}, \ldots, \{(x_j^2, y_j^2), \tau_j^2\}, \ldots, \{(x_{T_2}^2, y_{T_2}^2), 1\} \right\}$. Considering the piece-wise linear assumption, and constant speed between each pair of sampled points, it is possible to express the two trajectories with a common set of τ's by a linear interpolation. Once the two trajectories are registered such that they have the same normalized $L \in [min(T_1, T_2), (T_1 + T_2)]$ time-stamps we compute the squared Euclidean distance between P_1 and P_2 as follows:

$$\begin{aligned} d_E^2\left(P_1, P_2\right) &= \int_0^1 \left[(x_1(\tau) - x_2(\tau))^2 + (y_1(\tau) - y_2(\tau))^2 \right] d\tau = \\ &= \sum_{\ell=1}^{L} (\tau_\ell - \tau_{\ell-1}) \left\{ \begin{array}{l} |\bar{x}_1(\ell) - \bar{x}_2(\ell)|^2 + |\bar{y}_1(\ell) - \bar{y}_2(\ell)|^2 + \\ + \frac{1}{3} \left[|\dot{x}_1(\ell) - \dot{x}_2(\ell)|^2 + |\dot{y}_1(\ell) - \dot{y}_2(\ell)|^2 \right] \end{array} \right\} \end{aligned} \tag{1}$$

where:

- $\bar{x}_1(\ell) = \frac{x_1(\tau_\ell) + x_1(\tau_{\ell-1})}{2}$, $\bar{x}_2(\ell) = \frac{x_2(\tau_\ell) + x_2(\tau_{\ell-1})}{2}$, $\bar{y}_1(\ell) = \frac{y_1(\tau_\ell) + y_1(\tau_{\ell-1})}{2}$, and $\bar{y}_2(\ell) = \frac{y_2(\tau_\ell) + y_2(\tau_{\ell-1})}{2}$. The points $(\bar{x}_1(\ell), \bar{y}_1(\ell))$ and $(\bar{x}_2(\ell), \bar{y}_2(\ell))$ are, respectively, the centers of the segment that starts from $(x_1(\tau_{\ell-1}), y_1(\tau_{\ell-1}))$ and arrives at $(x_1(\tau_\ell)\ , y_1(\tau_\ell))$, respectively, the centers of the segment that starts from $(x_2(\tau_{\ell-1})\ , y_2(\tau_{\ell-1}))$ and arrives at $(x_2(\tau_\ell), y_2(\tau_\ell))$;
- $\dot{x}_1(\ell) = \frac{x_1(\tau_\ell) - x_1(\tau_{\ell-1})}{2}$, $\dot{x}_2(\ell) = \frac{x_2(\tau_\ell) - x_2(\tau_{\ell-1})}{2}$, $\dot{y}_1(\ell) = \frac{y_1(\tau_\ell) - y_1(\tau_{\ell-1})}{2}$, and $\dot{y}_2(\ell) = \frac{y_2(\tau_\ell) - y_2(\tau_{\ell-1})}{2}$. The value $(\dot{x}_1(\ell), \dot{y}_1(\ell))$ and $(\dot{x}_2(\ell)\ , \dot{y}_2(\ell))$ are, respectively, the pairs of the component-wise half widths of the segment that starts from $(x_1(\tau_{\ell-1})\ , y_1(\tau_{\ell-1}))$ and arrives at $(x_1(\tau_\ell)\ , y_1(\tau_\ell))$, respectively, of the segment that starts from $(x_2(\tau_{\ell-1})\ , y_2(\tau_{\ell-1}))$ and arrives at $(x_2(\tau_\ell)\ , y_2(\tau_\ell))$.

Distance in Eq. 1 can be naturally decomposed for sub-trajectories. In a k-means like algorithm, it is important to define an average object. Indeed, k-means algorithms rely on the definition of a within cluster homogeneity criterion that usually is expressed as a distance between objects and a representative of the cluster. In this case, being trajectory depending on time, it is possible that a misalignment occurs, biasing the average (prototype) object trajectory. In each representation step of the algorithm, we suggest computing the prototype after a recursive alignment of the trajectories belonging to the cluster and the average one such that a minimum DTW distance criterion is minimized.

The alignment step does not guarantee that the algorithm converges toward a minimum squared distance criterion (like in k-mean).

Using some benchmark data we show its empirical convergence and the obtained results. Some other warping methods will be discussed.

References

DIDAY, E., & GOVAERT, G. 1977. Classification Automatique avec Distances Adaptatives. *RAIRO Inf Comput Sci*, **11**(01), 329–349.

LEE, J.-G., HAN, J., & WHANG, K.-Y. 2007. Trajectory Clustering: A Partition-and-group Framework. *Pages 593–604 of: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD '07.

MORRIS, B., & TRIVEDI, M. 2009. Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. *In: Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*.

NIVAN, F., KLOSOWSKI, J. T., SCHEIDEGGER, C. E., & SILVA, C. T. 2013. Vector field k-means: Clustering trajectories by fitting multiple vector fields. *Computer Graphics Forum*, **32**(3 PART2), 201–210.

SANGALLI, L. M., SECCHI, P., VANTINI, S., & VITELLI, V. 2010. K-mean Alignment for Curve Clustering. *Comput. Stat. Data Anal.*, **54**(5), 1219–1233.

YUAN, G., SUN, P., ZHAO, J., LI, D., & WANG, C. 2017. A Review of Moving Object Trajectory Clustering Algorithms. *Artif. Intell. Rev.*, **47**(1), 123–144.

Sampling and Learning Mallows and Generalized Mallows Models Under the Cayley Distance: short paper

Ekhine Irurozki[1], Borja Calvo[1] and Jose A. Lozano[1, 2]

[1] Basque Center for Applied Mathematics, (e-mail: eirurozki@bcamath.org, jlozano@bcamath.org)

[2] Intelligent Systems Group, University of the Basque Country, (e-mail: borja.calvo@ehu.eus, ja.lozano@ehu.eus)

Abstract: The Mallows and Generalized Mallows models are compact yet powerful and natural ways of representing a probability distribution over the space of permutations. This short paper, which is a summary of the long paper of the same title, deals with the problems of sampling and learning such distributions when the metric on permutations is the Cayley distance. We propose new methods for both operations, and their performance is shown through several experiments. An application in the field of biology is given to motivate the interest of this model.

Keywords: permutations , Mallows model , sampling , learning , Cayley distance , Fisher-Yates-Knuth shuffle.

1 Introduction

The presence of data in the form of permutations or rankings of items is ubiquitous in many real world scenarios, from the computational social choice Brandt *et al.* , 2016 to preference learning Lu & Boutilier, 2011 or bioinformatics Critchlow, 1988. When it comes to handle uncertainty in permutation spaces the Mallows and the Generalized Mallows model are two of the the most popular alternatives Mallows, 1957; Critchlow *et al.* , 1991.

Both models rely on a distance for permutations and in this paper we focus on the Cayley distance*. It counts the number of swaps (not necessarily adjacent) to transform a given permutation into another one so it is closely related with the cyclic structure of permutations: the number of swaps to convert π into the identity permutation, and thus the Cayley distance $d(\pi)$, equals n minus the number of cycles of π.

*This is a difference with most literature on the topic, where the Kendall's-τ distance is usually considered

Most results in this paper are based on a decomposition of the Cayley distance of a permutation π, $d(\pi)^{\dagger}$ which is denoted as $\mathbf{X}(\pi)$. This $\mathbf{X}(\pi)$ is a vector of length $n-1$ where each position is defined as $X_j(\pi) = 0$ if j is the largest item in its cycle in π, and $X_j(\pi) = 1$ otherwise. Note that $d(\pi) = \sum_{j=1}^{n-1} X_j(\pi)$.

2 Mallows and Generalized Mallows models

The Mallows model is an exponential-location probability model for permutations based on distances. It is defined by a central permutation (the location parameter) denoted as σ_0 and the dispersion parameter, denoted θ. It can be expressed as follows:

$$p(\sigma) = \psi_j(\theta_j)^{-1} exp(-\theta d(\sigma, \sigma_0)). \qquad (1)$$

The GMM is defined on the distance decomposition vector for Cayley, $\mathbf{X}(\sigma)$. Specificly, for a central permutation σ_0 and dispersion parameter vector $(\theta_1, \ldots, \theta_{n-1})$ the GMM under the Cayley distance is defined as follows

$$p(\sigma) = \psi_j(\theta_j)^{-1} \prod_{j=1}^{n-1} exp(-\theta_j X_j(\sigma\sigma_0^{-1})). \qquad (2)$$

It is worth noticing that by setting every dispersion parameter θ_j to the same value we recover the MM. For both models, the mode is σ_0 provided that $\theta, \theta_j > 0$. The idea of the GMM is that the displacements at different positions should affect in a different way to the probability of a permutation, and this is controlled by setting different values to different dispersion parameters θ_j.

One of the best known references in the literature of statistical models on permutation data Critchlow *et al.* , 1991 shows how to exploit the properties on exponential models to obtain efficient expressions to work with these models. In particular, based on computable expressions for the moment generating function, the authors are able to reformulate $p(X_j)$ and the normalization constant ψ efficiently. In this paper, we extend their work and propose computationally efficient exact sampling and learning algorithms. We will illustrate the links between MM and GMM under the Cayley distance to other known models in the literature and adapt classical algorithms to these statistical problems by unraveling new properties of the algorithms.

†for notational convenience we use one of the permutations to be the identity, but these results apply in general.

3 Sampling

The first problem approached consists on obtaining a random permutation from a MM or a GMM. It is known that the probability of the distance decomposition vector can be expressed[‡] and sampled efficiently Critchlow *et al.* , 1991. Therefore, sampling a permutation can be done by (1) sampling a distance decomposition vector $\mathbf{X}$ and (2) obtaining a permutation σ such that $\mathbf{X}(\sigma) = \mathbf{X}$. Unfortunately, there exist possibly many permutations with this decomposition and obtaining uniformly at random one of those is not trivial.

In the long version of this paper we show how to use the Chinese Restaurant Process and the Fisher-Yates-Knuth (FYK) algorithms to sample permutations uniformly at random. We discuss which is the cyclic structure of the obtained permutations and consequently, Cayley distance decomposition vector. Finally, we propose an adaptation of the FYK algorithm to sample from a GMM in linear time, which is one of the main results of the paper. The experimental section compares the performance of our proposed sampler with an adaptation of a Markov chain Monte Carlo algorithm, on both time an quality results.

4 Learning

The learning task has been approached as a Maximum Likelihood Estimation of the parameters of a given sample of permutations. It can be shown that the MLE in MM can be broken in two stages, which are (1) finding the central permutation that minimizes the sum of the distances to the sample and then (2) computing the dispersion parameters. On the other hand, the learning process of the GMM cannot be broken in stages and it is done by looking for the permutation that maximizes the likelihood of the sample. However, both learning problems can be seen as looking for a permutation that optimizes a fitness function (sum of distances in MM and likelihood in GMM). Therefore, we refer to the learning as an optimization problem for the rest of the section, for which we propose two algorithms, one exact and one approximate for both MM and GMM.

The exact algorithm explores the tree of partial permutations looking for the permutation that optimizes the fitness function. The number of leaves in this tree is much larger than the number of permutations, so a raw search on this tree would be highly inefficient. However, for each partial permutation a

[‡] We should note that this paper corrects typos of the original version.

lower bound on the value of the objective function for every node in the brach can be computed efficiently. By making use of this clever lower bound, in practice we can prune the tree and search a large space efficiently. As usually occurs in this scenarios, the performance of the algorithm is highly increased if we consider a good initial candidate solution.

The experimental evaluation shows the performance of both methods for samples of various degrees of consensus. It concludes that as the sample differs from uniformity both algorithms quickly improve their performance: the quality in the case of the approximate and the time performance in the case of exact algorithm.

Acknowledgements

This research was supported by the Basque Government through the BERC 2018-2021 and Elkartek projects, the Spanish Ministry of Economy and Competitiveness MINECO and FEDER: BCAM Severo Ochoa excellence accreditation SEV-2017-0718 and MINECO TIN program.

References

BRANDT, F., CONITZER, V., ENDRISS, U., LANG, J., & PROCACCIA, ARIEL D. 2016. *Handbook of computational social choice*.

CRITCHLOW, D. E. 1988. Ulam's metric. *In Encyclopedia of Statistical Sciences*, **9**, 379–380.

CRITCHLOW, D. E., FLIGNER, M. A., & VERDUCCI, J. S. 1991. Probability Models on Rankings. *Journal of Mathematical Psychology*, **35**, 294–318.

LU, T., & BOUTILIER, C. 2011. Learning Mallows Models with Pairwise Preferences. *Pages 145–152 of: International Conference on Machine Learning (ICML)*.

MALLOWS, C. L. 1957. Non-null ranking models. *Biometrika*, **44**(1-2), 114–130.

THE GENDER PARITY INDEX FOR ACADEMIC STUDENTS PROGRESS

Aglaia Kalamatianou[1], Adele H. Marshall[2] and Mariangela Zenga[3]

[1] Department of Sociology, Panteion University, (e-mail: akalam@panteion.gr)

[2] School of Mathematics and Physics, Queen's University of Belfast, (e-mail: a.h.marshall@qub.ac.uk)

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

ABSTRACT: The research in this paper introduces the Gender Parity Index (GPI) to analyze gender differences in higher education. The GPI is applied to the time that it takes students to graduate beyond the recommended time period for a Greek University. Interesting insights from this analysis shows a significant difference in time to graduation for male and female students where female students, in general, have also obtained higher graduation marks.

KEYWORDS: Gender parity index, higher education, time to graduation.

1 Introduction

Gender is considered to have a fundamental influence on research in higher education. Access and enrolment to higher education have their own corresponding importance in higher education research involving gender. More reecently the research has focused on students' outcomes, where gender has its own relevance in terms of students' and institutions' success and performance or students' and institutions' efficacy, effectiveness and efficiency. Even though there is no consensus regarding the definition and measurement, those most commonly used fit into two categories; degree completion (percentage of degrees completed, non-completed, or rates of completion, drop-out rates) and time-to-degree, more generally considered as length of studies. The focus of this current research is on students' length of studies defined as the time duration between date of first enrolment to a university institution and up to the occurrence of an event that terminates studies in this same university. This paper draws on research of one individual level data set derived from social sciences-oriented departments in a University in Greece. In this institution 46 months is the minimum time for graduation but there is no maximum. In such data all possible situations of termination of study can occur for instance, if the student graduates on time, or drops out from their course. The paper is structured as follows. In the second section, we describe the GPI, while in Section 3 we report the results and section 4 presents the study's conclusions.

2 The Gender Parity Index

The Gender Parity Index (GPI) (UNESCO, 2017) is a socioeconomic index designed to measure the relative access to education of males and females. This index is commonly used by international organizations, such as UNESCO, but it is poorly mentioned by the literature accounting for gender differences (Hippe & Perrin, 2017). The GPI at t time is defined as follows:

$$GPI_t = \frac{Ind_{Ft}}{Ind_{Mt}} \tag{1}$$

where Ind_{Ft} is the female value of an indicator at t time, while Ind_{Mt} is the male value of the same indicator at t time. A GPI value equal to 1 indicates parity between females and males. In general, a value less than 1 for GPI indicates a disparity in favour of males and a value greater than 1 indicates a disparity in favour of females. The interpretation should be the other way round for indicators where normally, the approach to 0% is the ideal (e.g. repetition, dropout, illiteracy rates, etc). In these cases, a GPI of less than 1 indicates a disparity in favour of females and a value greater than 1 indicates a disparity in favour of males.

3 Data and results

The majority of university undergraduate curricula in Greece takes the form of four academic years; exceptions correspond to medicine engineering, veterinary science and agriculture. Graduation is possible at the end of the prescribed time interval if a certain number of course units have been successfully completed by the students. Students can graduate at exactly 46 months after the date of their first enrolment. Students who fail to do so can proceed to the next examination period for an unlimited number of times until the course unit condition is satisfied and they graduate. Student data are provided by Panteion University in Greece. The focus is on four cohorts of students who enrolled at the university for the first time from September 2000 to September 2003.

Enrolled students who transferred from other universities are excluded. Moreover, students who dropped out during the study time at the university are excluded. Every student was observed from the enrolment up to 40 months after the minimum legal duration of studies. The study data consist of 6219 students that are still enrolled at the beginning of the observation period. There is 70% of students who are female. It is interesting to consider the GPI in the following way:

$$GPI_t = \frac{S(t)_F}{S(t)_M} \tag{2}$$

with respect to the estimates of Kaplan-Meier survival function (Kaplan & Meier, 1958) for each time from after the minimum legal duration of studies. If GPI is less than 1, this indicates a disparity in favour of female. In fact, it means that at t time the proportion of ”survived” female students (still enrolled at the university) is lower than the same proportion for male students. On the contrary, if GPI has a value greater than 1, this indicates a disparity in favour of males. In Figure 1 Kaplan-Meier survival functions for male and female (left panel) and the GPI (right panel) are reported. The Kaplan-Meier estimates show that graduation rates are lower for male students for each $t$time, even if in the first months the graduation rates for male and female are very similar. The log-Rank test indicates to reject the hypothesis that the survival curves for females and males are identical ($z = -15.354$, $p - value < 0.00001$) hence suggesting a different student behaviour with respect to gender. The GPI assumes to always have a value lower than 1, underling a disparity in favour of females. The shape of the curve shows it is constantly decreasing in the first 20 months of TGaT then it decreases slowly and it seems to be constant at 0.6 after 30 months of TGaT, showing that the proportion of female students still enrolled to be lower than 40% with respect to the same enrolled male students.

4 Conclusions

This paper reports the analysis of the time it takes undergraduate students to complete their degree beyond that of the University’s expected time period. The analysis provides empirical evidence of gender gaps on length of studies for a Greek University in social sciences departments. Like most public institutions these universities have not publicly reported on whether gender is a factor in length of studies and timely graduation. This paper confirms through the use of GPI and survival functions that gender differences do exist.

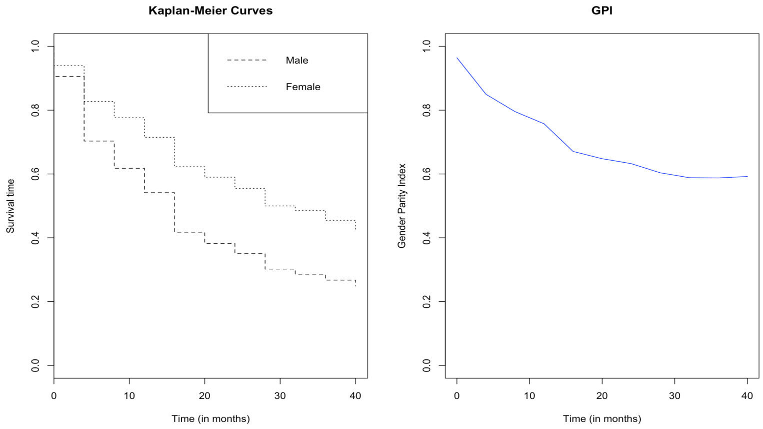


Figure 1. *Kaplan-Meier survival functions and GPI curve.*

References

KAPLAN, E., & MEIER, P. 1958. Nonparametric estimation from incomplete observations. *Journal or American Statistical Association.*, 53, 457–481.

HIPPE, R., & PERRIN, F. 2017. Gender equality in human capital and fertility in the European regions in the past. *Investigaciones de Historia Económica.*, 13(**3**), 166–179.

UNESCO. 2017. Accountability in education: meeting our commitments. *Global education monitoring report.*

SOME ASYMPTOTIC PROPERTIES OF MODEL SELECTION CRITERIA IN THE LATENT BLOCK MODEL

Christine Keribin[1, 2]

[1] Laboratoire de Mathématiques d'Orsay, Université Paris-Sud, CNRS, Université Paris-Saclay, (e-mail: christine.keribin@math.u-psud.fr)
[2] INRIA Saclay - Île de France, Équipe CELESTE

ABSTRACT: Co-clustering designs in a same exercise a simultaneous clustering of the rows and the columns of a data array. The Latent Block Model (LBM) is a probabilistic model for co-clustering, based on a generalized mixture model. LBM parameter estimation is a difficult problem as the likelihood is numerically untractable. However, deterministic or stochastic strategies have been designed and the consistency and asymptotic normality have been recently solved when the number of blocks is known. We address model selection for LBM and propose here a class of penalized log-likelihood criteria that are consistent to select the true number of blocks for LBM.

KEYWORDS: Latent block model, co-clustering, model selection, BIC, ICL.

1 Introduction

Clustering is an essential unsupervised tool to discover hidden structure from data by detecting groups of observations that are similar within a group and dissimilar from one group to another one. The challenge of modern data is to learn from observations $x_i \in R^d$ with a large number n of units observed on a large number d of variables, and the question is not only to cluster the observations, but also to cluster simultaneously the observations and the variables, leading to a tremendous parsimonious data representation.

This is called co-clustering and has many applications in many fields such as recommendation systems (to cluster simultaneously customers and goods), text mining (to co-cluster words and documents), genomics (to co-cluster genes and experimental conditions) for example. As for clustering, there are many ways to perform co-clustering, and we will focus here on the latent block model (LBM). We present the model and its asymptotical properties. In particular, we shall analyze the log-likelihood ratio under model order misspecifications, and derive a class of penalized log-likelihood criteria asymptotically consistent, results that are new for LBM.

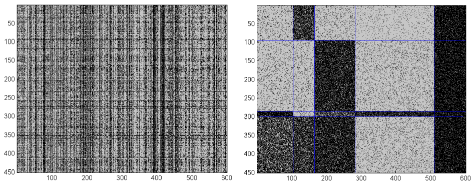


Figure 1. $n \times d = 450 \times 600$ *observations (left) and their reorganization according to the underlying structure in* 4×5 *blocks (right)*

2 The latent block model

LBM is a probabilistic model for co-clustering. Upon a data matrix $X = (x_{ij})$ of n rows and d columns, it defines a block clustering latent structure as the Cartesian product of a row partition $\mathbf{z}$ by a column partition $\mathbf{w}$ with three main assumptions:

- row assignments (or labels) $\mathbf{z}_i$, $i = 1, \ldots, n$, are independent from column assignments (or labels) $\mathbf{w}_j$, $j = 1, \ldots, d$: $p(\mathbf{z}, \mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$;
- row labels are independent, with a common multinomial distribution: $\mathbf{z}_i \sim \mathcal{M}(1, \pi = (\pi_1, \ldots, \pi_g))$; in the same way, column labels are i.i.d. multinomial variables: $\mathbf{w}_j \sim \mathcal{M}(1, \rho = (\rho_1, \ldots, \rho_m))$.
- conditionally to row and column assignments $(\mathbf{z}_1, \ldots, \mathbf{z}_n) \times (\mathbf{w}_1, \ldots, \mathbf{w}_d)$, the observed data X_{ij} are independent, and their (conditional) distribution $\varphi(., \alpha)$ belongs to the same parametric family, which parameter α only depends on the given block:

  $$X_{ij} | \{z_{ik} w_{j\ell} = 1\} \sim \varphi(., \alpha_{k\ell})$$

  where z_{ik} is the indicator membership variable of whether row i belongs to row-group k and $w_{j\ell}$ is the indicator variable of whether column j belongs to column-group ℓ.

Hence, the complete parameter set is $\theta = (\pi, \rho, \alpha)$, with $\alpha = (\alpha_{11}, \ldots, \alpha_{gm})$. With these assumptions, the likelihood of the *complete data* is

$$p(\mathbf{x}, \mathbf{z}, \mathbf{w}; \theta) = p(\mathbf{z}; \theta) p(\mathbf{w}, \theta) p(\mathbf{x} | \mathbf{z}, \mathbf{w}; \theta) = \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij}; \alpha_{k\ell})^{z_{ik} w_{j\ell}}$$

The labels are usually unobserved, and the *observed likelihood* is obtained by marginalization over all the label configurations:

$$p(\mathbf{x};\theta) = \sum_{\mathbf{z}\in\mathcal{Z},\mathbf{w}\in\mathcal{W}} \left(\prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi\left(x_{ij};\alpha_{k\ell}\right)^{z_{ik}w_{j\ell}} \right)$$

LBM deals with matrix of homogeneous data, such as binary (Govaert & Nadif, 2008), Gaussian (Lomet, 2012), categorical (Keribin *et al.*, 2015) or count (Govaert & Nadif, 2010) data. It involves a double missing data structure **z** for rows and **w** for columns, and the observed likelihood can not be factorized as a product of the mixing density as for simple mixture models. This implies that the likelihood is rapidly not tractable numerically even for few observations and few blocks, as the marginalization involves $k^n \times d^m$ terms. The estimation can however be performed either with numerical approximations (such as variational methods) or with Bayesian approaches (VBayes algorithm or Gibbs sampling).

3 Asymptotic properties

The double missing structure also leads to a very challenging and interesting study to state the asymptotic behavior of the maximum likelihood (MLE) and variational (VE) estimators. This question was first studied on the Stochastic Block Model (SBM) which is a LBM with the same statistical units in rows and columns, used to model graph adjacency matrices. In this case, there is only one set of latent variables **z**. Celisse *et al.*, 2012 first proved that under the true parameter value, the conditional distribution of the assignments of a binary SBM converges to a Dirac of the real assignments. Assuming the existence of an estimator of α converging at rate at least n^{-1}, they obtained the consistency of MLE and VE. Mariadassou & Matias, 2015 presented a unified framework for LBM and SBM for observations coming from an exponential family, but cannot get rid off the previous assumption to prove consistency. Using a different approach, Bickel *et al.*, 2013 showed for binary SBM (i) the consistency and aymptotic normality of the MLE in the complete model where the labels are known (ii) these properties can be transferred to the MLE of the observed model. Recently, Brault *et al.*, 2017 solved the consistency and the asymptotic normality of the MLE and VE for LBM observations coming from an exponential family.

These results were obtained when the true order $(K \times L)$ of the model is known. The question of the choice of K and L is crucial, and well-posed in the probability framework of LBM. Let K' (resp. L') be misspecifications of the number of row (resp. column) clusters. In this talk, we will study the likelihood ratio statistics

$$D_{KK',LL'} = \log \frac{\sup_{\theta \in \Theta_{K',L'}} p(\mathbf{x};\theta)}{\sup_{\theta \in \Theta_{K,L}} p(\mathbf{x};\theta)}$$

for $K' \neq K$ or $L' \neq L$ or both. Extending Wang *et al.*, 2017 methodology for SBM, we deal with the LBM double asymptotic in row and column to provide an appropriate penalty term and define a class of selection criteria asymptotically consistent.

References

BICKEL, P., CHOI, D., CHANG, X., ZHANG, H., *et al.* 2013. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, **41**(4), 1922–1943.

BRAULT, V., KERIBIN, C., & MARIADASSOU, M. 2017. Consistency and asymptotic normality of latent blocks model estimators. *arXiv preprint arXiv:1704.06629*.

CELISSE, A., DAUDIN, J.-J., PIERRE, L., *et al.* 2012. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, **6**, 1847–1899.

GOVAERT, G., & NADIF, M. 2008. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, **52**(6), 3233–3245.

GOVAERT, G., & NADIF, M. 2010. Latent block model for contingency table. *Communications in Statistics—Theory and Methods*, **39**(3), 416–425.

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.

LOMET, A. 2012. *Sélection de modèles pour la classification de données continues*. Ph.D. thesis, Université Technologique de Compiègne.

MARIADASSOU, M., & MATIAS, C. 2015. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, **21**(1), 537–573.

WANG, YXR., BICKEL, P., *et al.* 2017. Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, **45**(2), 500–528.

INVARIANT CONCEPT CLASSES FOR TRANSCRIPTOME CLASSIFICATION

Hans Kestler[1], Robin Szekely[1], Attila Klimmek[1] and Ludwig Lausser[1]

[1] Institute of Medical Systems Biology, Ulm University, Germany,
(e-mail: `firstname.lastname@uni-ulm.de`)

ABSTRACT: The field of classification is famous for its tremendous number of structural concept classes suitable for categorizing objects. Nevertheless, they are more likely to be chosen according to ad–hoc simulations than by sophisticated considerations on their theoretical properties. In this work, we discuss the idea of invariances properties as an a priori criterion for concept class selection. These invariances describe the data transformations that cannot affect the predictions of any member of the concept class.

As an example, we outline the landscape of linear classifiers for transcriptome classification and report four linked subclasses with distinct invariances. We show that the corresponding structural constraints may be incorporated in learning algorithms for general linear classifiers, such as linear support vector machines.

Surprisingly, we were able to attain comparable or even superior generalisation abilities to the linear one on the 27 investigated RNA-Seq and microarray data sets. This indicates that a-priori chosen invariant models can replace ad-hoc robustness analysis by interpretable and theoretically guaranteed properties in transcriptome categorization.

KEYWORDS: classification, invariance, linear classifier.

CLUSTERING OF TIES DEFINED AS SYMBOLIC DATA

Luka Kronegger[1]

[1] University of Ljubljana, Slovenia, (e-mail: luka.kronegger@fdv.uni-lj.si)

ABSTRACT: In the talk we are presenting the analysis of UK road network in which ties are defined as symbolic objects. The data descriptions of the units are called ”symbolic” when they are more complex than the standard ones due to the fact that they contain internal variation and are structured (Diday 2012). In our particular case the data are discrete distributions that present an overall annual traffic counts on road sections by vehicle types. In the analysis we used clamix package (Korenjak-Černe et.al 2011) available in R, to cluster ties into several categories applied to further analyzed and visualized road network.

KEYWORDS: network analysis, clustering, symbolic data, traffic.

APPLICATION OF DATA MINING IN THE HOUSING AFFORDABILITY ANALYSIS

Viera Labudová[1] and Ľubica Sipková[1]

[1] Faculty of Economic Informatics, University of Economics in Bratislava,
(e-mail: viera.labudova@euba.sk, lubica.sipkova@euba.sk)

ABSTRACT: Data mining is lately one of fastest growing new disciplines oriented on gaining knowledge from databases. Data mining uses artificial intelligence techniques, neural networks and advanced statistical tools (such as cluster analysis) to reveal trends, patterns and relationships, which might otherwise have remained undetected. This article describes the use of predictive data mining tools in housing affordability analysis.

KEYWORDS: EU SILC, housing affordability, logistic regression, decision trees, neural networks.

1 Introduction, Data and Methods

Housing affordability represents a challenge everyone faces when covering the costs of their current or potential housing and costs unrelated to their housing within the limits of their own income. One of the first definitions of housing affordability is provided by Howenstine: ”The ability of the household to acquire decent accommodation by the payment of a reasonable amount of its income on shelter”. In fequently cited definition of housing affordability by MacLennan and Williams affordability is concerned with securing some given standard of housing (or different standards) at a price or rent which does not impose, in the eyes of some third party (usually government) an unreasonable burden on household incomes. Wong and Sendi consider the lack of a definition of the term “unreasonable burden”. An explanation of the last term in the definition of “to be a detriment” is necessary to be expressed more accurately for measuring purposes.

The European Union uses an indicator-based approach to quantifying housing affordability, in which the household cost burden is calculated. The HCB (*household cost burden*) is defined as the ratio of housing costs (HH070*12 – annual total) less housing allowances (HY070G – annual total) to total available household income (HY020 – annual total), less housing allowances (in percentage after multiplying by 100):

$$HCB = \frac{HH070*12 - HY070G}{HY020 - HY070G}*100$$

The value of the variable HCB is assigned to every person living in the given household. The binary dependent variable has been created for modelling which equals ‘1’ if an person lives in a household where total housing costs (net of housing allowances) represents more than 40% of the household’s total disposable income (net of housing allowances) and ‘0’ if not. The aim of this paper is to analyse the relationship between the characteristics of individuals and household cost burden (HCB) in Slovak Republic. The analysis was carried out using an individual-level data extracted from EU SILC 2016 cross-sectional component provided by the Statistical Office of the Slovak Republic (EU SILC 2016, UDB 27/04/2017). In this article, we compare results from logistic regression and artificial neural networks with other popular classification algorithms from the data mining field, such as decision tree.

All analyses were carried out with SAS Enterprise Miner 12.1 software, which is SAS’ solution for data mining. Building models with SAS Enterprise Miner enables the analyst to access a comprehensive collection of data mining tools through a graphical user interface and to create process flow diagrams. Figure 1 shows the process flow for modelling on the EU SILC dataset containing data on 14,101 inhabitants aged 16 years and over.

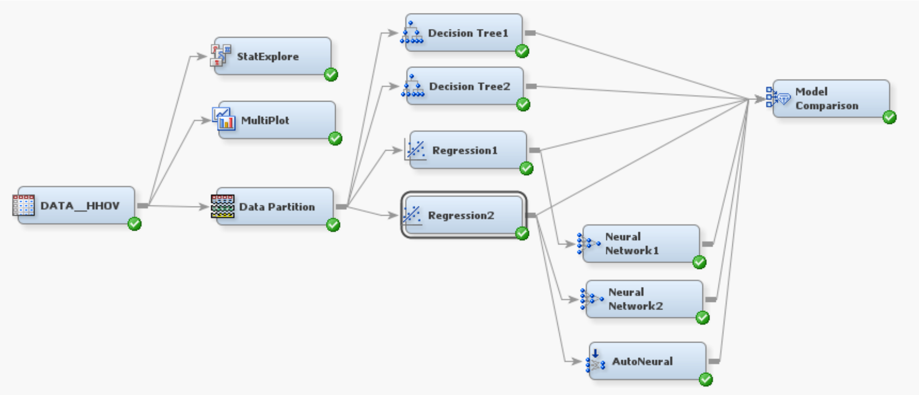


Figure 1. *Process flow diagram (Source: own elaboration)*

Train data subset (70% of the data) was used for preliminary model fitting. We tried to find the best model weights using this data set. The validation data set (30% of the data) was used to evaluate the adequacy of the model in the Model Comparison node.

Before creating neural network models, we reduced the number of input variables with stepwise elimination procedure in the logistic regression models (p-value > 0.05) (Hosmer & Lemeshow, 2004). We created two regression models Reg1 and Reg2. These models differ by using the AROPE variable. With the Reg1 model we selected these variables: AROPE (seven dummy variables indicating whether the person is either at risk of poverty, or severely materially deprived or living in a household with a very low work intensity; the reference category are persons not present in any sub-indicators), region (three dummy variables refers to the region of the residence of the household at the date of interview:

SK01/Bratislava Region, SK02/Western Slovakia and SK03/Central Slovakia; and SK04/Eastern Slovakia is our reference category), tenure status (four dummies indicating whether the person is owner paying mortgage, or tenant or subtenant paying rent at prevailing or market rate, or tenant or subtenant paying lower price than the market price, or tenant or subtenant who does not pay a rent; the reference category is outright owner), household type (eight dummy variables: single person, two adults younger than 65 years, households without dependent children, single person with dependent children, two adults with one dependent child, two adults with two dependent children, two adults with three or more dependent children, households with dependent children; two adults younger than 65 years is the reference category) and the logarithm of equalised household disposable income. In the second regression analysis (Reg2) we selected these variables: region, household type, tenure status and the logarithm of equalised household disposable income. The variable AROPE has been replaced by the following variables: poverty status (ARPT60i) (a dummy indicating whether the household's equalised disposable income (after social transfer) is above the at-risk-of-poverty threshold, which is set at 60 % of the national median equalised disposable income after social transfers), low work intensity of the household (LWI) (a dummy indicating whether the household's work intensity is not very low).

These variables have also been used in neural network models: (Neural network1 – with AROPE, Neural network2 and Autoneural without AROPE) (Kantardzic, 2003; Matignon, 2007).

In decision tree models, ordering the attributes for splitting is based on their entropy. For selection of variables it is important to work out how much the entropy of the entire training set would decrease if we choose each particular variable for the next classification step in a node of the tree (Kantardzic, 2003; Matignon, 2007). The most relevant variables in the first decision tree model (Decision Tree1) were: equalised household disposable income, household type and AROPE. In the second model (Decision Tree2) these were: equalised household disposable income, household type and low work intensity of the household.

2 Results

The Model Comparison tool selected the neural network model as the model with the smallest average squared error and the decision tree model with input variable AROPE as the model with the smallest validation misclassification rate (Table 1, Table2). Detected relationships facilitated identification of the factors with a significant influence on the housing cost burden for the inhabitants of Slovakia.

In addition to summary statistics, the Model Comparison tool also provides graphical model performance summaries. Plotting the trade-off between sensitivity and false positive fraction across all selected fractions of data creates a receiver operating characteristic (ROC) curve. For training and validation sample, ROC curves for logistic regression models, decision trees models and neural network models were analysed.The ROC chart on the validation data set showed the neural networks as the best models, followed by the regression models.

Table 1 *Fit Statistics Table for different models*

Fit Statistics (depth=10 %)	**Decision Tree1**	**Decision Tree2**	**Neural Network1**	**Neural Network2**	**ANN**	**Reg1**	**Reg2**
Average Squared Error	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Roc Index	0.88	0.87	0.94	0.94	0.93	0.93	0.93
Cumulative Percent Captured Response	70.57	70.07	68.48	70.29	67.75	69.57	67.39
Percent Captured Response	21.69	21.61	21.74	22.10	22.83	25.00	21.74
Misclassification Rate	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Mean Square Error	-	-	0.04	0.04	0.04	0.04	0.04
Cumulative Percent Response	45.93	45.61	44.58	45.75	44.1	45.28	43.87
Percent Response	28.24	28.14	28.3	28.77	29.72	32.55	28.3

Table 2 *Event Classification Table*

Model	**Data**	**FALSE Negative**	**TRUE Negative**	**FALSE Positive**	**TRUE Positive**	**Misclassification Rate**
Decission Tree1	TRAIN	365	9131	94	280	0.046505
Decission Tree1	VALID	156	3916	39	120	0.046088
Decission Tree2	TRAIN	367	9134	91	278	0.046403
Decission Tree2	VALID	158	3917	38	118	0.046325
Regression1	TRAIN	455	9140	85	190	0.05309
Regression1	VALID	183	3928	27	93	0.049634
Regression2	TRAIN	448	9142	83	197	0.053799
Regression2	VALID	187	3930	25	89	0.050106
Neural Network1	TRAIN	418	9119	106	227	0.05309
Neural Network1	VALID	169	3914	41	107	0.049634
Neural Network2	TRAIN	450	9151	74	195	0.05309
Neural Network2	VALID	182	3927	28	94	0.049634
AutoNeural	TRAIN	431	9133	92	214	0.052989
AutoNeural	VALID	177	3923	32	99	0.049397

This paper is a result of the research project VEGA 1/0770/17: Availability and affordability of housing in Slovakia.

References

HOSMER, D. W., & LEMESHOW, S. 2004. *Applied Logistic Regression*. USA: John Wiley & Sons, INC.

KANTARDZIC, M. 2003. *Data Mining: Concepts, Models, Methods, and Algorithms*. USA: J. Wiley and Sons.

MATIGNON, R. 2007. *Data Mining Using SAS Enterprise Miner*. USA: John Wiley & Sons, INC.

CYLINDRICAL HIDDEN MARKOV FIELDS*

Francesco Lagona[1]

[1] Department of Political Sciences, University of Roma Tre,
(e-mail: francesco.lagona@uniroma3.it)

ABSTRACT: Cylindrical hidden Markov fields are proposed as a parsimonious strategy to analyze spatial cylindrical data, i.e. bivariate spatial series of angles and intensities. These models are mixtures of copula-based bivariate densities, whose parameters vary across space according to a latent Markov random field. They enable segmentation of spatial cylindrical data within a finite number of latent classes that represent the conditional distributions of the data under specific environmental conditions, simultaneously accounting for spatial auto-correlation.

KEYWORDS: composite likelihood, copula, cylindrical data.

1 Introduction

Cylindrical spatial series are bivariate vectors of angles and intensities that are simultaneously observed at a number of sites in an area of interest. Their name is motivated by the special domain of these data, because the pair of an angle and an intensity can be described as a point on a cylinder. Cylindrical spatial series arise frequently in environmental and ecological studies. Examples include hurricane wind satellite data, wave directions and heights, speeds and directions of marine currents, as well as telemetry data of animal movement. The analysis of cylindrical spatial series is complicated by the cross-correlations between angular and linear measurements across space. Additional complications arise from the multimodality of the marginal distribution of the data, which are often observed under heterogeneous, space-varying conditions.

A cylindrical hidden Markov random field (MRF) model is proposed here to account for the specific features of cylindrical spatial series. The model is based on a mixture of copula-based cylindrical densities, whose parameters vary across space according to a latent Potts model. The Potts model is a categorical MRF, i.e. a multinomial process in discrete space, which fulfills a

*This work is supported by the 2015 PRIN project 'Environmental processes and human activities: capturing their interactions via statistical methods', funded by the Italian Ministry of Education, University and Scientific Research.

spatial Markovian property. It segments an area of interest according to an interaction parameter that captures the correlation between adjacent observations and controls the smoothness of the segmentation.

Hidden MRFs for data with circular components have been already proposed in the literature, by exploiting specific parametric distributions for circular and cylindrical data (Ranalli *et al.*, 2018; Ameijeiras-Alonso *et al.*, 2019). These proposals can be extended by considering copula-based cylindrical densities (Lagona, 2019). Copulas allow the marginal densities and the joint dependence structure to be modeled separately. As a result, they provide a general method for binding any pair of univariate marginal distributions together to form a bivariate distribution. This is particularly advantageous in the cylindrical setting, because a copula can be exploited to bind two marginal densities that do not necessarily have the same support.

2 A copula-based hidden Markov field

A cylindrical sample is a pair $\mathbf{z} = (x, y), x \in [0, 2\pi), y \in [0, +\infty)$. Let $f(x; \alpha)$ be a density on the circle, known up to a parameter α, with cumulative distribution function (cdf) $F(x; \alpha)$, defined with respect to a fixed, although arbitrary, origin. Moreover, let $f(y; \beta)$ be a density on the semi-line, known up to a parameter β, with cdf $F(y; \beta)$. Finally, let $g(u; \gamma), u \in [0, 2\pi)$ be a parametric circular density, known up to a parameter γ. Then,

$$f_q(\mathbf{z}; \theta) = 2\pi g\left(2\pi\left(F(x; \alpha) - qF(y; \beta)\right)\right) f(x; \alpha)) f(y; \beta)) \quad q = \pm 1 \qquad (1)$$

is a parametric cylindrical density with support $[0, 2\pi) \times (0, +\infty)$, known up to the parameter vector $\theta = (\alpha, \beta, \gamma)$, having the marginal densities $f(x; \alpha)$ and $f(y; \beta)$. Equation (1) is a typical example of a copula-based construction of a bivariate density, obtained by de-coupling the margins from the joint distribution. When the binding density g is the uniform circular distribution, say $g(x) = (2\pi)^{-1}$, then equation (1) reduces to the product of the marginal densities. Otherwise, the dependence between x and y is captured by the concentration of g: when g is highly concentrated, the dependence is high; when g is more diffuse, dependence is low. Finally, the constant $q = \pm 1$ determines whether the dependence between x and y is positive ($q = 1$) or negative ($q = -1$).

The Potts model is a multinomial process in discrete space with K classes. Given a lattice that divides an area of interest according to n observation sites $i = 1, \ldots, n$, a sample that is drawn from a spatial multinomial process is a segmentation of this area, obtained by associating each site with a segmentation label $k = 1,\ldots,K$. Formally, each observation site i is associated with a multinomial random variable $\mathbf{U}_i = (U_{i1},\ldots,U_{iK})$ with one trial and K classes. A specific segmentation of the area can be accordingly represented as a sample drawn from the multinomial process $\mathbf{U} = (\mathbf{U}_1,\ldots\mathbf{U}_n)$. Under a simple one-parameter specification, each segmentation $\mathbf{u}$ is associated with a single sufficient statistic $n(\mathbf{u})$ that indicates the number of neighboring sites which share the same class $k \neq K$. Accordingly, the probability of a specific segmentation $\mathbf{u}$ is known up to a single parameter ρ and it is given by

$$p(\mathbf{u};\rho) = \frac{\exp(\rho n(\mathbf{u}))}{W(\rho)}, \tag{2}$$

where $W(\rho)$ is the normalizing constant. The parameter ρ is an autocorrelation parameter: if it is positive (negative), then it penalizes segmentations with a few concordant (discordant) neighbors.

The specification of the cylindrical hidden MRF is completed by assuming that the cylindrical observations at the n sites of an areal partitioning are conditionally independent, given a segmentation generated by the Potts model. Formally, we assume that the conditional distribution of the observed process $\mathbf{z} = (\mathbf{z}_i, i = 1,\ldots,n)$, given the latent process, takes the form of a product density, say

$$f(\mathbf{z}|\mathbf{u};\theta) = \prod_{i=1}^{n}\prod_{k=1}^{K} f_q(\mathbf{z}_i;\theta_k)^{u_{ik}}, \tag{3}$$

where $\theta = (\theta_1 \ldots \theta_K)$ includes K label-specific parameters and $f_q(\mathbf{z};\theta_k), k = 1,\ldots,K$ are K copula-based densities defined in (1) and known up to the label-specific vector of parameters θ_k. Under this setting, the segmentation labels generated by the Potts model can be interpreted as latent classes, which cluster observation sites according to label-specific cylindrical distributions.

3 Parameter estimation and data segmentation

The maximum likelihood estimates, $\hat{\rho}$ and $\hat{\theta}$, of the parameters can be in principle obtained by maximizing the likelihood function

$$L(\rho,\theta;\mathbf{z}) = \sum_{\mathbf{u}} p(\mathbf{u};\rho) f(\mathbf{z} \mid \mathbf{u};\theta). \tag{4}$$

These parameter estimates can be usefully exploited to infer a posterior segmentation of the study area, by computing the posterior probabilities $p(u_{ik} =$

$1 \mid \mathbf{z};\hat{\rho},\hat{\theta})$ and exploiting a maximum-a-posterior (MAP) criterion: site i is associated to class k if

$$p(u_{ik} = 1 \mid \mathbf{z};\hat{\rho},\hat{\theta}) > p(u_{ih} = 1 \mid \mathbf{z};\hat{\rho},\hat{\theta})$$

for each $h \neq k$. According to this rule, data are clustered according to the latent class that is conditionally expected at each location, given the observed data and the estimated parameters.

When $\rho = 0$, data are independent and the proposed hidden MRF reduces to a latent class model that involves K cylindrical densities. In this setting, standard EM algorithms for mixture models can be exploited to maximize the likelihood function and maximum likelihood estimates can be exploited to compute posterior class membership probabilities. However, by assuming $\rho = 0$, we take a latent class approach to spatial segmentation and the cylindrical observations are clustered according to similarities in the variables space, i.e. the cylinder $[0,2\pi) \times (0,+\infty)$. More generally, by allowing $\rho \neq 0$, we account for the redundancy of the data which is due to spatial correlation. As a result, on the one side, taking a hidden MRF approach to segmentation, data clustering is not only based on similarities in the variables space, but also on similarities that occur in a spatial neighborhood. On the other side, assuming spatial dependence complicates maximum likelihood estimation and requires special approximation methods. We propose a computationally feasible EM algorithm to estimate the parameters of the model, by relying on composite likelihood methods that have been recently developed for hidden Markov fields (Ranalli *et al.*, 2018).

References

AMEIJEIRAS-ALONSO, J., LAGONA, F., RANALLI, M. & CRUJEIRAS, R. M. 2019. A circular nonhomogeneous hidden Markov field for the spatial segmentation of wildfire occurrences. *Environmetrics*, **30**(2), e2501.

LAGONA, F. 2019. Copula-based segmentation of cylindrical time series. *Statistics & Probability Letters*, **144**, 16 – 22. Advances in statistical methods and applications for Climate change and Environment.

RANALLI, M., LAGONA, F., PICONE M. & ZAMBIANCHI, E. 2018. Segmentation of sea current fields by cylindrical hidden Markov models: a composite likelihood approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**(3), 575–598.

COMPARING TREE KERNELS PERFORMANCES IN ARGUMENTATIVE EVIDENCE CLASSIFICATION

Davide Liga[1]

[1] CIRSFID, Università degli Studi di Bologna, (e-mail: davide.liga2@unibo.it)

ABSTRACT: The purpose of this study is to deploy a novel methodology for classifying argumentative support (or *evidence*) in arguments. The methodology shows that Tree Kernel can discriminate between different types of argumentative evidence with high scores, while keeping a good generalization. Moreover, the results of two different Tree Kernels are evaluated.

KEYWORDS: argument mining, argumentation, tree kernels, evidence detection.

1 Introduction, the Argument Mining pipeline

Argument Mining is relatively new field in the scientific community and several works have been written about this topic in the last few years (Cabrio & Villata, 2018, Lippi & Torroni, 2015). Broadly, its aim is to detect argumentative units from data and predict their relations. The achievement of this aim is not trivial and involves the resolution of multiple problems. In fact, Argument Mining can be seen as a a multifaceted problem and it is often considered as a pipeline. For example, Cabrio & Villata (Cabrio & Villata, 2018) described it as a pipeline composed of two steps, where the first step involves the identification of arguments and the second involves the prediction of argument relations. The first step includes both the classification *argumentative* vs *non-argumentative* and the identification of the arguments' components (claims, premises, etc.) along with their boundaries. The second step comprises predicting the nature of argument relations (e.g. *supports*, *attacks*) and the links between evidences and claims. The two steps are strictly dependent on the underlying argumentative model (e.g. the Waltonian claim/evidence dichotomy).

In this paper, a further step is considered, which involves fitting the achieved argumentative units into an Argument Schemes model, e.g. Walton's classification of Argument Schemes (Walton *et al.* , 2008). To achieve this aim, it is crucial to create classifiers capable of differentiating among different kinds of argumentative evidence (e.g. argument from Expert Opinion, argument from

Example). The proposed methodology is based on a Tree Kernel approach able to discriminate between different kinds of argumentative support.

2 Related Works

This work presents a method for classifying evidence typology within arguments using Tree Kernels (Moschitti, 2006), since being able to classify different kinds of support is crucial when dealing with Argument Schemes.

The advantage of Tree Kernels is the possibility to calculate similarities between different tree-structured data instead of engineering sophisticated features. Tree Kernels have already been used successfully in several NLP-related works. However, the application of Tree Kernel in the domain of Argument Mining has been relatively limited. One of the first implementations was presented by Rooney et Al. (Rooney *et al.* , 2012). Three years later, Lippi and Torroni suggested to exploit the ability of Tree Kernels of leveraging structural information to detect arguments (which can be considered the *first step* in the above-mentioned Argument Mining pipeline) (Lippi & Torroni, 2015).

This work is the continuation of a previous work (Liga, 2019) which aimed to classify argumentative support. A similar work (Liga & Palmirani, 2019) aimed to to classify argumentative opposition. Both these studies show that combining Tree Kernels and TFIDF vectorization can be a good strategy for this kind of classification. Particularly, the present approach tries to discriminate between two different kinds of evidence (or *premise*), comparing two different Tree Kernel functions.

3 Methodology

Following the method in Liga, 2019, two important Argument Mining datasets have been considered: the first (DS1) is taken from Al Khatib et al. (Al Khatib *et al.* , 2016) the second (DS2) from Aharoni et al. (Aharoni *et al.* , 2014). These two datasets have been built for different tasks but they share a very similar labelling system, which is the reason why the can be used jointly. More precisely, DS1 and DS2 classify argumentative texts depending on three common labels (i.e. Study/Statistics, Expert/Testimony, Anecdote). In particular, only the first two labels have been considered, with the aim of classifying evidences *from study* and *from expert*.

Two groups of classifiers were created using KeLP (Filice *et al.* , 2015), the first group was trained on DS1 while the second on DS2. For each classifier, a combination of a Linear Kernel and a Tree Kernel was employed, using a

GROUP 1	Performance on DS1					
	TFIDF + PTK			**TFIDF + SPTK**		
	P	R	F1	P	R	F1
Study	0.90	0.87	0.88	0.91	0.89	0.90
Expert	0.89	0.91	0.90	0.90	0.92	0.91
Average F1 (macro)	**0.89**			**0.91**		
Performance on DS2						
Study	0.74	0.68	0.71	0.78	0.66	0.72
Expert	0.76	0.80	0.78	0.75	0.85	0.80
Average F1 (macro)	**0.75**			**0.76**		

GROUP 2	Performance on DS2					
	TFIDF + PTK			**TFIDF + SPTK**		
	P	R	F1	P	R	F1
Study	0.69	0.69	0.69	0.69	0.69	0.69
Expert	0.74	0.74	0.74	0.74	0.74	0.74
Average F1 (macro)	**0.72**			**0.72**		
Performance on DS1						
Study	0.83	0.80	0.82	0.82	0.80	0.81
Expert	0.86	0.87	0.86	0.85	0.87	0.86
Average F1 (macro)	**0.84**			**0.84**		

Table 1. *Results of the two groups of classifiers (P=precision, R=recall, F1=F1 score)*

TFIDF vectorization and a GRCT (Grammatical Relation Centered Tree) representation. For the choice of the Tree Kernel function, two strategies have been attempted: the first deploys a Partial Tree Kernel (PTK, Moschitti, 2006), while the second deploys a Smoothed Partial Tree Kernel (SPTK, Croce *et al.* , 2011). To be sure that the classifiers were able to generalize, they were tested both on DS1 and on DS2 to detect whether sentences where an evidence from "Study/Statistics" or from "Testimony/Expert".

4 Results

As can be seen from Table 1, SPTKs outperform PTKs in group 1, while their performances in group 2 are mostly equal. Importantly, both Partial Tree Kernels and Smoothed Partial Tree Kernels keep a high degree of generalization, which is one of the main reasons why this methodology can be valuable for many classification problems in the Argumentation Mining pipeline.

References

AHARONI, E., POLNAROV, A., LAVEE, T., HERSHCOVICH, D., LEVY, R., RINOTT, R., GUTFREUND, D., & SLONIM, N. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. *Pages 64–68 of: Proceedings of the First Workshop on Argumentation Mining*.

AL KHATIB, K., WACHSMUTH, H., KIESEL, J., HAGEN, M., & STEIN, B. 2016. A news editorial corpus for mining argumentation strategies. *Pages 3433–3443 of: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

CABRIO, E., & VILLATA, S. 2018. Five Years of Argument Mining: a Data-driven Analysis. *Pages 5427–5433 of: IJCAI*.

CROCE, D., MOSCHITTI, A., & BASILI, R. 2011. Semantic convolution kernels over dependency trees: smoothed partial tree kernel. *Pages 2013–2016 of: Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM.

FILICE, S., CASTELLUCCI, G., CROCE, D., & BASILI, R. 2015. Kelp: a kernel-based learning platform for natural language processing. *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, 19–24.

LIGA, D. 2019. Argumentative Evidences Classification and Argument Scheme Detection Using Tree Kernels. *Pages 92–97 of: Proceedings of the 6th Workshop on Argument Mining*.

LIGA, D., & PALMIRANI, M. 2019. Detecting "Slippery Slope" and other argumentative stances of opposition using Tree Kernels in monologic discourse. *In: Rules and Reasoning. Third International Joint Conference, RuleML+RR 2019*.

LIPPI, M., & TORRONI, P. 2015. Argument mining: A machine learning perspective. *Pages 163–176 of: International Workshop on Theory and Applications of Formal Argumentation*. Springer.

MOSCHITTI, A. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. *Pages 318–329 of: European Conference on Machine Learning*. Springer.

ROONEY, N., WANG, H., & BROWNE, F. 2012. Applying kernel methods to argumentation mining. *In: Twenty-Fifth International FLAIRS Conference*.

WALTON, D., REED, C., & MACAGNO, F. 2008. *Argumentation schemes*. Cambridge University Press.

RECENT ADVANCEMENT IN NEURAL NETWORK ANALYSIS OF BIOMEDICAL BIG DATA

Pietro Liò[1], Giovanna Maria Dimitri[1] and Chiara Sopegno[2]

[1] Department of Computer Science and Technology, University of Cambridge,
(e-mail: pl219@cam.ac.uk, gmd43@cam.ac.uk)

[2] Department of Mathematical Sciences, Politecnico di Torino,
(e-mail: chiara.sopegno@studenti.polito.it)

ABSTRACT: This talk is divided in two parts and presents insights into methodological aspects of recent statistical machine learning developments and large scale comparison of different architectures. We will first discuss the use of graph neural networks to analyse dynamical systems, cross modal, space and temporal systems and image data. We will describe the effect of hierarchical multi co-attention for Interpretable architectures using neurological data. In the second part we will focus on recent work on generative autoencoders to integrate image and omics data towards better understanding of breast cancer. We will present several autoencoder architectures that integrate a variety of cancer patient data types (e.g., multi-omics and clinial data). We perform extensive analyses of these approaches and provide a clear methodological and computational framework for designing systems that enable clinicians to investigate cancer traits and translate the results into clinical applications.

KEYWORDS: deep learning, graph representation, biomedical data.

References

OPOLKA, F.L., SOLOMON, A., CANGEA, C., VELIČKOVIĆ, P., LIÒ, P & HJELM R.D. 2019. Spatio-Temporal Deep Graph Infomax. *arXiv* preprint arXiv:1904.06316.

ZHU, J., YANG, G. & LIÒ, P. 2019. How Can We Make GAN Perform Better in Single Medical Image Super-Resolution? A Lesion Focused Multi-Scale Approach. *arXiv* preprint arXiv:1901.03419.

CANGEA, C., VELIČKOVIĆ, P., JOVANOVIĆ, N., KIPF, T. & LIÒ P. 2018. Towards sparse hierarchical graph classifiers. *arXiv* preprint arXiv:1811.01287.

VELIČKOVIĆ, P., FEDUS, W., HAMILTON, W.L., LIÒ, P., BENGIO, Y. & HJELM R.D. 2018. Deep graph infomax. *arXiv* preprint arXiv:1809.1034.

BIAS REDUCTION FOR ESTIMATING FUNCTIONS AND PSEUDOLIKELIHOODS

Nicola Lunardon[1]

[1] Department of Economics, Quantitative Methods and Business Strategy, University of Milano Bicocca, (e-mail: nicola.lunardon@unimib.it)

ABSTRACT: Bias reduction is a methodology that aims at lowering the asymptotic bias of a reference estimator. The effectiveness of the approach hinges on the bias function of the estimator, which must be calculated at the actual, albeit unknown, underlying distribution. If the postulated distribution is misspecified, then the resulting bias function is in error and bias reduction becomes ineffective. To circumvent this problem, it is proposed an empirical approximation to the bias function which is fully driven by the data, so that its validity does not rely upon the knowledge of the underlying distribution.

KEYWORDS: bias reduction, estimating function, model misspecification, pseudo-likelihood.

1 Introduction

Let $y_1,\dots,y_n$ be realisations of the random vectors $Y_1,\dots,Y_n$ having common distribution function $G(y)$, $y \subseteq \mathbb{R}^d$, $d \geq 1$. Suppose to specify a working model for $G(y)$ in which the only unknown quantity is the parameter $\theta \subseteq \mathbb{R}^p$, $p \geq 1$. A point estimator $\hat{\theta}$ for θ can be defined as the root of either an estimating function $\Psi(\theta) = \sum_{i=1}^n \psi(\theta;y_i)$ or as the maximiser of a pseudo loglikelihood function $PL(\theta) = \sum_{i=1}^n pl(\theta;y_i)$, where $\psi(\cdot;\cdot) : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^p$ and $pl(\cdot;\cdot) : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ are known functions (Godambe, 1991, Besag, 1974). The estimator $\hat{\theta}$ is asymptotically unbiased as long as the first Bartlett's identity holds, i.e., either $\mathbb{E}_G\{\Psi(\theta)\} = 0$ or $\mathbb{E}_G\{\partial PL(\theta)/\partial\theta\} = 0$, where $\mathbb{E}_G(\cdot)$ denotes expectation under $G(y)$.

Estimating functions and pseudolikelihoods are appealing as they provide both valid inferential results under minimal assumptions about $G(y)$ and are computationally efficient. However, flexibility and minimal assumptions may reflect on the bias of $\hat{\theta}$, which may not negligible for small to moderate sample sizes. It is then possible to resort to the ideas underlying the bias reduction approach by Firth, 1993. The methodology was originally conceived for the maximum likelihood estimator so that, unless otherwise stated, in the sequel $PL(\theta)$ stands for the loglikelihood function and $\Psi(\theta) = \partial PL(\theta)/\partial\theta$ for the score function. Bias reduction defines an estimator $\tilde{\theta}$ as the root of the modified score function

$$\Psi(\theta) + \left\{\frac{\partial\Psi(\theta)}{\partial\theta^T}\right\} b(\theta), \tag{1}$$

where $b(\theta) \subseteq \mathbb{R}^p$ is the leading term of the bias expansion of $\hat{\theta}$, i.e.,

$$\mathbb{E}_G\{\hat{\theta} - \theta\} = b(\theta) + r(\theta).$$

Under the assumption that $b(\theta) = O(m^{-1})$ and $r(\theta) = O(m^{-2})$, it can be proved that the leading term in the bias expansion of $\tilde{\theta}$ is $\tilde{b}(\theta) = O(m^{-2})$, meaning that $\tilde{\theta}$ has smaller asymptotic bias than $\hat{\theta}$ (Firth, 1993); here m is an index of information about θ and does not necessarily coincide with the sample size n.

The core of bias reduction is $b(\theta)$ whose calculation involves expectations at the underlying distribution $G(y)$, implying that bias reduction comes along with shortcomings. From the one hand, even mild forms of model misspecification can affect the expression of $b(\theta)$: when estimating functions and pseudolikelihoods are considered, the second's Bartlett identity fails and the expression of $b(\theta)$ given by Firth, 1993 needs to be revised, otherwise the asymptotics justifying bias reduction are ruled out (see, e.g., Lunardon & Scharfstein, 2017). From the other hand, the applicability of bias reduction is hampered whenever the numerical evaluation of the working model is infeasible or a working model cannot be specified; max-stable processes provides such an instance (Padoan *et al.* , 2010, Genton *et al.* , 2011).

2 Proposal

Because of the highlighted problems and in order to extend bias reduction outside the maximum likelihood framework, e.g., to general estimating functions and pseudolikelihoods, we devise an empirical approximation to the second summand in (1). Let $\Psi(\theta)$ be an estimating function for θ satisfying $\mathbb{E}_G\{\Psi(\theta)\} = 0$ and differentiable up to fifth-order. The proposed approximation is

$$a(\theta) = \partial \log\{|I_p + h^{-1}(\theta) j(\theta)|\}/\partial\theta,$$

where I_p is the identity matrix of order p, $|\cdot|$ is the determinant operator, and

$$h(\theta) = \partial\Psi(\theta)/\partial\theta^T, \quad j(\theta) = \sum_{i=1}^{n} \psi(\theta; y_i)\psi(\theta; y_i)^T.$$

The function $a(\theta)$ satisfies $a(\theta) = \left\{\partial\Psi(\theta)/\partial\theta^T\right\} b(\theta) + o_p(1)$ so it can be used to define the modified estimating function

$$\Psi(\theta) + a(\theta). \tag{2}$$

The estimator $\bar{\theta}$, defined as the root of (2), is a bias reduced version of $\hat{\theta}$ in that $\mathbb{E}_G\{\bar{\theta} - \theta\} = O(m^{-3/2})$.

The approximation $a(\theta)$ suggests how to achieve bias reduction when inference is rooted on a pseudolikelihood function. Once $h(\theta)$ and $j(\theta)$ are redefined as $h(\theta) = \partial^2 PL(\theta)/(\partial\theta\partial\theta^T)$ and $j(\theta) = \sum_{i=1}^n \{\partial pl(\theta; y_i)/\partial\theta\}\{\partial pl(\theta; y_i)/\partial\theta\}^T$, it is possible to define the modified pseudo loglikelihood function

$$PL(\theta) + \log\{|I_p + h^{-1}(\theta) j(\theta)|\}. \tag{3}$$

Denoted by $\bar{\theta}$ the maximiser of (3), the relation $\mathbb{E}_G\{\bar{\theta} - \theta\} = O(m^{-3/2})$ still holds because the first partial derivative of (3) matches the structure of (2).

3 Simulation study - Gaussian max-stable process

The density function of a Gaussian max-stable process observed at K site locations in $\mathbb{R}^2$ involves a number of summands given by the K-th Bell number (Genton *et al.* , 2011). The evaluation of the loglikelihood is therefore viable for few site locations and a computationally appealing replacement can be a pairwise loglikelihood. Nonetheless, bias reduction would be computationally infeasible as the calculation of $b(\theta)$ still involves integrals with respect to the joint density for the K site locations. We aim to define a bias reduced estimator $\bar{\theta}$ through (3) when $PL(\theta)$ is the pairwise loglikelihood by Padoan *et al.* , 2010.

The simulation experiment resembles the one in Padoan *et al.* , 2010, Sect. 4, so it is supposed that n independent replicates of a Gaussian max-stable process are observed at $K = 50$ sites locations in $[0,40] \times [0,40]$ and $\theta = (\sigma_1^2, \sigma_2^2, \sigma_{12}^2)$, where the components control the range of the spatial dependence in the bivariate density functions. The parameter value is $\theta = (2000, 3000, 1500)$, the considered sample sizes are $n = \{20, 40, 80, 160\}$, and the corresponding number of Monte Carlo simulations is $n_j N$, $N = 500$, $j = 1, \ldots, 4$.

The logarithm of the theoretical asymptotic rate for the bias of $\hat{\theta}$ and $\bar{\theta}$ are respectively $-\log n$ and $-(3/2)\log n$. In Figure 1, we contrast these quantities with their Monte Carlo counterparts, i.e., $\log|\hat{\mathbb{E}}_G\{\hat{\theta} - \theta\}|$ and $\log|\hat{\mathbb{E}}_G\{\bar{\theta} - \theta\}|$, as functions of $\log n$. The figure confirms the $O(n^{-1})$ rate for the bias of $\hat{\theta}$ and reveals that the theoretical rate for the bias of $\bar{\theta}$ is slightly conservative.

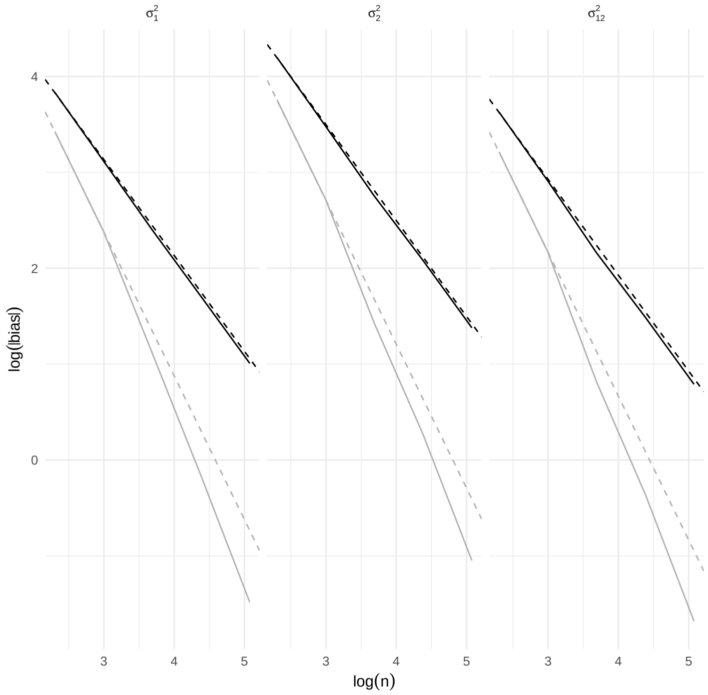


Figure 1. *Logarithm of the absolute bias of* $\hat{\theta}$ *(black) and* $\bar{\theta}$ *(gray): solid and dashed lines refer respectively to the Monte Carlo estimate and theoretical version.*

References

BESAG, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. B*, **36**, 192–225.

FIRTH, D. 1993. Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

GENTON, M., MA, Y., & SANG, H. 2011. On the likelihood function of Gaussian max-stable processes. *Biometrika*, 481–488.

GODAMBE, V. 1991. *Estimating functions*. Oxford University Press.

LUNARDON, N., & SCHARFSTEIN, D. 2017. Comment on Small sample GEE estimation of regression parameters for longitudinal data. *Statist. Med.*, **36**, 3596–3600.

PADOAN, S., RIBATET, M., & SISSON, S. 2010. Likelihood-based inference for max-stable processes. *J. Amer. Statist. Assoc.*, **105**, 263–277.

LARGE SCALE SOCIAL AND MULTILAYER NETWORKS

Matteo Magnani[1]

[1] InfoLab, Department of Information Technology, Uppsala University,
(e-mail: matteo.magnani@it.uu.se)

ABSTRACT: We present different ways of modelling information from social network sites based on a general data model known as multilayer network, and we discuss some approaches to identify communities in these networks using generalized clustering algorithms.

KEYWORDS: online social networks, multilayer networks, multiplex networks, temporal text networks, community detection.

1 Introduction and motivation

Social and information networks have been studied for a long time in disciplines such as social network analysis, and a core task in social network analysis for which clustering methods are commonly used is to identify communities, to explain why groups of entities (actors) belong together based on the explicit ties among them and/or the implicit ties induced by some similarity measures given some attributes of these entities. Since members of a community tend to generally share common properties, revealing the community structure in a network can provide a better understanding of the overall functioning of the network at large.

While social network analysis has often used simple graphs as a mathematical representation, reality is rarely mono-dimensional. A large amount of human-generated information is available online in the form of text exchanged between individuals at specific times, forming what we call human information networks. Examples include social network sites, online forums and emails.

This contribution focuses on the problem of clustering this complex information, that is, identifying parts of the data that are more similar or related to each other than to other parts of the data.

As a motivating example, consider Figure 1. One typical usage of social media data in research is to study how information propagates online. In one of the many studies on this topic, the authors have analyzed different aspects of the propagation process considering the online reactions generated by the death of a well-known Italian TV anchorman (Magnani *et al.*, 2010). In the figure

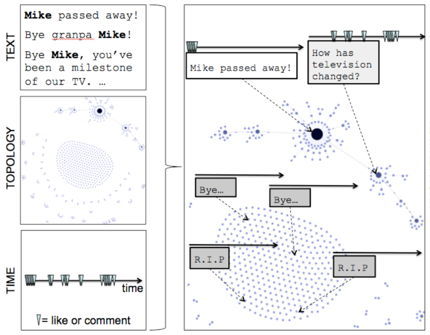


Figure 1. *Three aspects of a human information network: time, text and topology*

we have reproduced the text of some of the posts generated about this event, the information propagation network (topology), showing which posts contained information obtained by which others, and a temporal pattern indicating the number of comments per day. While each of these pieces of information alone reveals something, putting them together into a temporal text network (right-hand side) we obtain a much more comprehensive understanding of the process. On the one hand, we can see that for the posts representing explicit attempts to propagate information (e.g., *Mike passed away*) publication time is fundamental to determine their success, and only the first of this type of posts generated a large and sudden burst of reactions in a very short time; on the other hand, conversational posts evolving from it (e.g., *How has television changed?*) can appear later and still create long but less dense chains of reactions. Other posts not present in the information propagation network neither explicitly give the news nor ask for an answer, generating no or few reactions, but still have the role of re-activating the information cascade so that even the latecomers can find a trace of it; some of these posts (e.g., *Bye granpa Mike!*) form what has been called an online mourning ritual. In summary, time, text and topology together can lead to a deeper understanding of how this network evolved into its current status and how information propagated through it.

2 Multilayer networks: models, clustering and applications

To represent complex data such as the online human information networks mentioned above, several extended models have been proposed, such as networks of networks, multiplex networks, etc., leading to a fragmented literature on how to analyze complex network data. Recently, a model unifying several of the existing approaches has been defined, known as multilayer network model (Kivelä *et al.*, 2014; Dickison *et al.*, 2016). In this presentation we give a quick overview of these models, up to recent work on a data-cube-based version aiming at reconciling database models for graphs and the models used in network analysis.

Then we discuss different approaches to identify communities in multilayer networks used to represent human information networks. We start presenting a community detection method for multiplex networks (Tehrani *et al.*, 2018), showing how to extend an existing simple-graph method. The clique percolation method (Palla *et al.*, 2005) is based on the intuition that the presence of a community can be observed in a social network through the presence of cliques, that is, sets of actors who are all adjacent to each other. This method has a set of features that make it well-suited to the discovery of communities in social networks: (1) it allows to specify how much connectivity is necessary to recognize the presence of a community (minimum clique size k), (2) it allows the same actor to be present in multiple communities (overlapping), and (3) it does not force all actors to be part of a community (partial). Here we show how these features can be ported to multiplex networks. Then we discuss how to use multiplex network clustering approaches to include also the information exchanged among online users and the temporal traces of their interactions. In particular, we present the temporal text network model (Vega & Magnani, 2018) as a special type of multilayer network, showing that it captures the main features of existing approaches used in the literature, and we also show how clusters can be identified in this model.

We conclude presenting practical examples of the clustering approaches mentioned above when applied to Twitter data, to emphasize the challenges encountered when abstract algorithms are used in real contexts (Vega & Magnani, 2018; Hanteer *et al.*, 2018). Twitter data is significantly different from the data that can be extracted from other more structured social network platforms. On the one hand, on Twitter we can observe finer-grained social interactions: we can see if a user has mentioned another specific user, and we have also access to the text they exchanged. On the other hand, categorization of information on Twitter often relies on hashtags, that are not created through some centralized decision-making process. This can lead to different hashtags referring to the same topic, single hashtags used to refer to different topics in different tweets, as well as many tweets not explicitly mentioning any hashtag.

Ideally, we would like to use Twitter data to perform various types of analysis based on clustering: what is discussed in the tweets, where are the different topics appearing, when are they discussed (for example, when they emerge for the first time and when they reach a peak of popularity), who is leading the discussion on specific topics, and finally a joint analysis putting all these aspects together to map online conversations. In practice, some of these analyses are indeed possible given some assumptions and limitations, some are not, and to extract more knowledge from the data we may require frequent input from domain experts.

References

DICKISON, M. E., MAGNANI, M., & ROSSI, L.. 2016. *Multilayer Social Networks*. Cambridge University Press.

HANTEER, O., ROSSI, L., VEGA, D., & MAGNANI, M.. 2018. From Interaction to Participation: The Role of the Imagined Audience in Social Media Community Detection and an Application to Political Communication on Twitter. *Pages 531–534 of: ASONAM*. IEEE Computer Society.

KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y., PORTER, M. A. 2014. Multilayer Networks. *Journal of Complex Networks*, **2**(3), 203–271.

MAGNANI, M., MONTESI, D., & ROSSI, L.. 2010. Friendfeed breaking news: death of a public figure. *In: IEEE SocialCom*. IEEE Computer Society.

PALLA, G., DERÉNYI, I., FARKAS, I., & VICSEK, T.. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**(7043), 814–818.

AFSARMANESH, N., & MAGNANI, M.. 2018. Partial and overlapping community detection in multiplex social networks. *Pages 15–28 of: Social Informatics*. Lecture Notes in Computer Science, vol. 11186. Springer.

VEGA, D., & MAGNANI, M.. 2018. Foundations of Temporal Text Networks. *Applied Network Science*, **3**(1), 25:1–25:26.

Uncertainty in Statistical Matching by BNs

Daniela Marella[1], Paola Vicard[2] and Vincenzina Vitale[3]

[1] Dipartimento di Scienze della Formazione, University Roma TRE,
(e-mail: daniela.marella@uniroma3.it)

[2] Dipartimento di Economia, University Roma TRE,
(e-mail: paola.vicard@uniroma3.it)

[3] Dipartimento di Scienze Sociali ed Economiche, University Sapienza,
(e-mail: vincenzina.vitale@uniroma1.it)

Abstract: Statistical matching is a technique used for combining information when variables of interest are not jointly observed. In this paper we propose the use of Bayesian Networks to deal with the statistical matching problem. Bayesian networks admit a recursive factorization of the joint distribution useful both for data integration and for evaluating the statistical matching uncertainty in the multivariate context. The notion of uncertainty in statistical matching when BNs are used is discussed.

Keywords: Bayesian network, collapsibility, uncertainty.

1 Introduction

Let $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ be a multivariate random variable (rv) with joint discrete distribution P. Without loss of generality, let $\mathbf{X} = (X_1,\ldots,X_H)$, $\mathbf{Y} = (Y_1,\ldots,Y_K)$ and $\mathbf{Z} = (Z_1,\ldots,Z_T)$ be vectors of rvs of dimension H, K, T, respectively. Furthermore, let A and B be two independent samples of n_A and n_B independent and identically distributed records from $(\mathbf{X},\mathbf{Y},\mathbf{Z})$. Assume that $(\mathbf{X},\mathbf{Y})$ are observed in sample A while $(\mathbf{X},\mathbf{Z})$ are observed in sample B. Then, the units in A have $\mathbf{Z}$ missing values and the units in B have $\mathbf{Y}$ missing values.

Statistical matching aims at combining information obtained from different non-overlapping sample surveys. The main target is in estimating the joint probability distribution (pdf) of $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ from the samples A and B and constructing a complete synthetic data set where all the variables of interest are jointly observed, see D'Orazio *et al.*, 2006b.

The lack of joint observations on the variables $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ leads to uncertainty about the data generating model. In order to overcome this problem, three approaches can be distinguished. The first approach uses techniques based on the conditional independence assumption between $\mathbf{Y}$ and $\mathbf{Z}$ given

X (CIA assumption) see, e.g., Okner, 1972. The second approach uses techniques based on the external auxiliary information regarding the statistical relationship between **Y** and **Z**, as in Singh *et al.*, 1993. The third group of techniques addresses the so called *identification problem.* The sample information provided by A and B is actually unable to discriminate among a set of plausible models for $(\mathbf{X},\mathbf{Y},\mathbf{Z})$. Uncertainty in statistical matching is analyzed in Rässler, 2002, D'Orazio *et al.*, 2006a, Conti *et al.*, 2012, Conti *et al.*, 2013, Conti *et al.*, 2017 and Conti *et al.*, 2016 and references therein.

In this paper we propose the use of Bayesian networks (BNs) to deal with the statistical matching problem. The use of BNs is motivated by the following advantages: (i) BNs are widely used to describe dependencies among variables in multivariate distributions; (ii) BNs admit convenient recursive factorizations of their joint probability useful both for parameters estimation and for uncertainty evaluation in a multivariate context.

2 Statistical Matching by BNs

Bayesian networks (BN) are multivariate statistical models satisfying sets of conditional independence statements contained in a directed acyclic graph (DAG), see Pearl, 1995. Let $\mathbf{X} = (X_1,\ldots,X_H)$ be a random vector, then a BN specifies: (i) the set of conditional independence statements by means of a DAG and (ii) the set of conditional probability distributions associated to the nodes of the graph. The joint probability distribution can be factorized as follows:

$$P(x_1,\ldots,x_H) = \prod_{h=1}^{H} P\left(x_h|\mathrm{pa}(x_h)\right)$$

where $P\left(x_h|\mathrm{pa}(x_h)\right)$ is the probability distribution attached to node x_h given its parents $\mathrm{pa}(x_h)$, $h = 1,\ldots,d$, *i.e.* all the nodes linked to x_h by an arrow pointing to x_h.

As previously stressed, the statistical model of $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ is not identifiable on the basis of sample data, that is both the components of the BN (*i.e.* DAG and its parameters) can not be estimated from the available sample information. Then, the uncertainty can be decomposed as follows: (i) the uncertainty regarding the association structure, that is the presence or absence of an edge between the components of **Y** and **Z**; (ii) the uncertainty regarding the network parameters, that is the local probability distributions.

Let $P_{G_{XYZ}}$ be the pdf of $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ associated to the DAG G_{XYZ} and let us denote by $P_{G_{XY}}$ and $P_{G_{XZ}}$ the pdfs associated to G_{XY} and G_{XZ}, the DAGs estimated on sample A and B, respectively. First of all, the DAG G_X is estimated on the overall sample $A \cup B$. Secondly, given G_X, we proceed to estimate the association structure for $(\mathbf{X},\mathbf{Y})$ and $(\mathbf{X},\mathbf{Z})$ on the basis of sample data in A and B, respectively. As far as $P_{G_{XYZ}}$ is concerned, one can only say that it lies in the class of all joint probability distributions for $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ satisfying the estimate collapsibility over $\mathbf{Y}$ and $\mathbf{Z}$, respectively. Formally, we say that $P_{G_{XYZ}}$ is estimate collapsible over Z_t (a component of the vector $\mathbf{Z}$) if

$$\widehat{P}_{G_{XYZ}}(\mathbf{X},\mathbf{Y},\mathbf{Z}\backslash\{Z_t\}) = \widehat{P}_{G_{XYZ\backslash\{Z_t\}}}(\mathbf{X},\mathbf{Y},\mathbf{Z}\backslash\{Z_t\}). \quad (1)$$

That is, the estimate $\widehat{P}_{G_{XYZ}}(\mathbf{X},\mathbf{Y},\mathbf{Z}\backslash\{Z_t\})$ of $P_{G_{XYZ}}(\mathbf{X},\mathbf{Y},\mathbf{Z}\backslash\{Z_t\})$ obtained by marginalizing the maximum likelihood estimate (MLE) of $\widehat{P}_{G_{XYZ}}(\mathbf{X},\mathbf{Y},\mathbf{Z})$ under the original DAG model $(G_{XYZ}, P_{G_{XYZ}})$ coincides with the MLE under the DAG model $(G_{XYZ\backslash\{Z_t\}})$, see Kim & Kim, 2006. Estimate collapsibility over a set $\mathbf{Z}$ is defined similarly. Then, the class of plausible joint distributions for $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ can be described as follows

$$\mathcal{P}_{XYZ} = \{P_{G_{XYZ}} : \widehat{P}_{G_{XYZ}}(\mathbf{X},\mathbf{Y}) = \widehat{P}_{G_{XY}}(\mathbf{X},\mathbf{Y}), \widehat{P}_{G_{XYZ}}(\mathbf{X},\mathbf{Z}) = \widehat{P}_{G_{XZ}}(\mathbf{X},\mathbf{Z})\}2)$$

All the joint probability distributions in (2) are equally plausible. In terms of graphs, estimate collapsibility implies c-removability, such a concept allows to define the class $\mathcal{G}_{XYZ}$ of plausible DAGs given the available sample information.

Under the CIA, the class (2) is composed by a single pdf $P^{CIA}_{G_{XYZ}}$ corresponding to the DAG $G^{CIA}_{XYZ} = G_{XY} \bigcup G_{XZ}$ where $\mathbf{Y}$ and $\mathbf{Z}$ are d-separated by the set $\mathbf{X}$. Under the CIA, both the dependence structure and the BN parameters are estimable from the sample data. When the CIA does not hold, extra sample information or experts judgment can be used to choose a plausible joint probability distribution from the class (2).

Suppose that a DAG G^*_{XYZ} has been selected from the class of plausible DAGs $\mathcal{G}_{XYZ}$. Let $P^*_{G^*_{XYZ}}$ be the pdf associated to G^*_{XYZ}. According to G^*_{XYZ} the joint probability distribution $P^*_{G^*_{XYZ}}$ can be factorized into local probability distributions some of which can be estimated from the available sample information while other not due to the absence of joint observation on $\mathbf{Y}$ and $\mathbf{Z}$ variables.

In the case of categorical variables, uncertainty is dealt with in D'Orazio *et al.*, 2006a where parameters uncertainty is estimated according to the maximum likelihood principle and the set of maximum likelihood estimates is called likelihood ridge. In order to exclude some parameter estimates, it is important to introduce constraints characterizing the phenomenon under study.

These constraints can be defined in terms of structural zero and inequality constraints between pairs of distribution parameters, as specified in D'Orazio *et al.*, 2006a. Their introduction, as the introduction of auxiliary information regarding the association structure, is useful for reducing the overall parameter uncertainty.

References

CONTI, P.L., MARELLA, D., & SCANU, M. 2012. Uncertainty analysis in statistical matching. *Journal of Official Statistics.*, **28**, 69–88.

CONTI, P.L., MARELLA, D., & SCANU, M. 2013. Uncertainty Analysis for statistical matching of ordered categorical variables. *Computational Statistics & Data Analysis.*, **68**, 311–325.

CONTI, P.L., MARELLA, D., & SCANU, M. 2016. Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association.*, **111**, 1715–1725.

CONTI, P.L., MARELLA, D., & SCANU, M. 2017. How far from identifiability? A nonparametric approach to uncertainty in statistical matching under logical constraints. *Communication in Statistics: Theory and Methods.*, **46**, 967–994.

D'ORAZIO, M., DI ZIO, M., & SCANU, M. 2006a. Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints. *Journal of Offcial Statistics.*, **22**, 137–157.

D'ORAZIO, M., DI ZIO, M., & SCANU, M. 2006b. *Statistical Matching: Theory and Practice*. Chichester: Wiley.

KIM, S.H., & KIM, S.H. 2006. A Note on Collapsibility in DAG Models of Contingency Tables. *Journal of Official Statistics.*, **33**, 575–590.

OKNER, B. 1972. Constructing a new data base from existing microdata sets: the 1966 merge file. *Annals of Economic and Social Measurement.*, **1**, 325–342.

PEARL, J. 1995. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann Publishers Inc.

RÄSSLER, S. 2002. *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. New York: Springer.

SINGH, A.C., MANTEL, H., KINACK, M., & ROWE, G. 1993. Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. *Survey Methodology.*, **19**, 59–79.

EVALUATING THE RECRUITERS' GENDER BIAS IN GRADUATE COMPETENCIES

Paolo Mariani[1] and Andrea Marletta[1]

[1] Department of Economics Management and Statistics, University of Milano-Bicocca, (e-mail: paolo.mariani@unimib.it,andrea.marletta@unimib.it)

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

ABSTRACT: The study aims to observe the presence of different behaviours by entrepreneurs in the hiring processes of new graduates varying on the gender of the respondent. In particular, the analysis is based on the Education-for-Labour Elicitation from Companies' Attitudes towards University Studies Project involving 471 enterprises operating in Lombardy with 15 or more employees. The preference analysis of the recruiters is carried out using Conjoint Analysis.

KEYWORDS: Gender bias, graduates, conjoint analysis.

1 Introduction

In recent decades, tertiary-level Education has expanded rapidly across many countries, as well as in Italy. In general, the expectation is that higher education should prepare young people to become highly productive and successful in the labour market. Sometimes, the skills required of the graduates for the job do not coincide with the skills offered by the graduates applying, creating a mismatch between education and the labour market.

Beyond this mismatch, another bias could be generated by gender in the recruitment process. But, if there is a huge literature in gender gap during the recruitment process about politics or academic world (Hardin *et al.*, 2002, Van den Brink *et al.*, 2010), here the gender bias has been considered for the respondents and not for candidates. The aim of this paper is to understand if there exists a difference in the evaluation of possible candidates for a new hiring when the recruiter is male or female.

The paper is structured as follows: after this introduction, section 2 introduces the methodology of Conjoint Analysis, section 3 presents the results from the Electus research and section 4 gives the conclusions.

2 Conjoint Analysis

Conjoint Analysis (CA) is among the methods most used to analyse consumer choices and to assign consumers utility drawn from the properties of single characteristics of goods, services or, as in this application, jobs being offered on the market. In this paper, the conjoint rating response format is used to gather and use additional information about respondent's preferences. This preference model uses a part-worth utility linear function. Part-worth utilities are also assumed for each level of the various attributes estimated by using OLS multiple regression. In this formulation, attention is focused on a rating scale, opting for a very general preference model used in traditional CA. The utility function is defined as follows:

$$U_k = \sum_{i=0}^{n} \beta_{ij} x_{ik} \tag{1}$$

where x_0 is equal to 1 and n is the number of all levels of the attributes which define the combination of a given good. Each x_{ik} variable is a dichotomous variable, which refers to a specific attribute level. This variable equals 1 if the corresponding attribute level is present in the combination of attributes that describes the alternative k. Otherwise, that variable will be 0. As a result, the utility associated with the alternative k (U_k) is obtained by summing the terms β_{ij} x_{ik} over all attribute levels, where β_{ij} is the partial change in U_k for the presence of the attribute level i, with all other variables remaining constant.

The range of the utility values for each attribute from highest to lowest, provides an indicator of how important the attribute is compared to the others. The larger the utility ranges the more important is the role that the attributes play. For any attribute j, the relative importance can be computed by dividing its utility range by the sum of all utility ranges as follows:

$$I_j = \frac{\max(W_j) - \min(W_j)}{\sum_{j=1}^{J} [\max(W_j) - \min(W_j)]}, \tag{2}$$

where J is the number of attributes and W_j is the set of part-worth utilities referring to the various levels of attribute j. Usually, importance values are represented as percentages with a total score of one hundred.

3 Application and results

Data for this research concerns labour market comprehension policies for new graduates and the relationships among enterprises and universities. The study is based on the multi-centre research project, Education-for-Labour Elicitation from Companies' Attitudes towards University Studies (Fabbris & Scioni, 2015), which involved several Italian universities. Main results about this project have been already published (Mariani *et al.*, 2018b, Mariani *et al.*, 2018a). Here, the analysis is attempting to give a comparative view of the differences in terms of choice according the gender of the respondent to the questionnaire.

The survey was conducted in 2015 using Computer-Assisted Web Interviewing (CAWI). The questionnaire contained two sections: the first concerned the conjoint experiment for the five job positions and the second contained general information about the company. The five job positions considered for the new graduates, were Administration clerk, HR assistant, Marketing assistant, ICT professional and CRM assistant. Six attributes were used to specify the candidates' profile: Field of Study, Degree Level, Degree Mark, English Knowledge, Relevant Work Experience, Willingness to Travel. As far as the Milano-Bicocca research unit was concerned, there were 471 final respondents. The frequency distribution about gender was equally balanced with male (41%) and female (59%). For space reasons, here results about gender bias of the respondent are shown only for the Administration Clerk in table 1.

Table 1. *Best profile and importance indexes of competences for Administration Clerk*

	Respondent's gender			
	Male		**Female**	
Competence	**Best**	**Importance**	**Best**	**Importance**
Field of Study	Economics	40.47%	Economics	59.21%
English Knowledge	Suitable	20.81%	Suitable	12.62%
Relevant work experience	Regular	16.12%	Regular	13.72%
Degree Mark	High	14.30%	High	10.72%
Willingness to travel	Long	4.17%	Long	3.20%
Degree level	Bachelor	4.13%	Bachelor	0.53%

The gender bias is present not in the detection of the best profile but in terms of importance indexes. In fact, if on one hand, the best competencies for each attribute are equal generating the same best profile, on the other hand there is a lot of difference in computing the importance indexes of the attributes. When the respondent of the survey is a woman, Field of study has a very high importance index very close to 60%, this index is just over the 40% when the respondent is a man. Secondly, men tends to overestimate the importance of the English Knowledge, that is the second best competence with 20.81%. For women, the importance of this attribute is only 12.62%.

4 Conclusions

This paper proposed the use of Conjoint Analysis to measure a possible gender bias in the recruiters during a process of new hiring. Data referring to the Electus project applied in Lombardy, showed the existence of different kinds of attributes more or less important in addressing the choice of the candidate. Competencies were measured according to the perceived importance and *Field of Study* was proven to be the most relevant, whatever the gender of the respondent and Economics were preferred for the role of Administration Clerk. The high importance of *Field of Study* resulted different on the basis of the gender of the recruiter. Women overrated this importance with a value very close to 60%, and they underestimated the *English Knowledge* that was overtook by *Relevant work experience* as second best competence.

References

FABBRIS, L., & SCIONI, M. 2015. Dimensionality of scores obtained with a paired-comparison tournament system of questionnaire item. *Academic Proceedings of the 2015 University-Industry Interaction Conference*.

HARDIN, R, REDING, K, & STOCKS, M. 2002. The effect of gender on the recruitment of entry-level accountants. *Journal of Managerial Issues*, 251–266.

MARIANI, P, MARLETTA, A, MASSERINI, L, & ZENGA, M. 2018a. A latent class conjoint analysis for the Administrative Clerk figure: insights from ELECTUS. *Quality & Quantity*, 1–12.

MARIANI, P, MARLETTA, A, & ZENGA, M. 2018b. A New Relative Importance Index of Evaluation for Conjoint Analysis: Some Findings for CRM Assistant. *Social Indicators Research*, 1–14.

VAN DEN BRINK, M, BENSCHOP, Y, & JANSEN, W. 2010. Transparency in academic recruitment: a problematic tool for gender equality? *Organization Studies*, **31**(11), 1459–1483.

DYNAMIC CLUSTERING OF NETWORK DATA: A HYBRID MAXIMUM LIKELIHOOD APPROACH

Maria Francesca Marino[1] and Silvia Pandolfi[2]

[1] Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, (e-mail: mariafrancesca.marino@unifi.it)

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

ABSTRACT: Dynamic Stochastic Block Models (SBMs) represent an attractive field of research as they provide a flexible modeling tool for the dynamic clustering of network data. However, full maximum likelihood (ML) for this class of models is not a viable estimating strategy due to the intractability of the likelihood function: variational inference represents quite a standard alternative in the frequentist framework. However, despite its simplicity, it may lead to non-optimal estimators and may suffer from local maxima solutions. We extend the hybrid ML approach developed in the context of static SBMs to deal with dynamic networks also considering both weighted and unweighted relations as well as nodal attributes which may potentially affect the block structure.

KEYWORDS: stochastic blockmodels, latent Markov models, approximate inference.

1 Introduction

Stochastic Block Models (SBMs; e.g. Snijders & Nowicki, 1997) are widely employed in the social network literature when the focus is on clustering nodes with respect to their social behavior. Recently, a growing interest has focused on the evolution of networks over time. Dynamic SBMs (e.g., Yang *et al.*, 2011, Matias & Miele, 2017, Bartolucci *et al.*, 2018) assume that the probability of observing a connection between two nodes depends on the corresponding block membership only. The evolution of the latter over time is represented by a homogeneous, discrete, latent Markov chain. Nodes belonging to a given block (state) at a given occasion share similar social behaviors.

Besides the easiness of interpretation of the model, Maximum Likelihood (ML) inference remains problematic due to the intractability of the likelihood function. Among approximate solutions available in the statistical literature, the variational approach represents a quite standard choice in the frequentist framework (e.g. Matias & Miele, 2017).

Here, we extend the hybrid ML approach recently introduced by Bartolucci *et al.*, 2017 in the framework of static SBM to deal with its dynamic counterpart, also relaxing the homogeneity assumption underlying the latent structure of the model. Moreover, to account for both binary and valued relations, we further modify the estimation algorithm to accommodate different types of (conditional) response distributions.

2 Dynamic stochastic block models

For a network of n nodes observed at T occasions, let $\boldsymbol{Y}^{(t)}$ denote the $n \times n$ adjacency matrix observed at occasion t, whose generic element $Y_{ij}^{(t)}, i, j = 1, \ldots, n, j \neq i$, summarizes the relation existing between node i and j at occasion t. In the case of binary relations, $Y_{ij}^{(t)}$ will be a binary variable taking only zero/one values; in the case of valued relations, $Y_{ij}^{(t)}$ will be a count or a continuous variable. Without loss of generality, we focus on undirected relations without self-loops, so that $\boldsymbol{Y}^{(t)}$ denotes a symmetric matrix with missing values on the main diagonal. Finally, let $\mathcal{Y} = \{\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)}\}$.

Dynamic SBMs (e.g., Yang *et al.*, 2011, Matias & Miele, 2017, Bartolucci *et al.*, 2018) assume that each node in the network belongs to one of k distinct blocks identified by individual- and time-specific, discrete, latent variables $U_i^{(t)} \in \{1, \ldots, k\}$. These evolve over time according to a homogeneous latent Markov chain with initial probabilities λ_u, $u = 1, \ldots, k$ and transition probabilities $\pi_{u|v}$, $u, v = 1, \ldots, k$. Also, the dynamic SBMs postulate a *local independence assumption* between nodes: conditional on $U_i^{(t)} = u_1$ and $U_j^{(t)} = u_2$, responses $Y_{ij}^{(t)}$ are independent and identically distributed with distribution function depending on u_1 and u_2 only; that is,

$$\left[Y_{ij}^{(t)} \mid U_i^{(t)} = u_1, U_j^{(t)} = u_2\right] \sim p\left(y_{ij}^{(t)} \mid \boldsymbol{\psi}_{u_1 u_2}\right),$$

with $\boldsymbol{\psi}_{u_1 u_2}$ being a suitable vector of parameters. Therefore, at each occasion, the response variable distribution only depends on the block membership of nodes i and j at that occasion.

2.1 Assessing the impact of nodal attributes on block structure

In some cases, the homogeneity assumption underlying the latent structure of the model may be too restrictive. For instance, in a friendship network, individual features, such as gender or age, may play a role in determining the evolution of the block structure over time. To enhance the flexibility of the model, initial and transition probabilities defining the latent Markov chain may be parametrically specified as follows:

$$\log \frac{\lambda_{iu}}{\lambda_{i1}} = \tau_{u,\lambda} + \boldsymbol{x}'_{i,\lambda} \boldsymbol{\phi}_{u,\lambda}, \quad u = 2, \ldots, k,$$
$$\log \frac{\pi_{iu|v}}{\pi_{iu|u}} = \tau_{uv,\pi} + \boldsymbol{x}'_{i,\pi} \boldsymbol{\phi}_{uv,\pi}, \quad u, v = 1, \ldots, k, v \neq u.$$

Here, $\tau_{u,\lambda}$ and $\tau_{uv,\pi}$ denote state specific intercepts, while $\boldsymbol{\phi}_{u,\lambda}$ and $\boldsymbol{\phi}_{uv,\pi}$ denote the parameters measuring the impact of nodal attributes in $\boldsymbol{x}_{i,\lambda}$ and $\boldsymbol{x}_{i,\pi}$ on the initial and the transition probabilities of the latent Markov chain, respectively.

2.2 Model inference

Let $\mathcal{U} = \{\boldsymbol{U}_i, i = 1, \ldots, n\}$ denote the overall set of latent variables in the model, with $\boldsymbol{U}_i = (U_i^{(1)}, \ldots, U_i^{(T)})'$; the observed network distribution (i.e., the likelihood) is obtained by marginalizing out all these latent variables from the joint distribution of $\mathcal{Y}$ and $\mathcal{U}$. That is,

$$p(\mathcal{Y}) = \sum_{\mathcal{U}} \left\{ \prod_{i=1}^{n} \left[\lambda_{iu_i^{(1)}} \prod_{t=2}^{T} \pi_{iu_i^{(t)}|u_i^{(t-1)}} \prod_{t=1}^{T} \prod_{j>i} p(y_{ij}^{(t)} \mid U_i^{(t)} = u_i^{(t)}, U_j^{(t)} = u_j^{(t)}) \right] \right\}, \tag{1}$$

where $\sum_{\mathcal{U}}$ denotes the sum over the support of $\mathcal{U}$. Due to such a summation, computing the likelihood in eq. (1) becomes prohibitive even for small n and T. To overcome the issue, an alternative inferential procedure based on a variational approximation to the likelihood function is frequently considered in the literature (e.g., Matias & Miele, 2017). Although this method is well principled and computationally fast, it may lead to non-optimal estimators, and it may suffer from local maxima solutions.

3 Hybrid ML inference for dynamic SBMs

In this work, we extend the hybrid ML approach introduced by Bartolucci *et al.*, 2017 in the framework of static SBMs to make inference on model parameters. This lies in between a full ML and a classification likelihood inference. In this latter, the realization of the discrete latent variables in the models are considered as fixed parameters to be estimated. In this respect, the following hybrid log-likelihood function is defined:

$$\ell_{hyb} = \sum_{i=1}^{n} \left\{ \log \sum_{u_i^{(1)} \ldots u_i^{(T)}} \lambda_{iu_i^{(1)}} \prod_{t=2}^{T} \pi_{iu_i^{(t)}|u_i^{(t-1)}} \prod_{t=1}^{T} p(\boldsymbol{y}_i^{(t)} \mid U_i^{(t)} = u_i^{(t)}; \boldsymbol{U}_{(-i)}^{(t)} = \tilde{\boldsymbol{u}}_{(-i)}^{(t)}) \right\},$$

where $\boldsymbol{y}_i^{(t)}$ is the vector of all observed responses $y_{ij}^{(t)}$, $j \neq i$, for unit i at occasion t, $\boldsymbol{U}_{(-i)}^{(t)}$ denotes the vector of latent variables referring to time t and associated to all nodes in the network, but for the i-th one, and $\tilde{\boldsymbol{u}}^{(t)}$ denotes the corresponding realization. These latter are considered as fixed discrete parameters in the model taking values in the set $\{1,\ldots,k\}$. Last, $p(\boldsymbol{y}_i^{(t)} \mid U_i^{(t)} = u_i^{(t)}, \boldsymbol{U}_{(-i)}^{(t)} = \tilde{\boldsymbol{u}}_{(-i)}^{(t)}) = \prod_{j \neq i} p(y_{ij}^{(t)} \mid U_i^{(t)} = u_i^{(t)}; \boldsymbol{U}_{(-i)}^{(t)} = \tilde{\boldsymbol{u}}_{(-i)}^{(t)})$.

Let $\boldsymbol{\theta}$ denote the vector of all model parameters; this can be estimated by alternating the following steps until convergence: (*i*) *Classification step* – ℓ_{hyb} is maximized wrt $\tilde{\boldsymbol{u}}^{(t)}, t = 1,\ldots,T$, by identifying the configuration in the set $\{1,\ldots,k\}$ providing the highest log-likelihood value; (*ii*) *Expectation step* – for each i, the expected value of the individual complete-data (hybrid) log-likelihood is computed, given the current parameter estimates $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{u}}^{(t)}$; (*iii*) *Maximization step* – the expected value of the complete-data (hybrid) log-likelihood is maximized wrt $\boldsymbol{\theta}$.

References

BARTOLUCCI, F., MARINO, M.F., & PANDOLFI, S. 2017. Stochastic block models for social network data: inferential developments. *In: Proc. of the 32nd International Workshop on Statistical Modelling, Vol. I.*

BARTOLUCCI, F., MARINO, M.F., & PANDOLFI, S. 2018. Dealing with reciprocity in dynamic stochastic block models. *Computational Statistics & Data Analysis*, **123**, 86–100.

MATIAS, C., & MIELE, V. 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 1119–1141.

SNIJDERS, T.A., & NOWICKI, K. 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, **14**(1), 75–100.

YANG, T., CHI, Y., ZHU, S., GONG, Y., & JIN, R. 2011. Detecting communities and their evolutions in dynamic social networks - a Bayesian approach. *Machine Learning*, **82**, 157–189.

STABILITY OF JOINT DIMENSION REDUCTION AND CLUSTERING

Angelos Markos[1], Michel van de Velden[2] and Alfonso Iodice D'Enza[3]

[1] Department of Primary Education, Democritus University of Thrace,
(e-mail: amarkos@eled.duth.gr)

[2] Department of Econometrics, Erasmus University Rotterdam,
(e-mail: vandevelden@ese.eur.nl)

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

ABSTRACT: Several methods for joint dimension reduction and cluster analysis of categorical, continuous or mixed-type data have been proposed over time. These methods combine dimension reduction (PCA/MCA/PCAmix) with partitioning clustering (K-means) by optimizing a single objective function. Cluster stability assessment is a critical and inadequately discussed topic in the context of joint dimension reduction and clustering. We introduce a resampling scheme that combines bootstrapping and a measure of cluster agreement to assess global cluster stability of joint dimension reduction and clustering solutions and a Jaccard similarity approach for empirical evaluation of the stability of individual clusters. Both approaches are implemented in the R package `clustrd`.

KEYWORDS: dimension reduction, k-means, cluster stability, cluster validity.

1 Joint dimension reduction and clustering

Joint dimension reduction refers to a set of algorithmic or non-model based techniques aimining at simultaneously finding an optimal reduction of the variables and an optimal partitioning of the objects of a rectangular data set. Reduced K-means (De Soete & Carroll, 1994) and Factorial K-means (Vichi & Kiers, 2001) combine Principal Component Analysis (PCA) for dimension reduction with K-means for clustering and are suitable for data sets with continuous variables. In the case of categorical variables, MCA K-means (Hwang, Dillon & Takane, 2006), IFC-B (Iodice D'Enza & Palumbo, 2013) and Cluster Correspondence Analysis (van de Velden, Iodice D'Enza & Palumbo, 2017) a variant of Correspondence Analysis is used in the dimension reduction step and K-means is used for clustering. In the case of mixed-type data, that is when the data set contains both continuous and categorical variables, one can resort to GROUPALS (Van Buuren & Heiser, 1989) and Mixed Reduced/Factorial K-means (Vichi, Vicari & Kiers, 2009). These methods combine PCA for mixed data with K-means.

The general objective can be formulated as follows:

$$\min \phi_{\text{CDR}}(\mathbf{B}, \mathbf{Z}_K) = \alpha \left\|\mathbf{X} - \mathbf{X}\mathbf{B}\mathbf{B}'\right\|^2 + (1-\alpha)\left\|\mathbf{X}\mathbf{B} - \mathbf{P}\mathbf{X}\mathbf{B}\right\|^2 \tag{1}$$

where $\mathbf{X}$ is a $n \times Q$ data matrix, $\mathbf{B}$ a $Q \times d$ columnwise orthonormal loadings matrix, d is the user supplied dimensionality of the reduced space, $\mathbf{Z}_K$ a $n \times K$ binary matrix indicating cluster memberships of the n observations into the K clusters, $\mathbf{P} = \mathbf{Z}_K (\mathbf{Z}'_K \mathbf{Z}_K)^{-1} \mathbf{Z}'_K$ is a projection matrix, and $\mathbf{G} = \mathbf{PXB}$ a $K \times d$ cluster centroid matrix.

For categorical variables, the CDR objective can easily be adjusted by substituting $\mathbf{D}_z^{-1/2}\mathbf{MZ}$ for $\mathbf{X}$ in all equations. Similarly, for mixed-type data, $\mathbf{X}$ is set to $\begin{pmatrix}\mathbf{X} & \mathbf{D}_z^{-1/2}\mathbf{MZ}\end{pmatrix}$.

For given α, the following alternating least-squares algorithm is used to minimize the loss function in Eq.1:

1. Generate an initial cluster allocation $\mathbf{Z}_K$ (e.g., by randomly assigning subjects to clusters).
2. Find loadings $\mathbf{B}$ by taking the eigendecomposition of $\mathbf{X}^{*\prime}((1-\alpha)\mathbf{P} - (1-2\alpha)\mathbf{I})\mathbf{X}$.
3. Update the cluster allocation $\mathbf{Z}_K$ by applying K-means to the reduced space subject coordinates $\mathbf{XB}$.
4. Repeat the procedure (i.e., go back to step 2) using $\mathbf{Z}_K$ for the cluster allocation matrix, until convergence. That is, until $\mathbf{Z}_K$ remains constant.

Note that, for $\alpha = 1$ CDR reduces to PCAMIX, for $\alpha = 1/2$ we get mixed RKM method and for $\alpha = 0$ we have mixed FKM.

2 Global and local cluster stability via resampling

Cluster validation is important because cluster analysis presents clusters in almost any case. Here we focus on the stability of a partition in the case of joint dimension and clustering, that is, given a new sample from the same population, how likely is it to obtain a similar clustering? Stability can also be used to inform the selection of the number of clusters because if true clusters exist, the corresponding partition should have a high stability.

Resampling approaches (that is, bootstrapping, subsetting, replacement of points by noise) provide an elegant framework to computationally derive the distribution of interesting quantities describing the quality of a partition (Hennig 2007, Dolnicar & Leisch 2010). Simulations so far seem to suggest that resampling makes a lot of difference; the exact scheme used is not that important. Leisch (2015) provides a generic scheme for assessing cluster stability via resampling. Based on this scheme, we provide below two algorithms, one for assessing global stability, or the overall stability of a clustering partition, and one for assessing local or cluster-wise stability, or the stability of each one of the clusters in a given partition.

Algorithm GLOBAL STABILITY
Resampling: Draw bootstrap samples $\mathcal{S}^i$ and $\mathcal{T}^i$ of size n from the data and use the original data as evaluation set $\mathcal{E}^i = \mathbf{X}$. Apply a joint dimension reduction and clustering method to $\mathcal{S}^i$ and $\mathcal{T}^i$ and obtain $C^{\mathcal{S},i}$ and $C^{\mathcal{T},i}$.
Mapping: Assign each observation x_i to the closest centers of $C^{\mathcal{S},i}$ and $C^{\mathcal{T},i}$ using Euclidean distance, resulting in partitions $C^{X\mathcal{S},i}$ and $C^{X\mathcal{T},i}$, where $C^{X\mathcal{S},i}$ is the partition of the original data $\mathbf{X}$ predicted from clustering bootstrap sample $\mathcal{S}^i$ (same for $\mathcal{T}^i$ and $C^{X\mathcal{T},i}$).
Evaluation: Use the Adjusted Rand Index (ARI, Hubert & Arabie, 1985) or the Measure of Concordance (MOC, Pfitzner 2008) as measure of agreement and stability.
Inspect the distributions of ARI/MOC to assess the *global reproducibility* of the clustering solutions.

Algorithm LOCAL STABILITY
Resampling: Draw bootstrap samples $\mathcal{S}^i$ and $\mathcal{T}^i$ of size n from the data and use the original data as evaluation set $\mathcal{E}^i = \mathbf{X}$. Apply a joint dimension reduction and clustering method to $\mathcal{S}^i$ and $\mathcal{T}^i$ and obtain $C^{\mathcal{S},i}$ and $C^{\mathcal{T},i}$.
Mapping: Assign each observation x_i to the closest centers of $C^{\mathcal{S},i}$ and $C^{\mathcal{T},i}$ using Euclidean distance, resulting in partitions $C^{X\mathcal{S},i}$ and $C^{X\mathcal{T},i}$.
Evaluation: Obtain the maximum Jaccard agreement between each original cluster C_k and each one of the two bootstrap clusters, $C_{k'}^{X\mathcal{S},i}$ and $C_{k'}^{X\mathcal{T},i}$ as measure of agreement and stability, and take the average of each pair:

$$s_k^i = \Big(\max_{i \leq k' \leq K} \frac{C_k \cap C_{k'}^{X\mathcal{S},i}}{C_k \cup C_{k'}^{X\mathcal{S},i}} + \max_{i \leq k' \leq K} \frac{C_k \cap C_{k'}^{X\mathcal{T},i}}{C_k \cup C_{k'}^{X\mathcal{T},i}} \Big)/2$$

Inspect the distributions of s_k^i to assess the cluster level (local) stability of the solution.

The two algorithms are implemented in the `R` package `clustrd` via functions `global_bootclus()` and `local_bootclus()`, respectively.

3 Conclusions

Stability is an important aspect of clustering quality. Resampling approaches provide an elegant framework to assess global stability of Joint Dimension Reduction and Clustering solutions, as well as local quality of a cluster. However, maximizing stability for estimating the number of clusters amounts to implicitly defining the "true clustering" as the one with highest stability, which may not be appropriate. A comprehensive simulation study trying different combinations could offer guidance what works best in which situations.

References

DE SOETE, G., & CARROLL, J. D. 1994. K-means clustering in a low-dimensional Euclidean space. *New Approaches in Classification and Data Analysis*, 212–219.

DOLNICAR, S, & LEISCH, F. 2010. Evaluation of structure and reproducibility of cluster solutions using the bootstrap. *Marketing Letters*, **21**(1), 83–101.

HENNIG, C. 2007. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, **52**(1), 258–271.

HWANG, H, DILLON W, & TAKANE, Y. 2006. An Extension of Multiple Correspondence Analysis for Identifying Heterogenous Subgroups of Respondents. *Psychometrika*, **71**, 161–171.

IODICE D'ENZA, A., & PALUMBO, F. 2013. Iterative factor clustering of binary data. *Computational Statistics*, **28**(2), 789–807.

LEISCH, F. 2015. Resampling methods for exploring cluster stability. *Pages 658–673 of: Handbook of cluster analysis*. Chapman and Hall/CRC.

VAN BUUREN, S, & HEISER, WJ. 1989. Clustering *n* objects into *k* groups under optimal scaling of variables. *Psychometrika*, **54**(4), 699–706.

VAN DE VELDEN, M, IODICE D'ENZA A, & PALUMBO, F. 2017. Cluster Correspondence Analysis. *Psychometrika*, **82**(1), 158–185.

VICHI, M, & KIERS, H. 2001. Factorial *k*-means analysis for two-way data. *Computational Statistics & Data Analysis*, **37**(1), 49–64.

VICHI, M, VICARI D, & KIERS, H. 2019. Clustering and dimensional reduction for mixed variables. *Unpublished manuscript, to appear in Behaviormetrika 2018.*

HIDDEN MARKOV MODELS FOR CLUSTERING FUNCTIONAL DATA

Andrea Martino[1], Giuseppina Guatteri[1] and Anna Maria Paganoni[1]

[1] Department of Mathematics, Politecnico di Milano,
(e-mail: giuseppina.guatteri@polimi.it, andrea.martino@polimi.it, anna.paganoni@polimi.it)

ABSTRACT: Hidden Markov Models (HMMs) are a very popular tool used in many fields to model time series data. Usually, HMMs are used to model sequences of univariate or multivariate data. In this work, we extend the HMMs to the case of high dimensional data. Specifically, we focus on the case of functional data, by taking into consideration a sequence of multivariate curves that evolves in time. The functional observations are linked to the state of the HMM according to a similarity function, which depends on some metric in Hilbert spaces. After constructing a model that describes the time evolution of the functions, we apply the Viterbi algorithm to group the functional data into clusters. We assess our results in a simulation study, comparing our algorithm with a functional *k*-means.

KEYWORDS: clustering, functional data, hidden Markov models.

1 Introduction

Hidden Markov Models (HMMs) are a popular method for modeling time series. They consist of a Markov model in which the underlying states visited by the Markov process are unobservable (i.e. hidden) but the distribution that generates the output depends on the state (see Rabiner, 1989). In this work, we only consider models where the state space of the hidden variables is discrete. In the literature, HMMs usually consider an univariate or multivariate output; we extend the use of HMMs to functional observations and we use Viterbi algorithm to perform functional data clustering. In Section 2 we present the model, presenting some information about the theory of HMMs while in Section 3 we present a simulation study to assess its performance. All the analysis have been carried out using the statistical software R (R Core Team, 2017).

2 The model

The aim of this work is to develop a proper Hidden Markov Model (HMM) in the multivariate functional framework to cluster sequences of curves. Typically, HMMs are used to model univariate or multivariate data taking values in $\mathbb{R}^d$, $d \geq 1$. Let us consider a multivariate random curve $\mathbf{X} = \{\mathbf{X}(t)\}_{t \in I} = \{X_1(t), \dots, X_J(t)\}_{t \in I}$, with $J \geq 1$ and I compact interval of $\mathbb{R}$, as a random element of $L^2(I;\mathbb{R}^J)$ equipped with the Borel σ-algebra, such that $\{X_j(t)\}_{t \in I} \in L^2(I)$ for any $j \in \{1,..,J\}$.

We can define a Hidden Markov Model (see Cappé *et al.*, 2005) as a process $\{(Q_k, \{\mathbf{X}_k(t)\}_{t \in I})\}_{k \geq 0}$ on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\{\mathbf{X}_k(t)\}_{t \in I}$ is a multivariate random curve and $\{Q_k\}_{k \geq 0}$ is a Markov chain with a discrete and finite state space $\{s_1, \dots, s_N\}$, with $N \geq 1$, transition matrix $A = \{a_{ij}\} = \mathbb{P}(Q_k = s_j | Q_{k-1} = s_i)$ and initial distribution $\boldsymbol{\nu}$, where $\nu_i = \mathbb{P}(Q_0 = s_i)$. Given the process $\{Q_k\}_{k \geq 0}$, $\{\{\mathbf{X}_k(t)\}_{t \in I}\}_{k \geq 0}$ is a sequence of conditionally independent multivariate functions and $\{\mathbf{X}_k(t)\}_{t \in I}$ only depends on Q_k for each k . We denote the emission function of $\mathbf{X}_k$ conditionally on the event $\{Q_k = s_i\}$ with $b_i(\cdot;\boldsymbol{\mu}_i)$, for any $i = 1, \dots, N$, where $\boldsymbol{\mu}_i$ is a functional representative of state s_i; specifically, $b_i(\cdot;\boldsymbol{\mu}_i)$ is the likelihood that the function $\mathbf{X}_k$ is emitted from state s_i. We can completely define our HMM with the set of parameters $\lambda = (\boldsymbol{\nu}, A, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N)$. In this work, we use distances between functions to construct the emission functions $b_i(\cdot;\boldsymbol{\mu}_i)$, $i = 1, \dots, N$. Let us denote with d a generic distance in $L^2(I;\mathbb{R}^J)$; the likelihood that a realization of the stochastic process $\mathbf{X}$ is emitted from state s_i is $b_i(\cdot;\boldsymbol{\mu}_i) = h(d(\cdot, \boldsymbol{\mu}_i))$, where $h : \mathbb{R} \to \mathbb{R}$ is a function that transforms the distance into a similarity measure. In particular, we will use the L^2 distance.

Using the Baum-Welch algorithm, we are able to find the set of parameters $\lambda^* = \underset{\lambda}{\operatorname{argmax}} \mathcal{L}(\lambda|\mathbf{x})$ that maximizes the log-likelihood of our model

$$\begin{aligned}
\log(\mathcal{L}(\lambda|\mathbf{x})) = & \underbrace{\sum_{j=1}^{N} \gamma_1(j) \log \nu_j}_{\text{term 1}} + \underbrace{\sum_{i=1}^{N} \sum_{j=1}^{N} \left(\sum_{k=2}^{K} \xi_k(i,j) \right) \log a_{ij}}_{\text{term 2}} \\
& + \underbrace{\sum_{j=1}^{N} \sum_{k=1}^{K} \gamma_k(j) \log b_i(\mathbf{x}_k;\boldsymbol{\mu}_j)}_{\text{term 3}} .
\end{aligned} \tag{1}$$

where $\xi_k(i,j) = \mathbb{P}(Q_k = s_i, Q_{k+1} = s_j \mid X_1 = x_1, \dots, X_k = x_k, \lambda)$ and $\gamma_k(i) = \mathbb{P}(Q_k = s_i \mid X_1 = x_1, \dots, X_k = x_k, \lambda)$ (see Zucchini & Langrock, 2016 for further details). To estimate the functional representatives $(\mu_{ij})_{i=1,\dots,N;j=1,\dots,J}$ of the states of the HMM, we extend all the estimators commonly used in the functional data framework into the theory of functional HMM. Then, the estimator of $\boldsymbol{\mu}$ in the HMM framework is $\widehat{\boldsymbol{\mu}} = (\sum_{k=1}^{K} \gamma_k(j)\boldsymbol{X}_k)(\sum_{k=1}^{K} \gamma_k(j))^{-1}$.

Finally, we are able to cluster all the multivariate curves by finding the best state sequence Q. To solve this problem, we use the Viterbi algorithm. For every k, we want to find $\delta_k(i)$, i.e. the highest probability on a single path at time k the partial sequence $\mathbf{x}_1,\dots,\mathbf{x}_k$. By applying this algorithm, we are able to retrieve the best state sequence and obtain the clustering labels of the curves, by keeping track of the time evolution of the system.

3 Simulation Studies

We generate three samples of length n of i.i.d. realizations for three independent bivariate stochastic processes $\mathbf{X},\mathbf{Y},\mathbf{Z}$ in $L^2(I;\mathbb{R}^J)$, with $J=2$. Each sample is emitted from a different state of the following 3-state Markov Model:

- **State 1**: $\nu_1 = 1, a_{11} = 0.6, a_{12} = 0.3, a_{13} = 0.1$, $\mathbf{m}_1(t) = \begin{pmatrix} t(1-t) \\ 2t \end{pmatrix}$;
- **State 2**: $\nu_2 = 0, a_{21} = 0.1, a_{22} = 0.8, a_{23} = 0.1$, $\mathbf{m}_2(t) = \begin{pmatrix} t^2(1-t) \\ t^2 \end{pmatrix}$;
- **State 3**: $\nu_3 = 0, a_{31} = 0, a_{32} = 0, a_{33} = 1$, $\mathbf{m}_3(t) = \begin{pmatrix} t(1-t)^2 \\ \frac{1}{2}t^3 \end{pmatrix}$.

where $\boldsymbol{\nu} = (\nu_i)$ is the vector of the initial probabilities of the state, $A = (a_{ij})$ is the transition matrix and $\mathbf{m}_i(t), i = 1,\dots,N$, represent the real means of each sample. For each state, the sample is generated using the same exponential covariance kernel $C(s,t) = ae^{-b|s-t|}$, $a = 0.1$, $b = 0.3$. We choose the number of states, i.e. the number of clusters, by running our algorithm for $N = 2,\dots,5$ states and by computing each time the AIC and BIC criteria. Since both criteria reach the minimum value for $N = 3$, we choose this value as the "optimal" number of states for the HMM. After choosing the number of states, we summarize our results along 100 repetitions of our algorithm to estimate the parameters of the HMM. In Tab. 1 we can see the mean square error (MSE) and the standard deviation (SD) of the parameters along the repetitions. As we can see, all the parameters are very well estimated, both in terms of mean and standard deviation of the parameters.

Moreover, we can obtain some further information about the clustering structure of our data. Specifically, we use our model and apply the Viterbi algorithm on the output obtained from the Baum-Welch algorithm, to estimate

Parameter	MSE (SD)
a_{11}	$3.71 \cdot 10^{-2}$ $(9.23 \cdot 10^{-3})$
a_{12}	$8.30 \cdot 10^{-3}$ $(1.17 \cdot 10^{-2})$
a_{13}	$8.01 \cdot 10^{-2}$ $(4.10 \cdot 10^{-2})$
a_{21}	$2.89 \cdot 10^{-2}$ $(2.32 \cdot 10^{-3})$
a_{22}	$2.07 \cdot 10^{-3}$ $(1.70 \cdot 10^{-3})$
a_{23}	$9.72 \cdot 10^{-4}$ $(1.69 \cdot 10^{-3})$
a_{31}	$1.31 \cdot 10^{-3}$ $(9.29 \cdot 10^{-3})$
a_{32}	$7.10 \cdot 10^{-8}$ $(3.54 \cdot 10^{-7})$
a_{33}	$1.32 \cdot 10^{-3}$ $(9.31 \cdot 10^{-3})$
ν_1	$2.00 \cdot 10^{-2}$ $(1.41 \cdot 10^{-1})$
ν_2	$2.00 \cdot 10^{-2}$ $(1.41 \cdot 10^{-1})$
ν_3	$< 2 \cdot 10^{-16} (< 2 \cdot 10^{-16})$

Table 1. *MSE (SD) of the HMM parameters for 100 simulation runs of the Baum-Welch algorithm with $N = 3$ states for the HMM.*

the best state sequence and compare it with the output of the *k*-means algorithm (see Tarpey & Kinateder, 2003 for further details), based on the same distance. In particular, we obtain a Correct Classification Rate (CCR) of 0.857 by applying our method against a CCR of 0.591 by applying the functional *k*-means algorithm. We can conclude that, not only our method is able to detect the time structure behind the sequences of functional data and estimate all the parameters of the underlying hidden states but, by applying the Viterbi algorithm, we can also cluster the curves obtaining good values of accuracy.

References

CAPPÉ, O., MOULINES, E., & RYDEN, T. 2005. *Inference in hidden markov models*. New York: Springer.

R CORE TEAM. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RABINER, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286.

TARPEY, T., & KINATEDER, K. K. J. 2003. Clustering functional data. *Journal of classification*, **20**(1), 093–114.

ZUCCHINI, W., MACDONALD I. L., & LANGROCK, R. 2016. *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC.

COMPOSITE LIKELIHOOD INFERENCE FOR SIMULTANEOUS CLUSTERING AND DIMENSIONALITY REDUCTION OF MIXED-TYPE LONGITUDINAL DATA

Antonello Maruotti[1, 2], Monia Ranalli[3] and Roberto Rocci[3, 4]

[1] Dipartimento di Giurisprudenza, Economia, Politica e Lingue Moderne, Libera Università Maria Ss. Assunta, (e-mail: a.maruotti@lumsa.it)

[2] Department of Mathematics, University of Bergen,

[3] Dipartimento di Scienze Statistiche, Sapienza Università di Roma, (e-mail: monia.ranalli@uniroma1.it)

[4] Dipartimento di Economia e Finanza, Università di Tor Vergata,

ABSTRACT: We introduce a multivariate hidden Markov model (HMM) for mixed-type (continuous and ordinal) variables. As some of the considered variables may not contribute to the clustering structure, we built a hidden Markov-based model such that we are able to recognize discriminative and noise dimensions. The variables are considered to be linear combinations of two independent sets of latent factors where one contains the information about the cluster structure, following an HMM, and the other one contains noise dimensions distributed as a multivariate normal (and it does not change over time). The resulting model is parsimonious, but its computational burden may be cumbersome. To overcome any computational issue, a composite likelihood approach is introduced to estimate model parameters.

KEYWORDS: mixed-type data, data reduction, HMM, composite likelihood, EM algorithm.

1 Introduction

In this work we focus our attention on longitudinal multivariate-mixed type data (continuous and ordinal variables). This means there are three major dependency structures: correlation between multivariate variables, temporal dependence and heterogeneity. Furthermore, to be realistic, we assume the presence of dimensions (named noise) that are uninformative for capturing the heterogeneity over time and could obscure the true data structure. To simplify, the aim of the proposal is to recover the cluster structure underlying the data that varies over time through some discriminative factors. Following the the Underlying Response Variable (URV) (see e.g. Jöreskog, 1990, Lee *et al.* ,

1990) approach, both the continuous and the categorical ordinal variables follow a Gaussian mixture model (Mclachlan & Peel, 2000), where the ordinal variables are only partially observed through their ordinal counterparts. To take into account the temporal dependence, we assume that the Gaussian mixture changes over time according to the realizations of an homogeneous first order Markov chain. In other words we are assuming a partially observed hidden Markov model (HMM). This extends the mixture model for mixed-type data (Everitt, 1988; Ranalli & Rocci, b 2017) over time. As regards the presence of noise variables, in literature there are approaches based on a family of mixture models which fits the data into a common discriminative subspace (see e.g. Bouveyron & Brunet, 2012; Kumar & Andreou, 1998; Ranalli & Rocci, 2017). The key idea is to assume a common latent subspace to all latent states that is the most discriminative. This allows to project the data into a lower dimensional space preserving the clustering characteristics over time, leading to a better and more parsimonious visualization and interpretation of the underlying structure of the data. The model can be formulated as a HMM with a particular set of constraints on the latent state parameters. The parameter estimates is based on a composite likelihood approach (Lindsay, 1988). The material is organized as follows. In section 2, we present the model specification. In section 3, we outline the model parameter estimation. The EM-like algorithm and an example of application on real data showing the effectiveness of the proposal will be presented elsewhere for lack of space.

2 Model specification

Let $\mathbf{x}_t = [x_1, \ldots, x_O]'$ and $\mathbf{y}_t^{\bar{O}} = [y_{O+1}, \ldots, y_P]'$ be O ordinal and $\bar{O} = P - O$ continuous variables, respectively, with $t = 1, \ldots, T$. The associated categories for each ordinal variable are denoted by $c_i = 1, 2, \ldots, C_i$ with $i = 1, 2, \ldots, O$. Following the URV approach, the ordinal variables $\mathbf{x}$ are considered as a categorization of a continuous multivariate latent variable $\mathbf{y}_t^O = [y_1, \ldots, y_O]'$. We assume that the temporal evolution of these data is driven by a multinomial process in discrete time $\boldsymbol{\xi}_{1:T} = (\boldsymbol{\xi}_t, t = 1, \ldots, T)$, where $\boldsymbol{\xi}_t = (\xi_{t1}, \ldots, \xi_{tK})$ is a multinomial random variable with K classes. We specifically assume that such process is distributed as a homogeneous Markov chain, whose distribution, say $p(\boldsymbol{\xi}_{1:T}; \boldsymbol{p})$, is known up to a vector of parameters $\mathbf{p}$ that includes the initial probabilities and the transition probabilities of the chain. Conditionally on the value assumed each time by the Markov chain, the distribution of the data at time t depends on the specific component parameters of a partially observed multivariate normal. Formally, let define K initial probabilities as $p_k = P(\xi_{1k} = 1)$ with $\sum_{k=1}^{K} p_k = 1$ and K^2 transition probabilities as $p_{hk} = P(\xi_{tk} = 1 \mid \xi_{(t-1)h} = 1)$ with $h, k = 1, \ldots, K$ and $\sum_{h=1}^{K} p_{hk} = 1$. It follows that the Markov chain process is $p(\boldsymbol{\xi}_{1:T},\mathbf{p}) = \prod_{k=1}^{K} p_k^{\xi_{1k}} \prod_{t=1}^{T} \prod_{h=1}^{K} \prod_{k=1}^{K} p_{hk}^{\xi_{(t-1)h}\xi_{tk}}$. According to the URV, the joint distribution of $\mathbf{x}$ and $\mathbf{y}^{\bar{O}}$ can be constructed as follows. The latent relationship between $\mathbf{x}$ and $\mathbf{y}^{O}$ is explained by the threshold model, $x_i = c_i \Leftrightarrow \gamma_{c_i-1}^{(i)} \leq y_i < \gamma_{c_i}^{(i)}$, with $c_i = 1,\ldots,C_i$ and where $-\infty = \gamma_0^{(i)} < \gamma_1^{(i)} < \ldots < \gamma_{C_i-1}^{(i)} < \gamma_{C_i}^{(i)} = +\infty$ are the thresholds defining the C_i categories collected in a set $\boldsymbol{\Gamma}$ whose elements are given by the vectors $\boldsymbol{\gamma}^{(i)}$. To accommodate both cluster structure and dependence within the groups, we assume that the distribution $\mathbf{y}_t = [\mathbf{y}_t^{O\prime},\mathbf{y}_t^{\bar{O}\prime}]'$ given a particular point in time, say t and conditioning on $\boldsymbol{\xi}_t$, follows a partially observed multivariate normal, $f(\mathbf{y}_{nt} \mid \boldsymbol{\xi}_t) = \prod_{k=1}^{K} \phi_P(\mathbf{y}_{nt} \mid \boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)^{\xi_{nkt}}$, where the ξ_{nkt} is a Bernoulli variable that assumes value 1 if the $n-$th observation is classified in state k at time t, $\phi_P(\mathbf{y};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ is the density of a P-variate normal distribution with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

Let us set $\boldsymbol{\psi} = \{\mathbf{p},\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_K,\boldsymbol{\Gamma}\} \in \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the parameter space. For a random i.i.d. sample of size N, $(\mathbf{x}_1,\mathbf{y}_1^{\bar{Q}}),\ldots,(\mathbf{x}_N,\mathbf{y}_N^{\bar{Q}})$, the log-likelihood is

$$\ell(\boldsymbol{\psi}) \quad = \quad \sum_{n=1}^{N} \log\left[\sum_{\boldsymbol{\xi}_{1:T}} p(\boldsymbol{\xi}_t,\mathbf{p})\phi_{\bar{O}}(\mathbf{y}_{nt}^{\bar{O}} \mid \boldsymbol{\xi}_t,\boldsymbol{\mu}_k^{\bar{O}},\boldsymbol{\Sigma}_k^{\bar{O}})\pi_{nt}\left(\boldsymbol{\mu}_{nt;k}^{O|\bar{O}},\boldsymbol{\Sigma}_k^{O|\bar{O}},\boldsymbol{\Gamma},\boldsymbol{\xi}_t\right)\right], \qquad (1)$$

where, with obvious notation

$$\pi_{nt}\left(\boldsymbol{\mu}_{n;k}^{O|\bar{O}},\boldsymbol{\Sigma}_k^{O|\bar{O}},\boldsymbol{\Gamma},\boldsymbol{\xi}_t,\right) = \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \cdots \int_{\gamma_{c_O-1}^{(O)}}^{\gamma_{c_O}^{(O)}} \phi_O(\mathbf{u}_{nt};\boldsymbol{\mu}_{nt;k}^{O|\bar{O}},\boldsymbol{\Sigma}_k^{O|\bar{O}})d\mathbf{u}_{nt},$$

where $\pi_n\left(\boldsymbol{\mu}_{nt;k}^{O|\bar{O}},\boldsymbol{\Sigma}_k^{O|\bar{O}},\boldsymbol{\gamma}\right)$ is the conditional joint probability of response pattern $\mathbf{x}_{nt} = (c_1^{(1)},\ldots,c_O^{(O)})$ given the cluster k and the continuous variables $\mathbf{y}_{nt}^{\bar{O}}$. In order to identify the discriminative dimensions, it is assumed that there is a set of P latent factors $\tilde{\mathbf{y}}_t$, formed of two independent subsets.

In the first one, there are Q (with $Q \leq P$) factors that have some clustering information distributed as a mixture of Gaussians with class conditional means and variances equal to $E(\tilde{\mathbf{y}}^{Q} \mid k) = \boldsymbol{\eta}_k$ and $\mathrm{Cov}(\tilde{\mathbf{y}}^{Q} \mid k) = \boldsymbol{\Omega}_k$, respectively. In the second set there are $\bar{Q} = P - Q$ noise factors defining the so-called noise dimensions, that are independent of $\tilde{\mathbf{y}}^{Q}$ and their distribution does not vary from one class to another: $E(\tilde{\mathbf{y}}^{\bar{Q}} \mid k) = \boldsymbol{\eta}_0$ and $\mathrm{Cov}(\tilde{\mathbf{y}}^{\bar{Q}} \mid k) = \boldsymbol{\Omega}_0$. The link between $\tilde{\mathbf{y}}$ and $\mathbf{y}$ is given by a non-singular $P \times P$ matrix $\mathbf{A}$, as $\mathbf{y} = \mathbf{A}\tilde{\mathbf{y}}$. The final step is to identify the variables that could be considered as noise. Intuitively y_p is a noise variable if it is well explained by $\tilde{\mathbf{y}}^{\bar{Q}}$. Exploiting the independence between $\tilde{\mathbf{y}}^{Q}$ and $\tilde{\mathbf{y}}^{\bar{Q}}$, it is possible to compute proportions of each variable's variance that can be explained by the noise factors, and by one's complement, the proportions of each variable's variance that can be explained by the discriminative factors at each time point.

3 Construction of surrogate functions

The corresponding complete-data log likelihood involves multidimensional integrals that makes the maximum likelihood estimation computationally demanding and infeasible. To overcome this, we adopt a composite likelihood approach (Lindsay, 1988) based on $O(O-1)/2$ marginal distributions each of them composed of two ordinal variables and $\bar{O}$ continuous variables. The parameter estimates are carried out through an EM-like algorithm along with Baum-Welch recursion, that works in the same manner as the standard EM for HMMs.

References

BOUVEYRON, C., & BRUNET, C. 2012. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, **71**, 52–78.

EVERITT, B.S. 1988. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, **6**(5), 305–309.

JÖRESKOG, K. G. 1990. New developments in LISREL: analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, **24**(4), 387–404.

KUMAR, N., & ANDREOU, A.G. 1998. Heteroscedastic discriminant analysis and reduced rank {HMMs} for improved speech recognition. *Speech Communication*, **26**(4), 283 – 297.

LEE, S.-Y., POON, W.-Y., & BENTLER, P.M. 1990. Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability Letters*, **9**(1), 91–97.

LINDSAY, B. 1988. Composite likelihood methods. *Contemporary Mathematics*, **80**, 221–239.

MCLACHLAN, G., & PEEL, D. 2000. *Finite Mixture Models*. 1 edn. Wiley Series in Probability and Statistics. Wiley-Interscience.

RANALLI, M., & ROCCI, R. 2017. A Model-Based Approach to Simultaneous Clustering and Dimensional Reduction of Ordinal Data. *Psychometrika*.

RANALLI, M., & ROCCI, R. 2017. Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis*,**110**, 87–102.

BIVARIATE SEMI-PARAMETRIC MIXED-EFFECTS MODELS FOR CLASSIFYING THE EFFECTS OF ITALIAN CLASSES ON MULTIPLE STUDENT ACHIEVEMENTS

Chiara Masci[1], Francesca Ieva[1], Tommaso Agasisti[2]
and Anna Maria Paganoni[1]

[1] MOX - Modelling and Scientific Computing, Politecnico di Milano,
(e-mail: chiara.masci@polimi.it, francesca.ieva@polimi.it,
anna.paganoni@polimi.it)

[2] DIG - School of Management, Politecnico di Milano,
(e-mail: tommaso.agasisti@polimi.it)

ABSTRACT: We propose a bivariate semi-parametric mixed-effects model where the random effects are assumed to follow a discrete distribution with an unknown number of support points, together with an Expectation-Maximization algorithm to estimate its parameters - the BSPEM algorithm. This model for hierarchical data can be applied in many multivariate classification problems and enables the identification of subpopulations within the higher level of the hierarchy. In the case study, we apply the BSPEM algorithm to data about Italian middle schools, considering students nested within classes, and we identify subpopulations of classes that have different class effects on reading and mathematics student achievements.

KEYWORDS: EM algorithm, multivariate statistics, semi-parametric mixed-effects models, student achievements.

1 Introduction

In this work, exploiting previous results in the research about school and class value-added in the Italian education context, we propose a study that is innovative both from a methodological and an interpretative point of view. First of all, we develop a bivariate, i.e. for a bivariate response variable, semi-parametric mixed-effects linear model. Secondly, we show how this new method can be effective in the research about class effectiveness, by applying it in a case study that faces the new issue of the identification of clusters of Italian classes, standing on their joint effect on student achievement trends in reading and mathematics. The model that we propose is a bivariate two-level linear model where the coefficients of random effects, under semi-parametric assumptions, follow a bivariate discrete distribution with an unknown number of mass points. This model is the multivariate extension of the semi-parametric mixed-effects linear model proposed in **?**, that is totally new to the literature. In particular, it enters in the research line about the identification of subpopulations of the Growth Mixture Models (**?**) and of Latent Class Mixture Models (**?**), but with the novelty and the advantage that, contrarily to these existing methods, it does not need to fix a priori the number of latent subpopulations to be identified. In the application to the INVALSI data, the two-levels model, in which we consider students as first level and classes as second one, aims at identifying a latent clustering structure of classes where, within each cluster, the effect of the classes on their student achievement trends across years are similar.

2 The Dataset

The INVALSI database (`www.invalsi.it`) contains information about $18,242$ students attending the third year of junior secondary school in the year 2016/2017, nested within 1,082 classes. At pupil's level, we consider reading and mathematics INVALSI test scores at grade 8 (RS and MS); reading and mathematics INVALSI test scores at grade 5, three years before, of the same students; the socio-economic index (ESCS), the gender and the immigrant status of students. Moreover, INVALSI in the survey 2016/2017, by means of teacher questionnaires, collected information about teachers characteristics (age, education, gender...), teaching practices, class-body composition and geographical area.

3 Methodology

We consider the case of a bivariate semi-parametric two-level model with P fixed covariates, one random intercept and one random covariate. The model takes the following form:

$$\begin{gathered}
\mathbf{Y}_i = \begin{pmatrix}\mathbf{y}_{1,i} & \mathbf{y}_{2,i}\end{pmatrix} = \mathbf{X}_i \begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}^T + \mathbf{Z}_i \begin{pmatrix}\mathbf{c}_{1,m} \\ \mathbf{c}_{2,k}\end{pmatrix}^T + \varepsilon_i \\
i = 1,\dots,N \qquad m = 1,\dots,M \qquad k = 1,\dots,K \\
\varepsilon_i^T = \begin{pmatrix}\varepsilon_{1,i} \\ \varepsilon_{2,i}\end{pmatrix} \sim \mathscr{N}_2(\mathbf{0},\Sigma) \qquad ind.
\end{gathered} \tag{1}$$

where, in our application, N is the total number of classes; $\mathbf{Y}_i$ is the $(n_i \times 2)$-dimensional matrix of student achievements in reading and mathematics at grade 8 in class i; $\mathbf{X}_i$ is the $(n_i \times P)$-dimensional matrix of ESCS, gender and immigrant status of students in class i $(P = 3)$; $\mathbf{Z}_i$ is the $(n_i \times 2)$-dimensional matrix of the same students achievements in reading and mathematics at grade 5 (three years before) in class i and the intercepts. Σ is the (2×2)-dimensional variance-covariance matrix of the errors. $\beta = \begin{pmatrix}\beta_1 & \beta_2\end{pmatrix}$ is the $(P \times 2)$-dimensional matrix of coefficients of $\mathbf{X}_i$, while each $\mathbf{c}_{mk} = \begin{pmatrix}\mathbf{c}_{1,m} & \mathbf{c}_{2,k}\end{pmatrix}$ is the (2×2)-dimensional matrix of coefficients of $\mathbf{Z}_i$ and follows a discrete distribution P^* with $M \times K$ support points, where M and K are not known *a priori*. P^* can then be interpreted as the mixing distribution that generates the density of the stochastic model in (1). The ML estimator $\hat{P}^*$ of P^* can be expressed as a set of points $(\mathbf{c}_{11}, \dots, \mathbf{c}_{MK})$, where $M \leq N$, $K \leq N$ and $\mathbf{c}_{mk} \in R^4$ for $m = 1, \dots, M$, $k = 1, \dots, K$ and a set of weights $(w_{11}, \dots, w_{MK})$, where $\sum_{m=1}^{M} \sum_{k=1}^{K} w_{mk} = 1$ and $w_{mk} \geq 0$ for each $m = 1, \dots, M$ and $k = 1, \dots, K$. Given this, we develop an EM algorithm for the joint estimation of Σ, β, $(\mathbf{c}_{11}, \dots, \mathbf{c}_{MK})$ and $(w_{11}, \dots, w_{MK})$, that is performed through the maximization in closed form of the likelihood, mixture by the discrete distribution of the random effects,

$$
\begin{aligned}
L(\beta, \mathbf{c}_{mk}, \Sigma|\mathbf{y}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{w_{mk}}{\sqrt{|det(2\pi\Sigma)|^J}} \times \\
\times \exp\left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} -\frac{1}{2} \begin{pmatrix} y_{1,ij} - c_{1,1m} - \sum_{p=1}^{P} \beta_{1p} x_{1p,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \sum_{p=1}^{P} \beta_{2p} x_{2p,ij} - c_{2,2k} z_{2,ij} \end{pmatrix}^T \Sigma^{-1} \right. \qquad (2) \\
\left. \begin{pmatrix} y_{1,ij} - c_{1,1m} - \sum_{p=1}^{P} \beta_{1p} x_{1p,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \sum_{p=1}^{P} \beta_{2p} x_{2p,ij} - c_{2,2k} z_{2,ij} \end{pmatrix} \right\}
\end{aligned}
$$

with respect to Σ, β and $(\mathbf{c}_{mk}, w_{mk})$, for $m = 1, \dots, M$ and $k = 1, \dots, K$. Each class i, for $i = 1, \dots, N$ is therefore assigned to a cluster mk, for $m = 1, \dots, M$ and $k = 1, \dots, K$. $J = \sum_{i=1}^{N} n_i$. Moreover, given N starting support points, during the iterations of the EM algorithm, we reduce the support of the discrete distribution introducing a tuning parameter D: if two points are closer than D (in terms of Euclidean distance) they collapse to a unique point (e.g. two points $\mathbf{c}_{l*}$ and $\mathbf{c}_{m*}$ closer than D collapse to a unique point $\mathbf{c}_{(lm)*} = \frac{\mathbf{c}_{l*} + \mathbf{c}_{m*}}{2}$ with weight $w_{(lm)*} = w_{l*} + w_{m*}$).

4 Results

The BSPEM algorithm applied to INVALSI data identifies 5 subpopulations for the class effects in mathematics and 4 subpopulations for reading. Estimates of the parameters are shown in Table 1.

First response variable						
	$\hat{c}_{1,1}$ (intercept)	$\hat{c}_{1,2}$ (math5)	$\hat{w}_1$ (weight)	$\hat{\beta}_{11}$ (ESCS)	$\hat{\beta}_{12}$ (gender)	$\hat{\beta}_{13}$ (immigrant)
m=1	0.295	0.719	0.458	0.089	−0.055	0.048
m=2	−0.181	0.463	0.384			
m=3	0.762	0.463	0.025			
m=4	−1.301	0.112	0.064			
m=5	0.366	0.291	0.069			
Second response variable						
	$\hat{c}_{2,1}$ (intercept)	$\hat{c}_{2,2}$ (read5)	$\hat{w}_2$ (weight)	$\hat{\beta}_{21}$ (ESCS)	$\hat{\beta}_{22}$ (gender)	$\hat{\beta}_{23}$ (immigrant)
k=1	−2.848	−0.101	0.019	0.095	0.219	−0.083
k=2	−0.622	0.262	0.095			
k=3	−1.556	0.188	0.018			
k=4	0.054	0.544	0.868			

Table 1. *Estimates of the coefficients of Eq. (1) obtained by the BSPEM algorithm.*

References

MASCI C., IEVA, F., PAGANONI A. M., & AGASISTI T. 2019, in press. Semi-parametric mixedeffects models for the unsupervised classification of Italian schools. *Journal of the Royal Statistical Society - Series A*.

MUTHÉN, B. 2004. Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences*, **345**(368), 106–109.

MUTHÉN B., & ASPAROUHOV T. 2015. Growth mixture modeling with non-normal distributions. *Statistics in medicine*, **34**(6), 1041–1058.

MULTIVARIATE CHANGE-POINT ANALYSIS FOR CLIMATE TIME SERIES

Gianluca Mastrantonio[1], Giovanna Jona Lasinio[2], Alessio Pollice[3],
Giulia Capotorti[4], Lorenzo Teodonio[5] and Carlo Blasi[4]

[1] Department of Mathematical Sciences, Politecnico di Torino,
(e-mail: gianluca.mastrantonio@polito.it)

[2] Department of Statistical Sciences, Sapienza Università di Roma,
(e-mail: giovanna.jonalasinio@uniroma1.it)

[3] Department of Economics and Finance, Università di Bari Aldo Moro,
(e-mail: alessio.pollice@uniba.it)

[4] Department of Environmental Biology, Sapienza Università di Roma,
(e-mail: giulia.capotorti@uniroma1.it,carlo.blasi@uniroma1.it)

[5] ICRCPAL, Ministry of Cultural Heritage and Activities and Tourism, Roma,
(e-mail: lorenzoteodonio@gmail.com)

ABSTRACT: The aim of this work is to find individual and joint change-points in a large multivariate database of climate data. We model monthly values of precipitation, minimum and maximum temperature recorded in 360 stations covering all Italy for 60 years (12 × 60 months). The proposed three variate Gaussian change-point model lets us estimate a different change-point model for each station. As stations possibly share some of the parameters, this model framework provides an original definition of the change-points corresponding to changes in any subset of the 9 model parameters. In this paper, results for two stations in Southern Italy are shown as an example.

KEYWORDS: change-point model, hierarchical Dirichlet process, climate data.

1 Introduction

Climate elements and regimes, such as temperature, precipitation and their annual cycles, primarily affect the type and distribution of plants, animals, and soils as well as their combination in complex ecosystems. From a botanical perspective, the analysis of change-points allows to detect abrupt changes in the climatic behaviour and supports inferences of the potential effects of these changes on ecosystem composition, functionality, distribution, and dynamics at different spatial and time scales (Liu & Lei, 2015).

In order to simplify the joint distribution modeling of the climate variables, we standardize the temperatures and rescale the precipitation with its standard deviation; the latter is then seen as the realization of a latent variable belonging to the real line ($\mathbb{R}$) (Mastrantonio *et al.* , 2019) . At each station the trivariate time series can then be assumed to come from a change-point (CP) model with multivariate normal emission distribution, parametrized using 9 parameters: 3 intercepts, 3 variances and 3 correlations. We define the CP model using a modified version of the hierarchical Dirichlet process (Teh & Jordan, 2010), which allows that different time series can share a subset of the 9 parameters and that some or all of these parameters change at each change-point.

2 The model

We consider monthly records of precipitation and min/max temperature at 360 monitoring stations over 60 years (1951-2010). Almost all original time series are affected by variable amounts of missing data. The full database reports $360 \times 60 \times 12$ entries.

Let us denote with $y_{t,\mathbf{s},2}$ and $y_{t,\mathbf{s},3}$ the standardized minimum and maximum temperature, respectively, and $y_{t,\mathbf{s},1}$ be a "latent" standardized precipitation, assuming values in $\mathbb{R}$, where the negative values are associated to the event "no precipitation" (Mastrantonio *et al.* , 2019), where $\mathbf{s} \in \mathscr{S}$ are spatial coordinates and $t \in \mathscr{T}$ temporal indices. Our main idea is to model each trivariate time series $\mathbf{y_{t,s}}$ using a CP model with a multivariate normal emission distribution and the following features:

$$f(\mathbf{y}|\theta) = \prod_{\mathbf{s}\in\mathscr{S}} \prod_{t\in\mathscr{T}} \phi_3(\mathbf{y}_{t,\mathbf{s}}|\theta_{t,\mathbf{s}}), \tag{1}$$

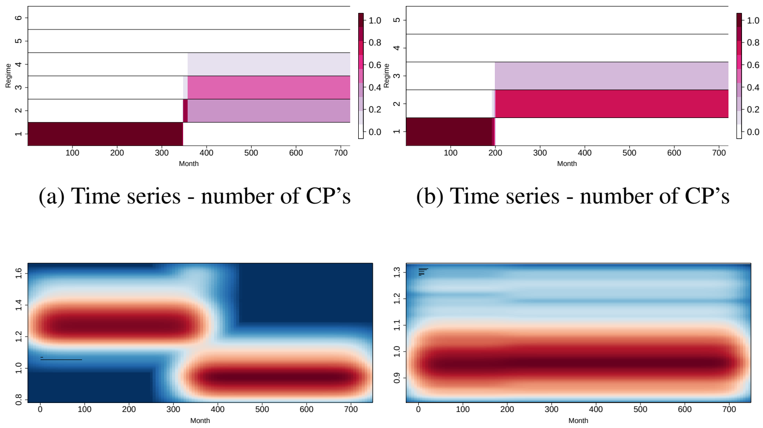
$$\theta_{t,\mathbf{s}}|(\theta_{t-1,\mathbf{s}} = \theta^*_{\mathbf{s},k}) \sim \sum_{j=k}^{\infty} \frac{\pi_{\mathbf{s},j}}{1-\sum_{h=1}^{k-1} \pi_{\mathbf{s},h}} \delta_{\theta^*_{\mathbf{s},j}}, \tag{2}$$

$$G_{\mathbf{s}} = \sum_{k\in\mathbb{Z}} \pi_{\mathbf{s},k} \delta_{\theta^*_{\mathbf{s},k}} \sim Dir(\alpha, G_0), \tag{3}$$

$$G_0 = \prod_{h=1}^{9} G_{\theta_h} \tag{4}$$

$$G_{\theta_h} = \sum_{k\in\mathbb{Z}} \xi_{\theta_h} \delta_{\theta^{**}_j} \sim Dir(\gamma, H_{\theta_j}), h = 1, \ldots, 9 \tag{5}$$

where $\phi_3()$ is a trivariate normal density, $\theta_{t,\mathbf{s}}$ contains the 3 regressive coefficients (β), 3 variances σ^2 and 3 correlations ρ that parametrize the likelihood, and $\delta_{\cdot}$ is the indicator function.



(a) Time series - number of CP's (b) Time series - number of CP's

(c) Posterior - number of CP's for $\beta_{t,s,11}$ (d) Posterior - number of CP's for $\beta_{t,s,11}$

Figure 1: Posterior distributions and time-series of the number of change-points for all parameters (a b) and for the intercept (c d) for the stations Lucera (a c) and Vieste (b d).

The trivariate normal density in (1) is parametrized using 9 parameters. For each of them we have a corresponding distribution drawn from a Dirichlet process, in (5).All these distributions are combined to obtain the discrete distribution G_0 in (4), with atoms given by (5) and weights obtained as the products of the ones of (5). As in the standard Hierarchical Dirichlet process, a common Dirichlet based distribution, here G_0, is used as base distribution for DP draws, see (3). The $G_{\mathbf{s}}$ distributions share the same set of atoms with different weights. Equations (1) -(2) define CP models for each station. At each time t, the values of the parameters are drawn from a discrete distribution which allows them to assume the same value of the previous time or a new one that has never been observed previously at the specific location $\mathbf{s}$. The occurrence of a change point does not imply that all 12 parameters change. Further, $G_{\mathbf{s}}$ and $G_{\mathbf{s}'}$ have the same set of atoms, implying that the two time series can have all or some of parameters in common.

3 Some Results

We describe part of the results for the stations *Lucera* and *Vieste*. The maximum a posteriori (MAP) number of change-points in Lucera is three, two CP's have high probability too, while in Vieste we observe two CP's. Figures 1 (a) and (b) show the probability that at a given time (horizontal) the time series is following a specific regime (vertical), the darker the color the largest the probability. We can observe that even if Lucera (a) registers three possible regimes, the second one lasts less than one year. Figures 1 (c) and (d) show the posterior densities of the considered parameter for each time point. The time points are represented by the x-axis, in the y-axis there are the values assumed by the parameter and the color represents the density, with blue equals to zero and the darkest red the maximum value. We can see a clear change of value at point 320, for the precipitation at Lucera while no strong evidence of change-point is obtained for the precipitation in Vieste. Notice that, starting from time 320 (when Lucera's change-point is found), there is a probability of 0.7 that the intercept of the precipitation has the same value in the two stations, as we can see from Figures 1 (c) and (d). The code is written in R/C++, and uses the openMP library to perform parallel computing. Our proposal allows for a very rich inference on the joint CP detection. Posterior estimates are obtained in 2 days, with 40000 iterations per day and 10 GB of ram usage.

Acknowledgement

The work of the first three authors is partially developed under the PRIN2015 supported-project EPHAStat (20154X8K23-SH3).

References

BLASI, C., CAPOTORTI, G., COPIZ, R., GUIDA, D., MOLLO, B., SMIRAGLIA, D., & ZAVATTERO, L. 2014. Classification and mapping of the ecoregions of Italy. *Plant biosystems - An International Journal Dealing with all Aspects of Plant Biology*, **148**(6), 1255–1345.

LIU, Y., & LEI, H. 2015. Responses of natural vegetation dynamics to climate drivers in China from 1982 to 2011. *Remote Sensing*, **7**, 10243–10268.

MASTRANTONIO, G., JONA LASINIO, G., POLLICE, A., CAPOTORTI, G., TEODONIO, L., GENOVA, G., & BLASI, C.. 2019. A hierarchical multivariate spatio-temporal model for clustered climate data with annual cycles. *Annals of Applied Statistics*, **13**, 797–823.

TEH, Y.W., & JORDAN, M.I. 2010. Hierarchical Bayesian Nonparametric Models with Applications. *In:* HJORT, N., HOLMES, C., MÜLLER, P., & WALKER, S. (eds), *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press.

A DYNAMIC STOCHASTIC BLOCK MODEL FOR LONGITUDINAL NETWORKS

Catherine Matias[1], Tabea Rebafka[1] and Fanny Villers[1]

[1] Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Centre National de la Recherche Scientifique, Université Paris Diderot, 4 Place Jussieu, 75252 Paris Cedex 05, France, (e-mail: catherine.matias@math.cnrs.fr, tabea.rebafka@sorbonne-universite.fr, fanny.villers@sorbonne-universite.fr)

ABSTRACT: We propose an extension of the stochastic block model for recurrent interaction events in continuous time, where every individual belongs to a latent group and conditional interactions between two individuals follow an inhomogeneous Poisson process with intensity driven by the individuals latent groups. We show that the model is identifiable and rely on a semiparametric variational expectation-maximization estimator. We develop two versions of the method, one using a nonparametric histogram approach with an adaptive choice of the partition size, and the other using kernel intensity estimators. The number of latent groups is selected by an integrated classification likelihood criterion. Synthetic experiments and two datasets illustrate the performance and utility of our approach, also compared with competing methods.

KEYWORDS: expectation-maximization algorithm, link streams, longitudinal network, variational approximation, temporal network.

1 Introduction

The past few years have seen a large increase in the interest for modelling dynamic interactions between individuals. Continuous-time information on interactions is often available, for example as email exchanges between employees in a company or face-to-face contacts between individuals measured by sensors, but most models use discrete time.

Clustering individuals based on interaction data is a well-established way to account for the intrinsic heterogeneity and to summarize information. For discrete-time sequences of graphs, many recent approaches propose generalizations of the stochastic block model to a dynamic context. Stochastic block models posit that all individuals belong to one of finitely many groups, and given these groups all pairs of interactions are independent. Stochastic block models induce more general clusterings than do community detection algorithms. Indeed, clusters are not necessarily characterized by intense within-group interaction and low interaction frequency towards other groups.

2 Method

We introduce a semiparametric stochastic block model for recurrent interaction events in continuous time, which we refer to as the Poisson process stochastic block model. Interactions are modelled by conditional inhomogeneous Poisson processes, whose intensities only depend on the latent groups of the interacting individuals. We do not rely on a parametric model where intensities are modulated by predefined network statistics; they are modelled and estimated in a nonparametric way. The model parameters are shown to be identifiable. Our estimation and clustering approach is a semiparametric ver sion of the variational expectation-maximization algorithm, where the maximization step is replaced by nonparametric estimators of the intensities. We propose two different estimators of the nonparametric part of the model: a histogram approach where the partition size is adaptively chosen, and a kernel estimator. With the histogram approach, an integrated classification likelihood criterion is proposed to select the number of latent groups.

Synthetic experiments and the analysis of two datasets illustrate the strengths and weaknesses of our approach. The first dataset uses the cycle hire usage data from the bike-sharing system of the city of London while the second is the Enron corpus containing emails exchanges among people working at Enron, during the period of the affair that led to the bankruptcy of the company.

3 Details

This talk is a presentation of the paper Matias *et al.* , 2018. The code is available in the R package `ppsbm`.

References

MATIAS, C., REBAFKA, T., & VILLERS, F. 2018. A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, **105**(3), 665680.

UNSUPERVISED FUZZY CLASSIFICATION FOR DETECTING SIMILAR FUNCTIONAL OBJECTS

Fabrizio Maturo[1], Francesca Fortuna[2] and Tonio Di Battista[3]

[1] University of Campania "L. Vanvitelli", Caserta,
(e-mail: fabrizio.maturo@unicampania.it)

[2] University "G. D'Annunzio" of Chieti-Pescara, Pescara,
(e-mail: francesca.fortuna@unich.it)

[3] University "G. D'Annunzio" of Chieti-Pescara, Pescara,
(e-mail: tonio.dibattista@unich.it)

ABSTRACT: In recent decades, clustering techniques based both on functional data analysis (FDA) and fuzzy logic theory have captivated the attention of many scholars. In these two areas of research, many metrics have been introduced over time to identify similarity among statistical units. This study underlines that, in the field of FDA, to the best of our knowledge, research has always focused on the so-called crisp clustering. The latter considers that a statistical unit can uniquely be associated with a single group. On the other hand, fuzzy clustering techniques, in a non-functional context, contemplate the feasibility that a statistical unit can belong to diverse groups at the same time. Therefore, the objective of this article is to blend the two strategies and offer an unsupervised fuzzy functional classification approach.

KEYWORDS: clustering, FDA, fuzzy clustering, k-means, fuzzy functional k-means.

1 Introduction

In contemporary statistical language, the term classification encompasses unsupervised (classification) that refers to clustering, where the class labels are latent, and supervised (classification) that indicates procedures such as linear discriminant analysis, logistic regression, and nearest neighbours where the group labels are known and can be adopted for learning classification rules to deal with further data. In addition to the classical statistical approaches, supervised and unsupervised classification methods can be based both on functional data analysis (FDA) and fuzzy logic theory. In these latter settings, classification techniques have also attracted the consideration of scholars from various research areas. Mainly, unsupervised classification techniques, i.e. clustering, have been very flourishing and led to copious utilisation in these two contexts.

Focusing on the k-means algorithm, to the best of our knowledge, FDA scholars have always concentrated on the so-called crisp functional clustering. The latter contemplates that a statistical unit can uniquely be associated with a single group.

On the other hand, fuzzy clustering techniques, in a non-functional context, consider the feasibility that a statistical unit can belong to distinct groups at the same time. The main reason behind this idea is that reality is shaded and, in some contexts, forcing a statistical unit to belong to a single group, as illustrated by Zadeh (1975) and his followers (Bezdek, 1981; Bora and Gupta, 2014; Ferraro and Giordani, 2015; Betti, 2016), provides a forced image of reality. It is sufficient to think of all those cases in which, by performing a k-means, we are obligated to provide a stopping rule because a statistical unit moves continuously from one group to another at each iteration. Sometimes, it is more reasonable to think that a statistical unit belongs to multiple sets simultaneously with a different degree of truth, which is the basic idea of fuzzy clustering, where the level of truth is the so-called membership function in [0,1].

Accordingly, the purpose of this study is to combine FDA and fuzzy logic to offer an unsupervised fuzzy functional classification procedure for clustering functional data.

2 Material and Methods

The fundamental concept of FDA is to manage data functions as single objects. However, in real applications, functional data are regularly observed as a sequence of point data. The first step in FDA is to transform the observed values $y_{i1}, y_{i2}, ..., y_{iT}$ for each unit $i = 1, 2, ..., N$ to a functional form computable at any desired point $x \in \Re$. To estimate the functional datum, several techniques are available, but the basis approximation is the most used (Ramsay, 2005):

$$x(t) = \sum_{j \in \mathbb{N}} c_j \phi_j(t) \approx \sum_{j=1}^{K} c_j \phi_j(t) \tag{1}$$

where c_j is the vector of coefficients defining the linear combination and $\phi_j(t)$ is the vector of basis functions.

The functional principal component decomposition (FPCA), is also widely adopted in the FDA framework. FPCA allows us representing the functions by a linear combination of a small number of functional principal components (FPCs). Thus, the functional data can be approximated by:

$$\hat{x}_i(t) = \sum_{i=1}^{K} \nu_{ik}\xi_k(t) \tag{2}$$

where ν_{ik} is the score of the generic FPC ξ_k for the generic function x_i $(i = 1, 2, ..., N)$. In the FDA context many metrics and semi-metrics have been proposed over time to group functional data using a crisp approach (see e.g. Ferraty and View, 2006; Febrero-Bande and de la Fuente, 2012; Jacques and Preda, 2014; Cuevas, 2014).

On the other hand, in a non-functional context, Bezdek (1981) introduced the fuzzy k-means, i.e. an algorithm that proceeds iteratively through the minimisation of the objective function:

$$J_m(U, v) = \sum_{g=1}^{G} \sum_{i=1}^{n} u_{ig}^m d_{ig}^2 \tag{3}$$

where $d_{ig} = |x_i - v_g|$ is a suitable norm on $\mathbb{R}^p$ for example the Euclidean norm, $x_i \in \mathbb{R}^p$ is the i-th component of units vector, $v_g \in \mathbb{R}^p$ is the g-th component of the centroid vector, U is the matrix of the degree of membership of dimension $n \times c$, and $m \in [1, +\infty)$.

In this setting, for each unit, it is determined the degree of membership to the G groups. The degree of membership of the i^{th} unit to the g^{th} group, denoted by u_{ig}, satisfies the constraints $0 \leq u_{ig} \leq 1$ and $\sum_{g=1}^{G} u_{ig} = 1$ where $i = 1, 2, ..., n$ and $g = 1, 2, ..., G$. The objective function depends on the distance d_{ig}, between the i-th unit and the centroid of the g-th group, and the parameter m which adjusts the level of fuzziness.

The basic idea of this study is to propose a fuzzy k-means clustering of functional objects according to the following procedure:

1. Smooth the original data using Equation 1. This phase is actually optional because it is possible to skip to the next phase. In reality, this phase makes it possible to exploit the representation of data to obtain additional information such as derivatives and other functional tools;
2. Perform the FPCs decomposition as in Equation 2;
3. Select the first S FPCs explaining 70%-80% of the total variability;
4. Get the first S scores ν_{ik} with $k = (1, 2, ..., S)$;
5. Use the scores to compute the objective function $J_m(U, v)$ as in Equation 3;
6. Obtain the membership functions for each group and for each functional object, so that each functional object belongs to all groups in a nuanced manner.

This procedure has interesting implications in real cases, where we deal with functional data with complex behaviours and thus, depending on the part of the domain considered, they might have a different tendency to belong to one group rather than another. Instead, forcing these functions to belong to a single group uniquely as in the classical k-means approach, could reduce the information available with respect to the proposed procedure.

References

BETTI, G. 2016. Fuzzy Measures of Quality of Life: a Multidimensional and Comparative Approach. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **24**(Suppl. 1), 25–37.

BEZDEK, J. C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer US.

BORA, D. J., & GUPTA, A. K. 2014. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. *International Journal of Computer Trends and Technology*, **10**(2), 108–113.

CUEVAS, A. 2014. A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, **147**, 1–23.

FEBRERO-BANDE, M., & DE LA FUENTE, M. 2012. Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software, Articles*, **51**(4), 1–28.

FERRARO, M. B., & GIORDANI, P. 2015. A toolbox for fuzzy clustering using the R programming language. *Fuzzy Sets and Systems*, **279**, 1–16.

FERRATY, F., & VIEU, P. 2006. *Nonparametric Functional Data Analysis*. Springer New York.

JACQUES, J., & PREDA, C. 2013. 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, 2nd edn*. New York: Springer.

ZADEH, L. A. 1975. The concept of a linguistic variable and its application to approximate reasoning. *Information Scienze*, **I**.

Mixture modelling with skew-symmetric component distributions

Geoff McLachlan[1]

[1] Department of Mathematics, University of Queensland,
(e-mail: g.mclachlan@uq.edu.au)

Abstract: In recent years there has been an increasing use of finite mixtures of skew distributions for the modelling and analysis of heterogeneous and nonnormal data. These models adopt component densities that offer a high degree of flexibility in distributional shapes. In particular, the skew-symmetric family of distributions, which include the classical skew normal and skew t-distributions, has become increasingly popular. But besides improving on existing models, there is also a need to provide a critical comparison of the available proposals to make them more accessible and easier to understand to people outside of the area, including to practitioners and researchers from other disciplines. We shall present a short review of the more commonly used skew distributions before focussing on recent proposals.

This is joint work with Sharon Lee

Keywords: multivariate skew normal mixtures, CFUST distributions, EM algorithm.

NEW DEVELOPMENTS IN APPLICATIONS OF PAIRWISE OVERLAP

Volodymyr Melnykov[1], Yana Melnykov[1], Domenico Perrotta[2], Marco Riani[3], Francesca Torti[2] and Yang Wang[1]

[1] Department of Information Systems, Statistics, and Management Science, University of Alabama, (e-mail: vmelnykov@cba.ua.edu, ymelnykov@cba.ua.edu, ywang311@crimson.ua.edu)

[2] European Commission, Joint Research Centre, (e-mail: domenico.perrotta@ec.europa.eu, francesca.torti@ec.europa.eu)

[3] Department of Economics, University of Parma, (e-mail: marco.riani@unipr.it)

ABSTRACT: Pairwise overlap between components is defined as the sum of two misclassification probabilities and can be used as a measure of their proximity. The clustering complexity of data can be assessed based on the magnitude of pairwise overlap values. In other words, components with low overlap produce well-separated clusters and those that exhibit high overlap are expected to yield more misclassifications. This can be used to simulate data that are either easy or difficult to group and study the systematic properties of clustering algorithms in various settings.

The efficient calculation of overlap values is available for Gaussian components. It is implemented in the R package MixSim that allows simulating clusters from Gaussian mixtures according to the pre-specified level of average or maximum overlap. However, Monte Carlo simulations are used for non-Gaussian mixtures. This affects the computing speed and makes the procedure for finding mixtures with pre-specified overlap level nearly impractical.

We propose novel methodology for the efficient calculation of overlap values in the case of mixtures of skewed components, heavy-tailed ones, as well as mixtures of regressions. This methodology can be used for simulating clusters with non-Gaussian characteristics and evaluating clustering algorithms in a broad range of settings.

KEYWORDS: pairwise overlap, misclassification probability, skewed components, data simulation, MixSim.

References

MAITRA, R., & MELNYKOV, V. 2010. Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics.*, **2**, 354–376.

MELNYKOV, V., CHEN, W.-C., & MAITRA, R. 2012. MixSim: R package for simulating datasets with pre-specified clustering complexity. *Journal of Statistical Software.*, **51**, 1–25.

RIANI, M., PERROTTA, D., & CERIOLI, A. 2015. Simulating mixtures of multivariate data with fixed cluster overlap in FSDA library. *Advances in Data Analysis and Classification.*, **9**, 461–481.

MODELLING UNOBSERVED HETEROGENEITY OF RANKING DATA WITH THE BAYESIAN MIXTURE OF EXTENDED PLACKETT-LUCE MODELS

Cristina Mollica[1] and Luca Tardella[2]

[1] Dipartimento di Metodi e Modelli per l'Economia, il Territorio e la Finanza, Sapienza Università di Roma, (e-mail: cristina.mollica@uniroma1.it)

[2] Dipartimento di Scienze Statistiche, Sapienza Università di Roma, (e-mail: luca.tardella@uniroma1.it)

ABSTRACT: The Plackett-Luce distribution (PL) is one of the most successful parametric options within the class of multistage ranking models to learn the preferences on a given set of items from a sample of ordered sequences. It postulates that the ranking process is carried out by sequentially assigning the positions according to the *forward order*, that is, from the top (most-liked) to the bottom (least-liked) alternative. This assumption has been relaxed with the *Extended Plackett-Luce model* (EPL), thanks to the introduction of the *reference order* parameter describing the rank attribution path. Starting from the recent formulation of the Bayesian EPL, in this work we investigate the further extension into the finite mixture approach as a method to explore the group structure of ranking data.

KEYWORDS: Ranking data, Plackett-Luce model, mixture model, Gibbs sampling, Metropolis-Hastings algorithm.

1 Introduction

A *ranking* is an ordered sequence resulting from the comparative evaluation of a given set of *items* according to a specific criterion. This framework is typical in several areas of research, involving surveys on preferences for consumer goods, psychological/behavioral studies on attitudes, voting systems and the competition/sport context, see Marden, 1995 for a broad review of the statistical literature on methods and models for analysing ranking data.

Formally, a ranking of K items is a vector $\pi = (\pi(1), \ldots, \pi(K))$, where the entry $\pi(i)$ indicates the position attributed to the i-th alternative. Data can be equivalently collected in the ordering format $\pi^{-1} = (\pi^{-1}(1), \ldots, \pi^{-1}(K))$, where the component $\pi^{-1}(j)$ denotes the item ranked in the j-th position. Thus, ranking data take values in the set of permutations $\mathcal{S}_K$ of the first K integers.

This work concentrates on the parametric family of stagewise models. In particular, our interest is in the *Extended Plackett-Luce model* (EPL), originally proposed by Mollica & Tardella, 2014, both in its basic form and into the finite mixture framework. In that work, inference on the EPL mixture was addressed in the frequentist domain via the EM algorithm. Starting from the recent contribution by Mollica & Tardella, 2019, here we explored the further extension of the Bayesian EPL into the finite mixture approach.

2 The Bayesian EPL mixture

2.1 The Extended Plackett-Luce model

The EPL proposed by Mollica & Tardella, 2014 relies on the relaxation of the conventional forward order assumption of the popular PL class through the introduction of the *reference order* parameter $\rho = (\rho(1), \ldots, \rho(K))$, indexing the position assignment order. So, the generic entry $\rho(t)$ indicates the rank attributed at the t-th stage of the ranking process and the entire vector ρ is a discrete parameter represented by a permutation of the first K integers. The probability of a generic ordering under the EPL can be written as

$$\mathbf{P}_{\text{EPL}}(\pi^{-1}|\rho, \underline{p}) = \mathbf{P}_{\text{PL}}(\pi^{-1}\rho|\underline{p}) = \prod_{t=1}^{K} \frac{p_{\pi^{-1}(\rho(t))}}{\sum_{v=t}^{K} p_{\pi^{-1}(\rho(v))}} \qquad \pi^{-1} \in \mathcal{S}_K.$$

The support parameters p_i's are proportional to the probabilities for each item to be selected at the first stage and, hence, to be ranked in the position indicated by the first entry of ρ.

2.2 Mixture model setup

In the EPL finite mixture scenario, one assumes that the random sample of N orderings $\underline{\pi}^{-1} = (\pi_1^{-1}, \ldots, \pi_N^{-1})$ is drawn from an *heterogenous population* represented by a convex combination of G *subpopulations* (or *groups*), each of which is modelled with a specific EPL distribution. Formally, we set

$$\pi_s^{-1}|\underline{\rho}, \underline{p}, \underline{\omega} \overset{\text{iid}}{\sim} \sum_{g=1}^{G} \omega_g \mathbf{P}_{\text{EPL}}(\pi_s^{-1}|\rho_g, \underline{p}_g),$$

where ρ_g, $\underline{p}_g$ and ω_g are, respectively, the reference order, the support parameters and the weight of the g-th mixture component.

In order to make Bayesian inference for the G-component EPL mixture analytically tractable, a joint data augmentation strategy combining two sets of latent variables has to be suitably introduced, specifically:

1. the unobserved group labels of each sample unit $s = 1, \ldots, N$

$$z_{sg} = \begin{cases} 1 & \text{if unit } s \text{ belongs to the } g\text{-th mixture component,} \\ 0 & \text{otherwise;} \end{cases}$$

2. the latent quantitative variables $\underline{y} = (y_{st})$ for $s = 1, \ldots, N$ and $t = 1, \ldots, K$, associated to each entry of the data matrix and linked to the component memberships $\underline{z}$ through the following parametric assumption

$$f(\underline{y}|\underline{\pi}^{-1}, \underline{z}, \underline{\rho}, \underline{p}) = \prod_{s=1}^{N} \prod_{t=1}^{K} f_{\text{Exp}} \left(y_{st} \middle| \prod_{g=1}^{G} \left(\sum_{\nu=t}^{K} p_{g\pi_s^{-1}(\rho_g(\nu))} \right)^{z_{sg}} \right).$$

Thus, the complete-data likelihood can be written as

$$L_c(\underline{\rho}, \underline{p}, \underline{\omega}, \underline{y}, \underline{z}) = \prod_{s=1}^{N} \prod_{g=1}^{G} \left(\omega_g \prod_{i=1}^{K} p_{gi} e^{-p_{gi} \sum_{t=1}^{K} \delta_{stig} y_{st}} \right)^{z_{sg}},$$

where

$$\delta_{stig} = \begin{cases} 1 & \text{if } i \in \{\pi_s^{-1}(\rho_g(t)), \ldots, \pi_s^{-1}(\rho_g(K))\}, \\ 0 & \text{otherwise.} \end{cases}$$

To complete the Bayesian model specification, we considered the following joint prior distribution for the unknown parameters $(\underline{\rho}, \underline{p}, \underline{\omega})$

$$\rho_g \overset{\text{iid}}{\sim} \text{Unif}\{\mathcal{S}_K\} \qquad p_{gi} \overset{\text{i}}{\sim} \text{Ga}(c_{gi}, d_g) \qquad \underline{\omega} \sim \text{Dir}(\alpha_1, \ldots, \alpha_G),$$

where the Gamma densities are indexed by the shape and rate parameters.

2.3 Estimation via MCMC methods

Under the Bayesian model setup described in Section 2.2, the MCMC method proposed by Mollica & Tardella, 2019 to estimate the basic EPL can be easily adapted for the G-component EPL mixture. The outline of the $(l+1)$-th iteration of the tuned joint Metropolis-Hasting within Gibbs sampling algorithm to approximate the posterior distribution turns out to be

$$\begin{aligned}
\underline{\omega}^{(l+1)}|\underline{z}^{(l)} &\sim \text{Dir}\left(\alpha_1+N_1^{(l)},\dots,\alpha_G+N_G^{(l)}\right),\\
\rho_g^{(l+1)},\underline{p}'_g|\underline{\pi}^{-1},\underline{z}^{(l)} &\sim \mathcal{K}_{\text{TJM}}\circ\mathcal{K}_{\text{SM}},\\
y_{st}^{(l+1)}|\pi_s^{-1},\underline{z}_s^{(l)},\underline{\rho}^{(l+1)},\underline{p}' &\sim \text{Exp}\left(\prod_{g=1}^{G}\left(\sum_{i=1}^{K}\delta_{stig}^{(l+1)}p'_{gi}\right)^{z_{sg}^{(l)}}\right),\\
p_{gi}^{(l+1)}|\underline{\pi}^{-1},\underline{y}^{(l+1)},\underline{z}^{(l)},\rho_g^{(l+1)} &\sim \text{Ga}\left(c_{gi}+N_g^{(l)},d_g+\sum_{s=1}^{N}z_{sg}^{(l)}\sum_{t=1}^{K}\delta_{stig}^{(l+1)}y_{st}^{(l+1)}\right),\\
\underline{z}_s^{(l+1)}|\pi_s^{-1},\underline{y}_s^{(l+1)},\underline{\rho}^{(l+1)},\underline{p}^{(l+1)},\underline{\omega}^{(l+1)} &\sim \text{Multinom}\left(1,\left(m_{s1}^{(l+1)},\dots,m_{sG}^{(l+1)}\right)\right),
\end{aligned}$$

where $N_g^{(l)}=\sum_{s=1}^{N}z_{sg}^{(l)}$ and

$$m_{sg}^{(l+1)}\propto\omega_g^{(l+1)}\prod_{i=1}^{K}p_{gi}^{(l+1)}e^{-p_{gi}^{(l+1)}\sum_{t=1}^{K}\delta_{stig}^{(l+1)}y_{st}^{(l+1)}}.$$

With $\mathcal{K}_{\text{TJM}}\circ\mathcal{K}_{\text{SM}}$ we denote the composition of two kernels, namely a tuned joint Metropolis (TJM) and a local swap move (SM) needed to solve the reference order simulation step and ensure an adequate mixing. For the mixture setting, the TJM and the SM are performed on the subsamples determined by the group memberships to iteratively draw the specific reference orders ρ_g.

The determination of the optimal number of mixture components can be addressed with the popular DIC (Spiegelhalter *et al.* , 2002).

References

MARDEN, J. I. 1995. *Analyzing and modeling rank data.* Monographs on Statistics and Applied Probability, vol. 64. Chapman & Hall.

MOLLICA, C., & TARDELLA, L. 2014. Epitope profiling via mixture modeling of ranked data. *Statistics in Medicine*, **33**(21), 3738–3758.

MOLLICA, C., & TARDELLA, L. 2017. Bayesian mixture of Plackett-Luce models for partially ranked data. *Psychometrika*, **82**(2), 442–458.

MOLLICA, C., & TARDELLA, L. 2019. Bayesian analysis of ranking data with the Extended Plackett-Luce model. *(submitted)*.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P., & VAN DER LINDE, A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.

ISSUES IN NONLINEAR TIME SERIES MODELING OF EUROPEAN IMPORT VOLUMES

Gianluca Morelli[1] and Francesca Torti[2]

[1] Department of Economics and Management, University of Parma, Parma,
(e-mail: gianluca.morelli@unipr.it)

[2] European Commission, Joint Research Centre, Ispra,
(e-mail: francesca.torti@ec.europa.eu)

ABSTRACT: This work is rooted in the statistical anti-fraud activity conducted by the European Commission for the protection of the financial interests of the European Union. We extend a recent framework for detecting outliers and level shifts in short time series that may have trend and seasonal patterns. Our contribution focuses on the inclusion of an autoregressive component in the model and on model selection. We also provide substantial evidence of the importance of fraud detection analysis for policy support.

KEYWORDS: Fraud detection, level shift, robust time series analysis.

We support the European Commission in the protection of the financial interests of the European Union, as foreseen by the founding treaties. We use statistical methods to address fraud control problems such as deflection of trade and mis-declaration of product and origin. The patterns to detect include *upward spikes* in trade flows and structural changes such as *level shifts*.

This contribution extends a new framework introduced by Rousseeuw et al. (2019) for detecting outliers (isolated or consecutive) and level shifts in short time series that may have trend and seasonal patterns, possibly relevant in our anti-fraud context. The original framework is based on a parametric approach to estimate level shifts that differs from the nonparametric smoothing methods in Fried and Gather (2007) or robust methods for REGARIMA models (Bianco et al.(2001)). The approach combines ideas from the FastLTS algorithm for robust regression with alternating least squares. Software for this framework is available in the MATLAB FSDA toolbox (http://fsda.jrc.ec.europa.eu or http://rosa.unipr.it/fsda.html).

The model works well on the typical trade time series, which are quite short: typically few tens of observations, one per month. When the time series are longer, the dependencies on the previous values cannot be neglected and an auto regressive component of some type needs to be included. This is our first proposal.

There are several thousands of relevant combinations of a product at fraud risk, a country of origin and a country of destination to analyze each month. This requires an automatic and computationally efficient approach that is able to report accurate information on outliers and the positions and amplitudes of level shifts. The precondition for this is to be able to identify the proper model for each case: our second proposal is a robust procedure conceived for the selection of the model components, in the spirit of Occam's razor.

References

Bianco, A.M., Garca Ben, M., Martnez, E.J. & Yohai, V.J. 2001. Outlier detection in regression models with arima errors using robust estimates. Journal of Forecasting 20(8):565579, DOI 10.1002/for.768, URL http://dx.doi.org/10.1002/for.768

Box, G.E.P. & Jenkins, G.M. 1976. Time Series Analysis: Forecasting and Control. Holden Day, San Francisco.

Fried, R. & Gather, U. 2007. On rank tests for shift detection in time series. Computational Statistics & Data Analysis 52(1):221233.

Rousseeuw, P.J., Perrotta, D., Riani, M. & Hubert, M. 2019. Robust Monitoring of Time Series with Application to Fraud Detection. Econometrics and Statistics, Volume 9, Pages 108-121.

Gaussian Parsimonious Clustering Models with Covariates and a Noise Component

Keefe Murphy[1] and Thomas Brendan Murphy[1]

[1] University College Dublin, (e-mail: keefe.murphy@ucdconnect.ie, brendan.murphy@ucd.ie)

Abstract: We consider model-based clustering methods for continuous, correlated data that account for external information available in the presence of mixed-type fixed covariates by proposing the MoEClust suite of models. These models allow different subsets of covariates to influence the component weights and/or component densities by modelling the parameters of the mixture as functions of the covariates. A familiar range of constrained eigen-decomposition parameterisations of the component covariance matrices are also accommodated. This paper thus addresses the equivalent aims of including covariates in Gaussian parsimonious clustering models and incorporating parsimonious covariance structures into all special cases of the Gaussian mixture of experts framework. The MoEClust models demonstrate significant improvement from both perspectives in applications to both univariate and data sets. Novel extensions to include a uniform noise component for capturing outliers and to address initialisation of the EM algorithm, model selection, and the visualisation of results are also proposed.

Keywords: Model-based clustering, mixtures of experts, multivariate response, covariates, noise component.

ILLUMINATION IN DEPTH ANALYSIS*

Stanislav Nagy[1]and Jiří Dvořák[1]

[1] Charles University, Prague, Faculty of Mathematics and Physics,
(e-mail: nagy@karlin.mff.cuni.cz)

ABSTRACT: Many classification procedures based on data depth suffer from the "outsider problem" — observations outside the convex hulls of the training data remain unclassified. We use the paradigm of illumination from convex geometry to solve this problem. Simultaneous use of the halfspace depth and illumination allows to devise affine invariant, highly robust classification rules that are asymptotically optimal in a broad class of scenarios.

KEYWORDS: depth, halfspace depth, illumination, optimality, Tukey depth.

1 Depth in classification

Data depth substitutes quantiles and order statistics when multivariate datasets, or more generally, multivariate probability measures, are involved. Proposed by Tukey, 1975, the (halfspace) depth of a point $x \in \mathbb{R}^d$ with $d \geq 1$ with respect to a Borel probability measure P on $\mathbb{R}^d$ is given as the minimum P-probability of a halfspace that contains x, that is

$$hD(x;P) = \inf_{u \in \mathbb{R}^d \setminus \{0\}} \mathsf{P}\left(\langle x, X\rangle \leq \langle x, u\rangle\right), \tag{1}$$

where X is a random vector with distribution P. $hD(\cdot;P)$ applied to $P = P_n$ the empirical measure of a random sample in $\mathbb{R}^d$ ranks the observations from the deepest ones (also called depth medians) to the least deep ones far from the central part of the dataset. For a recent survey see Nagy *et al.*, 2019.

Here, we focus on the use of the depth in classification. Given two independent random samples $P_{n_i,i}$ from unknown, different probability distributions P_i, and a single point x that was generated from one of P_i, $i = 1,2$, our task is to find which probability measure generated x. Equivalently, we want to determine to which of the data clouds $P_{n,i}$ the point x fits better. The *maximum depth classifier* assigns x to that group which maximizes $hD(x;P_{n,i})$. As shown by

*This work was supported by the grant 19-16097Y of the Czech Science Foundation, and by the PRIMUS/17/SCI/3 project of Charles University.

Ghosh & Chaudhuri, 2005, under appropriate conditions it is asymptotically optimal. Its performance can be substantially enhanced using more sophisticated depth-based techniques, see e.g. Li *et al.*, 2012.

The known depth-based classification rules are fully affine invariant, relatively fast and somewhat robust. They also perform very well in practice, with one notable exception that Lange *et al.*, 2014 called the *outsider problem*. It is easy to see that $hD(x; P_{n,i}) = 0$ for all x outside both convex hulls of the training samples. Such points naturally remain unclassified when depth is used. This appears to be a challenging problem, especially in higher dimensions d where many points remain unclassified due to the curse of dimensionality.

We solve this problem in a way that is conceptually and computationally simple, and at the same time affine invariant and highly robust. It is based on *illumination*, a tool dual to the halfspace depth. It allows to rank points outside the convex hulls of $P_{n,i}$ in a way analogous to the depth.

The results presented in this note with complete proofs will appear as Nagy & Dvořák, 2019.

2 Floating bodies and illumination

A *convex body* is a compact convex subset of $\mathbb{R}^d$ with non-empty interior. Convex bodies can be identified with uniform probability measures on those sets. Remarkably, the halfspace depth with respect to these uniform measures has been studied in geometry for decades — the *floating bodies*, known since 1820's, and their variants can be shown to be equivalent with the depth (1). For an account on the history and applications of floating bodies, and their connections to the depth, we refer to Nagy *et al.*, 2019.

Illumination, first advanced by Werner, 1994, serves in geometry as a notion complementary to floating bodies. While floating bodies rank points inside the convex body K, the illumination compares the fit of points outside of K. For a convex body K and $x \in \mathbb{R}^d$, the *illumination* of x onto K is

$$I(x; K) = \mathrm{vol}_d(\mathrm{co}(K \cup \{x\})) / \mathrm{vol}_d(K),$$

where $\mathrm{vol}_d(\cdot)$ is the Lebesgue measure, and $\mathrm{co}(\cdot)$ is the convex hull operator. Note that if $x \in K$, $I(x; K) = 1$. Many important properties of the illumination are known in geometry:

1. *illumination bodies* given as $\left\{x \in \mathbb{R}^d : I(x; K) \leq \delta\right\}$ with $\delta > 1$ are nested affine equivariant convex supersets of K. For K an ellipsoid they are also ellipsoids;

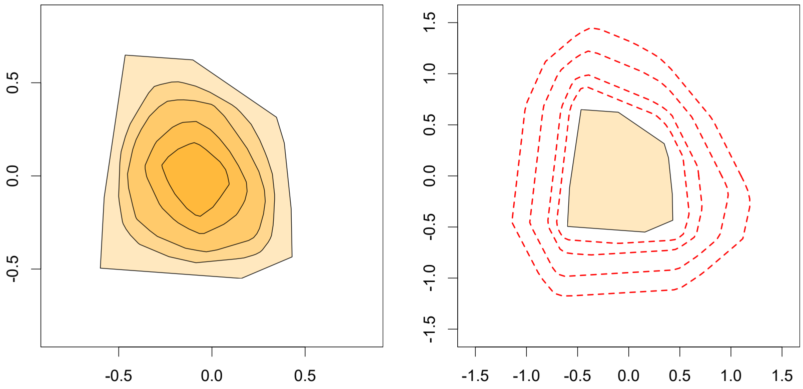


Figure 1. *A convex body and several of its halfspace depth central regions (left panel), and illumination bodies (right panel).*

2. the rate at which illumination bodies approach K as $\delta \to 1+$ from the outside is proportional to the rate at which (convex) floating bodies (and the depth) approach K as $\delta \to 0+$ from the inside;
3. illumination bodies and convex bodies are dual to each other when one takes into account polarity considerations.

Thus, the ranking of points according to their illumination is linked to that based on the depth. Consider now illumination for probability measures: for $x \in \mathbb{R}^d$, P a probability measure on $\mathbb{R}^d$ and $\alpha \in (0, \sup_{y \in \mathbb{R}^d} hD(y;P))$, the α-*illumination* of x onto P is given by

$$I_\alpha(x;P) = I(x;P_\alpha), \qquad \text{where } P_\alpha = \left\{ y \in \mathbb{R}^d : hD(y;P) \geq \alpha \right\} \tag{2}$$

is an upper level set of the depth (1). In other words, we illuminate onto a depth-central region P_α. The latter regions are known to be adequate representatives of both location and scatter of the majority of mass of P. In fact, for many distributions they are known to characterize P completely.

3 Robust LDA

Suppose we are given random samples of sizes $n = 500$ from bivariate normal distributions $P_i = \mathrm{N}_2(\mu_i, I_2)$ with $\mu_1 = (0,0)^\mathsf{T}$, $\mu_2 = (2,2)^\mathsf{T}$, P_1 contaminated with several outliers from $\mathrm{N}_2\left((20,20)^\mathsf{T}, I_2\right)$. Ghosh & Chaudhuri, 2005 demonstrated that when no outliers are present, the maximum depth classifier is asymptotically optimal. We propose to assign x into P_i if its illumination (2) with respect to $P_{n_i,i}$ is minimal. Only if $I_{\alpha_1}(x;P_{n,1}) = I_{\alpha_2}(x;P_{n,2})$ we use the maximum halfspace depth to classify x. The tuning parameters α_i are chosen so that the central regions in (2) contain 50 % of the data. In Table 1 we compare the maximum depth classifier (*hD*), LDA, and our new procedure (I), in terms of the average misclassification rate over 100 runs. In most settings, illumination convincingly outperforms both the maximum depth rule and LDA. Further experiments are described in Nagy & Dvořák, 2019.

	All points			Outsiders		
	I	LDA	*hD*	I	LDA	*hD*
0 %	0.079 (0.0062)	0.079 (0.0061)	0.086 (0.0069)	0.051 (0.030)	0.048 (0.030)	—
1 %	0.080 (0.0063)	0.083 (0.0064)	0.106 (0.0089)	0.042 (0.033)	0.053 (0.042)	—
5 %	0.080 (0.0060)	0.176 (0.0477)	0.168 (0.0151)	0.047 (0.042)	0.170 (0.135)	—
10 %	0.094 (0.0077)	0.493 (0.0633)	0.232 (0.0190)	0.054 (0.042)	0.520 (0.149)	—

Table 1. *Average and standard deviation (in brackets) of misclassification rates in a numerical experiment depending on the contamination levels. In the first part of the table, all testing data are considered. In the second part, only outsiders are taken.*

References

GHOSH, A. K., & CHAUDHURI, P. 2005. On maximum depth and related classifiers. *Scand. J. Statist.*, **32**(2), 327–350.

LANGE, T., MOSLER, K., & MOZHAROVSKYI, P. 2014. Fast nonparametric classification based on data depth. *Statist. Papers*, **55**(1), 49–69.

LI, J., CUESTA-ALBERTOS, J. A., & LIU, R. Y. 2012. *DD*-classifier: nonparametric classification procedure based on *DD*-plot. *J. Amer. Statist. Assoc.*, **107**(498), 737–753.

NAGY, S., & DVOŘÁK, J. 2019. *Illumination depth*. Under review.

NAGY, S., SCHÜTT, C., & WERNER, E. M. 2019. Halfspace depth and floating body. *Stat. Surv.*, **13**, 52–118.

TUKEY, J. W. 1975. Mathematics and the picturing of data. *Pages 523–531 of: Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2.* Canad. Math. Congress, Montreal, Que.

WERNER, E. M. 1994. Illumination bodies and affine surface area. *Studia Math.*, **110**(3), 257–269.

COPULA-BASED NON-METRIC UNFOLDING ON AUGMENTED DATA MATRIX

Marta Nai Ruscone[1] and Antonio D'Ambrosio[2]

[1] School of Economics and Management, LIUC Università Cattaneo, (e-mail: mnairuscone@liuc.it)

[2] Department of Economics and Statistics, University of Naples Federico II, (e-mail: antdambr@unina.it)

ABSTRACT: A multidimensional unfolding technique that is not prone to degenerate solutions and is based on multidimensional scaling of a complete data matrix is proposed. We adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, by using Copulas-based association measures among rankings (the individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings). The proposed technique leads to acceptable recovery of given preference structures.

KEYWORDS: copulas, unfolding, multidimensional scaling.

1 The copulas function

Copulas are functions that join multivariate distribution functions to their marginal distribution functions (Nelsen, 2013). They describe the dependence structure existing across pairwise marginal random variables. In this way we can consider bivariate distributions with dependency structures different from the linear one that characterizes the multivariate normal distribution.

A bivariate copula $C : I^2 \to I$, with $I^2 = [0,1] \times [0,1]$ and $I = [0,1]$, is the cumulative bivariate distribution function of a random variable (U_1, U_2) with uniform marginal random variables in [0,1]

$$C(u_1, u_2; \theta) = P(U_1 \leq u_1, U_2 \leq u_2; \theta), \quad 0 \leq u_1 \leq 1 \quad 0 \leq u_2 \leq 1 \tag{1}$$

where θ is a parameter measuring the dependence between U_1 and U_2.

The following theorem by Sklar (Nelsen, 2013) explains the use of the copula in the characterization of a joint distribution. Let (Y_1, Y_2) be a bivariate random variable with marginal cdfs $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ and joint cdf $F_{Y_1,Y_2}(y_1, y_2; \theta)$, then there always exists a copula function $C(\cdot, \cdot; \theta)$ with $C : I^2 \to I$ such that

$$F_{Y_1,Y_2}(y_1, y_2; \theta) = C\big(F_{Y_1}(y_1), F_{Y_2}(y_2); \theta\big), \quad y_1, y_2 \in \mathbb{R}. \tag{2}$$

Conversely, if $C(\cdot,\cdot;\theta)$ is a copula function and $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are marginal cdfs, then $F_{Y_1,Y_2}(y_1,y_2;\theta)$ is a joint cdf.
If $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are continuous functions then the copula $C(\cdot,\cdot;\theta)$ is unique. Moreover, if $F_{Y_1}(y_1)$ and $F_{Y_2}(y_2)$ are continuous the copula can be found by the inverse of (2):

$$C(u_1,u_2) = F_{Y_1,Y_2}(F_{Y_1}^{-1}(u_1), F_{Y_2}^{-1}(u_2)) \tag{3}$$

with $u_1 = F_{Y_1}(y_1)$ and $u_2 = F_{Y_2}(y_2)$. This theorem states that each joint distribution can be expressed in term of two separate but related issues, the marginal distributions and the dependence structures between them. The dependence structure is explained by the copula function $C(\cdot,\cdot;\theta)$. Moreover the (2) provides a general mechanism to construct new multivariate models in a straightforward manner. By changing the copula function we can construct new bivariate distributions with different dependence structures, with the association parameter indicating the strength of the dependence, also different from the linear one that characterizes the multivariate normal distribution.

Each copula is related to the most important measures of dependency: the Pearson correlation coefficient and the Spearman grade correlation coefficient. The Spearman grade correlation coefficient (see Nelsen, 2013 pp. 169-170 for the definition of the grade correlation coefficient for continuous random variables) measure the association between two variables and can be expressed as a function of the copula. More precisely, if two random variables are continuous and have copula C with parameter θ, then the Spearman grade correlation is

$$\rho_s(C) = 12\int_{I^2} C_\theta(u_1,u_2)du_1du_2 - 3. \tag{4}$$

For continuous random variables it is invariant with respect to the two marginal distributions, i.e. it can be expressed as a function of its copula. This property is also known as 'scale invariance'. Note that not all measures of association satisfy this property, e.g. Pearson's linear correlation coefficient (Embrechts *et al.* , 2002).

In the following, we focus on observations Y_{ik} of the latent continuous random variable Y^*_{ik}, describing the preference of the consumer i ($i \in N = \{1,...,n\}$) for the object k. Let $y_i = (y_{i1},...,y_{ik})$ be the vector of ranks of consumer i for the k objects, where y_{ik} is the rank of object k for the subject $i-th$. Be $U = F(Y^*_{ik})$ and $V = F(Y^*_{jk})$ the marginal cumulative distributions (cds). We assume that (Y_{ik}, Y_{jk}) correspond to the bivariate discrete random variable obtained by a discretization of the continuous latent variable $(U = F(Y^*_{ik}), V = F(Y^*_{jk}))$ with support $[0,1]\times[0,1]$ and cdf given by $C_\theta(\cdot,\cdot)$.

Let $A_{r,s} = [u_{r-1}, u_r] \times [v_{s-1}, v_s]$ $r,s = 1,...,k$ be rectangles defining the discretization. Let $p_{k,k}$ be the joint probabilities corresponding to the rectangle $A_{r,s}$ for $r,s = 1,...,k$ with value $1/k$ if the pair (y_{ik}, y_{jk}) is observed and 0 otherwise. Let $V_{C_\theta}(A_{11}), \ldots, V_{C_\theta}(A_{kk})$ be the volumes of the rectangles under the copula C_θ, then there exists a unique element in the family of copula for which the following relationship holds true:

$$(V_{C_\theta}(A_{11}), \ldots, V_{C_\theta}(A_{kk})) = (p_{11}, \ldots, p_{kk}). \tag{5}$$

Given the ranking of two subjects i and j, a $k \times k$ contingency table K_{rs} $(r,s = 1,...,k)$ is defined. A cell in this table takes value $1/k$ if (y_{ik}, y_{jk}) is observed and 0 otherwise. This contingency table provides the basis for our estimation procedure.

Fixed the copula C_θ and defined the Spearman grade correlation coefficients $\rho_s(C_\theta)$ (Nelsen, 2013) to each pair (Y_{ik}, Y_{jk}), $i \neq j$ with $i, j \in N$, we define the dissimilarity coefficient d_{ij}:

$$d_{ij} = \sqrt{1 - \frac{\rho_s + 1}{2}} \tag{6}$$

where ρ_s performs well in measuring the agreement between two rankings Y_{ik} Y_{jk}.

Notice that other ways of findings a correlation-type distance matrix have been provided in the literature (Kaufman & Rousseeuw, 2009). For instance, one may consider $d_{ij} = 1 - \rho$ or $d_{ij} = 1 - |\rho|$.

The parameter θ can be estimated via maximum likelihood. Estimating the value of the copula dependence parameter θ we obtain the grade of association between two rankings.

2 Unfolding as a special case of multidimensional scaling on Copulas based association between rankings

Unfolding applies multidimensional scaling (Cox & Cox, 2000) to an off-diagonal $n \times m$ matrix, usually representing the scores (or the rank) assigned to a set of m items by n individuals or judges (Borg & Groenen, 1997). The goal is to obtain two configuration of points representing the position of the judges (X) and the items (Y) in a reduced geometrical space. Each point representing the individuals is considered as an ideal point so that its distances to the object points correspond to the preference scores (Coombs, 1964). Unfolding can be seen as a special case of multidimensional scaling because the off-diagonal matrix is considered as a block of an ideal distance matrix in which both the within judges and the within items dissimilarities are missing. The presence of blocks of missing data causes the phenomenon of the so-called degenerate solutions, i.e., solutions that return excellent badness of fit measures but not graphically interpretable at all. To tackle the problem of degenerate solutions, several proposals have been presented in the literature (Borg & Groenen, 1997). By following the approach introduced by Van Deun *et al.* , 2007, we adopt the strategy of augmenting the data matrix, trying to build a complete dissimilarity matrix, and then applying any MDS algorithms. In order to augment the data matrix, we use Copulas-based association measures among rankings (the individuals), and between rankings and objects (namely, a rank-order representation of the objects through tied rankings). Both experimental evaluations and applications to well-known real data sets show that the proposed strategy produces non-degenerate non-metric unfolding solutions.

References

BORG, I., & GROENEN, P.J. 1997. *Modern multidimensional scaling*. Springer Series in Statistics. Springer-Verlag, New York. Theory and applications.

CHERUBINI, U., LUCIANO, E., & VECCHIATO, W. 2004. *Copula methods in finance*. John Wiley & Sons.

COOMBS, C.H. 1964. *A theory of data.* Wiley.

COX, T.F., & COX, M.A.A. 2000. *Multidimensional scaling*. Chapman and hall/CRC.

EMBRECHTS, P., MCNEIL, A., & STRAUMANN, D. 2002. Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, **1**, 176–223.

JOE, H. 1997. *Multivariate models and multivariate dependence concepts*. CRC Press.

KAUFMAN, L., & ROUSSEEUW, P. J. 2009. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

NELSEN, R.B. 2013. *An introduction to copulas*. Springer Science & Business Media.

VAN DEUN, K., HEISER, W.J., & DELBEKE, L. 2007. Multidimensional unfolding by nonmetric multidimensional scaling of Spearman distances in the extended permutation polytope. *Multivariate Behavioral Research*, **42**(1), 103–132.

A STATISTICAL MODEL FOR SOFTWARE RELEASES COMPLEXITY PREDICTION

Marco Ortu[1], Giuseppe Destefanis[2] and Roberto Tonelli[1]

[1] Department of Computer Science and Mathematics, University of Cagliari,
(e-mail: marco.ortu@unica.it, roberto.tonelli@dfs.unica.it)

[2] Brunel University London, London,
(e-mail: giuseppe.destefanis@brunel.ac.uk)

ABSTRACT: Micropatterns are a representation of design decisions in code, which can be detected from the source code with an automatic tool. These microstructures can help identify portions of code which should be improved (anti-patterns), or well-designed parts which need to be maintained. The definition of the concepts expressed in these design decisions is at class-level. In this paper, we present a longitudinal dataset of more than 200K software metrics, collected from 113 versions of Tomcat. The dataset can be used for various empirical research studies in software engineering, and we exploit the dataset to predict the complexity of a software release based on the release metrics. We used four different machine learning classifiers, and we found that the C5.0 algorithm is the best classifier with 0.96% of accuracy.

KEYWORDS: tomcat, micropatterns, software quality, prediction models.

1 Introduction

Software quality is a complex concept to measure, and efforts in this direction would be helpful for both software developers and managers in controlling and improving software development. Quality is quite an intangible concept, and when talking about software, given its immateriality, it becomes even more elusive. What are the characteristics of a high-quality software? We can start discussing meeting the requirements defined by the customer who commissioned the software, or about the results of testing activities, or the evolution of the software metrics throughout all the phases of the development. Numerous concepts have been introduced and used to control, measure and engineer the software development process, such as software metrics [Chidamber & Kemerer, 1994] and design patterns [Gamma, 1995], which represent a general concept or methodology used for designing a piece of software.

In this work, using our previous dataset [Destefanis *et al.* , 2018] as baseline, we present an enriched dataset containing information about micropatterns detected from 113 versions of Tomcat[*], an open source Java Servlet Container developed by the Apache Software Foundation (from version 3.3.2 to version 8.0.9) and largely used for empirical software engineering studies [Monperrus & Martinez, 2012, Destefanis *et al.* , 2017].

The dataset (in SQL format), which contains 113 SQL views with metrics and micropatterns for all the detected classes, is openly available at the following link *https://bitbucket.org/giuseppedestefanis/tomcatmpattern19*. We exploit the dataset to answer the following research question:

Is it possible to predict the Complexity of a release based on release metrics?

This paper is structured as follows. In Section II, we illustrate the process of construction of the complexity prediction algorithms and the comparison among the selected classification algorithm.

2 Results

Is it possible to predict the Complexity of a release based on release metrics?

To answer this question, we built four machine learning algorithms to evaluate the complexity of a software release based on other software release metrics. Tomcat Software is written in Java, which is an Object-Oriented language; thus, the whole software is organised in classes. For each class, the dataset provides the version along with 57 software metrics. We first evaluated the average of each metric per software version obtaining a dataset of 113 versions, and computed the *Complexity* metric of a version as a binary variable of the *Average Cyclomatic Complexity* [Watson *et al.* , 1996], 0 meaning less than the median, 1 otherwise. We used the median to obtain a balanced dataset. We then restricted the number of metrics to 35 using correlation analysis and filtering out those metrics with a Pearson correlation greater than ±0.7.

We then used a feature selection algorithm [Gevrey *et al.* , 2003] to refine the remaining features. We used the algorithm with a *general linear model* as predictor and a *repeated-cross-validation* as method to measure the variables' importance. Figure 1 shows the importance of the 35 metrics selected. An importance of 0.50 means that the predictor survived the filter in half of the re-samples, thus we kept only those feature with importance greater than 0.5.

We then selected four machine learning algorithms, *C5.0*, *SVM*, *Bayes Generalized Lineal Model* and *KNN*. We used a *repeated-cross-validation* method to evaluate the performance of selected algorithms. Figure 2 shows

[*] http://tomcat.apache.org

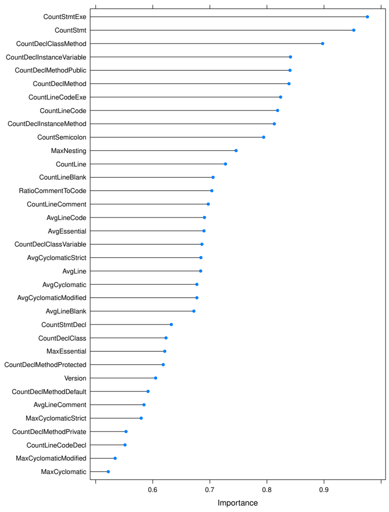


Figure 1. *Feature Selection*

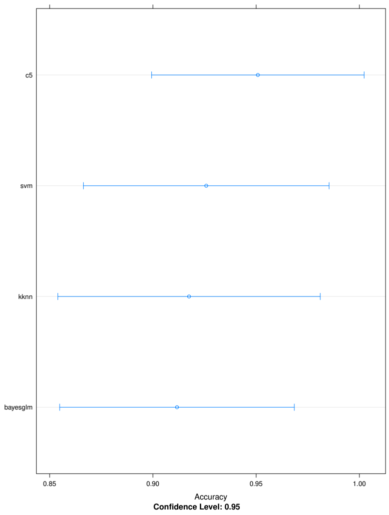


Figure 2. *Classification Comparison*

the results in terms of accuracy. The best classifier is the C5.0 algorithm with an average accuracy of 0.96%.

3 Conclusion

This paper provides a longitudinal micropatterns dataset collected from 113 versions of Tomcat (from version 3.3.2 to 8.0.9). We merged the micropatterns data with a previous Tomcat longitudinal dataset containing software metrics [Destefanis *et al.* , 2018]. We exploit this dataset to perform an analysis of release complexity prediction using the release metrics, we build four different machine learning classifiers, and we found that the best classifier is the C5.0 with and accuracy of 0.96%.

References

CHIDAMBER, S. R., & KEMERER, C. F. 1994. A metrics suite for object oriented design. *IEEE Transactions on software engineering*, **20**(6), 476–493.

DESTEFANIS, G., ORTU, M., COUNSELL, S., SWIFT, S., TONELLI, R., & MARCHESI, M. 2017. On the randomness and seasonality of affective metrics for software development. *Pages 1266–1271 of: Proceedings of the Symposium on Applied Computing*. ACM.

DESTEFANIS, G., ARZOKY, M., COUNSELL, S., SWIFT, S., ORTU, M., TONELLI, R., & MARCHESI, M. 2018. 113 times Tomcat: A dataset. *PeerJ Preprints*, **6**, e26491v1.

GAMMA, E. 1995. *Design patterns: elements of reusable object-oriented software*. Pearson Education India.

GEVREY, M., DIMOPOULOS, I., & LEK, S. 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, **160**(3), 249–264.

MONPERRUS, M., & MARTINEZ, M. 2012. Cvs-vintage: A dataset of 14 cvs repositories of java software.

WATSON, A. H., WALLACE, D. R., & MCCABE, T. J. 1996. *Structured testing: A testing methodology using the cyclomatic complexity metric*. Vol. 500. US Department of Commerce, Technology Administration, National Institute of Standards and Technology.

COMPARISON OF SERIOUS DISEASES MORTALITY IN REGIONS OF V4

Viera Pacáková[1] and Lucie Kopecká[1]

[1] Department of Mathematics and Quantitative methods, Faculty of Economics and Administration, University of Pardubice, (e-mail: viera.pacakova@upce.cz, lucie.kopecka1@student.upce.cz)

ABSTRACT: This article is about comparing the regions of V4 according to mortality caused by the most common serious diseases (cardiovascular and oncological diseases) and growing serious diseases mainly due to aging of population (mental diseases). Mortalities are indicators of health quality. The significant differences exist among individual countries of Europe or EU, mainly in case of Western and Eastern part. But the significant differences also exist among regions of these countries. This article is focused on the NUTS 2 regions of V4 countries (Poland, Czech Republic, Slovak Republic and Hungary). The main aim is to compare the V4 regions according to serious diseases mortality by using hybrid approach which combines multidimensional scaling (MDS) with linear ordering. The data were obtained from Eurostat database.

KEYWORDS: Serious diseases mortality, V4 regions, hybrid approach.

1 Introduction

The significant differences in health status of population exist among European countries. The health status in the Visegrad group (V4) involving the four following member countries the Czech Republic (CR), Hungary, Poland and Slovak Republic (SR) is not good in comparison with the most of European countries. According to Pacáková, Kopecká (2018) the countries such as Hungary and Slovak Republic did not record significant improvement in health status during the time period 2000 – 2015 in contrast with the CR and Poland. Inequalities in health status in European countries have been revealed by using multivariate statistical methods such as factor analysis, cluster analysis, multidimensional comparative analysis etc. The level of health status of population can be different not only among the countries but also among the individual regions. This is the reason for assessment of mortality due to serious diseases at regional level according to NUTS 2 classification in V4 countries based on Eurostat health database, 2018. According to NUTS 2 classification Poland has 16 regions, CR has 8 regions, SR has 4 regions and finally Hungary has 7 regions. The cardiovascular and oncological diseases are the most serious causes of death across all these countries. Also mental disorders are on the rise mainly due to aging of population in V4 by health profiles of countries (European Commission, 2017).

2 Data and Methods

The main objective of this article is to compare the 35 NUTS 2 regions of V4 countries by using hybrid approach based of variables which represent crude death rate per 100 000 population due to following 12 oncological, cardiovascular and mental diseases: *C1* – malignant neoplasm of stomach, *C2* – malignant neoplasm of colon and rectum, *C3* – malignant neoplasm of liver, *C4* – malignant neoplasm of pancreas, *C5* – malignant neoplasm of trachea, bronchus and lung, *C6* – malignant neoplasm of breast, *C7* – malignant neoplasm of cervix uteri, *C8* – leukaemia, *H1* – ischaemic heart diseases, *H2* – acute myocardial infarction, *H3* – cerebrovascular diseases, *M1* – mental and behavioural disorders.

The data source has been the database Eurostat, 2018.

Hybrid approach combines *multidimensional scaling* (MDS) with *linear ordering*, as describe Walesiak, 2016. In general, the goal of this analysis is to detect meaningful underlying dimensions that allow to explain observed similarities or dissimilarities (distances) between the investigated objects. It is possible to analyse any kind of similarity or dissimilarity matrix with MDS. The input matrix into MDS is for example matrix of distances between objects *n* x *n*, where *n* is number of objects. Then the matrix is analysed and it is specified that the distances will be reproduced in *R*-dimensions. In general then, MDS attempts to arrange objects in a space with a particular number of dimensions (two-dimensional is the most common) so as to reproduce the observed distances. The actual orientation of axes in the final solution is arbitrary. We could rotate the map in any way, because the distances between objects remain the same. Multidimensional scaling is a way to *"rearrange"* objects in an efficient manner, to reach configuration that best approximates the observed distances, as describes Schiffman et al., 1981.

Finally, the results of MDS (the coordinate system for the first and second dimension) are used for linear order of the objects. The objects are ordered according to *aggregate measure d_i*, which is given by formula (1)

$$d_i = 1 - \sqrt{\sum_{j=1}^{2}\left(v_{ij} - v_{+j}\right)^2} \Big/ \sqrt{\sum_{j=1}^{2}\left(v_{+j} - v_{-j}\right)^2}, \qquad (1)$$

where v_{ij} is *j*-th coordinate for the *i*-th object, v_{+j} is *j*-th coordinate for the Pattern object and v_{-j} is *j*-th coordinate for the Anti-pattern object.

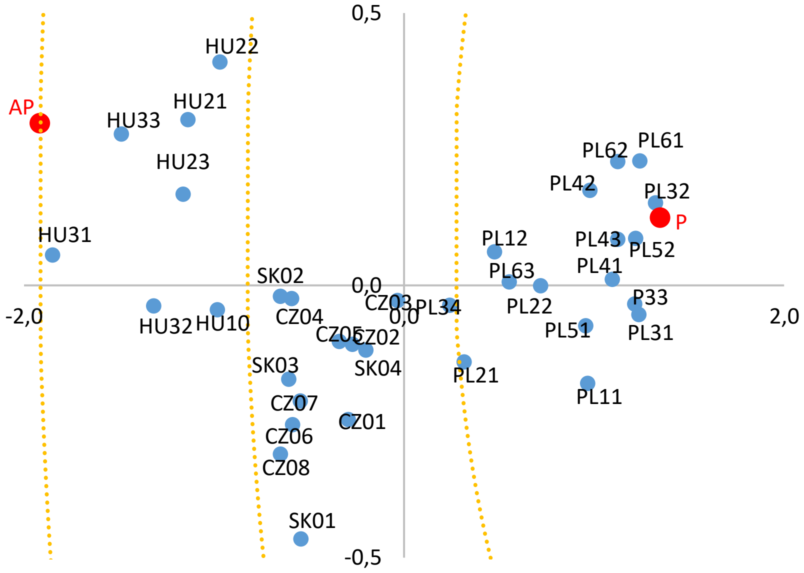
The high values of aggregate measure indicate low level of mortalities (Pattern object equals to 1) and low values of this measure indicate high level of mortalities (Anti-pattern object equals to 0), see Walesiak, 2016.

3 Results and Discussion

The input matrix into the hybrid approach is a matrix of Euclidean distances which is obtained from original data matrix. As mentioned above, original data matrix contains data of 12 mortality variables for 35 objects (regions). The size of Euclidean distance matrix is 37 x 37, because two "*artificial*" regions are added, namely *Pattern object* (P) and *Anti-pattern object* (AP). The Pattern object is created by the minimum values of all original variables and Anti-pattern object is constructed from their maximums.

Figure 1 displays the results of MDS in two dimensions. The largest distance exists between hypothetic “best” P and “worst” AP objects. It is possible to classify and to identify the regions with the similar level of mortalities but with different combinations based on Figure 1. Three circles which divide the distance between P and AP objects into the three equally large parts are able to create three logical groups of regions with the similar level of health quality. Quality of health measured by cardiovascular, oncological and mental mortality indicators is the worst in Hungary. The regions of Hungary are closer to AP object which has been created by maximum level of these mortalities. The middle annulus situated exactly between P and AP objects is created by the regions of the SR and the CR. Similar health situation exists among regions of the SR and Moravia, Silesia and Northwest regions of the CR. These regions are again closer to AP. The best quality of health according to mentioned mortality indicators is in regions of Poland in comparison with regions of CR, SR and Hungary. Only two polish regions, Podlaskie and Malopolskie, lies nearly in the same circle which is the closest to the P. This situation indicates nearly the same level of quality of health but with different combinations of mortalities.

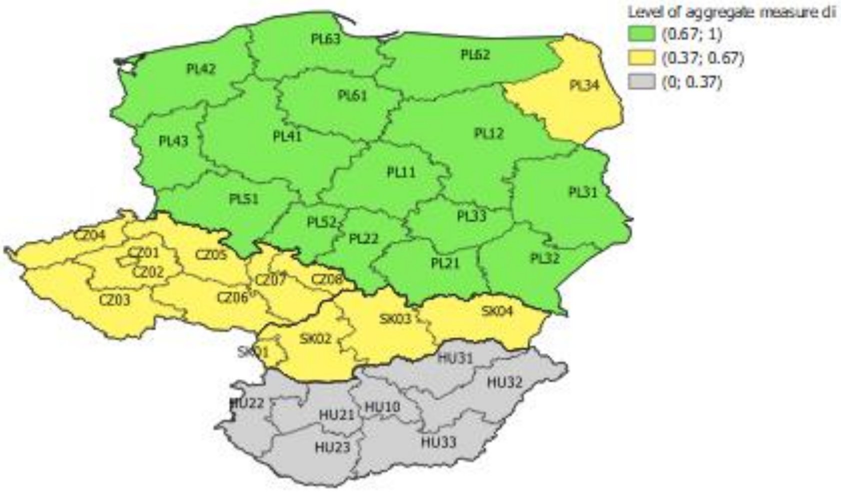
Figure 1: The results of MDS in two-dimensions



Source: own calculations (Eurostat, 2018)

Figure 2 displays the map of the three groups of V4 NUTS 2 regions according to aggregate measure by formula (1) and allows visual comparison of the health status.

Figure 2: Visualization of inequalities in NUTS 2 regions of V4 according to d_i



SOURCE: OWN CALCULATIONS (EUROSTAT, 2018)

4 Conclusion

The main aim of this article has been to compare the NUTS 2 regions of V4 by health status of population that as a multidimensional category has been specified by twelve indicators of mortality due to serious diseases. For the comparison, the hybrid approach has been chosen. The graphical outputs allows a visual overview of the population's health status in individual countries of the V4 grouping and theirs regions.

References

EUROPEAN COMMISION. 2017. *Country Health Profiles*. Available from: https://ec.europa.eu/health/state/country_profiles_en.

EUROSTAT. 2018. Database: *Regional Statistics by NUTS classification*. Available from: https://ec.europa.eu/eurostat/web/regions/data/database

SCHIFFMAN, S. S., REYNOLDS, M. L. & YOUNG, F. W. 1981. *Introduction to Multidimensional Scaling. Theory, Methods and Applications.* Emerald Group Publishing Limited.

PACÁKOVÁ, V., KOPECKÁ, L. 2018. Inequalities in Health Status Depending on Socio-economic Situation in the European Countries. *E+M Economics and Management*, **21**, 4-20.

WALESIAK, M. 2016. Visualization of linear ordering results for metric data with the application of multidimensional scaling. *Ekonometria,* **2**, 9-21.

Price and Product Design Strategies for Manufacturers of Electric Vehicle Batteries: Inferences from Latent Class Analysis

Friederike Paetz[1]

[1] Department of Economics and Marketing, Clausthal Technical University,
(e-mail: friederike.paetz@tu-clausthal.de)

Abstract: Nowadays, battery electric vehicles (BEVs) constitute prominent alternatives to vehicles with combustion motors. Since the competition between manufacturers of BEV-batteries increases, it is key for them to design batteries, which perfectly meet the needs of BEV manufacturers. However, the needs of BEV manufacturers are actually derivative needs of BEV customers. We, therefore, conduct an empirical discrete choice experiment with BEV customers in China and estimated Latent Class analysis. We found substantial preference heterogeneity of BEV customers, which transfers into varying needs of BEV manufacturers w.r.t. batteries. Using these results, we worked out price and product design strategies for manufacturers of BEV-batteries.

Keywords: latent class analysis, discrete choice experiment, derivative demand.

1 Motivation

Nowadays, battery electric vehicles (BEVs) constitute prominent alternatives to vehicles with combustion motors. As it is documented, the demand of BEVs has increased tremendously in recent years: While in 2014 approx. 750,000 BEVs were registered worldwide, it nearly quintupled to 3.2 mio in 2017 (Statista, 2019). Here, China yields the most impressive growth rate for electro mobility and has emerged as the most important market for BEVs.

With an increasing demand for BEVs, the demand for BEVs' batteries derivatively increases, too. The battery is the core component of a BEV and is one of the main BEV's cost drivers. Since the competition between manufacturers of BEV's batteries is prevalent, battery manufacturers have to accurately design their products and use sophisticated pricing strategies to gain a competitive advantage.

A battery manufacturer could increase its competitive ability, if its battery solves the components of the problem cluster of electro mobility, namely, driving range, (charging) infrastructure and purchase price. Therefore, battery manufacturers have to take into account the preferences of BEV customers, who, in turn, determine the preferences of BEV manufacturers, as it is shown in Figure 1.

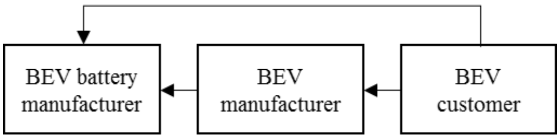


Figure 1: Preference relationships.

2 Empirical Analysis

In order to gain information on customers' preferences for BEVs, we use the data of a discrete choice experiment (DCE) that was conducted in China.[1] The final sample includes 194 respondents. The DCE was built up of 10 choice sets with three BEV alternatives and a 'no purchase' option. The BEVs were explained by the attributes driving range (150km, 250km, 350km), charging time (4h, 6h, 8h), purchase price (60,000¥, 160,000¥, 260,000¥) and car-body design (sedan, estate car, SUV). All attributes were chosen in accordance to the attributes in recent literature on DCE within the BEV category in China (Nie et al. 2018, Quia & Soopramanien, 2011). In addition, the first three attributes cover the problem cluster of electro mobility. The levels conform to the most prevalent realizations of the top 20 best-selling BEV models in China in 2017 (EV-Sales, 2017). However, we did not incorporate the models of Tesla, because Tesla's BEVs are much more expensive and yield a wider driving range and shorter charging time than **all** other top 20 BEVs in the Chinese market.

We used the data of eight choice sets for the estimation of Latent Class – Multinomial Logit (LC-MNL) models and considered two hold out choice sets. The estimation was performed with the Latent Class module of Sawtooth Software (Sawtooth Software, 2004). We considered effects-coding of the attributes and part-worth utilities for all attribute level. We estimated LC-MNL models for one up to eight segments and chose the segment-solution, that displays the best trade-off between fit (measured by ABIC and mean posterior (post.) segment memberships (memb.)) and predictive validity (measured by first choice (FC) hit rates). Table 1 displays the results of the information criteria and FC hit rates:

number of segments	**ABIC**	**Mean post. memb.**	**FC hit rates**
1	3780	100%	48%
2	3619	97%	52%
3	3552	91%	52%
4	3486	91%	58%
5	3435	90%	58%
6	3426	92%	59%
7	3429	90%	57%
8	3431	90%	56%

Table 1: Values of criteria for model selection

The six segment solution displays the best model fit, i.e. lowest ABIC value, and predictive validity, i.e., highest FC hit rate. In addition, the mean posteriori segment membership is markedly higher than those of the 5- or 7-segment-solution. Hence, we select the 6-segment-solution.

[1] We thank Yundi Cheng for collecting the data.

Table 2 contains the segment-specific part-worth utility estimates and the segment weights as well as the segment-specific relative attribute importances of the 6-segment-solution.

	seg. 1	seg. 2	seg. 3	seg. 4	seg. 5	seg. 6
weights	0.361	0.162	0.103	0.070	0.184	0.120
Part-worth utilities						
Driving range (in [km])						
150	-1.334	-2.044	-3.629	0.059	-0.224	-0.586
250	0.342	0.491	-0.162	-0.122	**0.152**	**0.322**
350	**0.992**	**1.553**	**3.792**	**0.062**	0.072	0.263
Charging time (in [h])						
4	0.221	**0.621**	**1.716**	-1.085	**0.471**	**0.135**
6	**0.244**	0.094	-0.353	0.193	-0.141	-0.158
8	-0.465	-0.715	-1.363	**0.892**	-0.330	0.023
Purchase price (in [¥])						
60,000	**0.856**	**0.664**	-1.157	-1.197	**0.329**	-0.561
160,000	0.222	0.072	0.102	**0.660**	0.223	0.168
260,000	-1.078	-0.735	**1.055**	0.537	-0.552	**0.393**
Car-body design						
estate car	0.175	-0.241	**0.153**	-0.225	-0.925	-0.142
sedan	-0.572	**0.448**	0.113	**0.309**	**1.015**	-1.564
SUV	**0.397**	-0.208	-0.266	-0.084	-0.090	**1.706**
Attribute importances (in [%])						
Driving range	**39.16**	**51.23**	**56.51**	4.05	9.40	16.73
Charging time	11.96	19.03	23.45	**43.41**	20.03	5.40
Purchase price	32.56	19.93	16.85	40.78	22.06	17.59
Car-body design	16.32	9.81	3.19	11.75	**48.51**	**60.28**

Highest segment-specific values are bold.

Table 2: Segment specific estimates

The inspection of Table 2 reveals, that all segments are of meaningful size. SUV customers attach the highest importance to the driving range and purchase price of the BEV (segment 1) or on car-body design (segment 6). Estate car customers (segment 3) attach the highest importance to the driving range and prefer higher prices, i.e., they view price as a quality signal. Sedan customers either exclusively care about the car-body design (segment 5) or about the attributes associated with the problem cluster of electro mobility (segment 2 and segment 4).

In order to derive inferences for a manufacturer of BEV-batteries, we could rely on the preferences of different BEV customers. Batteries build for SUVs and estate cars need a high reservoir capacity. Batteries for sedans must have a quick recharging technique as well as a high reservoir capacity. Customer of estate cars view a higher price as a quality signal, hence, batteries for estate cars could be offered by higher prices and, therefore, may inhibit higher costs to fulfil the preference towards a wide driving range and a quick charging time. For product design and price decisions, this means, that even identically constructed batteries could be offered by a higher price to manufacturers of electric estate cars, because they could smoothly offer higher prices to their BEVs' customers.

3 Conclusion

The increasing demand of BEVs derivatively increases the demand of batteries for BEVs. Since the competition is strong, battery manufacturers have to take the preferences of BEVs' manufacturers (and, therefore, the preferences of BEVs' customers) into account. Based on the data of a DCE, we estimated a Latent Class analysis and draw inferences concerning price and product design strategies for manufacturers of BEVs' batteries. We found that batteries build for SUVs, sedans and estate cars need a high reservoir capacity. In addition, batteries for sedans must have a quick recharging technique. Furthermore, we found customers of estate cars to view price as a quality signal. Therefore, batteries for electric estate cars could be more expensive (and could be sold at higher prices), because BEV manufacturers could smoothly pass their higher costs to their BEVs' customers.

References

EV SALES (2017) China September 2017. http://ev-sales.blogspot.com/2017/10/china-september-2017.html

NIE, Y., WANG, E. & GUO, Q. 2018. Examining Shanghai consumer preferences for electric vehicles and their attributes. *Discussion Paper Series DP2017-21, Research Institute for Economics & Business Administration*, Kobe University.

QUIAN, L., & SOOPRAMANIEN D. 2011. Heterogeneous consumer preferences for alternative fuel cars in China. *Transportation Research Part D: Transport and Environment*, **16**, 607-613.

SAWTOOTH SOFTWARE 2004. The CBC Latent Class Technical Paper. Version 3. https://www.sawtoothsoftware.com/download/techpap/lctech.pdf,

STATISTA 2019. Worldwide number of battery electric vehicles in use from 2012 to 2017, https://www.statista.com/statistics/270603/worldwide-number-of-hybrid-and-electric-vehicles-since-2009/

A MAHALANOBIS–LIKE DISTANCE FOR CYLINDRICAL DATA

Lucio Palazzo[1], Giovanni C. Porzio[2] and Giuseppe Pandolfo[3]

[1] Department of Economics and Statistics, University of Salerno,
(e-mail: lucio.palazzo@unicas.it)

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

[3] Department of Industrial Engineering, University of Naples Federico II,
(e-mail: giuseppe.pandolfo@unina.it)

ABSTRACT: A definition of a density–based distance function for data on the surface of cylinders is introduced. It is able to deal with the correlation structure of cylindrical data.

KEYWORDS: Mardia–Sutton distribution, linear–circular data.

1 Introduction

Cylindrical data arise from the joint distribution of a circular (i.e., bounded in $[0,2\pi)$) and a linear variable. This type of data arises in many scientific fields such as meteorology, geology and industry. One common example concerns the study of wind directions together with speed, or temperature.

Within this setting, little efforts have been dedicated to the problem of defining suitable distance measures. To clarify, the standard Euclidean distance on the cylinder does not take into account the data correlation structure. On the other hand, correlated cylindrical data are widely present in nature (see e.g. Lagona, 2018).

For these reasons, this work aims at introducing a new distance measure. More specifically, in analogy with the widely used Mahalanobis distance for data in linear multivariate spaces, a density based distance measure for cylindrical data is introduced. This way, the proposed distance will be able to deal with the specific structure and nature of such data.

Several distributions for cylindrical data can be found in the literature. The best known examples are probably the Mardia–Sutton distribution Mardia & Sutton, 1978, which is based on a conditioning argument from a trivariate normal distribution, and the Johnson–Wehrly cylindrical distribution Johnson & Wehrly, 1978, based on maximum entropy.

For the purpose of this work, the Mardia–Sutton cylindrical distribution is considered. In the following, we first briefly recall this distribution and then present our proposal.

2 The Mardia–Sutton distribution

The Mardia–Sutton cylindrical distribution is derived by a conditioning argument on a trivariate normal distribution, where the circular component is obtained by conditioning a bivariate normal. Its probability density function is given by

$$f(x,\theta) = \{2\pi I_0(\kappa)\}^{-1} \exp\{\kappa \cos(\theta - \mu_0)\} \left(2\pi\sigma_c^2\right)^{-\frac{1}{2}} \exp\left[-\left\{(x-\mu_c)^2/2\sigma_c^2\right\}\right],$$

where $-\infty < x < \infty$ is the linear component, $0 < \theta \leq 2\pi$ the angular component, and $\kappa > 0$ indicates the dispersion level of the circular component (larger values mean less variability). Furthermore, $I_0(\kappa)$ is the modified Bessel function of first kind and order zero, while

$$\begin{aligned} \mu_c &= \mu + \sigma\kappa^{\frac{1}{2}} \{\rho_1 (\cos\theta - \cos\mu_0) + \rho_2 (\cos\theta - \cos\mu_0)\}, \\ \sigma_c^2 &= \sigma^2\left(1-\rho^2\right), \quad \rho = \sqrt{\left(\rho_1^2 + \rho_2^2\right)} \quad 0 \leq \rho \leq 1. \end{aligned}$$

In this model, the circular variable Θ is distributed as a von Mises distribution $M(\mu_0, \kappa)$, while the conditional distribution of X given Θ is a normal distribution $N\left(\mu, \sigma^2\right)$ with the two correlation parameters $\rho_1 = corr(x, \cos\theta)$ and $\rho_2 = corr(x, \sin\theta)$.

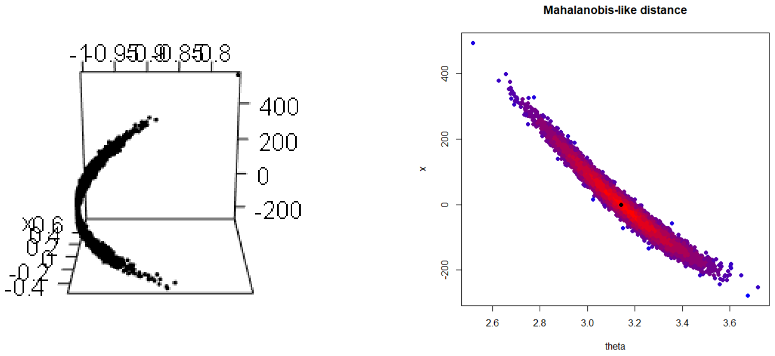
3 A Mahalanobis-like cylindrical distance

In the multivariate Euclidean space, given a center $\mu = (\mu_1, \mu_2, \ldots, \mu_n)'$ and a covariance structure Σ, the Mahalanobis distance of a point $x = (x_1, x_2, \ldots, x_n)'$ from the center is defined as:

$$d_{Mah}(x;\mu) := \sqrt{(x-\mu)' \Sigma^{-1} (x-\mu)}.$$

In analogy, within a cylindrical setting, and assuming a Mardia-Sutton probability structure, a Mahalanobis-like distance function for cylindrical data can be defined:

$$d_{Cyl}(x,\theta;\mu_0,\mu_c) := \sqrt{\kappa \cdot (1 - \cos(\theta - \mu_0)) + \frac{(x-\mu_c)^2}{2\sigma_c^2}}.$$



(a)

(b)

Figure 1: A sample drawn from a Mardia-Sutton distribution (a). The corresponding Mahalanobis-like cylindrical distance contours in the linear-angular space (b).

To illustrate, a 3D view of 10000 random points drawn from a Mardia-Sutton distribution with $\mu_0 = \pi$, $\kappa = 50$, $\mu = 0$, $\rho_1 = \rho_2 = 0.3$ and $\sigma^2 = 300$ is displayed in Figure 1, panel (a). The corresponding Mahalanobis-like distance contours are displayed through color intensity in panel (b), where the data have been mapped to the (θ, x) plane. It seems the distance contours are able to catch well the underlying data structure.

4 Remarks and further work

A new Mahalanobis–like distance measure for analyzing data on the surface of a cylinder is proposed. It appears to be able to capture the structure of data on the surface of a cylinder. Potentially, it can be exploited in many statistical contexts (e.g. cluster analysis, measures of variability, statistical tests etc.). Further work may include the definition of density based distances for poly–cylindrical data (see the new distribution introduced by Mastrantonio, 2018).

References

JOHNSON, R. A., & WEHRLY, T. E. 1978. Some angular-linear distributions and related regression models. *Journal of the American Statistical Association*, **73**(363), 602–606.

LAGONA, F. 2018. Correlated Cylindrical Data. *Chap. 3 of:* LEY, CHRISTOPHE, & VERDEBOUT, THOMAS (eds), *Applied Directional Statistics: Modern Methods and Case Studies*. Chapman and Hall/CRC.

MARDIA, K. V., & SUTTON, T. W. 1978. A model for cylindrical variables with applications. *Journal of the Royal Statistical Society: Series B (Methodological)*, **40**(2), 229–233.

MASTRANTONIO, G. 2018. The joint projected normal and skew-normal: A distribution for poly-cylindrical data. *Journal of Multivariate Analysis*, **165**, 14–26.

Archetypes, prototypes and other types

Francesco Palumbo[1], Giancarlo Ragozini[1] and Domenico Vistocco[1]

[1] Department of Political Science, Università di Napoli Federico II, (e-mail: [fpalumbo, giragoz, domenico.vistocco] @unina.it)

Abstract: Statistics and machine learning can significantly speed up human knowledge development, helping to determine the basic categories in a relatively short amount of time. The concept of categorization implies data summarization in a limited number of well-separated groups that must be maximally and internally homogeneous at the same time. This contribute presents a categorization approach that is based on the interval archetypal analysis.

Keywords: archetypes, prototypes, natural categories, interval-valued variables.

1 Introduction and motivation

Knowledge consists basically of categorizations: humans learn new concepts very fast by building complex relationships between a set of miscellaneous items or categories, as long as the total number of objects remains limited at most five/six objects (for example see Cowan, 2010). With the explosion of big data, the stored data represent an incredible source of knowledge, providing that they can be summarized in a (small) number of categories that are consistent with the human cognitive capabilities. However, in some conditions, data are not punctual and are naturally described by a complex data structure. Interval-valued data are probably one of the most widely considered kinds. They can result from several sources (Billard, 2008). A very interesting condition arises when data are naturally interval-valued; this is the situation of the European football league dataset considered in this paper. In football, as in all sports, tens of performance measuring variables accurately summarize any match. It is known and evident that teams (almost all) have a different game strategy playing at home and away. Averaging home and away data may cause relevant information losing, whereas considering home and away data as the two extremes of interval-valued variables allows important information recovering.

The present proposal aims to show how the archetypal analysis (AA) (Cutler & Breiman, 1994) can fruitfully contribute to summarize complex data in few categories that we call prototypes after the Rosch's definition (Rosch, 1973).

Because of the sake of space, the paper does not go into the details of methodological aspects and gives just a brief description of the analysis flow. Interested readers may refer to the literature. The real world dataset that considers the five major European football leagues will be instructive to understand the efficacy of the categorization process that exploits the archetypal analysis properties. The cognitive process of categorization through statistical learning techniques relying on the conceptual spaces framework is presented (Gardenfors, 2000).

2 Prototypes and ideal categorization: an integrated procedure

Initially introduced by E. Rosch (Rosch, 1973), in cognitive sciences as well as in statistical learning, the concept of the prototype is adopted to synthesize and represent categories. Prototypes are those elements (observed or unobserved) that better than others can represent a category. Their representativeness degree is measured using a distance function to a salient entity of the category (Fordellone & Palumbo, 2014; Ragozini *et al.* , 2017). Albeit, cluster analysis algorithms are the most used prototyping approaches, D'Esposito *et al.* , 2012 and Ragozini *et al.* , 2017 proposed the archetypal analysis (Cutler & Breiman, 1994) to identify the prototypes.

Archetypal analysis relies on the idea of "pure individual types" (the archetypes), a few points lying on the boundary of the data scatter and characterizing the archetypal pattern in the data. Let $\{\mathbf{x}_i, i = 1, \ldots, n\}$ be a set of multivariate data in $\Re^p$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$. Archetypal analysis looks for a set of m p-vectors $\left\{\mathbf{a}_j(m), j = 1, \ldots, m\right\}$ that are convex combinations of the input data $\mathbf{x}_i$'s and such that each data point is a convex combination of the vectors $\mathbf{a}_j$'s. Formally, given the data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, $\mathbf{X} \in \Re^{n \times p}$, the archetype matrix $\mathbf{A}(m) = (\mathbf{a}_1(m), \ldots, \mathbf{a}_m(m))'$, $\mathbf{A}(m) \in \Re^{m \times p}$, and the convex combination coefficients:

$$\beta_j(m) = (\beta_{j1}(m), \ldots, \beta_{jn}(m))' \quad \text{and} \quad \boldsymbol{\gamma}_i(m) = (\gamma_{i1}(m), \ldots, \gamma_{im}(m))',$$

the archetypes $\mathbf{a}_j(m), j = 1, \ldots, m$ are defined as the p-vectors that satisfy the following conditions:

$$\mathbf{a}'_j(m) = \beta'_j(m)\mathbf{X}, \quad j = 1, \ldots, m, \quad \beta_{ji}(m) \geq 0 \ \ \forall j, i, \quad \beta'_j(m)\mathbf{1} = 1 \ \ \forall j; \quad (1)$$

$$\mathbf{x}'_i = \gamma'_i(m)\mathbf{A}(m), \quad i = 1, \ldots, n, \quad \gamma_{ij}(m) \geq 0 \ \ \forall i, j, \quad \gamma'_i(m)\mathbf{1} = 1 \ \ \forall i. \quad (2)$$

The prototypes are defined starting from the archetypes (*step 1*), in order to exploit their properties of pureness, separability and strong characterization.

As the archetypes approximate the convex-hull of the data cloud, if $K > J$, they identify a convex region of the space (with K number of archetypes and J number of variables). Generally, for K relatively small, archetypes ensure a good approximation of the convex hull. Then, the center of the clusters around archetypes in the space spanned by the archetypes are computed (*step 2*). The first prototypes (*step 1*) and the second prototypes (*step 2*) are combined to get the final prototypes.

In the following, our interest focuses on the case when p interval-valued variables describe N statistical units. An interval-valued variable $\mathbb{X} \subset \mathbb{R}$ is represented by a series of sets of values delimited by ordered couples of bounds referred to as *minimum* and *maximum*: $\mathbb{X} = [\underline{\mathbf{X}}, \overline{\mathbf{X}}]$, where $\underline{\mathbf{X}} \leq \overline{\mathbf{X}}$. An equivalent description for $\mathbb{X}$ considers the midpoint $\mathbf{X}^c = \frac{1}{2}(\underline{\mathbf{X}} + \overline{\mathbf{X}})$ and the range $\mathbf{X}^r = \frac{1}{2}(\underline{\mathbf{X}} - \overline{\mathbf{X}})$. In analogy with the single value case, we define the *archetypes* $\mathbb{A}$ for *interval valued* data (D'Esposito *et al.*, 2012). Considering the midpoint and range spaces, two sets of archetypes, $\mathbf{A}^c$ and $\mathbf{A}^r$ are defined. Each data should be expressed as a unique convex combination of the interval data archetype in terms of midpoints and ranges. Therefore the mixture coefficients γ_i' are imposed to be the same in the two spaces. Hence the γ_i' coefficients represent the algebraic linkage of the two optimizations, and hence the linkage between the two spaces. Given the metric space provided by the Frobenius norm and the distance between interval matrices, for each m, the m interval valued archetypes $\mathbb{A}(m)$ can be determined by minimizing $\mathbb{X} - \tilde{\mathbb{X}}(m)$, and $\tilde{\mathbb{X}}(m) = \mathbf{\Gamma}(m)\mathbb{A}(m)$, $\tilde{\mathbb{X}}(m) \in \mathbb{IR}^{n \times p}$, *i.e.*, the data matrix reconstructed by m archetypal hyper-rectangle. The optimization procedure to derive interval archetypes is based on the use of Hausdorff distance and Frobenius norm and can be found in Corsaro & Marino, 2010.

3 Results and final remarks

The five most important European football leagues are in England, France, Germany, Italy and, Spain. They globally consist of 98 football teams that play in their respective national leagues consisting of 18 or 20 teams. Each team plays 34 or 38 matches (home and away) per season. Original data refer to five statistics (home and away) recorded for any single match, national championships 2016-17 and are available on the Kaggle web site `http://www.kaggle.com`. The analysis considers the seasonal arithmetic means of the following variables measured for the at home and away matches: possession (percentage), dribbles won (number per match), successful passes (number per match), key passes (number per match), aerials (number per match). Figure 3 shows the five archetypes and the 98 football team on the first Midpoint-Radii Principal Component Analysis (MR-PCA) (Palumbo & Lauro, 2003) factorial plan (81.5% of explained variance).

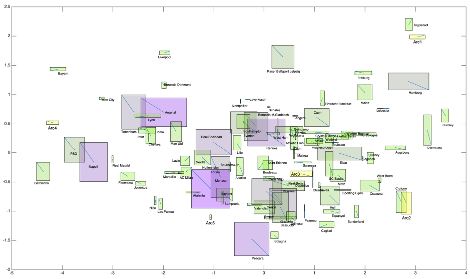


Figure 1. *Statistical units and archetypes on the first MR-PCA factorial plan.*

References

Billard, L. 2008. Some analyses of interval data. *CIT. Journal of Computing and Information Technology*, **16**(4), 225–233.

Corsaro, S., & Marino, M. 2010. Archetypal Analysis of Interval Data. *Reliable Computing*, **14**(1), 105–116.

Cowan, N. 2010. The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science.*, **19**, 51–57.

Cutler, Adele, & Breiman, Leo. 1994. Archetypal Analysis. *Technometrics*, **36**(4), 338–347.

D'Esposito, M.R., Palumbo, F., & Ragozini, G. 2012. Interval archetypes: a new tool for interval data analysis. *Statistical Analysis and Data Mining*, **5**(4), 322–335.

Fordellone, Mario, & Palumbo, Francesco. 2014. Prototypes definition through consensus analysis between Fuzzy c-Means and Archetypal Analysis. *Italian Journal of Applied Statistics*, **26**(2), 141–162.

Gardenfors, P. 2000. *Conceptual spaces: the geometry of thought. A bradford book*. Boston: MIT Press.

Palumbo, Francesco, & Lauro, Carlo N. 2003. A PCA for interval valued data based on midpoints and radii. *In:* Yanai, H., Okada, A., Shigemasu, K., Kano, Y., & Meulman, J.J (eds), *New developments in Psychometrics*. Tokyo: Springer-Verlag, for Psychometric Society.

Ragozini, G., Palumbo, F., & D'Esposito, M. R. 2017. Archetypal analysis for data-driven prototype identification. *Statistical Analysis and Data Mining: The ASA Data Science Journal.*, **10**, 6–20.

Rosch, E.H. 1973. Natural categories. *Cognitive psychology*, **4**(3), 328–350.

GENERALIZING THE SKEW-T MODEL USING COPULAS

Antonio Parisi[1] and Brunero Liseo[2]

[1] Department of Economics and Finance, University of Rome Tor Vergata,
(e-mail: antonio.parisi@uniroma2.it)

[2] MEMOTEF, Sapienza University of Rome,
(e-mail: brunero.liseo@uniroma1.it)

ABSTRACT: This work provides a fully Bayesian analysis of a copula model, in which both the dependence structure and the marginal variables have a skew-elliptical specification.

KEYWORDS: Skewness, kurtosis, Monte Carlo, copula models.

1 Introduction

Skewed Student-*t* distributions represent a very flexible parametric family of distributions (see, for example, Genton, 2004). In this paper, we will consider the *skew-t* distribution obtained by Azzalini & Capitanio, 2003.

Even if this model has been thoroughly studied both in a frequentist and in a Bayesian setup, an interesting approach to further generalize this family is represented by the construction of copula models involving a *skew-t* distribution for the marginal components or the dependency structure.

We propose a fully Bayesian analysis of a *p*-variate Gaussian copula model with *skew-t* margins. Notice that this model is not nested in the *p*-variate *skew-t* model.

The use of a copula representation of a multivariate distribution in our proposed model allows a large amount of flexibility because each single marginal may have its own number of degrees of freedom. On the other hand, the dependence structure is modeled in a different way. Although we restrict our attention to a Gaussian copula in this note, it is possible to implement a more general approach, as for example, in Wu *et al.* , 2015 where a non parametric mixture of Gaussian copulae is adopted.

The prior distribution of the model parameters have been selected in order to be minimimally informative.

Even if the structure of the model is conceptually simple, its estimation is hampered by several problems. Our approach offers a number of advantages with respect to existing procedures. In particular, as we use a Monte Carlo strategy we don't rely on convergence arguments; moreover, the evaluation of the Bayes factor comes as a simple by-product of the sampler.

2 The model

Given a sample $\boldsymbol{y}$ from a p-variate random variable $\boldsymbol{Y}$, the model likelihood is given by (see Smith, 2011)

$$\begin{aligned}
f(\boldsymbol{y}|\Theta,R) &= |R|^{-n/2}\prod_{i=1}^{n}\left(\exp\left\{-\frac{1}{2}\boldsymbol{x}_i'\left(R^{-1}-\mathbb{I}_p\right)\boldsymbol{x}_i\right\}\prod_{j=1}^{p}f_j(y_{ij}|\Theta_j)\right)\\
x_{ij} &= \Phi^{-1}(u_{ij})\\
u_{ij} &= F_{ST}(y_{ij}|\Theta_j)
\end{aligned}$$

where

- $f_j(y_{ij}|\Theta_j)$ denotes the distribution of the j-eth component of $\boldsymbol{Y}$, that is a univariate skew-t distribution,
- $\Theta = \{\Theta_j, j = 1,2,\ldots,p\}$ collects the parameters of the marginal distributions,
- R is a correlation matrix and $\mathbb{I}_p$ is the identity matrix of size p.

To fully specify a Bayesian model, we elicit a uniform prior on R, as in Joe, 2006, while we use the same priors of Parisi & Liseo, 2018a for the parameters of the marginal distributions. Rearranging terms, we can write

$$\begin{aligned}
\pi(\Theta,R|\boldsymbol{y}) &\propto \pi(R)\pi(\Theta)f(\boldsymbol{y}|\Theta,R)\\
&= \pi(R)|R|^{-n/2}\prod_{i=1}^{n}\left[\exp\left\{-\frac{1}{2}\boldsymbol{x}_i'\left(R^{-1}-\mathbb{I}_p\right)\boldsymbol{x}_i\right\}\right]\pi^{\star}(\Theta|\boldsymbol{y}),
\end{aligned}$$

Where

$$\begin{aligned}
\pi^{\star}(\Theta|y) &= \pi(\Theta)\prod_{i=1}^{n}\prod_{j=1}^{p}f_j(y_{ij}|\boldsymbol{\Theta}_j)\\
&= \prod_{j=1}^{p}\left[\pi(\Theta_j)\prod_{i=1}^{n}f_j(y_{ij}|\boldsymbol{\Theta}_j)\right].
\end{aligned}$$

In order to evaluate the integral

$$I = \int g(\Theta,R)\pi(\Theta,R|\boldsymbol{y})d\Theta dR$$

it is possible to implement a Bayesian version of the "inference for margin" procedure: we draw N values $\Theta^{(k)}$ from $\pi^\star(\Theta|y)$ and $R^{(k)}$ from $\pi(R)$ and evaluate the integral as

$$I \approx \sum_{k=1}^{N} g(\Theta^{(k)}, R^{(k)})\, \bar{w}^{(k)},$$

where $\bar{w}^{(k)}$, denotes the importance weights

$$\begin{aligned} w^{(k)} &= |R^{(k)}|^{-n/2} \prod_{i=1}^{n} \left(\exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i^{(k)})' \left((R^{(k)})^{-1} - \mathbb{I}_p \right) \boldsymbol{x}_i^{(k)} \right\} \right), \\ \bar{w}^{(k)} &= w^{(k)} / \sum (w^{(k)}) \end{aligned}$$

The procedure is divided in two steps

- in the first step, a PMC is implemented for each marginal component in order to draw N particles $\Theta^{(1)}, \ldots, \Theta^{(N)}$ from $\pi^\star(\Theta|y)$,
- in the second step, for each particle $\Theta^{(k)}$, draw a value $R^{(k)}$ from its prior distribution.

It is possibile to use the `mvst` package (Parisi & Liseo, 2018b) in order to obtain the particles from the first step. For the second step, a sampler is implemented in the function `rcorrmatrix` of the R package `clusterGeneration`.

3 Application

As a final illustration of the proposed algorithm, we analyze the same dataset used in Liseo & Parisi, 2013, namely the returns of two stocks in the NYSE composite index, namely the "ABM Industries Incorporated" and "The Boeing Company" (240 monthly observations).

The estimate of a bivariate *skew-t* model gives the following results

	Estimate	*Std.Error*	*Q5%*	*Me*	*Q95%*
ξ_1	0.0111	0.0039	0.0059	0.0102	0.0182
ξ_2	0.0167	0.0040	0.0104	0.0166	0.0221
$G_{1,1}$	0.0032	0.0004	0.0027	0.0032	0.0040
$G_{1,2}$	0.0008	0.0002	0.0004	0.0008	0.0011
$G_{2,2}$	0.0033	0.0004	0.0028	0.0032	0.0040
ψ_1	−0.0032	0.0027	−0.0076	−0.0026	−0.0002
ψ_2	−0.0019	0.0028	−0.0074	−0.0006	0.0004
ν	3.0890	0.2373	3.0000	3.0000	3.4945

with log-marginal likelihood equal to 462.2794.

The estimate of the copula model provides results which are consistent with the previous ones, both in terms of the skewness and kurtosis, but the additional flexibility allows a better fit. In fact, the log-marginal likelihood is evaluated as 473.2288. Hence, under equal prior model probabilities, the Bayes factor will prefer the copula model over the *skew-t* one.

References

AZZALINI, A., & CAPITANIO, A. 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *J. R. Statist. Soc. B*, **65**, 367–389.

GENTON, M.G. (ED.). 2004. *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. London: CRC/Chapman & Hall.

JOE, H. 2006. Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, **97**(10), 2177–2189.

LISEO, B., & PARISI, A. 2013. Bayesian inference for the multivariate skew-normal model: a population Monte Carlo approach. *Comput. Statist. Data Anal.*, **63**, 125–138.

PARISI, A., & LISEO, B. 2018a. Objective Bayesian analysis for the multivariate skew-t model. *Statistical Methods & Applications*, **27**(2), 277–295.

PARISI, A., & LISEO, B. 2018b. Statistical inference with skew t distributions: the mvst R package. *Annali del dipartimento di metodi e modelli per l'economia, il territorio e la finanza*, Nov, 97–115.

SMITH, M. 2011. Bayesian Approaches to Copula Modelling. *ERN: Bayesian Analysis (Topic)*, 12.

WU, J., WANG, X., & WALKER, S. G. 2015. Bayesian nonparametric estimation of a copula. *Journal of Statistical Computation and Simulation*, **85**(1), 103–116.

CONTAMINATION AND MANIPULATION OF TRADE DATA: THE TWO FACES OF CUSTOMS FRAUD

Domenico Perrotta[1], Andrea Cerasa[1], Lucio Barabesi[2], Mario Menegatti[3] and Andrea Cerioli[3]

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

[2] Department of Economics and Statistics, University of Siena, (e-mail: lucio.barabesi@unisi.it)

[3] Department of Economics and Management, University of Parma, Parma, (e-mail: andrea.cerioli@unipr.it, mario.menegatti@unipr.it)

ABSTRACT: We consider statistical tools for the detection of frauds in customs data collected in international trade, by developing a principled framework for goodness-of-fit testing of Benford's law. Our approach relies on a trader-specific contamination model, under which fraud detection has close connections with outlier testing. We also compare the performance of this approach with alternative tools based on robust statistics that rely on a different transaction-specific contamination model.

KEYWORDS: Benford's law, contamination, outlier detection.

1 Motivation

The contrast of fraud in international trade is a crucial task of modern economic regulations. For instance, import operations have a significant weight in the budget of the European Union (EU), through tax revenues that EU Member States receive from import duties, excise duties and VAT. The volumes involved are huge and typically correspond to a major share of the total own resources of the EU. Correspondingly, huge losses may occur when the value of imported goods is under-reported. Monitoring transactions in international trade is also important for the fight against criminal activities, as they are often used for illegal capital movements and for money laundering operations. As a consequence, there has been an increasing interest in the development of statistical procedures that could help to identify potential fraudsters among international traders, thus providing guidance to anti-fraud investigators (Cerioli & Perrotta, 2014; Barabesi *et al.*, 2016).

Outlier detection methods typically play a prominent role among statistical anti-fraud techniques for international transactions. The rationale is that the bulk of international trade data is made of legitimate transactions and major frauds may stand out as highly suspicious anomalies. In this work we focus on a peculiar framework for outlier identification that has recently attracted considerable interest for anti-fraud purposes and that we explore it in the case of international trade. The approach that we consider is based on testing conformance to Benford's law.

2 Benford's law

Benford's law (BL, for short) is a fascinating phenomenon which rules the pattern of the leading digits in many types of numerical data and mathematical sequences. Informally speaking, the law states that the digits are not uniformly scattered – as one may naively expect – but follow a logarithmic-type distribution in which the leading digit 1 is more likely to occur than the leading digit 2, the leading digit 2 is more likely than the leading digit 3, and so on. Indeed, the simplest form of BL (Benford, 1938) gives the probability that the first leading digit equals d_1, for $d_1 = 1, \ldots, 9$, as

$$\log_{10}\left(1+\frac{1}{d_1}\right). \tag{1}$$

In a probabilistic setting, a deep analysis of BL was first carried out by Hill, 1995, who proved a limit theorem for the significant-digit distribution. We refer to two recent books (Berger & Hill, 2015; Miller, 2015) for a thorough description of the mathematical properties of BL and for a survey of its main application areas. We also remark that scientists and practitioners have recently applied the law in diverse settings, fraud detection in business accounting being perhaps the most noticeable one (Nigrini, 2012).

3 Challenges in fraud detection

To our knowledge, most of the available applications of BL to anti-fraud problems fall into three main categories. The first group deals with aggregated data, such as those referring to the whole market for a specific product, or to the population of taxpayers (or electors) in a given country. In this instance, rejection of the hypothesis of conformance to BL often provides compelling evidence of fraud, but precise identification of the fraudsters must be left to further and possibly non-statistical investigations. In the second type of examples individual companies, or customers, are scrutinized by means of a formal test of conformance, relying on a large sample of reported digits. Therefore, the final decision rests upon a statistically principled criterion, but potential usefulness is limited to a restricted number of situations. The third case is that of suspicious individuals for which only a limited number of digits is available. This instance often takes a less formal route, ending up with visual inspection of the data and simple numerical comparisons between BL and the observed digit distribution. It is apparent that none of these strategies is suitable for routine analysis of international trade data, where many thousands of traders must be inspected under a wide variety of conditions on their number of transactions and market behavior. Therefore, the aim of our study is to fill the gap through a sound statistical methodology that might be eventually applied by anti-fraud and customs officers on the majority of traders operating in a given market, and even on those for which only a moderate number of transactions is available.

4 A contamination model for customs data

We phrase our BL anti-fraud approach within the framework of a trader-specific contamination model where each fraud corresponds to an outlier. We define $\pi_t(d_1,\ldots,d_k)$ to be the joint probability of observing the k-ple of significant digits $d_1,\ldots,d_k$ for trader t. Let T denote the total number of traders in the market. For $t = 1,\ldots,T$ and each $k \in \mathbb{Z}^+$, the general form of our contamination model is

$$\pi_t(d_1,\ldots,d_k) = (1-\tau_t)\Psi_t(d_1,\ldots,d_k) + \tau_t\Upsilon_t(d_1,\ldots,d_k), \tag{2}$$

where $\Psi_t(d_1,\ldots,d_k)$ is the probability of observing $d_1,\ldots,d_k$ for trader t in the absence of fraud, $\Upsilon_t(d_1,\ldots,d_k)$ is the probability of the same event for a manipulated transaction, and $0 \leq \tau_t \leq 1$ is the probability of fraud for trader t.

A remarkable feature of model (2) is that standard tests of the hypothesis $H_0 : \tau_t = 0$ are consistent even when the model holds with $\tau_t > 0$ and multiple outliers (i.e., frauds) are present in the data. This property arises from the fact that no parameter must be estimated when $\Psi_t(d_1,\ldots,d_k)$ is BL, for which (1) provides the one-digit marginal. Therefore, the usual tools of robust estimation are not needed to avoid masking in this context (Rousseeuw & Hubert, 2011).

5 Summary of results

In our work we address the following issues, which are crucial for effective implementation of BL in anti-fraud analysis of trade data.

- We explore the conditions under which $\Psi_t(d_1,\ldots,d_k)$ is BL for genuine customs data (Cerioli *et al.*, 2019; Lacasa, 2019).
- We address the problem of reducing the false-positive rate, both by adopting multiple-testing strategies (Barabesi *et al.*, 2018) and by defining tests based on alternative characterizations of BL.
- We compare the results obtained by testing $H_0 : \tau_t = 0$ in model (2) with those derived under a different, and more commonly applied, transaction-specific contamination model, in which outlier detection follows robust regression estimation of fair prices for the traded products.

References

BARABESI, L., CERASA, A., PERROTTA, D., & CERIOLI, A. 2016. Modeling international trade data with the Tweedie distribution for anti-fraud and policy support. *European Journal of Operational Research*, **248**, 1031–1043.

BARABESI, L., CERASA, A., CERIOLI, A., & PERROTTA, D. 2018. Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud. *Journal of Business and Economic Statistics*, **36**, 346–358.

BENFORD, F. 1938. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, **78**, 551–572.

BERGER, A., & HILL, T. P. 2015. *An Introduction to Benford's Law*. Princeton: Princeton Univ. Press.

CERIOLI, A., & PERROTTA, D. 2014. Robust clustering around regression lines with high density regions. *Advances in Data Analysis and Classification*, **8**, 5–26.

CERIOLI, A., BARABESI, L., CERASA, A., MENEGATTI, M., & PERROTTA, D. 2019. Newcomb-Benford law and the detection of frauds in international trade. *PNAS*, **116**, 106–115.

HILL, T. P. 1995. A statistical derivation of the significant-digit law. *Statistical Science*, **10**, 354–363.

LACASA, L. 2019. Newcomb-Benford law helps customs officers to detect fraud in international trade. *PNAS*, **116**, 11–13.

MILLER, S. J. (ed). 2015. *Benford's Law: Theory and Applications*. Princeton: Princeton Univ. Press.

NIGRINI, M. J. 2012. *Benford's Law*. Hoboken: Wiley.

ROUSSEEUW, P. J., & HUBERT, M. 2011. Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, **1**, 73–79.

Bayesian Clustering using Non-negative Matrix Factorization

Michael D. Porter[1] and Ketong Wang[2]

[1] Engineering Systems and Environment, University of Virginia,

(e-mail: mdporter@virginia.edu)

[2] Amazon, (e-mail: ketongw@amazon.com)

Abstract: Bayesian model-based clustering is a widely applied procedure for discovering groups of related observations in a dataset. These approaches use Bayesian mixture models, estimated with MCMC, which provide posterior samples of the model parameters and clustering partition. While inference on model parameters is well established, inference on the clustering partition is less developed. A new method is developed for estimating the optimal partition from the pairwise posterior similarity matrix generated by a Bayesian cluster model. This approach uses non-negative matrix factorization (NMF) to provide a low-rank approximation to the similarity matrix. The factorization permits hard or soft partitions and is shown to perform better than several popular alternatives under a variety of penalty functions.

Keywords: Bayesian clustering, non-negative matrix factorization (NMF).

1 Introduction

The goal of clustering is to discover partitions that assign observations into meaningful groups. A favorable property of Bayesian model-based clustering is that it provides versatile posterior uncertainty assessment on both the model parameters and cluster allocation estimates. However, while inference on model-specific parameters and mixing weights follow standard Bayesian practice, more development on estimating the clustering partition is needed.

To better clarify the notion of optimal partitioning, [Binder, 1978] introduced a loss function approach. This considers optimal clustering as a Bayesian action which attempts to minimize the expected loss of the partition. Under Binder's linear loss function, the problem reduces to searching for a partition $\boldsymbol{c}^*$ that produces a binary affinity matrix $\boldsymbol{\pi}^*$ nearest to the pairwise posterior similarity matrix $\boldsymbol{\pi}$ with $\pi_{ij} = p(c_i = c_j|\boldsymbol{y})$. [Fritsch & Ickstadt, 2009] showed that minimizing Binder's linear loss is equivalent to maximizing the Rand index. One extension, the adjusted Rand index, corrects for the chance of random agreement in the Rand index [Hubert & Arabie, 1985]. More recently, [Wade & Ghahramani, 2018] use a loss function based upon variation of information [Meilă, 2007] as well as providing a method of assessing the uncertainty in the estimated partition.

We find that there are several places where these current approaches can be improved. The optimization strategy of directly shuffling the clustering labels to find a *binary* matrix $\boldsymbol{\pi}^*$ closest to the posterior similarity $\boldsymbol{\pi}$ can be notoriously slow. Furthermore, conventional methods can only provide hard clustering by the nature of direct label manipulation. They possess no capability of accounting for ambiguous observations. Moreover, they tend to favor two extremes which either treat all uncertain points as singleton clusters or simply lump them into a major neighbor cluster.

This paper proposes the use of non-negative matrix factorization (NMF) to identify optimal partitions from Bayesian model-based clustering models. We find the NMF approaches not only outperform alternative methods on clustering accuracy but can also provide deeper interpretations of the partitioning results. Additionally, the clustering solutions produced by NMF are more compelling since they can carefully balance between the singleton-preferred and dominant-preferred extremes.

2 Methodology

A convenient way to summarize the partition information from a Bayesian mixture model, which is unaffected by label switching and can be used when the number of clusters K is not equal for all MCMC samples, is with the pairwise posterior similarity matrix. Intuitively, two data points are more likely to be members of the same cluster when they appear together frequently in the partitions $\boldsymbol{c}^{(1)}, \boldsymbol{c}^{(1)}, \ldots, \boldsymbol{c}^{(M)}$. To quantify this relationship, a pairwise posterior similarity matrix $\boldsymbol{\pi} = \left\{\pi_{ij}\right\}$ is defined by $\pi_{ij} = p(c_i = c_j | \boldsymbol{y})$, where c_i and c_j are the cluster assignments of observations y_i and y_j. When the true probabilities are unknown, the posterior similarity can be estimated from the MCMC samples

$$\hat{\pi}_{ij} = \hat{p}(c_i = c_j | \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{c_i^{(m)} = c_j^{(m)}\}, \tag{1}$$

where $\mathbb{1}\{c_i^{(m)} = c_j^{(m)}\}$ equals 1 if $c_i^{(m)} = c_j^{(m)}$ and 0 otherwise.

2.1 NMF Formulation

Nonnegative matrix factorization (NMF) decomposes a data matrix into lower-rank matrices which can help reveal underlying patterns of the data. The connections between NMF and clustering have been well established and successfully applied to many research fields [Hosseini-Asl & Zurada, 2014; Ding *et al.*, 2005; Kuang *et al.*, 2012; Wang & Zhang, 2013].

Let the posterior similarity matrix $\boldsymbol{\pi}$ be the data matrix to be approximated. In practice this would be the approximation $\hat{\boldsymbol{\pi}}$ from (1Methodologyequation.2.1). The basic NFM problem [Lee & Seung, 2001] is finding two lower-rank matrices that solve the optimization

$$(\widehat{\boldsymbol{W}}, \widehat{\boldsymbol{H}}) = \underset{\boldsymbol{W}, \boldsymbol{H} > 0}{\operatorname{argmin}} ||\boldsymbol{\pi} - \boldsymbol{W}\boldsymbol{H}||_F^2 , \tag{2}$$

where F indicates the Frobenius norm, $\boldsymbol{\pi} \in \mathbb{R}_+^{n \times n}$, $\boldsymbol{W} \in \mathbb{R}_+^{n \times K}$ and $\boldsymbol{H} \in \mathbb{R}_+^{K \times n}$. The *non-negativity* in NMF indicates that matrix elements must be non-negative.

This least-squares type of NMF problem has a natural clustering interpretation for nonnegative data because the columns of $\widehat{\boldsymbol{W}}$ can be considered centroids and the columns of $\widehat{\boldsymbol{H}}$ are the cluster weights of the observations [Kuang *et al.*, 2012]. This provides a close connection with the K-means algorithm [Ding *et al.*, 2005]. While K-means restricts $\widehat{\boldsymbol{H}}$ to be a binary matrix which implies hard assignment of cluster allocation, the NMF method results in a real-valued positive matrix $\widehat{\boldsymbol{H}}$ which contains both hard and soft clustering information. Specifically, the hard cluster allocation of observation i can be estimated using the following equation

$$\hat{c}_i = \underset{k=1:K}{\operatorname{argmax}} \widehat{H}_{ki} , \tag{3}$$

which is similar to MAP method. In other words, the row index of the maximum value of a column represents the cluster membership of the corresponding observation. In addition, the soft label assignment of i^{th} observation can be obtained by standardizing the columns of $\widehat{\boldsymbol{H}}$

$$\hat{\boldsymbol{c}}_i^{\texttt{soft}} = \frac{\widehat{\boldsymbol{H}}_i}{\sum_{k=1}^K \widehat{H}_{ki}} , \tag{4}$$

where $\widehat{\boldsymbol{H}}_i$ is the i^{th} column of matrix $\widehat{\boldsymbol{H}}$. To interpret the soft clustering, the vector $\hat{\boldsymbol{c}}_i^{\texttt{soft}}$ describes the confidence levels of assigning observation i into the clusters. The higher the value of $\hat{\boldsymbol{c}}_{ik}^{\texttt{soft}}$, for example, the more confidence there is in assigning observation i into cluster k. To further understand the soft clustering nature of NMF, one can follow the claim by [Ding *et al.*, 2005] that a strictly orthogonal $\boldsymbol{H}$ matrix leads to hardened partitions and a near-orthogonal $\boldsymbol{H}$ provide soft clustering which can be interpreted as posterior cluster probabilities.

3 Conclusions

This paper addresses the problem of estimating the optimal partition from Bayesian cluster models. Particularly, a non-negative matrix factorization (NMF) framework has been proposed that uses the pairwise posterior similarity matrix constructed from the MCMC clustering samples to obtain the most appropriate clustering estimates under various penalty functions. Instead of a direct optimization using label manipulation, the NMF approach implicitly induces the partition from the weight matrix $\boldsymbol{H}$ supplied by the flexible low-rank approximation $\boldsymbol{\pi} \approx \boldsymbol{WH}$. Another favorable property of the NMF models is a soft/probabilistic interpretation of the clustering label assignment which can be helpful in understanding the uncertainty in the partition.

References

BINDER, D.A. 1978. Bayesian cluster analysis. *Biometrika*, **65**(1), 31–38.

DING, C., HE, X., & SIMON, H.D. 2005. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. *Pages 606–610 of: SIAM International Conference on Data Mining*, vol. 5. SIAM.

FRITSCH, A., & ICKSTADT, K. 2009. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian analysis*, **4**(2), 367–391.

HOSSEINI-ASL, E., & ZURADA, J.M. 2014. Nonnegative Matrix Factorization for Document Clustering: A Survey. *Pages 726–737 of: Artificial Intelligence and Soft Computing*. Springer International Publishing.

HUBERT, L., & ARABIE, P. 1985. Comparing partitions. *Journal of classification*, **2**(1), 193–218.

KUANG, D., DING, C., & PARK, H. 2012. Symmetric Nonnegative Matrix Factorization for Graph Clustering. Philadelphia: Society for Industrial and Applied Mathematics.

LEE, D.D., & SEUNG, H.S. 2001. Algorithms for non-negative matrix factorization. *Pages 556–562 of: Advances in neural information processing systems*.

MEILĂ, M. 2007. Comparing clusterings–an information based distance. *Journal of Multivariate Analysis*, **98**(5), 873–895.

WADE, S., & GHAHRAMANI, Z. 2018. Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*, **13**(3), 559–626.

WANG, Y., & ZHANG, Y. 2013. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, **25**(6), 1336–1353.

Exploring Gender Gap in International Mobility Flows through a Network Analysis Approach

Ilaria Primerano[1] and Marialuisa Restaino[1]

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

Abstract: The study of international mobility flows across different European countries has become an important research topic due to the relevance of internationalisation process in the university context. The analysis of the factors pulling and pushing students and/or university academic staff in a foreign country in higher education is, indeed, a key feature for the implementation of university policies. The present contribution aims at analysing the student and staff mobility flows by considering a network analysis approach (Derszi, 2011, Breznik, 2015, Barnett, 2016). In particular, the main purposes are to discover if a gender gap exists across countries/universities and subject areas. Thanks to the European Union Open Data Portal (EU ODP), a statistical overview of Erasmus mobility of students and academic staff from the period 2008/09 to 2013/14 is obtained with respect to the gender. At macro-level perspective, temporal network data structures are defined in which the nodes are the countries and the links represent the students and staff mobility exchanges between countries with a weight proportional to the number of students and staff involved. Hence, the directed networks are built considering the outgoing and the incoming subjects.

Keywords: Gender gap, Erasmus student mobility, European open data, social network analysis.

References

Barnett, G.A., Ke Jiang M.L. Park H.W. 2016. The flow of international students from a macro perspective: a network analysis. *Compare: A Journal of Comparative and International Education*, **46**, 533–559.

Breznik, K., Ragozini G. 2015. Exploring the italian erasmus agreements by a network analysis perspective. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).*

DERSZI, A., DERSZY N. KAPTALAN E. NEDA Z. 2011. Topology of the Erasmus student mobility network. *Physica A: Statistical Mechanics and its Applications*, **390**, 2601–2610.

CLUSTERING TWO-MODE BINARY NETWORK DATA WITH OVERLAPPING MIXTURE MODEL AND COVARIATES INFORMATION

Saverio Ranciati[1], Veronica Vinciotti[2], Ernst C. Wit[3]
and Giuliano Galimberti[1]

[1] Department of Statistical Sciences, Università di Bologna,
(e-mail: saverio.ranciati2@unibo.it, giuliano.galimberti@unibo.it)

[2] Department of Mathematics, Brunel University London,
(e-mail: veronica.vinciotti@brunel.ac.uk)

[3] Institute of Computational Science, Università della Svizzera italiana,
(e-mail: wite@usi.ch)

ABSTRACT: Network data may be collected as *actor-event* information, where two types of nodes describe the interactions in the network itself: *actors* are units - individuals - recorded as attending *events*. Same examples are people voting in an election, users likes and dislikes, and so forth. Discovering communities, also called clusters, of units by exploiting their patterns of attendance to the events is an intuitive and reasonable criterion to define such groups. Here we propose an extension of the model called `manet`: our contribution includes covariates in the model, while retaining the characteristic of parsimony and interpretability of the original model. We assess the performance of our approach in a simulated comparative environment.

KEYWORDS: Bayesian inference, two-mode network, MCMC, probit regression.

1 Introduction

Methods and approaches for network data have witnessed an increasing usage and demand, following the growing interest of many practitioners in the appealing capability of network analysis to model complex interactions. Some of these methods and models are reviewed in Kolaczyk, 2009. Among data types in network analysis, some are coded as a set of interactions between two types of nodes forming a network structure, i.e. units - actors - attending events. These are called two-mode networks, bimodal networks, or affiliation networks (Wasserman & Faust, 1994, Chapter 8). A recent approach for model-based clustering of binary two-mode networks was proposed by Ranciati *et al.*, 2017. We build from that model to incorporate covariate informations pertaining to both types of nodes, thus: (i) extending `manet` (Ranciati *et al.*, 2017) through the use of covariates; (ii) and providing comparative results on data from a simulated environment.

2 Mixture model for two-mode binary network data

Two-mode network data are organized in an $n \times d$ matrix Y of observations y_{ij}, recording attendances of $i = 1, \ldots, n$ units - *actors* - to $j = 1, \ldots, d$ events. For binary data, each y_{ij} codes individual i attending event j if equal to 1, and zero otherwise. The idea is to cluster these n actors based on their attendances via a mixture model, which is a prime tool in the framework of model-based clustering (Frühwirth-Schnatter, 2006). Traditionally, clusters are assumed to be mutually exclusive, with prior sizes $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$; also, two conditions hold: (i) $\alpha_k \geq 0$, for each k; (ii) $\sum_{k=1}^{K} \alpha_k = 1$. We start from and build upon the model proposed in Ranciati *et al.*, 2017, where the authors detail a Bayesian **m**ultiple **a**llocation model for **net**work data (`manet`): the model relaxes conditions on the cluster sizes $\boldsymbol{\alpha}$, and allows $\{\mathbf{z}_i\}$ to have multiple elements equal to 1. The number of all possible group-allocating configurations is equal to $K^\star = 2^K$, the cardinality of the set containing sequences of 1 and 0 values for latent vector $\{\mathbf{z}_i\}$. A new $K^\star$-dimensional allocation vector $\mathbf{z}_i^\star$ is defined for each i, satisfying $\sum_{h=1}^{K^\star} z_{ih}^\star = 1$, with a 1-to-1 correspondence between $\boldsymbol{z}_i$, allocating actors into overlapping *parent* clusters, and $\boldsymbol{z}_i^\star$, allocating actors into non-overlapping *heir* clusters. This re-parametrization corresponds to the following hierarchical model $\boldsymbol{\alpha}^\star \sim \mathrm{Dir}(\boldsymbol{\alpha}^\star; a_1, \ldots, a_{K^\star})$, $\mathbf{z}_i^\star | \boldsymbol{\alpha}^\star \sim \mathrm{Multinom}(\mathbf{z}_i^\star; \alpha_1^\star, \ldots, \alpha_{K^\star}^\star)$, $\mathbf{y}_i | \mathbf{z}_i^\star, \boldsymbol{\pi} \sim \prod_{h=1}^{K^\star} \prod_{j=1}^{d} \left[\mathrm{Ber}\big(y_{ij}; \pi_{hij}^\star\big)\right]^{z_{ih}^\star}$, with prior on the original parameters $\pi_{kij} \sim \mathrm{Beta}(\pi_{kij}; b_1, b_2)$, and (b_1, b_2) hyper-parameters. The quantities $\{\pi_{hij}^\star\}$ are not additional parameters to be sampled, but probabilities of attendances derived from $\boldsymbol{\pi}$: for each actor i and event j, they are computed via a function $\psi\big(\boldsymbol{\pi}_{\cdot ij}, \mathbf{z}_i\big)$. We focus on introducing covariates into `manet`, retaining parsimony from the original formulation. Covariates could be characteristic related to an actor, such as, gender, age, etc, or features of an event, i.e. type of event, date, duration, and so forth. We focus on the case with only actor-specific covariates, as it is straightforward to include also event covariates. We define $\mathbf{x}_{i\cdot}$ to be the L-dimensional vector of covariates for actor i. Covariates enter the model through a *link* function as in the generalized linear models context (McCulloch & Neuhaus, 2001). To the hierarchical model, we add $\boldsymbol{\mu}_k \sim \mathrm{N}(\mu_k; 0, \sigma_\mu^2)$, $\boldsymbol{\beta}_k \sim \mathrm{N}_L(\boldsymbol{\beta}_k; \mathbf{0}_L, \sigma_\beta^2 I_L)$, $\eta_{ki} = \mu_k + \sum_{l=1}^{L} \beta_{kl} x_{il}$, $\pi_{ki}\big(\mathbf{x}_{i\cdot}\big) = \Phi(\eta_{kij})$, where: η_{ki} is the linear predictor; $\Phi(\cdot)$ is the gaussian cumulative function; $(\sigma_\mu^2, \sigma_\beta^2)$ as hyper-parameters. We define $\pi_{hi}^\star = \Phi[\psi(\boldsymbol{\mu}, \boldsymbol{\beta}, x_i, \mathbf{z}_i)] =$

$\Phi\left[\frac{\mathbf{z}_i\boldsymbol{\mu}}{||\mathbf{z}_i||_1}+\left(\frac{\mathbf{z}_i\boldsymbol{\beta}}{||\mathbf{z}_i||_1}\right)x_i\right]$, with $||\mathbf{z}_i||_1$ being the sum of the elements of $\mathbf{z}_i$. We can sample the parameters of interest $\{\boldsymbol{\mu},\boldsymbol{\beta}\}$ as in a single probit regression model, and use them to compute $\{\pi^\star_{hi}\}$ for evaluating the likelihood. We let $\tilde{\mathbf{y}}$ be an $\tilde{n}\times 1$ column vector obtained by stacking columns of data matrix Y, with $\tilde{n}=n\cdot d$; we stack together the cluster-specific vector of intercepts and regression coefficients into $\tilde{\boldsymbol{\beta}}=[\mu_1\ \boldsymbol{\beta}_1\ \mu_2\ \boldsymbol{\beta}_2\ \dots\ \mu_K\ \boldsymbol{\beta}_K]'$. We define a probit regression formulation, where $f\big(\tilde{y}_i|\tilde{\boldsymbol{\beta}},\mathbf{z}_i\big)=\text{Ber}\big(\tilde{y}_i;\Phi(\tilde{\eta}_i)\big)$, with $\tilde{\eta}_i$ being an element of $\tilde{\boldsymbol{\eta}}=\tilde{X}\tilde{\boldsymbol{\beta}}$. The new design matrix $\tilde{X}$ is built conditional on the cluster allocations $\{\mathbf{z}_i\}$. For example, when $K=2$, we have

$$\tilde{X}=\begin{bmatrix} X^{[1]} & X^{[2]} & \mathbf{0}_{1+L} & \frac{1}{2}X^{[4]} \\ X^{[1]} & \mathbf{0}_{1+L} & X^{[3]} & \frac{1}{2}X^{[4]} \end{bmatrix}^\top$$

where $X^{[h]}=\{\mathbf{x}_i,\forall i:\mathbf{z}_i=\mathbf{u}_h\}$. We use the Bayesian probit regression framework from Holmes & Held, 2006 to implement an MCMC scheme.

3 Simulation study

We investigate the performance of our proposal `manet+cov` by simulating 25 independent datasets, and we average the results across the replicates. Data are simulated from `manet` with $K=2$ overlapping clusters. We consider 5 scenarios with different type of covariates, sample size n and number of events d. Performances are measured via: (i) the misclassification error rate (*MER*), fraction of wrongly allocated units, and (ii) Adjusted Rand Index (*ARI*), measure of agreement between true and estimated labels - with 100 as best value. Benchmarks values for *MER* and *ARI* are, respectively, 0.685 and 0. Results are reported in Table 1. Except for the scenario with $n=50$ and $d=5$, we are able to obtain satisfactory values for both *MER* and *ARI*, as we can see from Table 1. This is indeed expected, given we are simulating from the very same model we are trying to fit. It is worth mentioning that in these simulations results, when *actors* were not correctly assigned to the heir cluster, they were at least allocated to one of the two correct parent clusters.

4 Conclusions

In this manuscript we proposed an extension of a model for clustering binary two-mode network data introduced by Ranciati *et al.*, 2017: our contribution allows the use of covariates information through a probit regression framework, while still retaining parameter parsimony. We evaluated the performance of our proposal in a simulated environment, and we provided insights on how

Covariates	# of actors/events	Misclasf. Err. Rate	Adj. Rand Index
actor	$n = 50,\ d = 5$	42.08	19.52
actor	$n = 50,\ d = 15$	14.72	68.51
event	$n = 50,\ d = 15$	15.60	67.62
event	$n = 150,\ d = 15$	13.73	70.61
actor, *event*	$n = 250,\ d = 21$	18.00	66.00

Table 1. *Misclassification error rate and Adjusted Rand index, averaged across the 25 replicated datasets (reported as percentages); data from* `manet+cov`.

to interpret these results. To better understand the potential of using covariates information to aid the clustering task, we aim to apply our model to a real world dataset, in order to measure and understand the impact of *actor* and *event* characteristics on the pattern of attendances. We started to analyze data on Supreme Court votings (Doreian *et al.*, 2004). Preliminary results seem to show the existence of an overlapping pattern of decisions among the $n = 9$ members of the Supreme Court on $d = 26$ topics, categorized by an *event* covariate. Indeed, with $K = 2$, two of the justices are allocated into the heir cluster $\mathbf{z}_i = (1,1)$, while the rest are split into the two parent clusters 4 units in $\mathbf{z}_i = (1,0)$ and 3 units in $\mathbf{z}_i = (0,1)$.

References

DOREIAN, P., BATAGELJ, V., & FERLIGOJ, A. 2004. Generalized blockmodeling of two-mode network data. *Social networks*, **26**(1), 29–53.

FRÜHWIRTH-SCHNATTER, S. 2006. *Finite Mixture and Markov Switching Models*. Springer Science & Business Media, New York.

HOLMES, C. C., & HELD, L. 2006. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, **1**(1), 145–168.

KOLACZYK, E. D. 2009. *Statistical Analysis of Network Data: Methods and Models*. New York: Springer.

MCCULLOCH, C. E., & NEUHAUS, J. M. 2001. *Generalized linear mixed models*. Wiley Online Library.

RANCIATI, S., VINCIOTTI, V., & WIT, E. C. 2017. Identifying overlapping terrorist cells from the Noordin Top actor-event network. *arXiv preprint arXiv:1710.10319*.

WASSERMAN, S., & FAUST, K. 1994. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press.

A STOCHASTIC BLOCKMODEL FOR NETWORK INTERACTION LENGTHS OVER CONTINUOUS TIME

Riccardo Rastelli[1] and Michael Fop[1]

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

ABSTRACT: We introduce a new stochastic blockmodel for the analysis of binary network data evolving over time in a continuous fashion. Data of this type is particularly common in a variety of fields and may describe proximity interactions within a community of individuals, or visual contact between the participants at an event. The model does not rely on a discretization of the time dimension and focuses on the analysis of interaction lengths in the network. The framework assumes a clustering structure on the nodes, where two nodes belonging to the same cluster tend to create interactions and non-interactions of similar lengths. An efficient variational expectation-maximization algorithm is used to perform inference, while the Integrated Completed Likelihood is adopted to select the number of clusters.

KEYWORDS: mixture model, model-based clustering, statistical network analysis, stochastic blockmodel, variational EM algorithm.

1 Introduction

In recent years, a number of network models have been introduced in the literature to study how binary interactions between entities evolve over time. One common approach relies on the discretization of the time dimension: once an appropriate time grid is specified, the continuous data are essentially transformed into a collection of static network snapshots. This approach has facilitated the extension of many static network models to a dynamic framework. The Stochastic Block Model (SBM) of Wang & Wong, 1987 has been recently adapted to the dynamic case by Yang *et al.*, 2011 and Matias & Miele, 2017. However, the approach based on the discretization of the time dimension has been recently criticized, due to the effects that the data transformation may have on the results. In fact, the discretization process always involves a certain level of arbitrariness, either due to the data being collected at specific given times, or because of a post-collection transformation. Indeed, in many data analysis applications, the interactions evolve over time in a continuous

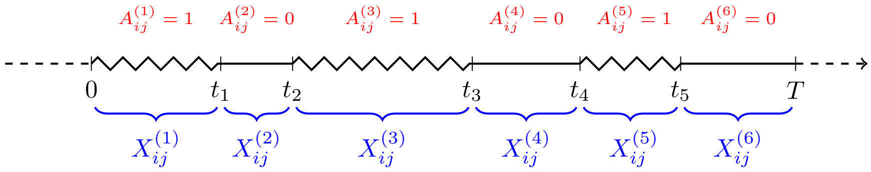


Figure 1. *Timeline representation of the interactions between two arbitrary nodes (interactions are shown with a jagged line). In this case, since $M_{ij} = 6$, there are 4 embedded sub-segments which yield the exponential interaction lengths $X_{ij}^{(3)}$ and $X_{ij}^{(5)}$, and the non-interaction lengths $X_{ij}^{(2)}$ and $X_{ij}^{(4)}$. The interaction length $X_{ij}^{(1)}$ and non-interaction length $X_{ij}^{(6)}$ are truncated from the left and from the right, respectively.*

fashion. For example, in data concerning phone call networks, visual contact networks, speech networks, or proximity networks, the interactions among a collection of entities may be protracted over time, and the object of analysis can be to model for how long these entities interact (and conversely do not interact) within an observed time period.

In this work we introduce a SBM for continuous time network interaction data with the goal of directly modelling the lengths of the observed binary interactions. The SBM structure postulates that the nodes are characterized by a cluster membership variable, which determines both the lengths of the interactions and the lengths of the non-interactions. This framework allows the allocation of nodes into clusters characterized by different interaction and non-interaction rates.

2 Interaction length network data

Consider a collection of nodes $\mathcal{N} = \{1, \ldots, N\}$, which are interacting in a time period of length T. Define the indicator $\mathcal{E}_{ij}(t) = 1$ if nodes i and j are interacting at time t, and $\mathcal{E}_{ij}(t) = 0$ otherwise, for any pair of nodes i and j and for any $t \in [0, T]$. One can represent the observed interaction length data as a collection of M_{ij} segment lengths $\mathbf{X}_{ij}$ and binary indicators $\mathbf{A}_{ij}$, where the m-th value $X_{ij}^{(m)}$ corresponds to the length of either an interaction or non-interaction between nodes i and j, whereas $A_{ij}^{(m)} = 1$ (resp. $A_{ij}^{(m)} = 0$) indicates that the segment corresponds to an interaction (resp. non-interaction). Figure 1 shows the timeline for the interactions between two given nodes: here, the Xs indicate the lengths of each segment, whereas the As denote whether the segment corresponds to an interaction or non-interaction. In practice, the data correspond to an alternating sequence of interactions and non-interactions in the interval $[0,T]$ and is fully characterized by the collection $X_{ij}^{(m)}$ and the initial values $A_{ij}^{(1)}$.

3 Stochastic Blockmodel for interaction length data

We propose a Stochastic Blockmodel for such network interaction data and to perform clustering of the nodes according to their interaction and non-interaction time lengths. The model assumes a mixture distribution with K components, with prior probabilities τ_k. A latent allocation variable Z_{ik} is assigned to each of the nodes, to indicate which one of the K groups node i belongs to, such that $Z_{ik} = 1$ if node i arises from cluster k, 0 otherwise. Interaction and non interaction lengths $X_{ij}^{(m)}$ are modelled assuming a collection of exponential distributions, with parameters μ_{kh} and λ_{kh} respectively. Hence, the parameter μ_{kh} denotes the interaction length rate between a node in cluster k and a node in cluster h; similarly, λ_{kh} denotes the rate for the non-interaction lengths.

Conditionally on the allocation variables, we have the log-likelihood:

$$\log p(\mathbf{X},\mathbf{A}\,|\,\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\lambda}) = \sum_{i\neq j}^{N}\sum_{g=1}^{K}\sum_{h=1}^{K} Z_{ig}Z_{jh}\log p\left(\mathbf{X}_{ij},\mathbf{A}_{ij}\,\middle|\,\mu_{gh},\lambda_{gh}\right),$$

with the term $p\left(\mathbf{X}_{ij},\mathbf{A}_{ij}\,\middle|\,\mu_{gh},\lambda_{gh}\right)$ given by

$$\begin{aligned} p\left(\mathbf{X}_{ij},\mathbf{A}_{ij}\,\middle|\,\mu_{gh},\lambda_{gh}\right) = {} & \left[1-F\left(X_{ij}^{(1)};\mu_{gh}\right)\right]^{A_{ij}^{(1)}}\left[1-F\left(X_{ij}^{(1)};\lambda_{gh}\right)\right]^{1-A_{ij}^{(1)}} \\ & \times\prod_{m=2}^{M_{ij}-1} f\left(X_{ij}^{(m)};\mu_{gh}\right)^{A_{ij}^{(m)}} f\left(X_{ij}^{(m)};\lambda_{gh}\right)^{1-A_{ij}^{(m)}} \\ & \times\left[1-F\left(X_{ij}^{(M_{ij})};\mu_{gh}\right)\right]^{A_{ij}^{(M_{ij})}}\left[1-F\left(X_{ij}^{(M_{ij})};\lambda_{gh}\right)\right]^{1-A_{ij}^{(M_{ij})}}, \end{aligned}$$

where $f(\cdot;\theta)$ and $F(\cdot;\theta)$ are the pdf and cdf of an exponential variable with rate θ, respectively. The terms involving the cdf take into account the fact that the observations at the extremities of the time interval are truncated and interaction (or non-interaction) lengths $X_{ij}^{(1)}$ and $X_{ij}^{(M_{ij})}$ only provide a lower bound for the actual non-observed lengths.

4 Inference

As is usual in model-based clustering, we perform inference for this model by maximizing the marginal log-likelihood $\log p(\mathbf{X},\mathbf{A}\,|\,\boldsymbol{\mu},\boldsymbol{\lambda})$ with respect to the model parameters using an EM algorithm. However, integrating out the allocations $\mathbf{Z}$ is not computationally feasible and the posterior distribution $p(\mathbf{Z}\,|\,\mathbf{X},\mathbf{A})$ does not factorize into a simple form. As a consequence, the E-step cannot be performed exactly, due to the higher computational costs, and this makes the standard EM algorithm not applicable. To overcome this limitation and perform inference, we resort to a variational EM algorithm (Daudin *et al.*, 2008), where a variational approximation is used to replace the posterior distribution on the allocations by a more tractable one, which allows an efficient use of the EM algorithm.

In this framework, model selection corresponds to selection of the optimal number of clusters K, which is often not known and needs be estimated from the data. For this task, we adapt the Integrated Completed Likelihood (ICL) criterion, which has been widely used to perform model choice for mixture models (Biernacki *et al.*, 2000; Côme & Latouche, 2015).

References

BIERNACKI, C., CELEUX, G., & GOVAERT, G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, **22**(7), 719–725.

CÔME, E., & LATOUCHE, P. 2015. Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling*, **15**(6), 564–589.

DAUDIN, J. J., PICARD, F., & ROBIN, S. 2008. A mixture model for random graphs. *Statistics and computing*, **18**(2), 173–183.

MATIAS, C., & MIELE, V. 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**(4), 1119–1141.

WANG, Y. J., & WONG, G. Y. 1987. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, **82**(397), 8–19.

YANG, T., CHI, Y., ZHU, S., GONG, Y., & JIN, R. 2011. Detecting communities and their evolutions in dynamic social networks – a Bayesian approach. *Machine learning*, **82**(2), 157–189.

COMPUTATIONALLY EFFICIENT INFERENCE FOR LATENT POSITION NETWORK MODELS

Riccardo Rastelli[1], Floran Maire[2] and Nial Friel[1, 3]

[1] School of Mathematics and Statistics, University College Dublin,
(e-mail: riccardo.rastelli@ucd.ie, nial.friel@ucd.ie)

[2] Department of Mathematics and Statistics, University of Montréal,
(e-mail: maire@dms.umontreal.ca)

[3] Insight Centre for Data Anlytics, University College Dublin,

ABSTRACT: Statistical inference for latent position network models generally require a computational cost which grows with the square of the number of nodes in the graph. This makes the analysis of large social networks impractical. In this paper, we propose a new method characterised by a linear computational complexity. Our approach relies on an approximation of the likelihood function, where the amount of noise introduced can be arbitrarily reduced at the expense of computational efficiency. We establish several theoretical results that show how the likelihood error propagates to the invariant distribution of the Markov chain Monte Carlo sampler. In particular, we illustrate that one can achieve a substantial reduction in computing time and still obtain a reasonably good estimation of the latent structure.

KEYWORDS: latent position models, noisy Markov chain Monte Carlo, social networks, bayesian inference.

1 Latent Position Models

A random graph is an object $\mathcal{G} = \left\{\mathcal{V}, \mathcal{E}\right\}$ where $\mathcal{V} = \{1, \dots, N\}$ is a fixed set of labels for the nodes and $\mathcal{E}$ is a list of the randomly realised edges. The observed data can be represented by an adjacency matrix $\mathcal{Y}$, where the generic entry y_{ij} is 1 if an edge between i and j appears, or 0 otherwise, with $j > i$.

The nodes are characterised by a latent position, generically denoted $\mathbf{z} \in \mathbb{R}^2$, which determines their social behaviour. The probability of an edge appearing is determined by the positions of the nodes at its extremes and by other global parameters:

$$\log\left(\frac{p\left(\mathbf{z}_i, \mathbf{z}_j; \psi\right)}{1 - p\left(\mathbf{z}_i, \mathbf{z}_j; \psi\right)}\right) := \psi - d\left(\mathbf{z}_i, \mathbf{z}_j\right); \quad (1)$$

where $d(\mathbf{z}_i, \mathbf{z}_j)$ denotes the Euclidean distance between the two nodes, and $\psi \in \mathbb{R}$ is simply an intercept parameter. The likelihood function reads as follows:

$$\mathcal{L}_{\mathcal{Y}}(\mathcal{Z}, \psi) = \prod_{\{i \in \mathcal{V}\}} \prod_{\{j \in \mathcal{V} \backslash i\}} \left\{ [p(\mathbf{z}_i, \mathbf{z}_j; \psi)]^{y_{ij}} [1 - p(\mathbf{z}_i, \mathbf{z}_j; \psi)]^{1-y_{ij}} \right\} \tag{2}$$

and inference is generally carried out through Markov chain Monte Carlo sampling from the posterior distribution:

$$\pi(\mathcal{Z}, \psi | \mathcal{Y}) \propto \mathcal{L}_{\mathcal{Y}}(\mathcal{Z}, \psi) \pi(\mathcal{Z}) \pi(\psi). \tag{3}$$

Since the likelihood depends on all the pairwise distances between the nodes, the number of operations required for inference grows with N^2.

2 Grid approximation of the latent distances

Following an approach similar to that of Parsonage & Roughan, 2017, we create a partitioning of the latent positions $\mathcal{Z}$ using a grid in $\mathbb{R}^2$. The grid is made of adjacent squares (called boxes hereafter) of side length $b > 0$, each having both sides aligned to the axes, as in Figure 1. If we use the centre of the box $\mathbf{c}[g,h]$ as a proxy for the positions of the nodes contained in the same box, the likelihood defined in (2) may be replaced by the following *noisy* likelihood:

$$\tilde{\mathcal{L}}_{\mathcal{Y}}(\mathcal{Z}, \psi) := \left\{ \prod_{i,g,h} [p(\mathbf{z}_i, \mathbf{c}[g,h]; \psi)]^{\xi_i[g,h]} [1 - p(\mathbf{z}_i, \mathbf{c}[g,h]; \psi)]^{\zeta_i[g,h]} \right\}^{1/2} \tag{4}$$

where $\xi_i[g,h]$ (resp. $\zeta_i[g,h]$) is the number of edges (resp. missing edges) between node i and the nodes allocated in $B[g,h]$. The noisy likelihood has a linear computational cost in N.

3 Noisy algorithm

We denote with `NoisyLPM` a Metropolis-within-Gibbs sampler where the likelihood $\mathcal{L}_{\mathcal{Y}}(\mathcal{Z}, \psi)$ is replaced by its proxy $\tilde{\mathcal{L}}_{\mathcal{Y}}(\mathcal{Z}, \psi)$, for some grid parameter b. The approximate Metropolis-within-Gibbs acceptance ratios imply that the stationary distribution of `NoisyLPM` may not coincide anymore with the posterior distribution in 3. In Rastelli *et al.* , 2018, we show that the `NoisyLPM` generates a sequence of random variables whose distribution can be made arbitrarily close to the true posterior $\pi(\cdot | \mathcal{Y})$.

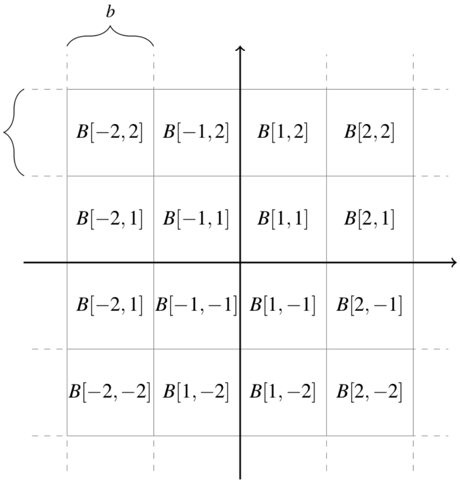

b
b
$B[-2,2]$ $B[-1,2]$ $B[1,2]$ $B[2,2]$
$B[-2,1]$ $B[-1,1]$ $B[1,1]$ $B[2,1]$
$B[-2,1]$ $B[-1,-1]$ $B[1,-1]$ $B[2,-1]$
$B[-2,-2]$ $B[1,-2]$ $B[1,-2]$ $B[2,-2]$

Figure 1. *Grid partitioning the latent space.*

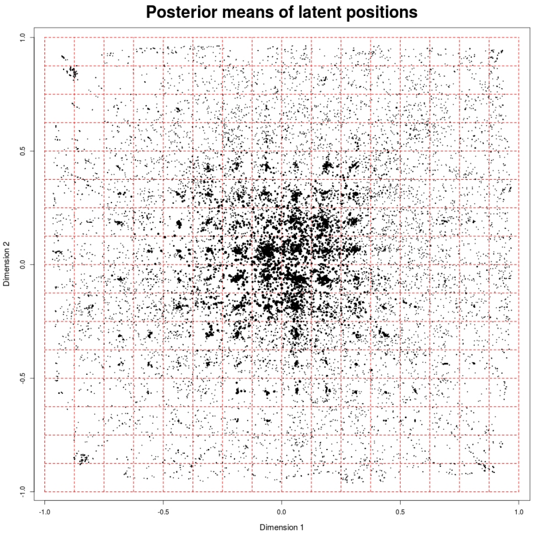


Figure 2. ***Astrophysics***. *Average latent positions of the nodes with circle size proportional to node degree. The grid in dashed red line corresponds to the partitioning imposed.*

4 Coauthorship in astrophysics

The coauthorship network studied in this section was first analysed by Leskovec *et al.* , 2007. The nodes correspond to 18,872 authors, whereas the presence of an edge between two nodes means that the two researchers appear as coauthors on a paper submitted to arXiv, in the astrophysics category.

Figure 2 shows the average latent positions for all of the nodes in the network. We point out that the nodes have a tendency to be distributed close to the centre of each box. This is a natural consequence of our construction, since the centre of the boxes is used as a proxy to calculate the latent distances. We argue that, while the overall macro-structure of the network (i.e. the association of nodes to boxes) is properly recovered, the micro-structure, given by the positions of the nodes within each box, may not necessarily be accurate. The computing time required to obtain the sample was about 46 hours (3.3 seconds per iteration). The non-noisy sampler required an average of 453 seconds per iteration, corresponding to a theoretical 262 days of computations for the full sample.

References

LESKOVEC, J., KLEINBERG, J., & FALOUTSOS, C. 2007. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, **1**(1), 2.

PARSONAGE, E., & ROUGHAN, M. 2017. Fast generation of spatially embedded random networks. *IEEE Transactions on Network Science and Engineering*, **4**(2), 112–119.

RASTELLI, R., MAIRE, F., & FRIEL, N. 2018. Computationally efficient inference for latent position network models. *arXiv preprint arXiv:1804.02274*.

Clustering of complex data stream based on Barycentric Coordinates

Parisa Rastin[1], Basarab Matei[1] and Guénaël Cabanes[1]

[1] Université Paris 13, Sorbonne Paris Cité LIPN - CNRS UMR 7030,
(e-mails: firstname.lastname@lipn.univ-paris13.fr)

Abstract: This paper presents a clustering approach adapted to big and dynamic relational data. The main idea is to use a set of support points chosen among the objects of the data set, independently from the clusters, and use these support points as a basis for the definition of a representation space, using the Barycentric Coordinates formalism. This dynamic approach is applied on a real data set to detect and follow the dynamic of areas of interest over time in user's web navigation.

Keywords: data stream, clustering, relational data, barycentric coordinates.

1 Introduction

Unsupervised learning allows the computation of a model of the data structure when no other information is known. The objects in the data set are grouped in "clusters", based on their similarities. Several applications have been proposed is many domains, such as marketing (Bigné *et al.*, 2010) or climatology (Ramadas *et al.*, 2017). In many cases, the data sets are in perpetual evolution, characterized by a variable structure over time, as new information is constantly appearing. However, this is a difficult problem because of the calculation and storage costs associated with the volumes involved. Indeed, the stream of information represents usually an enormous mass of data to deal with. In addition, the probability distribution associated with the data may change over time ("concept drift", Zhang *et al.*, 2017).

In this paper, we propose a new approach based on the Barycentric Coordinate formalism (Hille, 2005) adapted to complex data streams, allowing creation and suppression of prototypes to follow the dynamic of the data structure. This approach is applied to real data in order to analyze and follow the evolution of areas of interest over time in user's web navigation. Our practical motivation is to perform real-time profiling of connected users, i.e. recognizing the "mindset" of users through their navigation on various websites or their interaction with digital "touch points" (Ahmad *et al.*, 2017).

2 Proposed approach

In this study, a new algorithm based on barycentric coordinates approach to deal with complex data streams is presented.

In the barycentric Coordinate system, the representation space is defined by a unique set of P support points chosen among the objects O. These support points can be any objects chosen randomly from O and represent a virtual space of dimension $P-1$. We aim at representing each cluster by a prototype $\{\mu^1, \mu^2, ..., \mu^K\}$ with K is the number of prototypes. We define the set of support points $O_S = o^i, i \in S \subset O$ associated to an unknown representation in X by $X_S = \{\mathbf{s}^i; \mathbf{s}^i = \mathbf{x}^i, i \in S\} \subset X$. The prototype μ^k of cluster k is defined as a convex combination of the support points:

$$\mu^k = \sum_{p=1}^{P} \beta_p^k \cdot \mathbf{s}^p, \text{ where } \beta^k = (\beta_1^k, \dots, \beta_p^k)^T \in \mathbb{R}^p, \text{ with} \sum_{p=1}^{P} \beta_p^k = 1. \tag{1}$$

This is also the definition of the barycentric coordinate of an object in the space defined by the support points. In other words, β^k are the barycentric coordinates of μ^k with respect to the system of support points X_S. Any object o in the database can also be defined using barycentric coordinates: $o^i = \sum_{p=1}^{P} \beta_p^i \mathbf{s}^p$ with the coordinates β^i satisfying $\sum_{p=1}^{P} \beta_p^i = 1$.

The following metric is defined to evaluate the distance between an object o^i and a prototype μ^k:

$$d^2(o^i, \mu^k) = -\frac{1}{2}(\beta^i - \beta^k)^T \cdot D_S \cdot (\beta^i - \beta^k), \tag{2}$$

where $D_S = (d(o^i, o^j))_{i,j \in S}$ is the dissimilarity matrix between the support points.

In order to obtain the coordinates β^i of an object o^i, with respect to the system of support points O_S, we consider the following matrices:

$$A = \begin{pmatrix} d(\mathbf{s}^1,\mathbf{s}^1)\text{-}d(\mathbf{s}^2,\mathbf{s}^1) & ... & d(\mathbf{s}^1,\mathbf{s}^P)\text{-}d(\mathbf{s}^2,\mathbf{s}^P) \\ . & ... & . \\ . & ... & . \\ . & ... & . \\ d(\mathbf{s}^1,\mathbf{s}^1)\text{-}d(\mathbf{s}^P,\mathbf{s}^1) & ... & d(\mathbf{s}^1,\mathbf{s}^P)\text{-}d(\mathbf{s}^P,\mathbf{s}^P) \\ 1 & ... & 1 \end{pmatrix}, \quad J^i = \begin{pmatrix} d(o^i,\mathbf{s}^1)\text{-}d(o^i,\mathbf{s}^2) \\ . \\ . \\ . \\ d(o^i,\mathbf{s}^1)\text{-}d(o^i,\mathbf{s}^P) \\ 1 \end{pmatrix}. \tag{3}$$

By using the symmetry of D_S, we obtain β^i as solution of the following linear system:

$$A \cdot \beta^i = J^i \Rightarrow \beta^i = A^{-1} \cdot J^i. \tag{4}$$

The problem to optimize in order to find the coordinates of each prototype is a minimization of inertia. The algorithm must update incrementally the barycentric coordinates of each prototype μ^k with respect to the support points X_S for each object of the stream presented to the system if this object belongs to cluster k. As we can compute the barycentric coordinates of o^i in terms of the support points O_S (equation (4)), the update rule of β^k can be written as:

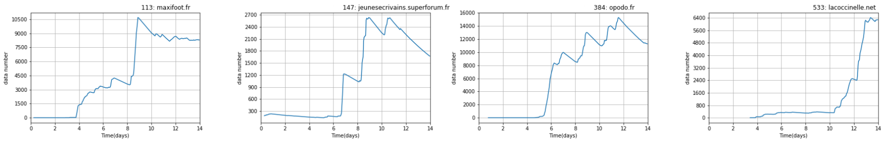
$$\beta^k{}_{t+1} = \beta^k_t - \gamma(\beta^i - \beta^k_t). \tag{5}$$

where γ is the weight (or learning rate) defining the importance of o^i in the new barycentric coordinates.

Finally, in order to take into account the variation over time in the distribution of data in the stream, for each object o^i, after projection in the barycentric space, if the distance between o^i and the closest prototype μ is higher than a maximum radius, a new prototype is created with coordinates β^i. Otherwise, the new object is assigned to the nearest prototype. An "*age*" parameter is associated to each prototype and increased over time. Each time a prototype is the closest to a new object, its "age" is set to 0. If a prototype's age reaches a defined threshold, it is removed from the model.

3 Results

The proposed algorithm was applied to study the dynamic of web navigation behavior, as recorded during two weeks in August 2017 among French Internet users. The evolution of the stream structure has been analyzed in order to highlight trends and variation in the users' behaviors over time. Two measures of similarity were used to compare URLs, using semantic and contextual information associated to each URL. Our objective was to produce a dynamic clustering of this data, in order to monitor the general users' behavior and interest and follow their evolution over time. Figures 1 are examples of results obtained with the proposed algorithm. Such results are very interesting for online marketing companies which constantly need to adapt their advertising strategy to user's "mindset".



a. maxifoot.fr b. jeunesecrivans.fr c. opodo.fr d. lacoccinelle.net

Figure 1. *Time (days) in function of the number of visiting users in each cluster. The closest domain name is given for each cluster.*

4 Conclusion

In this paper, we proposed a new approach able to deal with complex data stream. This algorithm was applied to a real stream of user's web-pages navigation, in order to analyze the structure and dynamics of user's area of interest over time. The results are convincing and encouraging, the clusters are homogeneous with clear associated topics and the evolution of user's interest can be recorded and visualized for each cluster.

References

AHMAD, S., LAVIN, A., PURDY, S., & AGHA, Z.. 2017. Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, **262**, 134 – 147.

BIGNÉ, E., ALDAS-MANZANO, J., KÜSTER, I., & VILA, N.. 2010. Mature market segmentation: a comparison of artificial neural networks and traditional methods. *Neural Computing and Applications*, **19**(1), 1–11.

HILLE, E. 2005. *Analytic Function Theory*. AMS Chelsea Publishing Series, no. vol.2. AMS Chelsea Publishing.

RAMADAS, M., PANT, M., ABRAHAM, A., & KUMAR, S.. 2017. Segmentation of weather radar image based on hazard severity using RDE: reconstructed mutation strategy for differential evolution algorithm. *Neural Computing and Applications,2017*, Jun.

ZHANG, Y., CHU, G., LI, P., HU, X., & WU, X.. 2017. Three-layer concept drifting detection in text data streams. *Neurocomputing*, **260**, 393 – 403.

AN INDSCAL BASED MIXTURE MODEL TO CLUSTER MIXED-TYPE OF DATA

Roberto Rocci[1, 2] and Monia Ranalli[1]

[1] Department of Statistical Sciences, Sapienza University of Rome,
(e-mail: monia.ranalli@uniroma1.it)

[2] Department of Economics and Finance, University of Rome Tor Vergata,
(e-mail: roberto.rocci@uniroma2.it)

ABSTRACT: A new parsimonious model to cluster mixed-type of data is presented. Continuous and ordinal data are modeled by a mixture of Gaussians partially observed. To be parsimonious, it is used a reparameterization of the covariance matrices of the multivariate Gaussians. This permits to control for the number of parameters and simplifies the interpretation of the results.

KEYWORDS: mixture models, mixed-type data, EM algorithm, parsimonious modelling.

1 Introduction

To cluster mixed-type data, i.e. ordinal and continuous variables (Everitt, 1988 and Ranalli & Rocci, 2017a), ordinal variables are assumed to be a discretization of some latent continuous variables jointly distributed with the continuous ones as a Gaussian mixture model (Mclachlan & Peel, 2000). However, a large number of parameters have to be estimated, especially when covariance matrices change over components. Several authors have proposed parsimonious reparametrizations, mainly for continuous data. For example, some constrain the eigenvalues and/or the eigenvectors of the covariance matrices to be the same across the groups (Banfield, 1993), while others reduce the number of parameters by using a factor analysis model for each covariance matrix (McLachlan *et al.* , 2003). In the same context, we find proposals where mixtures of factor analyzers are used to obtain different parsimonious models (McNicholas & Murphy, 2008). A different approach has been developed for continuous and ordinal data (see by Kumar & Andreou, 1998 and Ranalli & Rocci, 2017b, respectively). They assume that there exist two within uncorrelated sets of factors that generate the variables as linear combinations, whose distributions have group specific parameters only for the first set. In this way the reduction is not only in the number of parameters but also in the dimensionality of the data. In fact, the component variables of the second set, without class specific parameters, can be considered noise dimensions. In this framework, we propose a new parsimonious reparameterization based on the assumption that the variables are linear combinations of within uncorrelated latent variables where only some of them are characterized by class specific parameters. The material is organized as follows. In section 2, we present the model specification. In section 3, we outline the model parameter estimation. Finally some remarks and considerations are discussed in section 4. The EM-like algorithm and an example of application on real data showing the effectiveness of the proposal will be presented elsewhere for lack of space.

2 Model specification

Let $\mathbf{x} = [x_1, \ldots, x_O]'$ and $\mathbf{y}^{\bar{O}} = [y_{O+1}, \ldots, y_P]'$ be O ordinal and $\bar{O} = P - O$ continuous variables. The associated categories for each ordinal variable are denoted by $c_i = 1, 2, \ldots, C_i$ with $i = 1, 2, \ldots, O$. Following the Underlying Response Variable approach (Muthén, 1984), the ordinal variables $\mathbf{x}$ are considered as a categorization of a continuous multivariate latent variable $\mathbf{y}^O = [y_1, \ldots, y_O]'$. The latent relationship between $\mathbf{x}$ and $\mathbf{y}^O$ is explained by the threshold model, $x_i = c_i \Leftrightarrow \gamma^{(i)}_{c_i - 1} \leq y_i < \gamma^{(i)}_{c_i}$, where $-\infty = \gamma^{(i)}_0 < \gamma^{(i)}_1 < \ldots < \gamma^{(i)}_{C_i - 1} < \gamma^{(i)}_{C_i} = +\infty$ are the thresholds defining the C_i categories collected in a set $\mathbf{\Gamma}$ whose elements are given by the vectors $\boldsymbol{\gamma}^{(i)}$. We assume that $\mathbf{y} = [\mathbf{y}^{O\prime}, \mathbf{y}^{\bar{O}\prime}]'$ follows a heteroscedastic Gaussian mixture, $f(\mathbf{y}) = \sum_{g=1}^{G} p_g \phi_P \left(\mathbf{y}; \boldsymbol{\mu}_g, \mathbf{\Sigma}_g\right)$, where the p_g's are the mixing weights and $\phi_P \left(\mathbf{y}; \boldsymbol{\mu}_g, \mathbf{\Sigma}_g\right)$ is the density of a P-variate normal distribution with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\mathbf{\Sigma}_g$. All the parameters are contained in $\boldsymbol{\psi}$.
For a random i.i.d. sample of size N, $(\mathbf{x}_1, \mathbf{y}^{\bar{O}}_1), \ldots, (\mathbf{x}_N, \mathbf{y}^{\bar{O}}_N)$, the log-likelihood is

$$\ell(\boldsymbol{\psi}) \quad = \quad \sum_{n=1}^{N} \log \left[\sum_{g=1}^{G} p_g \phi_{\bar{O}}(\mathbf{y}^{\bar{O}}_n; \boldsymbol{\mu}^{\bar{O}}_g, \mathbf{\Sigma}^{\bar{O}}_g) \pi_n \left(\boldsymbol{\mu}^{O|\bar{O}}_{n;g}, \mathbf{\Sigma}^{O|\bar{O}}_g, \mathbf{\Gamma} \right) \right], \qquad (1)$$

where, with obvious notation $\pi_n \left(\boldsymbol{\mu}^{O|\bar{O}}_{n;g}, \mathbf{\Sigma}^{O|\bar{O}}_g, \mathbf{\Gamma} \right) = \int_{\gamma^{(1)}_{c_1 - 1}}^{\gamma^{(1)}_{c_1}} \cdots \int_{\gamma^{(O)}_{c_O - 1}}^{\gamma^{(O)}_{c_O}} \phi_O(\mathbf{u}; \boldsymbol{\mu}^{O|\bar{O}}_{n;g}, \mathbf{\Sigma}^{O|\bar{O}}_g) d\mathbf{u}$
where, $\pi_n \left(\boldsymbol{\mu}^{O|\bar{O}}_{n;g}, \mathbf{\Sigma}^{O|\bar{O}}_g, \boldsymbol{\gamma} \right)$ is the conditional joint probability of response pattern $\mathbf{x}_n = (c^{(1)}_1, \ldots, c^{(O)}_O)$ given the cluster g and the continuous variables $\mathbf{y}^{\bar{O}}_n$. Finally p_g is the probability of belonging to group g subject to $p_g > 0$ and $\sum_{g=1}^{G} p_g = 1$. We assume that in each class the P variables are linear combinations of the same P latent factors, which are uncorrelated and change, from one

cluster to another, only in the means and variances. In formulas, if observation n comes from the subpopulation g ($g = 1,\ldots,G$), then the following model holds

$$\mathbf{y}_n = \mathbf{B}(\boldsymbol{\eta}_g + \mathbf{L}_g^{1/2}\mathbf{f}_n) \tag{2}$$

where $\mathbf{B}$ is a full rank $(P \times P)$ matrix of component loadings, $\mathbf{f}_n$ is a random vector of P latent variables normally distributed with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_P$ and $\boldsymbol{\eta}_g$ and $\mathbf{L}_g$ are a column vector and a positive definite diagonal matrix, respectively. This model implies that in component g-th the P observed variables are linear combination of P latent factors having $\boldsymbol{\eta}_g$ and $\mathbf{L}_g$ as mean vector and covariance matrix, respectively. The density of $\mathbf{y}_n$, given that observation n comes from the g-th subpopulation, is multivariate normal with mean $\boldsymbol{\mu}_g = \mathbf{B}\boldsymbol{\eta}_g$ and covariance matrix $\boldsymbol{\Sigma}_g = \mathbf{B}\mathbf{L}_g\mathbf{B}'$, obtaining a reparameterization of the covariance matrices well-known in the multidimensional scaling literature under the name INDSCAL (Carroll & Chang, 1970).

3 Model Estimation

To overcome the presence of multidimensional integrals, here, the full log-likelihood is replaced by a composite likelihood (Lindsay, 1988) formed of $O(O-1)/2$ marginal distributions each of them composed of two ordinal variables and the $\bar{O}$ continuous variables. This leads to the following surrogate function

$$c\ell(\boldsymbol{\psi}) = \sum_{n=1}^{N}\sum_{i=1}^{O-1}\sum_{j=i+1}^{O}\sum_{c_i=1}^{C_i}\sum_{c_j=1}^{C_j} \delta_{nc_ic_j}^{(ij)} \log\left[\sum_{g=1}^{G} p_g \pi_{c_ic_j}^{(ij|\bar{O})}(\boldsymbol{\mu}_g^{(ij|\bar{O})}, \boldsymbol{\Sigma}_g^{(ij|\bar{O})}, \boldsymbol{\Gamma}^{(ij)})\phi_{\bar{O}}(\mathbf{y}_n^{\bar{O}}; \boldsymbol{\mu}_g^{\bar{O}}, \boldsymbol{\Sigma}_g^{\bar{O}\bar{O}})\right],$$

where $\delta_{nc_ic_j}^{(ij)}$ is a dummy variable assuming 1 if the n-th observation presents the combination of categories c_i and c_j for variables x_i and x_j respectively, 0 otherwise; $\pi_{c_ic_j}^{(ij|\bar{O})}(\boldsymbol{\mu}_g^{(ij|\bar{O})}, \boldsymbol{\Sigma}_g^{(ij|\bar{O})}, \boldsymbol{\Gamma}^{(ij)})$ is the conditional probability of variables x_j and x_i of being in category c_i and c_j respectively, given all the continuous variables $\mathbf{y}^{\bar{Q}}$. The parameter estimates are carried out through an EM-like algorithm, that works in the same manner as the standard EM on a composite complete log-likelihood.

4 Discussion

In this work we propose a parsimonious version of mixture of Gaussian partially observed based on a specific decomposition of the covariance matrices. This is always true when $G = 2$ but does not necessarily hold for G $\geq$ 3. In this case, if needed, we can circumvent the lack of fit by relaxing some of the constraints, for example by requiring a block diagonal form rather than simply diagonal for the matrices $\mathbf{L}_g$, or by assuming that the number of factors is greater than P, i.e. the number of manifest variables (assuming $\mathbf{B}$ to be rectangular).

References

BANFIELD, J. D .AND RAFTERY, A. E. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803–821.

CARROLL, J. D., & CHANG, J.-J. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, **35**(3), 283–319.

EVERITT, B.S. 1988. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, **6**(5), 305–309.

KUMAR, N., & ANDREOU, A.G. 1998. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, **26**(4), 283 – 297.

LINDSAY, B. 1988. Composite likelihood methods. *Contemporary Mathematics*, **80**, 221–239.

MCLACHLAN, G., & PEEL, D. 2000. *Finite Mixture Models*. 1 edn. Wiley Series in Probability and Statistics. Wiley-Interscience.

MCLACHLAN, G. J., PEEL, D., & BEAN, R.W. 2003. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, **41**(3-4), 379–388.

MCNICHOLAS, P. D., & MURPHY, T. B. 2008. Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**(3), 285–296.

MUTHÉN, B. 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49**(1), 115–132.

RANALLI, M., & ROCCI, R. 2017a. Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis*, **110**(C), 87–102.

RANALLI, M., & ROCCI, R. 2017b. A Model-Based Approach to Simultaneous Clustering and Dimensional Reduction of Ordinal Data. *Psychometrika*.

TOPOLOGICAL STOCHASTIC NEIGHBOR EMBEDDING

Nicoleta Rogovschi[1], Nistor Grozavu[2], Basarab Matei[2],
Younès Bennani[2] and Seiichi Ozawa[3]

[1] LIPADE, University of Paris 5,
(e-mail: nicoleta.rogovschi@parisdescartes.fr)

[2] LIPN-UMR CNRS 7030, Université de Paris 13,
(e-mail: name.surname@lipn.univ-paris13.fr)

[2] Graduate School of Engineering, Kobe University,
(e-mail: ozawa@eedept.kobe-u.ac.jp)

ABSTRACT: This paper introduces a new topological machine learning model in order to project high dimensional datasets without loosing the structure of the data. The model is based on SNE (Stochastic Neighbor Embedding) dimensionality reduction method and Self-Organizing Maps (SOM). The SNE method which performs good results for visulaization allows a projection of the dataset in low dimensional spaces that make it easy to use for very large datasets. Using SNE during the learning process will allow to reduce the dimensionality and to preserve the topology of the dataset by increasing the clustering accuracy.

KEYWORDS: Stochastic Neighbor Embedding; Self-Organizing Maps, Clustering, Visualization.

1 Introduction

Topological learning is a recent direction in Machine Learning which aims to develop methods grounded on statistics to recover the topological invariants from the observed data points. Most of the existed topological learning approaches are based on graph theory or graph-based clustering methods.

The main purpose of unsupervised learning methods is to extract generally useful features from unlabelled data, to detect and remove input redundancies, and to preserve only essential aspects of the data in robust and discriminative representations. Unsupervised methods have been routinely used in many scientific and industrial applications.

Unsupervised feature learning algorithms aim to find good representations for data, which can be used for different tasks i.e. classification, clustering, reconstruction, visualization,... Recently, the SNE (Hinton & Roweis, 2003) a method has shown high feature learning performance used for dimensionality reduction and visualization (Kitazono *et al.* , 2016).

Given a data matrix represented as vectors of variables (p observations and n features), the goal of the unsupervised transformation of feature space is to produce another data matrix of dimension (p,n') (the transformed representation of n' new latent variables) or a similarity matrix between the data of size (p,p). Applying a meodel on the transformed matrix should provide better results compared to the original dataset.

In this paper we have focused on models that are based on topological unsupervised learning and the SNE (Stochastic Neighbor Embedding) in order to reduce the data dimensionality of the data and to take advantage from the topological preservation of information.

2 Proposed model

Tha main principle of the Stochastic Neighbor Embedding (SNE) is to convert the high-dimensional Euclidean distances between datapoints into conditional probabilities that represent similarities. The similarity of datapoint x_j to datapoint x_i is the conditional probability, $p_{j|i}$ defined as follows:

$$p_{j|i} = \frac{\exp(-\frac{\|x_i - x_j\|^2}{2\delta_i^2})}{\sum_{k\neq i}\exp(-\frac{\|x_i - x_k\|^2}{2\delta_i^2})} \tag{1}$$

where δ_i is the variance of the Gaussian that is centered on datapoint x_i.

For the low-dimensional data y_i and y_j corresponding to high-dimensional datapoints x_i and x_j, a conditional probability denote by $q_{j|i}$ is computed as follows:

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k\neq i}\exp(-\|y_i - y_k\|^2)} \tag{2}$$

The Kullback-Leibler divergence is used as a measure of the faithfulness with which $q_{j|i}$ models $p_{j|i}$. SNE minimizes the sum of Kullback-Leibler divergences over all datapoints using a gradient descent method given by the cost function:

$$C_{SNE} = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{3}$$

wher P_i represents the conditional probability distribution over all other datapoints given datapoint x_i, and Q_i represents the conditional probability distribution over all other map points given map point y_i.

Nexxt, the SOM model is used by minimizing the objective function:

$$R(\chi, \mathcal{W}) = \sum_{i=1}^{N} \sum_{j=1}^{|\mathcal{W}|} \mathcal{K}_{j,\chi(\mathbf{x}_i)} \|\mathbf{x}_i - \mathbf{w}_j\|^2 \tag{4}$$

where χ assigns each observation $\mathbf{x}_i$ to a single cell in the map $\mathcal{C}$. This cost function can be minimized using both stochastic and batch techniques T., 2001.

Algorithm 1 Topological Stochastic Neighbor Embedding

```
Input: $n$ data points $x_1, x_2, ..., x_n \in \mathbb{R}^m$; Cluster number $k$ ;
Output: $k$ clusters;

1. Compute pairwise affinities $p_{j|i}$ (using Equation 1)
2. for i=1 to Iter do
       compute low-dimensional affinities $q_{ij}$
       compute the low dimensional data ($Y$) using the gradient $\frac{\delta_C}{\delta_Y}$
3. end for
4. Compute the prototypes matrix $W$ using the SOM algorithm on the
   low dimensions $Y$
5. Cluster each cell ( prototype ) from $W$ into $k$ clusters via Hierarchical
   Clustering algorithm
```

In the Algorithm 1 we present the proposed model which allows to cluster a dataset by preserving the topological structure of the data.

To evaluate the proposed method, we used several datasets of different size and complexity presented in table 1 by comparing with k-means and spectral clustering. We performed several experiments on diferent problems from the UCI Repository of machine learning databases (Asuncion & Newman, 2007).

The obtained results shows what the proposed model outperforms the k-means and spectral clustering in terms of the Accuracy and Rand index. For

MNIS daaset the obtained Rand index is 0,95 for the proposed model compared to 0,92 for spectral clustering.

For other datasets we can note that our method always outperforms the classical k-means and the spectral clustering, but we have to note here that the goal is also to preserve the topological structure of the data for visualization.

3 Conclusions

In this study we proposed a new topological unsupervised learning model which allows to cluster a large dataset by preserving the local structure of the data. The proposed method use the Self-Organizing Maps by reducing the dimensionality using the SNE model. The obtained results show that the proposed method improves the clustering results in term of external indexes. For future work, the spectral topological clustering can be used to improve the clustering results, and to adapt this method for multi-view datasets.

References

ASUNCION, A., & NEWMAN, D.J. 2007. *UCI Machine Learning Repository.*

HINTON, GEOFFREY, & ROWEIS, SAM. 2003. Stochastic Neighbor Embedding. *Advances in neural information processing systems*, **15**, 833–840.

KITAZONO, JUN, GROZAVU, NISTOR, ROGOVSCHI, NICOLETA, OMORI, TOSHIAKI, & OZAWA, SEIICHI. 2016. t-Distributed Stochastic Neighbor Embedding with Inhomogeneous Degrees of Freedom. *Pages 119–128 of: Neural Information Processing - 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16-21, 2016, Proceedings, Part III.*

T., KOHONEN. 2001. *Self-organizing Maps.* Springer Berlin.

VAN DER MAATEN, L.J.P., & HINTON, G.E. 2008. Visualizing High-Dimensional Data Using t-SNE.

FUNCTIONAL DATA ANALYSIS FOR SPATIAL AGGREGATED POINT PATTERNS IN SEISMIC SCIENCE

Elvira Romano[1], Jonatan González Monsalve[2], Francisco Javier Rodríguez Cortés[3]
and Jorge Mateu[2]

[1] Department of Mathematics and Physics, University of Campania Luigi Vanvitelli,
(e-mail: elvira.romano@inicampania.it)

[2] Department of Mathematics, University of Jaume I,
(e-mail: jmonsalv@uji.es, mateu@mat.uji.es)

[3] Escuela de estadística, Universidad Nacional de Colombia, Medellín,
(e-mail: frrodriguezc@unal.edu.co)

ABSTRACT: Earthquakes can be seen as realization of a spatial, temporal, or spatiotemporal point process. Given a dataset of earthquakes in a mixed geographical region, a scientific question that naturally arises is whether can we separate the earthquakes in two fundamental disjoint sets: triggered (sequential) and background (complete random). Such a separation becomes quite important as background earthquakes are basically blurring main spots of triggered ones. We consider LISA functions as functional marks attached to the points in the spatial point pattern of the earthquakes. We then classify the points through Aitchison distance and subsequent multivariate classification techniques. The performance of our method is demonstrated by simulation.

KEYWORDS: clustering, spatial point pattern, functional data, LISA functions.

References

ANSELIN, K. 1995. Local indicators of spatial association-LISA. *Geographical Analysis*, 27, 93–115.

CRESSIE, N., COLLINS, B. 2001. Analysis of Spatial Point Patterns Using Bundles of Product Density LISA Functions. *Journal of Agricultural, Biological, and Environmental Statistics 6.1,118-135.*

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

ROC CURVES WITH BINARY MULTIVARIATE DATA

Lidia Sacchetto[1] and Mauro Gasparini[1]

[1] Department of Mathematical Sciences, Politecnico di Torino,
(e-mail: lidia.sacchetto@polito.it, mauro.gasparini@polito.it)

ABSTRACT: Binary unsupervised classification of n units based on a set of r binary items is considered. Classical results on likelihood ratio based procedures are revisited and discussed, together with some recent results. Hard assignment is considered, as opposed to the soft assignment preferred by the latent class (or mixture) authors.

KEYWORDS: concentration function, unsupervised clustering, hard assignment.

1 Introduction

This is a review of some recent and some ongoing works on a basic model-based approach to ROC curves. Recently, whole books have been devoted to the study of ROC curves (e.g. Pepe, 2003, Krzanowski & Hand, 2009, Zou *et al.*, 2011), but we would like to give further contribution to the study of their properties in the presence of *multivariate binary data*.

Our multivariate binary data, such as those generated by polls, questionnaires and online automated interviews, are binary words representing the *profile* of a statistical unit about which several binary questions have been answered; each bit in the profile binary word then represents a binary answer. The goal is to classify the sampled units into one of two populations, say underachievers and overachievers (in Education), healthy and diseased (in Medicine), or right-wing and left-wing (in Political Science) and so on. In this work the two populations (or, equivalently, their probability measures) are indicated as P_+ and P_-.

Traditionally, in the presence of such data, a substantive researcher (such as a Pedagogist, a Physician or a Political Scientist in the above mentioned examples) would focus on certain clearly interpretable key profiles of interest, make up a distance of the observed profile from the few key profiles, then proceed with distance-based classification methods. We would like instead to evaluate up to which point, with modern technology, we can use basic methods such as *maximum likelihood* and *hard assignment* towards automated identification of the two classes and of the class labels of the sampled subjects.

2 Likelihood Ratio based ROC curves and concentration

As it is well known since the early developments in classification, likelihood ratio (LR from now on) based methods are optimal. A general treatment of this topic is contained in Sacchetto & Gasparini, 2019, whence three very general results can be stated:

1. LR is a scalar score which can be constructed as long as two compatible probability measures are assigned to the two competing populations;
2. the ROC curve based on the LR is proper;
3. the ROC curve is in a one-to-one correspondence to a general measure of concentration portrayed in Cifarelli & Regazzini, 1987 according to the following formula: $\mathrm{ROC}(x) = 1 - \varphi(1-x) \quad \forall 0 \leq x \leq 1$, where $\mathrm{ROC}(\cdot)$ is the ROC for the LR-based procedure for P_+ *vs.* P_- and $\varphi(\cdot)$ if the concentration function due to Cifarelli & Regazzini, 1987, a generalization of the Lorenz curve to very general abstract measures on P_+ and P_-.

Based on Result 1 above, we now proceed to study some properties of the ROC curve when the data are multivariate binary.

3 Hard assignment versus soft assignment

When performing binary classification with multivariate binary data, one is confronted with a database, such as the one in the first two columns of Table 1 in the example presented in Section 5. The statistical units, i.e. the rows of the database, have to be assigned to P_+ or P_-; in other words, the unknown unit labels $\gamma_1, \gamma_2, \ldots, \gamma_n$ have to be inferred. Notice this is, again, a basic way to do classification, which is sometimes called *hard assignment* as opposed to *soft assignment*, such as the one used in the mixture modelling literature. Nowadays, the latter is prevalent and it is implemented in popular softwares, as, for example, the `poLCA` R library (Linzer & Lewis, 2011). With hard assignment, $\gamma_1, \gamma_2, \ldots, \gamma_n$ are parameters themselves, together with the unknown probabilities of success in the two populations $\pi^+ = (\pi_1^+, \ldots, \pi_r^+)$ and $\pi^- = (\pi_1^-, \ldots, \pi_r^-)$, where r is the number of items, i.e. the length of the binary profile. Maximum likelihood estimates are then the maximizer of the likelihood.

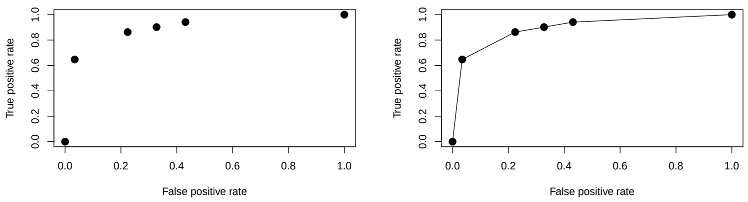


Figure 1. *The usual definition of a ROC based on discrete data (left) and the ROC on the same data completed via a randomization device (right).*

4 Discrete data and ROC curves

A technical issue is that, when the LR takes on a discrete set of values, as it happens with binary data, the ROC curve is only a discrete set of points, as in Figure 1, left panel. A byproduct of Result 3 above, highlighted in Sacchetto & Gasparini, 2019, is that if we give a general definition of ROC curve including a randomization device (similar to randomized Neyman-Pearson tests), then the discrete points can be joined to form a *concave* and *proper* ROC curve, as in Figure 1, right panel.

5 An example from the literature

Bartholomew *et al.*, 2008 described latent class models for binary data using an educational assessment dataset, containing answers to $r = 4$ binary questions provided by $n = 142$ subjects. The information was collected to study the learning process in children and, in particular, to classify students in two groups: “masters” students were expected to answer correctly to most questions (majority of 1’s in the profile), while mostly 0’s were expected from “non-masters”. However, students could also answer in the right way by chance, as well as give the wrong answer due to oversight.We compared the results obtained via `poLCA` with our hard assignment approach. Results are shown in Table 1. 20 out of 142 subjects are classified differently by the two methods: the reasons for the observed differences are a matter of discussion. In particular, one should further investigate profiles with high posterior probabilities in the latent variable approach which are classified differently by hard assignment. Future work will consist of simulations aimed at evaluating the quality of the classification via appropriate indices such as sensitivity, specificity and accuracy of the two partitions with respect to the true one.

Table 1. *Latent classes analysis (`poLCA` Ass) versus hard assignment (Hard Ass) on Macready & Dayton, 1977 data. Posterior probabilities calculated using `poLCA`.*

Profiles	Frequency	Prob. $\in P_-$	Prob. $\in P_+$	`poLCA` Ass.	Hard Ass.
1111	15	0.000	1.000	+	+
1101	23	0.002	0.998	+	+
1110	7	0.002	0.998	+	+
0111	4	0.001	0.999	+	+
1011	1	0.003	0.997	+	-
1100	7	0.087	0.913	+	+
1001	6	0.095	0.905	+	-
0101	5	0.025	0.975	+	+
1010	3	0.100	0.900	+	-
0110	2	0.026	0.974	+	+
0011	4	0.029	0.971	+	-
1000	13	0.822	0.178	-	-
0100	6	0.526	0.474	-	+
0001	4	0.550	0.450	-	-
0010	1	0.563	0.437	-	-
0000	41	0.982	0.018	-	-

References

Bartholomew, D.J., Steele, F., & Moustaki, I. 2008. *Analysis of multivariate social science data*. Chapman and Hall/CRC.

Cifarelli, D.M., & Regazzini, E. 1987. On a general definition of concentration function. *Sankhyā, Series B*, 307–319.

Krzanowski, W.J., & Hand, D.J. 2009. *ROC curves for continuous data*. Chapman and Hall/CRC.

Linzer, D.A., & Lewis, J.B. 2011. poLCA: An R Package for Polytomous Variable Latent Class Analysis. *J. Stat. Soft.*, **42**(10), 1–29.

Macready, G.B., & Dayton, C.M. 1977. The use of probabilistic models in the assessment of mastery. *J. Educat. Stat.*, **2**(2), 99–120.

Pepe, M.S. 2003. *The statistical evaluation of medical tests for classification and prediction*. Medicine.

Sacchetto, L., & Gasparini, M. 2019. Proper likelihood ratio based ROC curves for general binary classification problems. *arXiv:1809.00694*.

Zou, K.H., Liu, A., Bandos, A.I., Ohno-Machado, L., & Rockette, H.E. 2011. *Statistical evaluation of diagnostic performance: topics in ROC analysis*. CRC Press.

SILHOUETTE-BASED METHOD FOR PORTFOLIO SELECTION

Marco Scaglione[1], Carmela Iorio[2] and Antonio D'Ambrosio[1]

[1] Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Napoli Federico II, (e-mail: marco.scaglione@unina.it, antdambr@unina.it)

[2] Dipartimento di Ingegneria Industriale, Università degli Studi di Napoli Federico II, (e-mail: carmela.iorio@unina.it)

ABSTRACT: Many new methods have been proposed for improving protfolio selection process, applying clustering techniques to stock price time series. We apply a recently proposed clustering method that employs stock price's time series P-Spline coefficients, instead of the original time series, in order to speed up and stabilize the clustering process. Then we propose a cluster validation benchmark technique that enable us to automatically build financial portfolio that complies with some strategic constraints.

KEYWORDS: Silhouette, P-splines, PAM, portfolio selection.

1 Introduction

The problem of portfolio optimization is one of the most important issues in asset management. Since the seminal works of Harry Markowitz in the fifties (Markowitz, 1952; Markovitz, 1959), many other researches have been focused on several aspects of portfolio optimization both from an applied and from a theoretical point of view.

In the last years the attention brought by the scientific community to the field of statistical learning, pushed many clustering techniques in building optimized portfolio (Tola et al., 2008) or achieving some tracking results (Dose and Cincotti, 2005). A review of financial application of data mining techniques is provided in Hajizadeh et al. (2010).

Unsupervised learning techniques have yielded some good result in this direction, for example mining stock categories in stock exchanges (see, for example, Liao et al., 2008; Nanda et al., 2010), building stable portfolios (Zhang and Maringer, 2009) and showed good effects on portfolio formation and risk analysis (Lemieux et al., 2014).

The purpose of many unsupervised learning techniques is to group data maximizing the similarity within the groups and minimizing the similarity between

them, and this lead to a natural link with the object of portfolio management and risk diversification.
We propose a cluster validation benchmark aiming to portfolio selection based on cluster silhouette method. In section 2 we explain the main feature of the method and how to achieve the portfolio through the silhouette statistic. In section three we give some remarks about the proposal and point some conclusions.

2 P-spline clustering and silhouette based portfolio building procedure

Recently, Iorio et al. (2016) exposed the benefits in terms of computational saving and data pre-processing of clustering time series through the coefficients of P-spline smoothers. Iorio et al. (2018) provide an extensive application of the method to a portfolio selection problem.
Our proposal is to cluster time series of stock prices by the Penrose shape distance (Penrose, 1952) of P-spline coefficients, through a Partition Around Medoids algorithm (Kaufman and Rousseeuw, 1987), attempting to cluster them by their shape. Penrose shape distance is a distance in R^n, defined as

$$D_{xy} = \sqrt{\sum((x_i - \bar{x}) - (y_i - \bar{y}))^2}. \tag{1}$$

2.1 P-spline in a nutshell

A Spline is a function defined by many polynomial functions, linked in many points, named knots, endowed with an high degree of uniformity. Its generalized expression is:

$$B_{ik} = \sum_{j=i}^{i+k-1} b_{jk} X_j \tag{2}$$

The goal of estimating the best fitting function for data without introducing variability in the estimation, is achieved minimizing the objective function expressed as:

$$S_\lambda = (y - Bc)^T (y - Bc) + \lambda a^T Pa, \tag{3}$$

where B is the basis matrix, P is the penalty matrix and c is the vector of coefficients. The problem is solved through a least square procedure, producing the following expression:

$$(B^T B + \lambda P)\hat{a} = B^T y. \tag{4}$$

This allows curve interpolation avoiding overfitting, so that when we have an high number of knots we don't have interpolated curves showing too much variability.

2.2 Silhouette for portfolio selection

After the clustering of the data, validated through the use of the silhouette statistics (Rousseeuw, 1987), we compute the silhouettes of each series observed and clustered, then we proceed to build the portfolio. First, we compute the average silhouette of each group found by the clustering procedure and select only those laying on the right tail of each cluster average. Each selected stock enters the portfolio attending the following weighting scheme:

$$W_{ij} = \frac{S_{ij} - \mu_{Cj}}{\sum_{i \in C_j}(S_{ij} - \mu_{Cj})}, \tag{5}$$

where S_{ij} is the silhouette value of the i-th series in the j-th cluster and μ_{C_j} is the average silhouette value for the series in cluster j. In this way we impose restrictions on the portfolio composition. Considering only those having silhouette values higher than the cluster average silhouette we are not allowing for any negative weighting scheme to be considered. Moreover, we ensure to invest all our wealth in the portfolio.

3 Conclusions

We propose a new strategy for building portfolio with some pre-specified features, through the cluster analysis of time series' P-splines coefficients, working on their silhouettes, that automatically achieves the feasibility constraints and some degrees of diversification or some benchmark level.

References

Dose, C. and Cincotti, S. (2005). Clustering of financial time series with application to index and enhanced index tracking portfolio. *Physica A: Statistical Mechanics and its Applications*, 355(1):145–151.

Hajizadeh, E., Ardakani, H. D., and Shahrabi, J. (2010). Application of data mining techniques in stock markets: A survey. *Journal of Economics and International Finance*, 2(7):109–118.

Iorio, C., Frasso, G., D'Ambrosio, A., and Siciliano, R. (2016). Parsimonious time series clustering using p-splines. *Expert Systems with Applications*, 52:26–38.

Iorio, C., Frasso, G., D'Ambrosio, A., and Siciliano, R. (2018). A p-spline based clustering approach for portfolio selection. *Expert Systems with Applications*, 95:88–103.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. statistical data analysis based on the l1 norm. *Y. Dodge, Ed*, pages 405–416.

Lemieux, V., S. Rahmdel, P., Walker, R., Wong, B., and Flood, M. (2014). Clustering techniques and their effect on portfolio formation and risk analysis. pages 1–6.

Liao, S.-H., Ho, H.-h., and Lin, H.-w. (2008). Mining stock category association and cluster on taiwan stock market. *Expert Systems with Applications*, 35(1-2):19–29.

Markovitz, H. (1959). Portfolio selection: Efficient diversification of investments. *The Journal of Finance*.

Markowitz, H. (1952). Portfolio selection. *The journal of finance*, 7(1):77–91.

Nanda, S., Mahanty, B., and Tiwari, M. (2010). Clustering indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12):8793–8798.

Penrose, L. S. (1952). Distance, size and shape. *Annals of Eugenics*, 17(1):337–343.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Tola, V., Lillo, F., Gallegati, M., and Mantegna, R. N. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258.

Zhang, J. and Maringer, D. (2009). Improving sharpe ratios and stability of portfolios by using a clustering technique. In *Proceedings of the World Congress on Engineering*, volume 1.

ITEM WEIGHTED KEMENY DISTANCE FOR PREFERENCE DATA

Mariangela Sciandra[1], Simona Buscemi[1] and Antonella Plaia[1]

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

ABSTRACT: Preference data represent a particular type of ranking data where a group of people gives their preferences over a set of alternatives. The traditional metrics between rankings don't take into account that the importance of elements can be not uniform. In this paper the item weighted Kemeny distance is introduced and its properties demonstrated.

KEYWORDS: Preference data, item importance, distances.

1 Introduction

Ranking is one of the most simplified cognitive processes used by people to handle many aspects of their lives. When some subjects are asked to indicate their preferences over a set of alternatives (items), ranking data are called preference data. Therefore, ranking data arise when a group of n individuals (judges, experts, voters, raters, etc.) shows their preferences for a finite set of items (m different alternatives of objects, like movies, activities and so on). The two representations of a ranking are the rank vector and the order vector. The rank vector lists the ranks given to the objects, the order vector lists the true order of objects in order from best to worst. It is possible to switch from orderings to rankings and vice-versa, and in this paper, we will refer to orderings. If the m items, labeled $1,\dots m$, are ranked in m distinguishable ranks, a complete (full) ranking or linear ordering is achieved (Cook, 2006): this ranking a is a mapping function from the set of items $\{1,\dots,m\}$ to the set of ranks $\{1,\dots,m\}$, endowed with the natural ordering of integers, where $a(i)$ is the rank given by the judge to item i. Ranking a is, in this case, one of the $m!$ possible permutations of m elements. When some items receive the same preference, then a tied ranking or a weak ordering is obtained. In real situations, sometimes not all items are ranked and we talk of partial rankings, when judges are asked to rank only a subset of the whole set of items, and incomplete rankings, when judges can freely choose to rank only some items. In order to obtain homogeneous groups of subjects with similar preferences, it is natural to measure the spread between rankings through dissimilarity or distance measures d between two rankings, a non-negative value ranging in $0-Dmax$. In this sense, a consensus is defined as the ranking that is closest (i.e. with the minimum distance) to the whole set of preferences. Another possible way for measuring (dis)-agreement between rankings is in terms of a correlation coefficient: rankings in full agreement are assigned a correlation of $+1$, those in full disagreement are assigned a correlation of -1, and all others lie in between. Kumar and Vassilvitskii (2010) introduced two essential aspects for many applications involving distances between rankings: positional weights and element weights. In brief, i) the importance given to swapping elements near the head of a ranking could be higher than the importance attributed to elements belonging to the tail of the list or ii) changing the ranking of important items should be less penalized than changing the ranking of important ones The first aspect has been widely addressed in literature. Recently Plaia et al (2019a, 2019b) proposed a new position weighted correlation coefficient for linear and weak orderings. Differently, the aspect of element weights is less explored. As Kumar and Vassilvitskii (2010) say, item weights are important, for example, when swapping similar elements should be less penalized than swapping dissimilar elements. To illustrate the idea, when ranking politicians, we should take into account if candidates belong to the same or to different parties: if two rankings differ for the position of two candidates from the same party, it should be reasonable to assume that the distance between these two rankings must be lower than the one between two rankings that differ for the position of candidates that belong to different parties.

In order to take this aspect into account, in this paper we introduce the item weighted Kemeny distance.

2 Distances for ranking data: item weighting

In order to get homogeneous groups of subjects having similar preferences, it is natural to measure the spread between rankings through dissimilarity or distance measures among them. Among the metrics proposed in the literature for computing distances between rankings, we choose to consider the Kemeny distance (Kemeny and Snell, 1962) that, with reference to two rankings a and b, is a city-block distance defined as:

$$K(a,b)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}|a_{ij}-b_{ij}|. \tag{1}$$

a_{ij} and b_{ij} are the generic elements of the $m\times m$ score matrices associated to a and b, respectively, assuming a value equal of 1 if item i is preferred to item j, -1 if item j is preferred to item i and 0 if the two items are tied or if $i = j$.

The choice of the Kemeny's axiomatic framework (Kemeny and Snell 1962) is justified beacuse we consider the possibility of ties, thus the geometrical space of preference rankings is the generalized permutation polytope (D'Ambrosio and Heiser, 2016), for which the natural distance measure is the Kemeny distance.

In order to consider the possibility that the items are not equally important, we introduce a vector of weights $w = (w_1, w_2, \ldots, w_m)$, with $w_i \geq 0$, whose elements represent the weight (i.e. the importance) we give to each item. The item weighted Kemeny distance is defined as:

$$d_K^{iw}(a,b) = \sum_{i<j}^{m} \frac{w_i + w_j}{2} |a_{ij} - b_{ij}| \tag{2}$$

It is easily demonstrated that the maximum value of eq. (2) is $d_{max} = (m-1)\sum_{i=1}^{m} w_i$.

3 Distance properties

We will prove that eq (2) meet the usual properties of a distance function; given two rankings, a and b:

1. Non negativity: $d(a,b) \geq 0$ and equality hold if and only if $a = b$ limited to items corresponding to weights $w_i > 0$,
2. Symmetry: $d(a,b) = d(b,a)$,
3. Triangle inequality: $d(a,b) \leq d(a,c) + d(c,b)$ if b is between a and c (in case of a metric).

Moreover, a desirable property of any distance is its invariance toward a renumbering of the elements (the so-called label invariance, right invariance or equivariance).

Proof

1. Eq (2) is a sum of absolute values, hence it cannot be negative. If $a \neq b$ at least for the items with corresponding weights greater than 0, then the distance is positive. At the same time, if $a = b$ at least for the items with corresponding weights greater than 0, then the distance is null.
2. Symmetry occurs since $|a_{ij} - b_{ij}| = |b_{ij} - a_{ij}|$.

3. Given i and j, the triangular inequality reduces to:

$$\frac{w_i + w_j}{2}|a_{ij} - b_{ij}| \leq \frac{w_i + w_j}{2}|a_{ij} - c_{ij}| + \frac{w_i + w_j}{2}|c_{ij} - b_{ij}|$$

   and dividing by $\frac{w_i+w_j}{2}$ we return to the known Kemeny distance that, as demonstrated by Kemeny and Snell, meets the inequality if c is between a and b.

Finally, since a permutation of items simply rearranges the rows and columns of the score matrix, if a' results from a by a permutation, and b' results from b by the same permutation, then $d_K^{iw}(a', b')$ is the sum of the same terms as $d_K^{iw}(a, b)$, with the terms occurring in a different order: hence the label invariance holds.

References

COOK, W. D. 2006. Distance based and ad hoc consensus models in ordinal preference ranking. *European Journal of Operational Research 369–385*, **172**, 369–385.

D'AMBROSIO, A., HEISER W.J. 2016. A recursive partitioning method for the prediction of preference rankings based upon Kemeny distances. *Psychometrika*, **81**(3), 774–794.

EDMOND, E. J., & MASON, D. W. 2002. A new rank correlation coefficient with application to the concensus ranking problem. *Journal of Multicriteria decision analysis*.

GARCIA-LAPRESTA, J. L., & PÉREZ-ROMÁN, D. 2010. Consensus measures generated by weighted Kemeny distances on weak orders. *In: Procceedings of the 10th International Conference on Intelligent Systems Design and Applications, Cairo*.

KEMENY, J. G., & SNELL, J. L. 1962. *Preference rankings an axiomatic approach.* MIT Press.

KUMAR, R., & VASSILVITSKII, S. 2010. Generalized Distances Between Rankings. *Pages 571–580 of: Proceedings of the 19th International Conference on World Wide Web*. WWW '10. New York, NY, USA: ACM.

PLAIA, A., BUSCEMI, S., & SCIANDRA, M. 2019a. A new position weight correlation coefficient for consensus ranking process without ties. *in press*.

PLAIA, A., BUSCEMI, S., & SCIANDRA, M. 2019b. Consensus among preference rankings: a new weighted correlation coefficient for linear and weak orderings. *submitted*.

A FAST AND EFFICIENT MODAL EM ALGORITHM FOR GAUSSIAN MIXTURES

Luca Scrucca[1]

[1] Department of Economics, Università degli Studi di Perugia, (e-mail: luca.scrucca@unipg.it)

ABSTRACT: In modal clustering, clusters are defined in terms of local maxima of the underlying probability density function. Therefore, clusters are closely related to certain regions around the density modes. An estimate of the density function can be obtained either nonparametrically or by using finite mixture models. A Modal EM algorithm can be used to identify the local maxima of a density function, so that every cluster corresponds to a bump of the density. In this contribution, we propose a fast and efficient Modal EM algorithm when the density function is estimated through Gaussian mixture models with parsimonious covariance structures.

KEYWORDS: Modal EM algorithm, model-based density estimation, finite mixture of Gaussians, density modes.

1 Introduction

In *model-based clustering* each component of a mixture distribution is associated to a cluster (McLachlan & Peel, 2000; Fraley & Raftery, 2002). Thus, observations are allocated to the cluster with maximal density among the components. *Modal clustering* is another density-based approach to clustering that directly looks for "[...] regions of high density separated from other such regions by regions of low density" (Hartigan, 1975, p. 205). Several mode-seeking algorithms have been proposed in the literature, such as the mean-shift algorithm (for a recent review see Menardi, 2016).

Let $f(\boldsymbol{x}) = \sum_{k=1}^{G} \pi_k f_k(\boldsymbol{x})$ be the mixture density for $\boldsymbol{x} \in \mathbb{R}^d$, where π_k is the mixing probability of component k with density function $f_k(\boldsymbol{x})$. Modal EM (MEM) is an iterative algorithm that aims to identify the local maxima of a density function (Li *et al.* , 2007). Given an initial starting point $\boldsymbol{x}^{(0)}$, the following steps are iteratively executed until a stopping criterion is met:

$$\text{E-step:} \quad p_k = \pi_k f_k(\boldsymbol{x}^{(t)})/f(\boldsymbol{x}) \quad \text{for } k = 1,\dots,G;$$

$$\text{M-step:} \quad \boldsymbol{x}^{(t+1)} = \arg\max_{\boldsymbol{x}} \sum_{k=1}^{G} p_k \log f_k(\boldsymbol{x})$$

The objective function in the M-step has a unique maximum if the $f_k(\boldsymbol{x})$ are Gaussian densities (Li *et al.* , 2007). Furthermore, a closed-form solution is only available in case of GMMs with common covariance matrix, and numerical procedures are required for the M-step if the covariance matrices are different across components.

In this contribution we propose a fast and efficient MEM algorithm for densities estimated by finite mixture of multivariate Gaussians having any of the parsimonious covariance structures available in the `mclust` R package (Scrucca *et al.* , 2016).

2 Modal EM algorithm for Gaussian mixtures

Assume that the components of the mixture are multivariate Gaussians with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$, i.e. $f_k(\boldsymbol{x}) \equiv \phi(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Furthermore, assume that the mixing proportion π_k, the mean vector $\boldsymbol{\mu}_k$, and the covariance matrix $\boldsymbol{\Sigma}_k$ are given (either estimated or known) for all $k = 1, \dots, G$. Thus, the mixture density for any data point $\boldsymbol{x}_i$ can be written as

$$f(\boldsymbol{x}_i) = \sum_{k=1}^{G} \pi_k \phi(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

The MEM algorithm starts with $t = 0$ and initial points $\boldsymbol{x}_i^{(0)} = \boldsymbol{x}_i$, for $i = 1, \dots, n$. At iteration t, MEM performs the following steps:

- Set $t = t + 1$.
- E-step: update the posterior conditional probability of the current data point x_i to belong to the kth mixture component:

  $$z_{ik}^{(t)} = \frac{\pi_k \phi(\boldsymbol{x}_i^{(t-1)} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{g=1}^{G} \pi_g \phi(\boldsymbol{x}_i^{(t-1)} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)},$$

  for all $i = 1, \dots, n$, and $k = 1, \dots, G$.
- M-step: update the current value of $\boldsymbol{x}_i$ by solving the optimisation problem:

  $$\boldsymbol{x}_i^{(t)} = \arg\max_{\boldsymbol{x}_i} \sum_{k=1}^{G} z_{ik}^{(t)} \log \phi(\boldsymbol{x}_i^{(t-1)} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- Iterate the above steps until a stopping criterion is satisfied, for instance $\max(|(\boldsymbol{x}_i^{(t)} - \boldsymbol{x}_i^{(t-1)})/\boldsymbol{x}_i^{(t-1)}|) < \varepsilon$, or a pre-specified maximum number of iterations is reached.

By the ascending property of the MEM algorithm, at convergence the value of $\boldsymbol{x}_i^{(t)}$ is the mode associated with data point $\boldsymbol{x}_i$ (Li *et al.* , 2007, Appendix A).

To obtain a fast version of the above MEM algorithm, note that the objective function in the M-step can be written as

$$Q(\boldsymbol{x}_i) = \sum_{k=1}^{G} z_{ik} \log \phi(\boldsymbol{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

The gradient and Hessian of this function with respect to the observed vector $\boldsymbol{x}_i$ (assuming the mixture parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{G}$ as fixed) are, respectively,

$$\nabla Q(\boldsymbol{x}_i) = -\sum_{k=1}^{G} z_{ik} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k),$$

and

$$\nabla^2 Q(\boldsymbol{x}_i) = -\sum_{k=1}^{G} z_{ik} \boldsymbol{\Sigma}_k^{-1}.$$

Note that, because all covariance matrices $\boldsymbol{\Sigma}_k$ are positive definite and $z_{ik} > 0$ for all k and i, the Hessian is negative definite. Thus, maximisation of the Q-function can be pursued by equating the gradient to zero, and then solving for $\boldsymbol{x}_i$ we obtain

$$\boldsymbol{x}_i^* = \left(\sum_{k=1}^{G} z_{ik} \boldsymbol{\Sigma}_k^{-1} \right)^{-1} \sum_{k=1}^{G} z_{ik} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k.$$

Through an appropriate use of the Kronecker product, the solution of the optimisation problem in the M-step can be efficiently computed in a single step for all data points. However, since at each step of the MEM algorithm the conditional probabilities z_{ik} are updated, we would like to avoid large jumps that may miss the closest mode in the neighbourhood of $\boldsymbol{x}_i$. For this reason, in practice, we suggest to compute the update at iteration t as

$$\boldsymbol{x}_i^{(t)} = (1-\alpha)\boldsymbol{x}_i^{(t-1)} + \alpha \boldsymbol{x}_i^*,$$

where $\alpha = t/(t+1)$ is a parameter that controls the step size. At earlier iterations the updated value $\boldsymbol{x}_i^{(t)}$ is obtained as convex linear combination between the previous value and the proposed value, with the weight of the latter that converges to one as the number of iterations increase.

3 Example

The Old Faithful dataset provides the data on the duration (in minutes) and the waiting time (in minutes) for 272 eruptions of the Old Faithful geyser in the Yellowstone National Park. Figure 1 shows the MEM paths for some selected data points to their associated modes (left panel), and the final estimated modes (right panel).

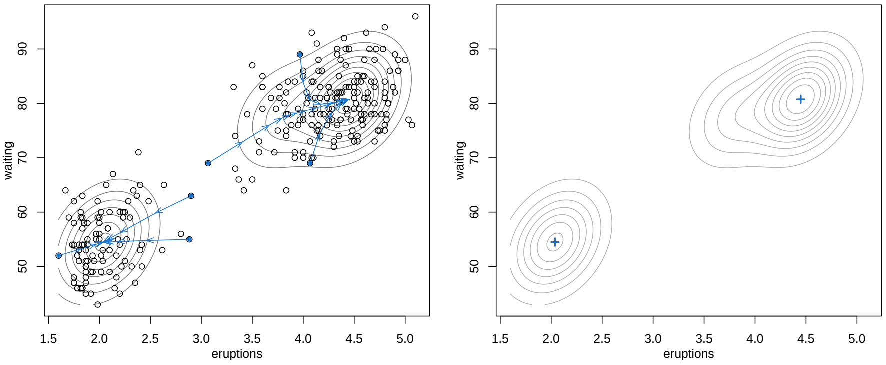


Figure 1. *Old Faithful data: MEM paths for some selected data points (left panel); mixture density contours with modes estimated by MEM algorithm (right panel).*

References

FRALEY, C., & RAFTERY, A. E. 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.

HARTIGAN, J. A. 1975. *Clustering Algorithms*. New York: John Wiley & Sons.

LI, J., RAY, S., & LINDSAY, B. G. 2007. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, **8**(Aug), 1687–1723.

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

MENARDI, G. 2016. A review on modal clustering. *International Statistical Review*, **84**(3), 413–433.

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), 205–233.

Probabilistic archetypal analysis

Sohan Seth[1]

[1] School of Informatics, University of Edinburgh,
(e-mail: sseth@staffmail.ed.ac.uk)

Abstract: Archetypal analysis represents a set of observations as convex combinations of pure patterns, or archetypes. It approximates the convex hull of the observations and assumes them to be real–valued. Probabilistic archetypal analysis accommodates other observation types such as integers, binary, and probability vectors. An appropriate visualization tool will be presented to summarize the archetypal analysis solution.

Keywords: archetypes, prototype, matrix factorization, majorization-minimization.

1 Introduction and motivation

Given a set of observations, archetypal analysis finds 'extreme' examples, i.e., archetypes that represent the observations well. Following the geometric formulation proposed by Cutler and Breiman (1994) this is achieved by approximating the convex hull of the set of observations with the archetypes such that the observations can be explained as convex combinations of the archetypes; an analogy being the colors red, green and blue that can explain the color spectrum as convex combinations of these archetypal colors. Archetypal analysis can be seen as a matrix factorization problem, and is closely related to other 'prototype' finding approaches, e.g., k-means clustering and topic modelling.

The standard approach of finding archetypes assumes that the observations are real valued, which, unfortunately, is not compatible with many practical situations. For example, one may ask to find archetypal responses for a set of binary questions, or archetypal document given a set of word count vectors of a set of documents. In this contribution, I will revisit archetypal analysis from the basic principles, and discuss a probabilistic framework that accommodates these scenarios, i.e., data types such as integers, categorical, and stochastic vector. This formulation is equivalent to performing archetypal analysis in the continuous parameter space of the probability distribution than in the discrete observation space, and for a range of exponential family distributions, such as Bernoulli, Poisson, and multinomial, the resulting optimization problem can be efficiently solved using majorization-minimization. For categorical variables, e.g., multiple-option questions, I will introduce an extension of this approach to a generative framework using Dirichlet prior over the mixing parameters for which the approximate posterior distribution can be efficiently inferred using variational Bayes', and associated hyperparameters help finding a suitable number of archetypes.

I will show the application of these formulations for finding archetypal tourists based on binary survey data, archetypal disaster-affected countries based on disaster count data, archetypal customers using German credit data, archetypal images using SUN image attribute data, and archetypal behaviour from Big Five personality data. I will also present an appropriate visualization tool to summarize the archetypal analysis solution, and address some recent developments in this area and some open questions.

References

CUTLER, A. & BREIMAN, L. 1994. Archetypal analysis. *Technometrics*, **36 (4)**, 338-347.

EUGSTER, M.J.A. & LEISCH, F. 2011. Weighted and robust archetypal analysis. *Computational Statistics and Data Analysis*, **55(3)**, 1215-1225.

LEE, D.D., & SEUNG, H.S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, **401(6755)**, 788-791.

SETH, S. & EUGSTER, M.J.A. 2016. Probabilistic archetypal analysis. *Machine Learning*, **102**, 85-113.

MULTILINEAR TESTS OF ASSOCIATION BETWEEN NETWORKS

Daniel K. Sewell[1]

[1] Department of Biostatistics, University of Iowa,
(e-mail: daniel-sewell@uiowa.edu)

ABSTRACT: It is often of interest to determine whether or not two networks measured on the same set of actors are associated with each other. Existing tests of association are all permutation tests. This paper proposes an alternative type of association test between two networks. Our test relies on a multilinear representation of the two networks to be compared and is motivated in large part by latent space network models. We demonstrate that the proposed test accurately controls the Type I error rate while maintaining power comparable to that of permutation methods.

KEYWORDS: latent space models, linear mixed models, singular value decomposition, social network analysis.

1 Introduction

One of the first steps of any data analysis is often to test whether or not two variables are associated. This problem becomes challenging when these variables correspond to relationships between a set of actors, i.e., we wish to compare two networks measured over the same set of actors. Existing methods rely on permutation tests, typically focusing on the quadratic assignment procedure (QAP) (Krackardt, 1987). This paper proposes a multilinear test of association (MLT) which provides a unique approach for testing the association between two networks measured on the same set of actors. The proposed approach can be seen as having been motivated from a geometrical point of view, or alternatively from a latent space modeling framework.

2 Approach

A weighted digraph is defined by a set of actors and a set of weighted directed edges, and may be represented as a square $n \times n$ adjacency matrix Y, where the i^{th} row j^{th} column entry Y_{ij} corresponds to the weight of the directed edge from actor i to actor j, and n is the number of actors in the network. In our context, we consider a second network over the same set of actors represented by the adjacency matrix X. The goal is to determine if Y and X are associated. The QAP is a nonparametric permutation test designed with this goal in mind. In contrast, our proposed approach is likelihood-based.

The singular value decomposition (SVD) of X is $U\Sigma V'$, where U and V are orthonormal matrices, and Σ is a diagonal matrix with the non-negative singular values $\underline{\sigma} = (\sigma_1, \ldots, \sigma_n)$ along the diagonal. This SVD of X can be viewed as a matrix transformation consisting of an aligning rotation (V), a scaling along the axes (Σ), and another rotation (U). If X and Y are indeed associated, then it should be reasonable to assume that their SVDs should be similar. We assume that Y represents a similar matrix transformation with an adjusted scaling step plus noise. That is,

$$Y = U\mathrm{Diag}(f(\sigma_1), f(\sigma_2), \ldots, f(\sigma_n))V' + E \tag{1}$$

for some function $f : [0, \infty) \mapsto [0, \infty)$ and $n \times n$ matrix of white noise E. In practice, of course, the form of f is unknown and must be estimated. If we assume that $f \in \mathcal{C}^d$, the space of d-differentiable continuous functions, then we may well approximate the function f using splines. Under this approximation we may write

$$f(\sigma) = \begin{pmatrix} f_0(\sigma) & f_1(\sigma) & \cdots & f_K(\sigma) \end{pmatrix} \underline{\gamma}, \tag{2}$$

where $\underline{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_K)'$ is a vector of unknown coefficients to be estimated.

This framework is motivated in large part not just by geometrical considerations but also on a large body of literature on latent space models. Hoff (2008, 2009) generalized these models with the "eigenmodel". Motivated by important theorems proved in Aldous (1981), his proposed approach estimates latent factors that take on the same form as SVD, i.e., $U\Sigma V'$.

The intuition of the multilinear model is that U and V represent feature vectors corresponding to the actors, and Σ determines how these features relate to each other to form edges (e.g., similarity in some features may promote edge formation, while similarity in others may discourage edge formation). By using the singular vectors of X, we are assuming that the features of the actors are constant regardless of what we are measuring, but how these factors relate to each other to form edges depends on which network type, Y or X, we are considering.

Importantly, combining (1) and (2) yields the linear model

$$Y_{ij} = \sum_{k=0}^{K} \gamma_k U_i \mathrm{Diag}(\mathbf{f}_0(\underline{\sigma})) V_j' + E_{ij} \tag{3}$$

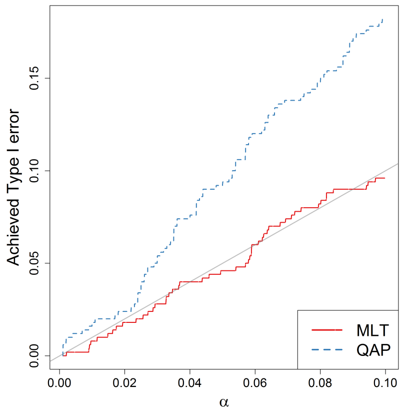


Figure 1a.

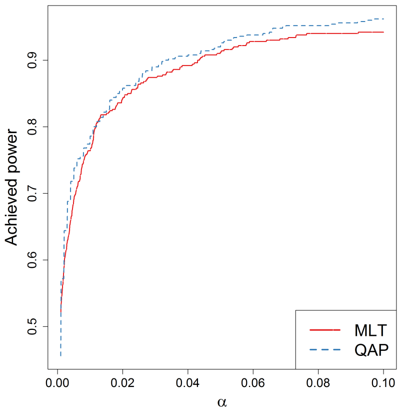


Figure 1b.

Following Hoff (2009), we assume that $(E_{ij}, E_{ji})' \overset{iid}{\sim} N(\underline{0}, \Omega)$. Estimation can then be made within a linear mixed model (LMM) framework. Specifically, we can use a likelihood ratio test under $H_0 : \underline{\gamma} = 0$ to examine if X and Y are related. If they are not, there is only a $100\alpha\%$ chance of rejecting this test.

3 Simulation Study

To evaluate the type I error and power of our approach we performed a simulation study. For each simulation we drew X from a standard normal and drew Y according to (3) when $\underline{\gamma} = \underline{0}$ as well as when $\underline{\gamma}$ was equal to $(0.4, -0.4, 0.8, 1.2)$ corresponding to a spline basis with 4 degrees of freedom. We ran 500 simulations under both scenarios and tested for an association using both MLT and QAP. Figure 1a shows the results for the context where $\underline{\gamma} = \underline{0}$. The horizontal axis corresponds to the level-α test, and the vertical axis corresponds to the achieved Type I error. It is evident that the QAP has greatly inflated Type I errors, while our proposed approach maintains this error at the correct level. Figure 1b shows the results for the context where $\underline{\gamma} \neq \underline{0}$. Again the horizontal axis corresponds to the level-α test and the vertical axis corresponds to the power. We see that although MLT controlled the Type I error much more effectively than QAP, the MLT obtained comparable power.

4 Correlates of War data analysis

We analyzed data from the Correlates of War project (Barbieri *et al.*, 2009) to determine if the network of formal alliances between countries were associated with the trade network between countries as measured by imports/exports. The p-value associated with the MLT was < 0.001. We thus conclude that there is in fact a relationship between economic ties and defense alliances.

5 Discussion

Our proposed multilinear test provides a novel alternative to permutation tests and has strong motivation based on the latent space network literature. The MLT, however, is limited to contexts in which the two networks under consideration can be thought to be expressed as functions of latent features of the network actors and hence would not be appropriate to test an association between, e.g., a friendship network and geospatial distances between the actors. Additionally, a failure to reject H_0 does not necessarily imply that there is no relationship between the two networks, as there may be a different specification between them; however, if the test is rejected one may feel confident that there is in fact a relationship between the two networks.

References

ALDOUS, D. J. 1981. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, **11**(4), 581–598.

BARBIERI, K., KESHK, O. M. G., & POLLINS, B. 2009. Trading Data: Evaluating our assumptions and coding rules. *Conflict Management and Peace Science*, **26**(5), 471–491.

HOFF, P. D. 2008. Modeling homophily and stochastic equivalence in symmetric relational data. *Pages 657–664 of:* PLATT, J. C., KOLLER, D., SINGER, Y., & ROWEIS, S. T. (eds), *Advances in Neural Information Processing Systems 20.* Curran Associates, Inc.

HOFF, P. D. 2009. Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, **15**(4), 261–272.

KRACKARDT, D. 1987. QAP partialling as a test of spuriousness. *Social Networks*, **9**(2), 171 – 186.

USE OF MULTI-STATE MODELS TO MAXIMISE INFORMATION IN PRESSURE ULCER PREVENTION TRIALS

Linda Sharples[1, 2], Isabelle Smith[2] and Jane Nixon[2]

[1] Department of Medical Statistics, London School of Hygiene and Tropical Medicine, (e-mail: linda.sharples@lshtm.ac.uk)

[2] Leeds Institute of Clinical Trials Research, Clinical Trials Research Unit, University of Leeds, (e-mail: i.l.smith@leeds.ac.uk, j.e.nixon@leeds.ac.uk)

ABSTRACT: Large, tightly controlled, randomised controlled trials are the most reliable method of establishing a causal effect between interventions and outcomes. The most common challenge for trials is the need to recruit patients from a number of healthcare providers and within the busy environment of routine health service delivery. As a result trials often exceed the planned period of recruitment and patients do not always remain in the trial for the full follow up period. This study aims to assess whether more efficient trials, with smaller sample size/greater power, can be designed based on multistate models that maximise the use of measurements taken during the trial. The effect of design features (frequency of measurement, intervals between measurements, duration of follow-up and misclassification) on sample size estimates are explored and estimands that may be of interest in this context are clarified. The methods are applied to pressure ulcer prevention trials in patients admitted to hospital, but the methods are general and can be applied to any conditions that are well represented by multi-state models.

KEYWORDS: clinical trials, multi-state models, efficient design.

1 Background

Long stay in hospital and poor mobility put people at risk of pressure ulcers (PU) at a number of areas of the body (buttocks, heels etc). Once developed PUs result in prolonged hospital stay, poor quality of life and significant costs.

PUs are classified on a 4 point ordered scale from 1-4. In RCTs skin assessment for onset or progression of PUs takes place at a number of fixed time points, resulting in hierarchical and longitudinal measurements of PU categories at up to 14 skin sites. Thus, each patient has 50-100 PU assessments during trial follow-up. The process is not observed continuously, resulting in panel data and interval censoring. Moreover, due to administrative and patient-related events, scheduled measurements may be missed or only partially completed. This results in observation times that are different for different patients and intervals between assessments may vary.

Often, the primary outcome for PU prevention trials is the time from randomisation to the first category 2 PU at any skin site, so that the 50-100 assessments per patient are reduced to a single outcome measurement. This outcome is inefficient in that it ignores the information from longitudinal measurements and multiple skin sites; it may also be biased due to interval censoring between observations and missed assessments. Thus sample sizes for PU prevention trials may be larger than necessary.

2 Aim

In this presentation we investigate the use of multi-state models in disease prevention trials, in order to provide less biased and more efficient estimates of treatment effects.

3 Methods and results

We show how to design a PU prevention trial and analyse resulting data. Multi-state models that incorporate longitudinal data on disease categories are developed. Assumptions that are required for different models, their implications and their validity in this context are presented, as are methods for estimation within this framework. Re-analysis of data from 1846 patients from the PRESSURE2 prevention trial compares differences between commonly used estimands (odds ratios and hazard ratios) and multi-state model outputs and demonstrates how fixed covariates (e.g. treatment group and stratification factors) can be incorporated into the analysis. For a range or realistic model parameters, efficiency of multi-state models compared to incidence of PU and time to PU outcomes is explored using simulation studies. The extent of improvement in power for multi-state models depends on duration of follow up and frequency of assessments, as well as sample size.

4 Conclusion

Given difficulties in recruiting to RCTs it is important to make best use of the rich data that accrue during trials. Reductions in sample size for PU trials may be possible if all available observations are included in the analysis, but this depends on the estimand of interest.

PARTIAL LEAST SQUARES FOR COMPOSITIONAL CANONICAL CORRELATION

Violetta Simonacci[1], Massimo Guarino[1] and Michele Gallo[1]

[1] Department of Human and Social Sciences, University of Naples L'Orientale,
(e-mail: vsimonacci@unior.it, mguarino@unior.it, mgallo@unior.it)

ABSTRACT: Compositional data are quantitative descriptions of the parts of some whole, conveying relative information. The relationship between two sets of compositional descriptors can be explored by use of Canonical Correlation analysis with a procedure based on Partial Least Squares (PLS). This method offers a way to deal with matrix singularity in an efficient fashion and presents the further advantage of being easy to interpret. In order to fully explore the potential of PLS for analyzing the relationships between two sets of compositions, the performances of the NIPALS, SIMPLS and Kernel algorithms are compared on simulated data.

KEYWORDS: Compositional data, log-ratio transformation, simpls, nipals, kernel.

1 Introduction

Canonical correlation analysis (CCA) is a method proposed by Hotelling (1936) for exploring the relationships between two groups of variables. Let us consider a set of variables $\mathbf{X}$ $(I \times J)$ with dispersion matrix Σ_X, a second set $\mathbf{Y}$ $(I \times K)$ with dispersion matrix Σ_Y and the covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$ denoted by Σ_{XY}. The main aim of CCA is to search for canonical variates $\mathbf{A} = \mathbf{aX}$ and $\mathbf{B} = \mathbf{bY}$ that have maximal correlation:

$$\underset{\mathrm{var}(\mathbf{A})=\mathrm{var}(\mathbf{B})=1}{\arg\max} \frac{\mathrm{cov}(\mathbf{A},\mathbf{B})^2}{\mathrm{var}(\mathbf{A})\mathrm{var}(\mathbf{B})} = \underset{\mathbf{a}^t\Sigma_X\mathbf{a}=\mathbf{b}^t\Sigma_Y\mathbf{b}=1}{\arg\max} \frac{\mathbf{a}^t\Sigma_{XY}\Sigma_{YX}\mathbf{b}}{(\mathbf{a}^t\Sigma_X\mathbf{a})(\mathbf{b}^t\Sigma_Y\mathbf{b})}, \qquad (1)$$

where $\mathbf{a}$ and $\mathbf{b}$ are linear combination vectors. If the dispersion matrices can be inverted, the solution of equation (1) is efficiently calculated by using singular value decomposition. When one or both sets of variables are compositions, however, Σ_X and /or Σ_Y are singular and each row will sum to 0, forcing at least one covariance term to be negative. Thus, classical CCA is not able to determine canonical variates and its correlations. In order to be able to deal with the purely multicollinear structure and the negative bias that characterize compositional data, in this work an approach based on log-ratio preprocessing and Partial Least Squares (PLS - Rayens, 2000) is proposed.

2 Theory

A compositional matrix $\mathbf{X}$ $(I \times J)$ has all non-negative elements and its row vectors present a biased covariance structure due to an implicit or explicit sum constraint, i.e. $x_{i1} + \cdots + x_{iJ} = \kappa$, where κ is a positive constant. This bounded covariance imposes a purely multicollinear structure to the data since the elements of a compositional vector are not linearly independent and thus Σ_X will be singular.

Geometrically, $\mathbf{X}$ describes I points bounded in a subspace of $\Re_+^{I \times J}$ known as simplex and defined as:

$$S^{I \times J} = \{(x_{i1}, \ldots, x_{iJ}) : x_{i1} \geq 0, \ldots, x_{iJ} \geq 0; x_{i1} + \cdots + x_{iJ} = \kappa; i = 1, \ldots, I\}.$$

The simplex is characterized by its own geometric rules, called Aitchison geometry, thus standard statistical methods designed to operate within a Euclidean framework cannot be applied without distortions (Aitchison, 1986; Pawlowsky-Glahn *et al.* , 2015).

Compositional vectors can, however, be converted into Euclidean space coordinates by using log-ratio transformations: pairwise, centered, additive or isometric. For the purpose of this contribution we will only be referring to centered log-ratio (clr) coordinates which can be expressed as:

$$[log(x_{i1}/g(\mathbf{x}_i)), \ldots, log(x_{iJ}/g(\mathbf{x}_i))],$$

where $g(\mathbf{x}_i)$ is the geometric mean of the parts of the composition $\mathbf{x}_i$.

After performing this preprocessing step, standard statistical tools can be applied as long as results are interpreted in compositional terms. It is important to note that clr-coordinates by providing an $S^{I \times J}$ to $\Re^{I \times J}$ projection, do not remove the collinearity problem.

Hinkle (1995) and Wang *et al.* , 2010 examined the problems that occur when one performs a PLS analysis on compositional data and suggested the use of clr preprocessing. Gallo (2010) proposed the application of this approach to discriminate compositions.

Given that $f + 1 \leq \min(\text{rank}(\mathbf{X}), \text{rank}(\mathbf{Y}))$, PLS can be defined as:

$$\underset{\mathbf{u}^t\mathbf{U}=\mathbf{0}^t;\mathbf{u}\mathbf{1}=0}{\arg\max} \left\{ \frac{\text{cov}(\mathbf{Xu}, \mathbf{Yv})^2)}{(\mathbf{u}^t\mathbf{u})(\mathbf{v}^t\mathbf{v})} \right\} = \{\mathbf{u}_{f+1}, \mathbf{v}_{f+1}\}, \tag{2}$$

where $\mathbf{u}_{f+1}$ is the eigenvector of $\Sigma_{XY}\Sigma_{YX}$ corresponding to the $(f+1)$-th largest eigenvalue, and $\mathbf{v}_{f+1} = \Sigma YX\mathbf{u}_{f+1}$. The usual constraints of PLS $\mathbf{U} =$

$[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_f]$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_f]$ define the directions constrained to be orthogonal in the **X** -space and in the **Y**-space respectively; while the additional constraint **u1**=0 can be viewed as an orthogonality constraint with respect to a redefined inner product.

This is not the only definition of PLS, however, it has the intuitive advantage of being cosmetically very similar to the CCA. In fact, it is possible to note that $\text{cov}(\mathbf{Xa}, \mathbf{Yb})^2 = \text{var}(\mathbf{Xa})\text{corr}(\mathbf{Xa}, \mathbf{Yb})^2\text{var}(\mathbf{Yb})$.

From equation (1) the $(f+1)$-th pair of canonical variates is given by $A_{f+1} = \mathbf{X}a_{f+1} = \dot{u}_{f+1}\Sigma_X\mathbf{X}$ and $B_{f+1} = \mathbf{Y}b_{f+1} = \dot{v}_{f+1}\Sigma_Y\mathbf{Y}$, where $\dot{u}_{f+1}$ and $\dot{v}_{f+1}$ are the $(f+1)$-th eigenvector of $\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\Sigma_X^{-1/2}$ and $\Sigma_Y^{-1/2}\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\Sigma_Y^{-1/2}$ respectively. While the eigenvalue of $\Sigma_X^{-1/2}\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX}\Sigma_X^{-1/2}$ is the squared correlation between the canonical variates $\mathbf{A}_{f+1}$, $\mathbf{B}_{f+1}$.

In other words, PLS can be interpreted as a penalized CCA, with basically a PCA in the **X** space and a PCA in the **Y** space providing the penalties.

Now, if **X** and/or **Y** are clr it is clear that it is not possible to calculate the canonical variates and the canonical correlation by equation (1), but it is possible to find them with an algorithmic solution of equation (2) just replace the original compositional data with their corresponding centered log-ratio.

3 Conclusion

Two alternative methods have been proposed in the literature to use CCA for studying the relationships between two sets of compositional variables. The first one is based on the use of log-ratio transformations which do not inherit pure collinearity, namely the isometric log-ratio (Filzmoser & Hron, 2009).

In a second approach, correlation matrices are inverted by use of the generalized inverse in order to handle perfect multicollinearity so that the data can be more intuitively transformed in centered or pairwise log-ratios (Graelman *et al.* , 2017). This approach allows to deal with singularity while avoiding the loss of simplicity and interpretability which using other transformation may entail. However, it is very time consuming when the matrix is large.

Following Gallo (2010), we suggest a third strategy based on the use of PLS for performing CCA. Thanks to its algorithmic nature, this procedure offers a way to deal with matrix singularity without compromising interpretability or computational efficiency. PLS can be estimated with different algorithms, i.e. NIPALS, SIMPLS and Kernel, thus for a complete assessment of the proposed methodology the performance of these different procedures is compared on artificial data. In this manner, based on the characteristics and dimensionality of compositional data sets, it will be possible to identify and select the algorithm which is faster and easier to interpret.

In the simulation design, score and loading matrices for both **X** and **Y** are artificially created. In particular, matrices are randomly generated from a uniform distribution then a given level of canonical correlation is imposed among the score matrix of **X** and the one of **Y**. Loading matrices are adjusted to have columns summing to 0 in order to ensure a compositional structure. Afterwards, **X** and **Y** are reconstructed and mild homoscedastic noise is added. Different parameters (correlation, noise, dimensionality) are considered throughout the study. Performance is assessed on the basis of accurate estimation of correlation and efficiency. All calculations are carried out using R language, version 3.5.0, processor 2, 3 GHz Intel Core i7.

References

.

AITCHISON, J. 1986. *The statistical analysis of compositional data.* Boca Raton: Chapman & Hall.

FILZMOSER, P., & HRON, K. 2009. Correlation analysis for compositional data. *Mathematical Geosciences*, **41**, 905.

GALLO, M. 2010. Discriminant partial least squares analysis on compositional data. *Statistical Modelling*, **10**, 41–56.

GRAELMAN, J., PAWLOWSKY-GLAHN, V., EGOZCUE, JJ., & A., BUCCIANTI. 2017. Compositional canonical correlation analysis. *bioRxiv*, **144584**, 1–39.

HINKLE, J., & RAYENS, W. 1995. Partial least squares and compositional data: problems and alternatives. *Chemometrics and Intelligent Laboratory Systems*, **30**, 159–172.

HOTELLING, H. 1936. Relations between two sets of variates. *Biometrika*, **28**, 321–377.

PAWLOWSKY-GLAHN, V., EGOZCUE, JJ., & TOLOSANA-DELGADO, R. 2015. *Modeling and analysis of compositional data.* Hoboken: Wiley.

RAYENS, W. 2000. *The art of maximizing covariance.* University of Kentucky Technical: Report 383.

WANG, H., MENG, J., & TENENHAUS, M. 2010. *Regression Modelling Analysis on Compositional Data.* Berlin: In: Esposito Vinzi V., Chin W., Henseler J., Wang H. (eds) Handbook of Partial Least Squares. Springer Handbooks of Computational Statistics.

DYNAMIC MODELLING OF PRICE EXPECTATIONS

Rosaria Simone[1], Domenico Piccolo[1] and Marcella Corduas[1]

[1] Department of Political Sciences, Universita` degli Studi di Napoli Federico II, (e-mail: rosaria.simone@unina.it, domenico.piccolo@unina.it, marcella.corduas@unina.it)

ABSTRACT: Evaluation surveys are often repeated over time in order to check for trends in subjects' behaviors and opinions. The paper proposes a dynamic model for the serial correlation of ratings' intrinsic components, which is discussed on the basis of time series of price expectations in Italy collected within a survey organized by ISTAT.

KEYWORDS: Rating data, Seemingly unrelated regressions, Price Expectation, Mixture models.

1 Motivation and methods

Assume that, for each time $t = 1,\dots,n$, N_t subjects are asked to rate their perception on a given topic on a scale with $m > 3$ ordered categories. For each t and each subject $i = 1,\dots,N_t$, the response process R_{it} can be described in terms of a suitable family of probability distributions with parameters $\boldsymbol{\theta}_t$. Then, modelling $\boldsymbol{\theta}_t$ over time provides a parametric analysis of the dynamic evolution of the rating process.

In case of the expectation of price levels, literature and data agree in the existence of an "excess of frequencies" in some categories. Thus, if R_{it} is the response of the i-th subject interviewed at time t, a stochastic process based on CUB models with *shelter* effect can be introduced:

$$Pr(R_{it} = r \mid \boldsymbol{\theta}_t) = \pi_t^{(1)} \binom{m-1}{r-1} \xi_t^{m-r} (1-\xi_t)^{r-1} + \pi_t^{(2)} \frac{1}{m} + (1-\pi_t^{(1)} - \pi_t^{(2)}) D_r^{(c)}, \tag{1}$$

$r = 1,\dots,m$, where CUB stands for **C**ombination of a discrete **U**niform and shifted **B**inomial (see Piccolo & Simone, 2019, for an updated discussion of CUB models and Proietti, 2019, for the original proposal of dynamic CUB modelling). Trajectories of $\boldsymbol{\theta}_t = (\pi_t^{(2)}, \xi_t)'$, for varying t, will give dynamic measures of the weight of uncertainty/heterogeneity of the responses and of the degree of feeling towards the item, respectively. For short, we will set $\pi_t = \pi_t^{(2)}$; when the shelter effect is not significant, the baseline CUB model is considered and π_t is the weight of the Uniform distribution. In absence of covariates, the sufficient statistics to infer on the parameter vector (π_t, ξ_t), for each t, are just the (absolute) frequencies $(n_{1,t}, \ldots, n_{m-1,t})'$, $t = 1, \ldots, n$, where n_{jt} is the number of interviewees who selected the j-th category at time t, and $n_{m,t} = n - \sum_{j=1}^{m-1} n_{j,t}$.

Since both π_t and ξ_t are compelled to the unit range, a more suitable model may be defined in terms of the logit transformations:

$$x_t = \log\left(\frac{\pi_t}{1-\pi_t}\right); \qquad y_t = \log\left(\frac{\xi_t}{1-\xi_t}\right), \qquad t = 1, \ldots, n. \tag{2}$$

Then, after that data have been smoothed by replacing x_t with the average of x_{t-1},x_t, and x_{t+1} (similarly for y_t), a bivariate model is introduced. If time dependence is limited to lag $p \geq 1$, say, the specification of a dynamic model can be based on seemingly unrelated regressions (SUR: see Greene, 2008):

$$\begin{cases} x_t &= \alpha_x + \sum\limits_{l=1}^{p} \beta_l x_{t-l} + a_t; \\ y_t &= \alpha_y + \sum\limits_{l=1}^{p} \gamma_l y_{t-l} + b_t, \end{cases} \tag{3}$$

for $t = p+1, \ldots, n$. Here, $\boldsymbol{U}_t = (a_t, b_t)'$ is a bivariate white noise (WN) process with zero mean vector and variance-covariance matrix $\boldsymbol{V}$. Thus, unspecified interactions between x_t and y_t are accounted by the residuals' correlation.

Summarizing, the proposed modelling foresees a two-step estimation procedure: first, ML estimates $(\hat{\pi}_t, \hat{\xi}_t)$ are computed at each time point by means of the devoted EM algorithm, then the model (3) is fitted on their logit values (2).

2 Price expectation

As part of a EU project tailored to measure consumers' opinions towards aspects of economic conditions, here the focus will be on monthly expectations about price levels for the next 12 months in Italy, as collected by ISTAT (www.istat.it). For each month, about 2000 sample observations are available as time series of frequencies for each response category, ranging from $R = 1$ ('will not at all increase') up to $R = 5$ ('it will definitely increase'). Starting from January 1994, three different periods are considered determined by the introduction of Euro (2002) and the beginning of the 2008 economic crisis.

Since data confirm substantial correlation between WN components, SUR models are more efficient than separate OLS methods for estimating the models in (3): the R package `systemfit` (Henningsen & Hamann, 2007) has been considered. Results for lag $p \leq 2$ are hereafter reported with residual variance and correlation estimates, and McElroy R^2 as a global fitting measure:

- January 1994-December 2001:

$$\begin{cases} x_t &= \underset{(0.090)}{-0.348} + \underset{(0.086)}{1.170} x_{t-1} - \underset{(0.086)}{0.352} x_{t-2} + a_t\,; \\ y_t &= \underset{(0.073)}{-0.271} + \underset{(0.090)}{0.973} y_{t-1} - \underset{(0.090)}{0.206} y_{t-2} + b_t\,. \end{cases} \tag{4}$$

$$\hat{\sigma}_a^2 = 0.068\,; \quad \hat{\sigma}_b^2 = 0.019\,; \quad \hat{\rho}_{ab} = 0.582\,; \qquad R^2 = 0.738\,.$$

- January 2002-December 2007:

$$\begin{cases} x_t &= \underset{(0.183)}{-0.473} + \underset{(0.105)}{1.380} x_{t-1} - \underset{(0.107)}{0.496} x_{t-2} + a_t\,; \\ y_t &= \underset{(0.019)}{-0.010} + \underset{(0.041)}{0.910} y_{t-1} + b_t\,. \end{cases} \tag{5}$$

$$\hat{\sigma}_a^2 = 0.386\,; \quad \hat{\sigma}_b^2 = 0.021\,; \quad \hat{\rho}_{ab} = 0.226\,; \qquad R^2 = 0.879\,.$$

- January 2008-January 2019:

$$\begin{cases} x_t &= \underset{(0.046)}{-0.207} + \underset{(0.066)}{1.211} x_{t-1} - \underset{(0.063)}{0.374} x_{t-2} + a_t\,; \\ y_t &= \underset{(0.029)}{-0.045} + \underset{(0.075)}{1.129} y_{t-1} - \underset{(0.075)}{0.274} y_{t-2} + b_t\,. \end{cases} \tag{6}$$

$$\hat{\sigma}_a^2 = 0.144\,; \quad \hat{\sigma}_b^2 = 0.096\,; \quad \hat{\rho}_{ab} = 0.675\,; \qquad R^2 = 0.857\,.$$

In addition, Table 1 compares the variance-covariance matrices of the bivariate processes (x_t, y_t) and $(\hat{a}_t, \hat{b}_t)$ in terms of determinant and trace: a substantial reduction in variability is implied by the estimated models.

We mention that the spectral features of each estimated model emphasize a stochastic periodicity in the logits of π_t series (about 3 years, as a consequence of the complex roots of the characteristic equation for all the periods), whereas the spectra of the logits of ξ_t series show a stochastic trend (confirmed by real roots near the unit circle in all the periods).

Table 1. *Comparison of variability of stochastic processes*

Periods	Bivariate Process (x_t, y_t)		Bivariate Process $(\hat{a}_t, \hat{b}_t)$	
	Determinant	*Trace*	*Determinant*	*Trace*
Jan.1994-Dec.2001	0.00977	0.311	0.00085	0.087
Jan.2002-Dec.2007	0.55946	3.233	0.00804	0.407
Jan.2008-Dec.2001	0.39916	1.767	0.01324	0.240

3 Concluding remarks

Dynamic modelling the intrinsic components of the rating process, as postulated by CUB , can support predictive analysis and understanding of subjects' behaviours. Further, temporal interdependencies among parameters characterizing the rating distributions can be analysed also by means of VAR models. From the computational point of view, we acknowledge that the estimation procedure could be refined by pursuing a simultaneous estimation of the mixture parameters (π_t, ξ_t) and regression parameters for the SUR model (3): future developments will address this task.

References

GREENE, W.H. 2008. *Econometric analysis*. New Jersey: Pearson Prentice Hall.

HENNINGSEN, A., & HAMANN, J.D. 2007. systemfit: A Package for Estimating Systems of Simultaneous Equations in R. *Journal of Statistical Software*, **23**(4), 1–40.

PICCOLO, D., & SIMONE, R. 2019. The class of CUB models: Statistical foundations, inferential issues and empirical evidence. *Statistical Methods and Applications*. doi:10.1007/s10260-019-00461-1.

PROIETTI, T. 2019. Discussion of: The class of CUB models: Statistical foundations, inferential issues and empirical evidence, by Piccolo D. & Simone R. *Statistical Methods and Applications*. (forthcoming).

TOWARDS AXIOMS FOR HIERARCHICAL CLUSTERING OF MEASURES

Philipp Thomann[1], Ingo Steinwart[1] and Nico Schmid[1]

[1] Institute for Stochastics and Applications, University of Stuttgart,
(e-mail: philipp.thomann@mathematik.uni-stuttgart.de,
ingo.steinwart@mathematik.uni-stuttgart.de,
nico.schmid@mathematik.uni-stuttgart.de)

ABSTRACT: Clustering is an important area of machine learning. Yet there is no common precise notion of its general objectives. Establishing axioms for clustering hence is a first step towards a mathematical theory of clustering. Furthermore, clustering has to be understood not only for finite samples but also for entire probability distributions. We propose a novel approach to axiomatic clustering. As in any axiomatic system there is a user choice: On one hand the clusters for some elementary measures are stipulated by the user. On the other hand a topological separation relation has to be specified. Then two additivity and one continuity axiom are shown to yield a unique notion of clustering for a large set of distributions. Note that this is done without the need of any notion of metric, similarity or dissimilarity and it is completely parameter-free.

KEYWORDS: clustering axioms, hierarchical clustering, separation.

INFLUENCE OF OUTLIERS ON CLUSTER CORRESPONDENCE ANALYSIS

Michel van de Velden[1], Alfonso Iodice D'Enza[2] and Lisa Schut[1]

[1] Department of Econometrics, Erasmus University of Rotterdam,
(e-mail: vandevelden@ese.eur.nl)

[2] Department of Political Sciences, Università degli Studi di Napoli Federico II,
(e-mail: iodicede@unina.it)

ABSTRACT: This paper focuses on determining the influence of outliers on a joint dimension reduction and clustering method for categorical data, namely Cluster Correspondence Analysis (CCA). Joint methods, such as CCA, solutions consist of both a cluster membership vector and a set of low dimensional scores for observations and attributes. We evaluate the impact of outliers on the identification of the cluster structure. As a benchmark, we use the tandem approach, which is a sequential application of multiple correspondence analysis followed by K-means clustering. The appraisal is based on synthetic data and outliers generated using an evolutionary algorithm that provides data with a user-defined cluster structure.

KEYWORDS: clustering, dimension reduction, outliers.

1 Introduction

Clustering is an unsupervised learning method to allocate observations to groups (clusters) that are internally homogeneous with respect to the observed set of attributes. A set of popular clustering methods are distance-based, meaning that, upon defining an appropriate distance measure, observations closer to each other are assigned to a same cluster. The choice of the distance measure is crucial and it depends on the nature of the attributes (continuous, categorical or mixed). When the observations are described by several attributes, the identification of the underlying cluster structure becomes increasingly difficult. This is due to both the possible presence of noise (attributes that do not discriminate among clusters) and to the so-called course of dimensionality: pairwise-distances between observations tend to converge as the considered number of attributes increases. Furthermore, the computational burden increases with the dimensionality of the data. To overcome these issues, practitioners often perform a dimension reduction prior to the clustering. In particular, principal components methods are used to define a reduced set of linear combinations (components) of the starting attributes, then the observations are clustered using components-based distances. Such sequential approach is referred to as tandem approach. While tandem approach may provide viable solutions, it may also fail, as the dimension reduction can miss or even hide the cluster structure. This cluster-masking problem has been pointed out in the literature (see, e.g., Vichi & Kiers, 2001), and it depends on the fact that the target function of dimension reduction is optimised irrespective to the following clustering step. A class of methods have been proposed in the literature that seek for an optimal solution for both the dimension reduction and clustering steps. Joint dimension reduction and clustering methods have been proposed for continuous (Vichi & Kiers, 2001; De Soete & Carroll, 1994), categorical (van de Velden *et al.* , 2017; Hwang *et al.* , 2006) and mixed data sets (Vichi *et al.* , 2019; van de Velden *et al.* , 2019).

Outliers, also known as anomalies, are often present in observed data. They may disrupt the underlying structure of data and can have adverse effects on the quality of analysis if ignored (Aggarwal, 2015). In this paper we study the influence of outliers on cluster correspondence analysis (CCA, van de Velden *et al.* , 2017) solutions in comparison with tandem analysis. In Section 2 we briefly define CCA whereas in Section 3 we describe the simulation setup and report the main results.

2 Cluster Correspondence Analysis

Consider a set of n observations described by p categorical variables, each with q_j categories, $j = 1, \ldots, p$. The corresponding indicator matrix $\mathbf{Z}$ is $n \times Q$, with $Q = \sum_{j=1}^{p} q_j$. The cluster membership can also be represented by a $n \times K$ indicator matrix $\mathbf{Z}_k$, where K is the user defined number of clusters. Furthermore, let $\mathbf{D}_z = \mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ and $\mathbf{D}_k = \mathbf{Z}_k^{\mathsf{T}}\mathbf{Z}_k$ be diagonal matrices with $\mathbf{Z}$ and $\mathbf{Z}_k$ column margins, respectively, and let $\mathbf{B}$ the $Q \times d$ matrix containing the d considered components (weights of the linear combinations). The objective function of CCA is, therefore,

$$\max_{\mathbf{Z}_k, \mathbf{B}^*} \quad \phi(\mathbf{Z}_k, \mathbf{B}^*) = \frac{1}{p} trace\left(\mathbf{B}^{*\mathsf{T}}\mathbf{D}_z^{-1/2}\mathbf{Z}^{\mathsf{T}}\mathbf{M}\mathbf{Z}_k\mathbf{D}_k^{-1}\mathbf{Z}_k^{\mathsf{T}}\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-1/2}\mathbf{B}^*\right)$$

$$s.t. \mathbf{B}^{*\mathsf{T}}\mathbf{B}^* = \mathbf{I}_d,$$

where $\mathbf{M}$ is a cantering operator and $\mathbf{B}^* = \frac{1}{\sqrt{np}}\mathbf{D}^{1/2}\mathbf{B}$.

The solution is found by iterating two steps. In the first step $\mathbf{B}^*$ is obtained, for fixed $\mathbf{Z}_k$ (initialised via random allocation), using the following eigenvalue

Table 1 *Averages over 50 trials of ARI of the solutions with and without outliers. The entries denoted in bold are significantly smaller than the ARI of the cluster formation with 0 outliers at 5 % confidence level. The standard deviations are provided in the parentheses below the averages.* .

	Association Strength 0.50				Association Strength 0.70			
	Noise		No Noise		Noise		No Noise	
No. Outliers	CCA	Tandem	CCA	Tandem	CCA	Tandem	CCA	Tandem
0	0.82 (0.02)	0.48 (0.03)	0.88 (0.01)	0.89 (0.01)	0.95 (0.02)	0.92 (0.02)	0.95 (0.02)	0.98 (0.01)
10	0.77 (0.03)	**0.39** (0.03)	0.84 (0.02)	**0.62** (0.03)	0.93 (0.02)	0.90 (0.02)	0.92 (0.02)	0.94 (0.02)
20	0.78 (0.02)	**0.39** (0.03)	**0.81** (0.02)	**0.63** (0.03)	0.94 (0.02)	0.87 (0.03)	0.90 (0.02)	0.96 (0.02)
30	0.79 (0.02)	**0.37** (0.03)	**0.79** (0.02)	**0.59** (0.03)	0.94 (0.02)	0.87 (0.02)	0.92 (0.02)	0.95 (0.02)

decomposition

$$\frac{1}{p}\mathbf{D}_z^{-1/2}\mathbf{Z}^{\mathsf{T}}\mathbf{M}\mathbf{Z}_k\mathbf{D}_k^{-1}\mathbf{Z}_k^{\mathsf{T}}\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-1/2} = \mathbf{B}^{*\mathsf{T}}\Lambda\mathbf{B}^{*}.$$

In the second step, $\mathbf{Z}_k$ is obtained, for fixed $\mathbf{B}^*$, by applying a K-means clustering procedure on the observations low dimensional scores

$$\mathbf{Y} = \sqrt{\frac{n}{p}}\mathbf{M}\mathbf{Z}\mathbf{D}_z^{-1}\mathbf{B}^{*}.$$

Both $\mathbf{B}^*$ and $\mathbf{Z}_k$ are updated at each iteration until the value of the objective function stops increasing.

3 Results

To appraise the effect of the presence of outliers on CCA and tandem analysis, we generated structured categorical data sets using the evolutionary algorithm (EA) described in Van de Velden *et al.* (2017). The data-generating strategy is: generate a cluster membership attribute, that is, the *true* cluster allocation; use the EA to generate a set of further attributes being associated with some defined strength to the cluster membership; finally, generate a set of noise variables with low-to-none association with the cluster membership.

An observation can be considered an outlier if it is characterised by the less-occurring categories of each attribute in the considered data set (He *et al.* , 2005). Then outliers have been generated according to such characteristics. In the experiments, 1000 observations of 12 active and 12 noise categorical attributes have been considered. Also, we referred to two strength levels of clustering structure, measured by the association of the active variables to the true allocation in the $K = 4$ considered clusters (each of different size). The performance of the tandem approach is more sensitive to outliers than CCA (see results in Table 1). In particular, we observe that in the case of a low association strength and no noise, there is a significant drop in the performance of the tandem approach.

References

AGGARWAL, C. C. 2015. Outlier analysis. *Pages 237–263 of: Data mining*. Springer.

DE SOETE, G. & CARROLL, J.D. 1994. K-means clustering in a low-dimensional Euclidean space. *Pages 212–219 of: New approaches in classification and data analysis*. Springer.

HE, Z., DENG, S. & XU, X. 2005. An optimization model for outlier detection in categorical data. *Pages 400–409 of: International Conference on Intelligent Computing*. Springer.

HWANG, H., DILLON, W.R. & TAKANE, Y. 2006. An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, **71**(1), 161–171.

VAN DE VELDEN, M., IODICE D'ENZA, A. & PALUMBO, F. 2017. Cluster correspondence analysis. *Psychometrika*, **82**(1), 158–185.

VAN DE VELDEN, M., IODICE D'ENZA, A. & MARKOS, A. 2019. Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, **11**(3), e1456.

VICHI, M. & KIERS, H.A.L. 2001. Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, **37**(1), 49–64.

VICHI, M., VICARI, D., & KIERS, H.A.L. 2019. Clustering and dimension reduction for mixed variables. *Behaviormetrika*, 1–27.

EARTHQUAKE CLUSTERING AND CENTRALITY MEASURES

Elisa Varini[1], Antonella Peresan[2] and Jiancang Zhuang[3]

[1] Institute of Applied Mathematics and Information Technologies *E. Magenes*, National Research Council, Milano, (e-mail: elisa@mi.imati.cnr.it)

[2] National Institute of Oceanography and Experimental Geophysics. CRS-OGS, Udine, (e-mail: aperesan@inogs.it)

[3] Institute of Statistical Mathematics, Research Organization of Information and Systems, Tokyo, (e-mail: zhuangjc@ism.ac.jp)

ABSTRACT: Earthquake clustering is a relevant feature of seismic catalogs, both in time and space. Several methodologies for earthquake cluster identification have been proposed in the literature in order to characterize geophysical clustering properties and to analyze background seismicity. We consider two recent data-driven declustering techniques, one is based on nearest-neighbor distance and the other on a point process model. Since the different assumptions underlying each method may lead to different classifications of earthquakes into main events and secondary events, we investigate the classification similarities by exploiting graph representations of earthquake clusters and tools from Network analysis.

KEYWORDS: earthquake clustering, rooted trees, centrality measures.

1 Two declustering algorithms

Declustering algorithms perform the partition of an earthquake catalog in two subsets of events, respectively named background events and secondary events. Background seismicity is intended to include spontaneous and independent events, whose occurrence rate is approximately constant. Secondary events are triggered by other events, e.g. foreshocks, aftershocks, seismic swarms; when secondary events appear, the occurrence rate is greater than usual. In order to indentify these events, we consider two data-driven declustering algorithms which also provide the topological structure of clusters.

Let's assume a catalog that includes n earthquakes, each of which is denoted by its occurrence time t_i, magnitude m_i, and epicentral location (x_i, y_i) $(i = 1, ..., n)$.

Nearest-neighbor (NN) algorithm (Zaliapin & Ben-Zion, 2016). This approach is based on the nearest-neighbor distance between two earthquakes in the space-time-energy domain:

$$\eta_{ij} = (t_j - t_i) r_{ij}^{d} 10^{-bm_i} \tag{1}$$

where $t_i < t_j$ and r_{ij} is the spatial distance between events i and j; this distance combines the inter-occurrence time, the fractal dimension of the hypocentres distribution, and the Gutenberg–Richter law. There are only two unknown parameters, namely fractal dimension d and b-value, which are jointly and robustly identified by the Unified Scaling Law for Earthquakes (USLE) method; a separation distance η_0 is also estimated in order to identify background and secondary events (details in Peresan & Gentili, 2018). Each event i is connected to its nearest-neighbor $j = argmin_k \eta_{ik}$. Then, by removing all connections η_{ij} such that $\eta_{ij} > \eta_0$, earthquake clusters and background events are unambiguously identified.

Stochastic declustering (SD) algorithm (Zhuang, 2006 and references therein). The approach is based on the space-time ETAS (epidemic-type aftershock sequence) model, a branching point process defined by its intensity function conditional on the observation history $\mathcal{H}_t$:

$$\lambda(t,x,y \mid \mathcal{H}_t) = \mu(x,y) + \sum_{k:t_k<t} \nu(t-t_k, x-x_k, y-y_k \mid m_k) \tag{2}$$

where $\mu(x,y)$ is the background rate of a time-homogeneous Poisson process and, at time t, $\nu(t-t_k, x-x_k, y-y_k, m_k)$ is the contribution to seismic hazard due to triggered events. According to point process theory, the probability that event j is generated by the background process is $\varphi_j = \mu(x_j,y_j)/\lambda(t_j,x_j,y_j \mid \mathcal{H}_{t_j})$, and the probability that it is triggered from previous event i is $\rho_{ij} = \nu(t_j - t_i, x_j - x_i, y_j - y_i, m_i)/\lambda(t_j,x_j,y_j \mid \mathcal{H}_{t_j})$. Thinning the process according to these probabilities allows splitting the catalog into background events and triggered events, and also setting connections between triggering and triggered events. Unlike NN method, SD algorithm can provide many declustered catalogs by simulation.

2 Cluster analysis of seismicity in North-Eastern Italy

We consider the earthquake bulletins for North-Eastern Italy and Western Slovenia compiled at the National Institute of Oceanography and Experimental Geophysics since 1977. This catalog is statistically complete from 1994 to 2018 for events having magnitude at least 2.0. We apply both NN and SD algorithms to this dataset. Hereafter, we show the results obtained from SD method by retaining the most probable connections between any pair of events according to the estimated probabilities $\hat{\varphi}_j$ and $\hat{\rho}_{ij}$ $(i, j = 1, ..., n)$.

The declustered catalogs turn out to be organized in rooted time-oriented trees, where tree roots and nodes are background events and triggered events, respectively. We aim at characterizing common features among clusters as well as comparing the results obtained from the two declustering algorithms. We address the question whether some measures of the network topology may characterize the spatio-temporal properties of earthquake clustering in the region under study. To this end, we exploit tools from the R package *igraph* (Csardi & Nepusz, 2006) for data visualization and analysis.

Fig.1 shows the rooted trees of the cluster which includes the strongest earthquake in the catalog, the 1998/04/12 M5.6 earthquake. NN-cluster contains 720 events (left panel), SD-cluster has 697 events (right panel), and even 677 events are associated with the 1998 cluster by both methods. Despite the large number of events identified by both methods, Fig.1 clearly shows that the hierarchical structure of the SD-cluster is more complex than that obtained from NN method.

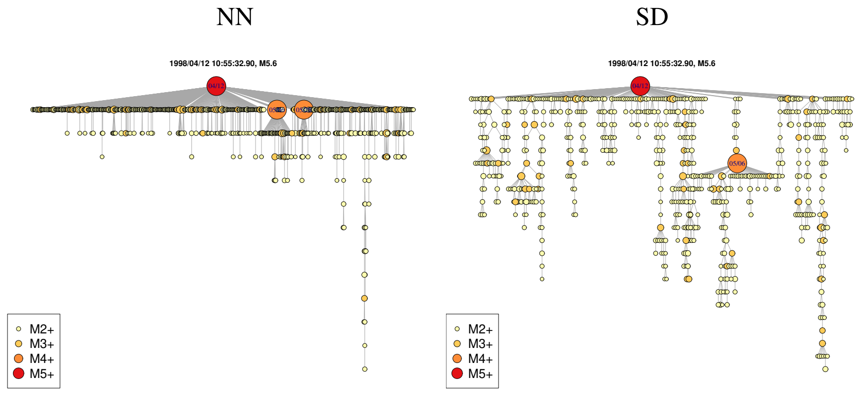


Figure 1. *Tree representations of the cluster related to the 1998/04/12 M5.6 earthquake obtained from NN algorithm (left) and SD algorithm (right). Earthquake magnitude is denoted by different colors and size.*

We focus on centrality measures which should express the way earthquakes (nodes) get organized in clusters (trees). In this paper, for simplicity, we only mention closeness centrality, a measure of the distance of each node from every other nodes. Closeness centrality of node x_i is defined by

$$close(x_i) = \frac{n-1}{\sum_{x_j} d(x_i, x_j)} \tag{3}$$

where $d(x_i, x_j)$ is the geodesic (shortest path) distance from x_i to x_j; if node x_j is not reachable from x_i, geodesic distance $d(x_i, x_j)$ is set equal to n. It is noted that the numerator $n-1$ is the minimum value the denominator can have, so that closeness centrality ranges in $[0,1]$. High closeness values are associated with the most central nodes, those closest to other nodes on average. Centralization is a global index for the entire cluster based on closeness centrality values of its nodes; it is defined as the average discrepancy between the centrality of most central node x^* and that of all other nodes: $C = \sum_x [close(x^*) - close(x)] / (n-1)$. It also ranges in $[0,1]$. High C values indicate simple structures inside the cluster, in which few nodes dominate others. On the contrary, small C values denote more complex hierarchical structures. As for the 1998 cluster in Fig.1, closeness centralization is 0.63 for the NN-cluster (simple tree) and 0.19 for the SD-cluster (complex tree).

Acknowledgement

This work is supported by National grant MIUR, PRIN-2015 program, Prot. 20157PRZC4: Complex space-time modeling and functional analysis for probabilistic forecast of seismic events. We also acknowledge financial support from Civil Defence of the Friuli Venezia Giulia Region.

References

CSARDI, G., & NEPUSZ, T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems*.

PERESAN, A., & GENTILI, S. 2018. Seismic clusters analysis in Northeastern Italy by the nearest-neighbor approach. *Physics of the Earth and Planetary Interiors*, 274, 88–104.

ZALIAPIN, I., & BEN-ZION, Y. 2016. A global classification and characterization of earthquake clusters. *Geophysical Journal International*, 207, 608–634.

ZHUANG, J. 2006. Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *Journal of the Royal Statistical Society: Series B*, 68, 635–653.

CO-CLUSTERING HIGH DIMENSIONAL TEMPORAL SEQUENCES SUMMARIZED BY HISTOGRAMS

Rosanna Verde[1], Antonio Irpino[1] and Antonio Balzanella[1]

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

ABSTRACT: This paper proposes a co-clustering method for data streams summarized by histogram series. We assume that a set of sensors record huge amounts of data over time so that we are interested in discovering a partitioning of the sensors and to understand how different time periods impact on such partition. To reach our aim, we summarize the incoming data split into non overlapping windows, by histograms. The latter, become the input of an online procedure which finds, both, a partition of the streams and a partition of time intervals according to sensed data.

KEYWORDS: data stream mining, histogram data, co-clustering.

1 Introduction

Massive datasets, having the form of continuous streams with no fixed length, are becoming very common due to the technological developments in recent years. Typical data sources include: sensor networks performing, at a very high frequency, repeated measurements of environmental variables as temperature, sound, pollution, humidity; real-time data recorded by surveillance systems; data recorded by vehicle traffic monitoring systems; electricity consumption recording; network traffic monitoring.

The statistical analysis in these applicative fields is a very challenging task since online collected data quickly become too large to fit in main memory so that random access, which is commonly used in traditional data mining, is prohibitively expensive. Moreover, since online monitoring can concern highly evolving scenarios, appropriate methods able to incorporate the new available information but also eliminate the effects of outdated data, are needed.

The data stream mining framework offers a wide range of specific tools for dealing with these potentially infinite and online arriving data.

A common practice in the Data stream mining literature is to cope with the high velocity and huge volume of data through appropriate synopsis. In this paper we use histograms as synopsis of non overlapping batches of each data stream, that is, similarly to Arroyo & Maté, 2009, we represent each data stream as a series of histograms. A histogram keeps a detailed view of data, reducing memory occupation and supporting fast computation. It records information about the moments of data as well as the quantiles thus, it is a more informative tool than simpler aggregates such as the average of data subsequences.

On this kind of data representation we introduce a co-clustering algorithm. It is a two-step strategy based on the classic Double k-means proposed in Vichi, 2001 however, it is adapted to process online arriving data summarized by histograms. Our aim is to obtain a partition of the streams and a partition of the time windows. By means of the partition of the streams we can understand which sensors record similar observations over time. The partition of the window allows to understand how different time windows contribute the the similarity among sensors.

2 Online co-clustering on data streams

Let $Y = \{Y_1, \ldots, Y_i, \ldots, Y_n\}$ be a set of n data streams $Y_i = (y_i^1, t_1), \ldots, (y_i^j, t_j), \ldots$ made by real valued observations y_i^j on a discrete time grid $T = \{t_1, ..., t_j, ...\}$, with $t_j \subseteq \Re$ and $t_j > t_{j-1}$.

We propose to split the flowing data into non overlapping windows, identified by $w = 1, \ldots, \infty$.

A window is an ordered subset of T, having size b, which frames a data batch $Y^w = \{Y_1^w, \ldots, Y_i^w, \ldots, Y_n^w\}$, where $Y_i^w = \left\{(y_i^j, t_j), \ldots, (y_i^{j+b}, t_{j+b})\right\}$ is a subsequence of Y_i.

The objective is to get a partition P of the Y_i (with $i = 1, \ldots, n$) in K homogeneous clusters C_k, and a partition G of the data batches Y^w (with $w = 1, \ldots, \infty$) into H clusters.

As shown in the previous section, we represent the input data by histograms. That is, every time a new batch of data is available, the observations of each subsequence Y_i^w are represented by a histogram H_i^w, formalized as follows:

$$H_i^w = \{(I_1, \pi_1), \ldots, (I_l, \pi_l), \ldots, (I_L, \pi_L)\}$$

.

where I_l for $(l = 1, \ldots, L)$ are L consecutive intervals (bins) associated to the π_l weights (relative frequencies), summing to 1 on the all bins.

In order to reach our aim we introduce an algorithm based on two steps. The first step performs the analysis of data of a set of time windows in order to provide the partition P of the streams. The second step, starts from the partition P and performs the partitioning of data batches to discover the partition G.

The whole strategy is based on the L^2 Wasserstein distance to compare histograms.

Focusing on the first step, the procedure analyzes the set of time windows $w = j, \ldots, j'$ (with $j' > j$) in order to get the partition P of the sensors minimizing the following criterion function:

$$\Delta(P) = \sum_{k=1}^{K} \sum_{i,m \in C_K} d(Y_i, Y_m) \tag{1}$$

where d is computed through the L^2-Wasserstein distance between the histograms associated to Y_i and Y_m.

The algorithm proposed to address the previous optimization problem is detailed in Balzanella & Verde, 2019; Balzanella & Verde, 2013.

Once we have the partition P for a time period, we propose to analyze the time windows $w > j'$ in order to get the partition G. To cluster each data batch Y^w, we consider that its subsequences Y_i^w have been allocated to clusters C_k of P by the step 1 of the proposed procedure. This allows to compute a centroid for each cluster C_k, which is the minimizer of the distances inside the cluster. The idea is to summarize each data batch Y^w by means of a set of K centroids B_k^w. The clustering step we perform for getting the partition G of the data batches is based on minimizing the following criterion:

$$\Delta(G) = \sum_{h=1}^{H} \sum_{w} \sum_{k=1}^{K} d(B_k^w, \overline{B_k^h}) \tag{2}$$

where $\overline{B_k^h}$, is a set of centroids for the cluster G_h.

To minimize this optimization function, we use the algorithm proposed in Balzanella & Irpino, 2019.

With the flowing of data, the first step and second step are alternated in order to keep a detailed summary of the partitioning structure of data as well as, to deal with data evolution.

3 Conclusions

In this paper we have introduced a co-clustering algorithm able to analyze data streams recorded by sensors. It is able to cope with the high dimensionality of data and their online arriving nature. Preliminary results on environmental data confirm the effectiveness of our proposal.

References

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

BALZANELLA, A., & IRPINO, A. 2019. Spatial prediction and spatial dependence monitoring on georeferenced data streams. *Statistical Methods & Applications*, Apr.

BALZANELLA, A., & VERDE, R. 2013. Clustering and Change Detection in Multiple Streaming Time Series. *Pages 1–14 of:* KOŁODZIEJ, JOANNA, DI MARTINO, BENIAMINO, TALIA, DOMENICO, & XIONG, KAIQI (eds), *Algorithms and Architectures for Parallel Processing*. Cham: Springer International Publishing.

BALZANELLA, A., & VERDE, R. 2019. Histogram-based clustering of multiple data streams. *Knowledge and Information Systems*, Mar.

VICHI, M. 2001. Double k-means Clustering for Simultaneous Classification of Objects and Variables. *Pages 43–52 of:* BORRA, SIMONE, ROCCI, ROBERTO, VICHI, MAURIZIO, & SCHADER, MARTIN (eds), *Advances in Classification and Data Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg.

STATISTICAL ANALYSIS OF ITEM PRE-KNOWLEDGE IN EDUCATIONAL TESTS: LATENT VARIABLE MODELLING AND OPTIMAL STATISTICAL DECISION

Chen Yunxiao[1], Lu Yan[1] and Irini Moustaki[1]

[1] Department of Statistics, London School of Economics, London,
(e-mail: y.chen186@lse.ac.uk, y.lu62@lse.ac.uk,
i.moustaki@lse.ac.uk)

ABSTRACT: Fairness is essential to the validity of educational tests. Test scores are no longer a fair indicator of examinees' true ability if some test questions favour some test takers over others. For this reason, it is important to ensure the fairness of educational tests. This paper concerns the issue of item pre-knowledge in educational tests. That is, test takers cheat by gaining prior access to leaked items. As a result, they have inflated performance on the set of leaked items. We develop methods for simultaneous detecting test takers who cheat and compromised items based on item response data from a single test administration, without knowing any specific subsets of cheaters and compromised items.

Latent variable models are proposed for the modelling of (1) data consisting only of item-level binary scores and (2) data consisting of both item-level binary scores and response time, where the former is commonly available in paper-and-pencil tests and the latter is widely encountered in computer-based tests. The proposed model adds a latent class model component upon a latent factor model (also known as item response theory model) component, where the latent factor model component captures normal item response behaviour and the latent class model component captures response patterns due to item pre-knowledge. We further formulate the detections of cheaters and compromised items under a statistical decision framework, and propose Bayesian decision rules and compound decision rules that control local false discovery rate or local false non-discovery rate. Statistical inference is carried out under the Bayesian framework. The proposed method is applied to data from a computer-based nonadaptive licensure assessment.

KEYWORDS: Item response theory, latent factor model, bayesian hierarchical modelling, false discovery rate, test security.

References

CHO, S.-J., SUH, Y. & LEE, W.-y. 2016. An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, **35**, 48-61.

EVALUATION OF THE WEB USABILITY OF THE UNIVERSITY OF CAGLIARI PORTAL: AN EYE TRACKING STUDY

Gianpaolo Zammarchi[1] and Francesco Mola[1]

[1] Department of Economics and Business Sciences, University of Cagliari,
(e-mail: gp.zammarchi@unica.it, mola@unica.it)

ABSTRACT: A web portal is one of the main tools used by companies, institutions and individual citizens to make information available to anyone. Designing a portal that has good usability means allowing an average user to find the information he needs as soon as possible. The objective of this work is to evaluate the web usability of the portal of the University of Cagliari, using the eye tracking technology. High school and university students were asked to perform specific tasks within the portal. The results were evaluated through a quantitative analysis of the time and number of fixations required to complete each task, as well as a qualitative analysis of heat maps and gaze plots representing participants' fixations. The analysis has allowed to (i) detect a high efficiency for most of the web pages, (ii) highlight the most critical elements of the portal and (iii) suggest the most appropriate changes to be made.

KEYWORDS: eye tracking, web usability, heat map, gaze plot.

1 Introduction

Nowadays a web portal is one of the main tools used by companies, institutions and individual citizens to make useful information available to anyone. Designing a portal that has good or excellent usability means allowing an average user to find the information he/she needs as soon as possible. In order to assess whether the interface of a web portal is intuitive and easy to use, most studies use a measure defined as web usability, which is often evaluated exclusively through the administration of questionnaires to users. The use of the eye tracking technology allows to define web usability in a more objective way through the analysis of ocular movements during visualization of images, texts or other visual stimuli (Jacob & Karnet, 2003; Goldberg & Kotval, 1999). The eye tracking technology has been increasingly applied to the study of web usability in different fields such as tourism (Scott et al., 2017) and e-commerce (Bach, 2018; Hwang & Lee, 2017).

The main objective of this study is to evaluate the web usability of the web portal of the University of Cagliari (*www.unica.it*) using eye tracking technology, in order to improve the user experience, including the experience of future students using the site for the first time. The new portal of the University of Cagliari was launched in 2017 and has been the first portal of an Italian university to meet the requirements of the Agenzia Italiana digitale del Consiglio dei ministri (Agid).

2 Materials and methods

We carried out a study to assess the efficiency of the web portal of the University of Cagliari through the execution of ten different tasks (e.g. find the library section, WiFi instructions, deadline for enrolment, university fee regulations and so on). The tasks were executed by two groups of participants: high school students and university students. Objective of the analysis was to collect information about the behavior of a group of experienced users (students already enrolled in the University) as well as of non-experienced users (high school students). In light of the exploratory nature of the study, for the first group we randomly selected a group of students present in group study rooms of different departments (choosing different days of the week and different times). For the second group we randomly selected students from Sardinian high schools who attended the Unica University Fair. For each participant, we collected information on age, gender, high school institute and university course. These characteristics were compared between the two groups using chi-squared test or Student's t test.

Throughout the execution of the tasks, the exact position of the eyes has been detected through a Tobii X2-60 Compact eye tracker. Different eye movement classification algorithms can be used to identify various types of eye movements (Komogortsev et al., 2010). The fixation is the most commonly studied type of eye movement in human research since fixations are usually connected to the moment in which information are registered by the brain (van der Lans et al., 2011). Among available fixation classification algorithms, the Velocity-Threshold Identification (I-VT) algorithm classifies eye movements based on the velocity of the directional shifts of the eye (Salvucci and Goldberg, 2000). We applied this filter to extract fixations using the Tobii studio software version 3.3.1. Data for different metrics, including time to completion of the task and number of fixations for the whole page, as well as for specific areas of interest (AOI), were collected. These data were also used to produce two main typologies of graphical outputs: heat map (a graphical representation of the data where the individual values contained in a matrix are represented as colors) and gaze plot (a map showing gaze fixations on a webpage in the order in which they occur) (Dong et al., 2014). The tasks have been defined as efficient or not efficient. Specifically, relative efficiency in terms of different metrics (e.g. time to completion, number of fixations) has been assessed comparing each task to a threshold value established through evaluation of all the other tasks executed by the two groups of participants. The tasks defined as not efficient in both groups were further evaluated through a quantitative analysis of the main efficiency indicators as well as a qualitative analysis of heat maps and gaze plots representing participants' fixations. Analyses have been conducted using R v. 3.5.0 (R Core Team, 2018).

3 Results

Data for 100 participants (Group 1: 46 high school students and Group 2: 54 university students) were analyzed. The two groups did not differ in terms of gender (chi-squared: p = 0.45) or high school institute (chi-squared: p = 0.46), while mean age was higher in the group of university students (t-test: p < 0.001).

The analysis allowed to detect a high efficiency for most of the evaluated pages. In particular, the tasks classified as efficient for both high school and university students allowed to highlight how the site is easily accessible even by those who have used it a few times. However, the tasks classified as less efficient in both groups allowed to highlight some aspects that might be improved.

For instance, the quantitative analysis of the number of fixations in the different AOIs as well as the qualitative analysis of heat maps and gaze plots showed that the large majority of observations was focused on the upper part of a web page (Figure 1). Therefore, information that needs to be noticed by a large number of users should not be placed at the bottom of a page.

Moreover, we observed that in some cases the participants were not able to understand the meaning of specific links at first sight or failed to retrieve the information required to complete the task even after reaching the correct page.

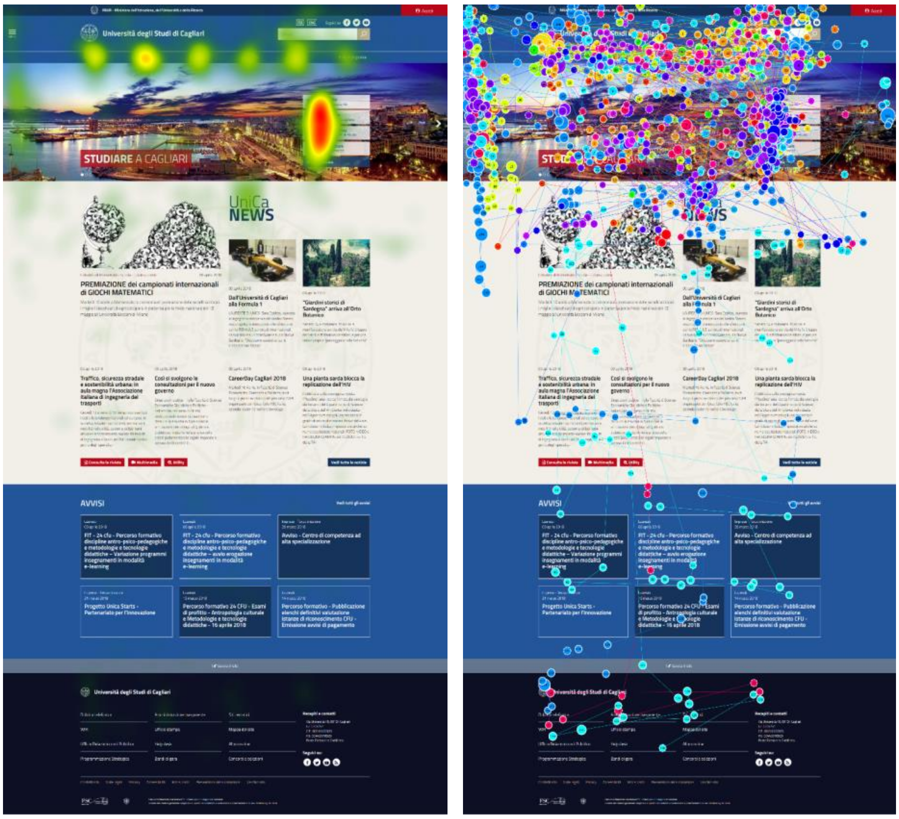

Università degli Studi di Cagliari
STUDIARE A CAGLIARI
UniCa NEWS
PREMIAZIONE dei campionati internazionali di GIOCHI MATEMATICI
AVVISI

Figure 1. Heat map (on the left) and gaze plot (on the right) of the Home Page of the web portal of the University of Cagliari

4 Conclusions

The objective of this work is to evaluate the web usability of the portal of the University of Cagliari, using the eye tracking technology. The analysis has allowed to (i) detect a high efficiency for most of the web pages examined, (ii) highlight the most critical elements of the portal and (iii) suggest the most appropriate changes to be made. The identified critical aspects would have been difficult to detect without the eye tracking, which allowed to highlight the areas of the pages that received the greatest number of fixations. These results could help to further improve the web usability of the University of Cagliari's website.

References

BACH, M. P. 2018. Usage of social neuroscience in E-Commerce research – Current research and future opportunities. *Journal of Theoretical and Applied Electronic Commerce Researchers*, **13**, I-IX.

DONG, W., LIAO, H., ROTH, R. & WANG, S. 2014. Eye Tracking to Explore the Potential of Enhanced Imagery Basemaps in Web Mapping. *The Cartographic Journal*. **51**, 313-329.

GOLDBERG, J. H., & KOTVAL, X. P. 1999. Computer interface evaluation using eye movements: Methods and constructs, in: *International Journal of Industrial Ergonomics*, **24**, 631-645.

HWANG, Y. M., & LEE, K. C. 2017. Using an eye tracking approach to explore gender differences in visual attention and shopping attitudes in an online shopping environment. *International Journal of Human-Computer Interaction,* **34**, 15-24.

JACOB, R. J. K., & KARN, K. S. 2003. Eye tracking in human computer interaction and usability research: Ready to deliver the promises (Section commentary), in: *The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements,* Oxford: Elsevier Science.

KOMOGORTSEV, O. V., GOBERT, D. V., JAYARATHNA, S., KOH, D. H., & GOWDA, S. M. 2010. Standardization of Automated Analyses of Oculomotor Fixation and Saccadic Behaviors. *Biomedical Engineering, IEEE Transactions,* **57**, 2635-45.

R CORE TEAM. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

SALVUCCI, D. D., & GOLDBERG, J. H. 2000. Identifying fixations and saccades in eye-tracking protocols, in Proceedings of the symposium on Eye tracking research & applications - ETRA '00, Palm Beach Gardens, Florida, United States, 71-78.

SCOTT, N., ZHANG, R., DUNG L., & BRENT, M. 2017. A review of eyetracking research in tourism, *Current Issues in Tourism,* **22**, 1244-1261.

VAN DER LANS, R., WEDEL, M., & PIETERS, R. 2011. Defining eye-fixation sequences across individuals and tasks: the Binocular-Individual Threshold (BIT) algorithm. *Behav Res Methods,* **43**, 239-257.

APPLICATION OF SURVIVAL ANALYSIS TO CRITICAL ILLNESS INSURANCE DATA

David Zapletal[1] and Lucie Kopecká[1]

[1] Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administrations, University of Pardubice, (e-mail: david.zapletal@upce.cz, lucie.kopecka1@student.upce.cz)

ABSTRACT: The data set of critical illness insurance policies from the commercial insurance company is studied by survival analysis. Specifically, the Cox proportional hazard model is used to investigate the influence of the gender, age and the place of residence of the policyholder on the time to occurrence of the insured event. Beside the investigation of gender influence, the approximately homogeneous age groups and the groups of the regions due to the risk of occurrence of the insured event are created.

KEYWORDS: Critical illness, insurance, survival analysis, Cox proportional hazard model.

1 Introduction

There are many professional books and research articles dealing with the modelling survival data especially in medical research. Presented paper focuses on application of the Cox proportional hazard model on data arising from critical illness insurance. The proportional hazard model with the unspecified baseline hazard function was proposed by Cox (1972). This paper introduced the notion of partial likelihood, which was subsequently considered in great detail by Cox (1975). A detailed review of the model and its extension is contained in Therneau and Grambsch (2000).

Scientific articles on the application of survival models to insurance data are far from as much as in the case of medical research. Car insurance data was analysed by various statistical methods, including survival analysis, by Beirlant et al. (1992). The Cox proportional hazard model to estimate transition intensities in long-term care insurance in Germany was used by Czado and Rudolph (2002). A study which incorporates the survival analysis of unemployment duration into pricing of Taiwan's unemployment insurance program was done by Chuang and Yu (2010).

Probably the most common use of survival analysis in insurance is a problem of the cancellation of insurance contracts. The analysis of customer survival time in the insurance company after a policy cancellation was introduced by Guillen et al. (2003). The Cox proportional hazard model was used by Ho and Su (2006) to investigate China's residential mortgage life insurance prepayment risk behavior. The data set of Danish households possessing multiple insurance policies was studied by Brockett et al. (2008). Haugen and Moger (2016) investigated corporate customers holding multiple car contracts with the same insurance company. The shared gamma frailty model was presented by them in order to study time to lapse of single car policies.

2 Data and Model

We obtained data on critical illness insurance from one of the commercial insurance company operating in the Czech Republic. The set contains data for 231,046 persons for the period from July 1, 1997 to April 30, 2017. The number of insured events in the monitored period was 1,045. For each person we have information about an age at the commencement of the policy, age at eventual occurrence of the insured event, respectively the termination of the policy, the information about a region where the insured person lives, and about a gender of the insured person.

The influence of the gender of the insured person, the age at which the person entered into insurance contract, and the region where the insured person lives on time to occurrence of the insured event is investigated. To do this, the Cox proportional hazard model is fitted. The model includes following explanatory variables: gender of insured person, age at which the person entered into the insurance contract, and region where the insured person lives. Beside the investigation of gender influence, the approximately homogeneous age groups and the groups of the regions due to the risk of occurrence of the insured event are created.

The Cox model of the hazard at time t for the i-th individual is given by the equation

$$h_i(t) = \exp(\beta_1 Gender_i + \beta_2 Age_i + \beta_3 Region_i)\, h_0(t),$$

where $h_0(t)$ is the baseline hazard function of unspecified form. For great details see Cox (1972) or, for example, Thernau and Grambsch (2000).

A crucial assumption made when using the Cox model is that of proportional hazards. Hazards are said to be proportional if the ratios of hazards are independent of the time. The hazard proportionality assumption of the Cox model has been tested by so called zph test based on Schoenfeld residuals which was developed by Grambsch and Thernau (1994). The assumption of hazards proportionality has not been rejected in case of variables Gender and Region but this crucial assumption has been rejected for the Age variable. That is why the stratification in two parts of the data set was done at age 18. Due to limited space only the results for individuals over or equal 18 years of age are presented here. This subset contain 139,963 persons of whom 931 occurred an insured event.

Because of categorical explanatory variables each parameter β_i for i = 1, 2, 3 is represented by q-1 estimated parameters, where q means the number of categories of corresponding explanatory variable. The variable *Gender*, of course, includes two categories (q = 2), for the variables *Age* and *Region* their categories were designed to create groups with a different rate of risk of occurrence of the insured event, i.e. significantly different hazard ratio. Therefore variable *Age* is made up of four categories (q = 4): 18-30; 31-40; 41-50 and over 50 years. Regarding the variable *Region*, first it should be noted that Czech Republic is divided into fourteen territorial administration units, called regions. According to the rate of hazard, these regions were divided into following three groups (q = 3): 1st group - containing the regions Liberec, Pardubice, Prague and Zlin where the hazard is the lowest; 2nd group containing nine regions (Central Bohemia, Hradec Kralove, Moravia-Silesia, Olomouc, Plzen, South Bohemia, South Moravia, Vysocina and Usti nad Labem) and 3rd group contains only one region Karlovy Vary with the highest hazard of occurrence of the insured event.

The categories of corresponding explanatory variables with the lowest hazard were determined as reference categories, i.e. male for variable *Gender*, 18-30 years for variable *Age* and the 1st group for variable *Region*.

3 Results and Discussion

Estimations of coefficients of the Cox proportional hazard model simultaneously with their statistical significance (p-values), hazard ratios and corresponding confidence intervals are shown in Tab. 1.

Tab. 1: Estimations of Cox proportional hazard model

Variable	Level of Effect	Parameter Estimate	p-value	Hazard Ratio	95% Lower CI	95% Upper CI
Gender	female	0.027	0.678	1.028	0.903	1.170
Age	31-40	0.822	0.000	2.274	1.806	2.864
Age	41-50	1.680	0.000	5.367	4.307	6.687
Age	over 50	2.305	0.000	10.023	7.933	12.664
Region	2nd group	0.240	0.008	1.271	1.064	1.519
Region	3rd group	0.521	0.000	1.684	1.286	2.207

We can see from Tab. 1 that for the time to occurrence of the insured event is not statistically significant (p-value 0.678 is greater than the significance level 0.05) if the insured person is male or female. It means that the risk of critical illness is comparable for men and women, hazard ratio equals 1.028 and confidence interval contains one.

As expected, the situation is different for explanatory variables *Age* and *Region*. In particular, age plays very significant role in the risk of critical illness. Based on the hazard ratios, we can say that the age category 31–40 years has more than twice as larger risk of occurrence of the insured event (i.e. occurrence of the critical illness) in comparison with the reference age category 18–30 years. For persons entering into an insurance contract in age from 41 to 50 years, the risk is more than five times greater and for the people over fifty, it is more than ten times higher.

In the case of the regions, the differences are not so great. The best situation is in the regions: Liberec, Pardubice, Prague and Zlin, which form the reference category marked as 1st group. Interestingly, these are not neighboring regions. On the other hand, the worst situation is in Karlovy Vary region (3th group) where the risk of occurrence of critical illness is almost twice as large in comparison with the lowest hazard regions. The other nine regions (2nd group) have a comparable risk of occurrence of the insured event. But this risk is significantly higher in comparison the lowest hazard regions, approximately 1.3 times.

4 Conclusions

The Cox proportional hazard model containing three explanatory variables (gender, age and region) based on data of critical illness insurance was fitted. The age categories and the groups of regions with significantly different hazard ratios of occurrence of the insured event were constructed. On the other hand, the statistical significance of gender has not been demonstrated.

References

BEIRLANT, J. et al. 1992. Statistical risk-evaluation applied to (Belgian) car insurance. *Insurance Mathematics & Economics*. **10**, 289-302.

BROCKETT, P.L. et al. 2008. Survival analysis of a houshold portfolio of insurance policies: How much time do you have to stop total customer defection? *Journal of Risk and Insurance*. **75**, 713-737.

CHUANG, H.L., & YU, M.T. 2010. Pricing unemployment insurance – an unemployment-duration-adjusted approach. *Astin Bulletin*. **40**, 519-545.

COX, D. R. 1972. Regression models and life Tables (with discussion). *Journal of the Royal Statistical Society, B.* **34**, 187-220.

COX, D. R. 1975. Partial likelihood. *Biometrika*. **62**, 269-276.

CZADO, C., & RUDOLPH, F. 2002. Application of survival analysis methods to long-term care insurance. *Insurance Mathematics & Economics*. **31**, 359-413.

GUILLEN, M. et al. 2003. The analysis of customer survival time in the insurance company after a policy cancellation. *Insurance Mathematics & Economics*. **33**, 434-434.

GRAMBSCH, P.M., & THERNAU, T.M. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515-526.

HAUGEN, M., & MORGER, T.A. 2016. Frailty modelling of time-to-lapse of single policies for customers holding multiple car contracts. *Scandinavian Actuarial Journal*. **6**, 489-501.

HO, K.H., & SU, H.Y. 2006. Structural prepayment risk behavior of the underlying mortgages for residential mortgage life insurance in a developing market. *Journal of Housing Economics*. **15**, 257-278.

KAPLAN, E. L. & MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. **53**, 457–481.

THERNAU, T.M., & GRAMBSCH, P.M. 2000. *Modelling Survival Data: Extending the Cox Model.* New York: Springer.

CLADAG 2019
Cassino (ITALY)
11-13 September, 2019

The CLAssification and Data Analysis Group of the Italian Statistical Society (SIS) promotes advanced methodological research in multivariate statistics with a special vocation in Data Analysis and Classification.

CLADAG supports the interchange of ideas in these fields of research, including the dissemination of concepts, numerical methods, algorithms, computational and applied results.

CLADAG is a member of the International Federation of Classification Societies (IFCS).

Among its activities, CLADAG organizes a biennial international scientific meeting, schools related to classification and data analysis, publishes a newsletter, and cooperates with other member societies of the IFCS to the organization of their conferences.

Founded in 1985, the IFCS is a federation of national, regional, and linguistically-based classification societies. It is a non-profit, nonpolitical scientific organization, whose aims are to further classification research.

SIS Società Italiana di Statistica

UNICAS 1979>2019
40 YEARS FROM FOUNDATION
UNIVERSITÀ DI CASSINO E DEL LAZIO MERIDIONALE