

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



CLADAG 2019

11-13 SEPTEMBER 2019
CASSINO

```
def business_model()  
  arr = [ ]  
  items = a, b, c  
  items >> arr  
  return arr  
end
```



Book of Short Papers

Giovanni C. Porzio
Francesca Greselin
Simona Balzano
Editors

12-TH SCIENTIFIC MEETING
CLASSIFICATION AND DATA ANALYSIS



Società
Italiana di
Statistica

© CC – Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
<https://creativecommons.org/licenses/by-nc/4.0/>

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 <i>Christophe Biernacki</i>	3
Statistics with a human face <i>Adrian Bowman</i>	4
Bayesian model-based clustering with flexible and sparse priors <i>Bettina Grün</i>	5
Grinding massive information into feasible statistics: current challenges and opportunities for data scientists <i>Francesco Mola</i>	6
Statistical challenges in the analysis of complex responses in biomedicine <i>Sylvia Richardson</i>	7

Invited and contributed sessions

Model-based clustering of time series data: a flexible approach using nonparametric state-switching quantile regression models <i>Timo Adam, Roland Langrock, Thomas Kneib</i>	8
Some issues in generalized linear modeling <i>Alan Agresti</i>	12
Assessing social interest in burnout using functional data analysis through google trends <i>Ana M. Aguilera, Francesca Fortuna, Manuel Escabias</i>	16
Measuring equitable and sustainable well-being in Italian regions: a non- aggregative approach <i>Leonardo Salvatore Alaimo, Filomena Maggino</i>	20
Bootstrap inference for missing data reconstruction <i>Giuseppina Albano, Michele La Rocca, Maria Lucia Parrella, Cira Perna</i>	22
Archetypal contour shapes <i>Aleix Alcacer, Irene Epifanio, M. Victoria Ibáñez, Amelia Simó</i>	26

Random projections of variables and units <i>Laura Anderlucci, Roberta Falcone, Angela Montanari</i>	30
Sparse linear regression via random projections ensembles <i>Laura Anderlucci, Matteo Farnè, Giuliano Galimberti, Angela Montanari</i>	34
High-dimensional model-based clustering via random projections <i>Laura Anderlucci, Francesca Fortunato, Angela Montanari</i>	38
Multivariate outlier detection in high reliability standards fields using ICS <i>Aurore Archimbaud, Klaus Nordhausen, Anne Ruiz-Gazen</i>	42
Evaluating the school effect: adjusting for pre-test or using gain scores? <i>Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto, Carla Rampichini</i>	45
ACE, AVAS and robust data transformations <i>Anthony Atkinson</i>	49
Mixtures of multivariate leptokurtic Normal distributions <i>Luca Bagnato, Antonio Punzo, Maria Grazia Zoia</i>	53
Detecting and interpreting the consensus ranking based on the weighted Kemeny distance <i>Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio</i>	57
Predictive principal components analysis <i>Simona Balzano, Maja Bozic, Laura Marcis, Renato Salvatore</i>	61
Flexible model-based trees for count data <i>Federico Banchelli</i>	63
Euclidean distance as a measure of conformity to Benford's law in digital analysis for fraud detection <i>Mateusz Baryła, Józef Pocięcha</i>	67
The evolution of the purchase behavior of sparkling wines in the Italian market <i>Francesca Bassi, Fulvia Pennoni, Luca Rossetto</i>	71
Modern likelihood-frequentist inference at work <i>Ruggero Bellio, Donald A. Pierce</i>	75
Ontology-based classification of multilingual corpuses of documents <i>Sergey Belov, Salvatore Ingrassia, Zoran Kalinić, Paweł Lula</i>	79
Modeling heterogeneity in clustered data using recursive partitioning <i>Moritz Berger, Gerhard Tutz</i>	83

Mixtures of experts with flexible concomitant covariate effects: a bayesian solution <i>Marco Berrettini, Giuliano Galimberti, Thomas Brendan Murphy, Saverio Ranciati</i>	87
Sampling properties of an ordinal measure of interrater absolute agreement <i>Giuseppe Bove, Pier Luigi Conti, Daniela Marella</i>	91
Tensor analysis can give better insight <i>Rasmus Bro</i>	95
A boxplot for spherical data <i>Davide Buttarazzi, Giuseppe Pandolfo, Giovanni C. Porzio, Christophe Ley</i>	97
Machine learning models for forecasting stock trends <i>Giacomo Camba, Claudio Conversano</i>	99
Tree modeling ordinal responses: CUBREMOT and its applications <i>Carmela Cappelli, Rosaria Simone, Francesca Di Iorio</i>	103
Supervised learning in presence of outliers, label noise and unobserved classes <i>Andrea Cappozzo, Francesca Greselin, Thomas Brendan Murphy</i>	104
Asymptotics for bandwidth selection in nonparametric clustering <i>Alessandro Casa, José E. Chacón, Giovanna Menardi</i>	108
Foreign immigration and pull factors in Italy: a spatial approach <i>Oliviero Casacchia, Luisa Natale, Francesco Giovanni Truglia</i>	112
Dimensionality reduction via hierarchical factorial structure <i>Carlo Cavicchia, Maurizio Vichi, Giorgia Zaccaria</i>	116
Likelihood-type methods for comparing clustering solutions <i>Luca Coraggio, Pietro Coretto</i>	120
Labour market analysis through transformations and robust multilevel models <i>Aldo Corbellini, Marco Magnani, Gianluca Morelli</i>	124
Modelling consumers' qualitative perceptions of inflation <i>Marcella Corduas, Rosaria Simone, Domenico Piccolo</i>	128
Noise resistant clustering of high-dimensional gene expression data <i>Pietro Coretto, Angela Serra, Roberto Tagliaferri</i>	132
Classify X-ray images using convolutional neural networks <i>Federica Crobu, Agostino Di Ciaccio</i>	136

A compositional analysis approach assessing the spatial distribution of trees in Guadalajara, Mexico <i>Marco Antonio Cruz, Maribel Ortego, Elisabet Roca</i>	140
Joining factorial methods and blockmodeling for the analysis of affiliation networks <i>Daniela D'Ambrosio, Marco Serino, Giancarlo Ragozini</i>	142
A latent space model for clustering in multiplex data <i>Silvia D'Angelo, Michael Fop</i>	146
Post processing of two dimensional road profiles: variogram scheme application and sectioning procedure <i>Mauro D'Apuzzo, Rose-Line Spacagna, Azzurra Evangelisti, Daniela Santilli, Vittorio Nicolosi</i>	150
A new approach to preference mapping through quantile regression <i>Cristina Davino, Tormod Naes, Rosaria Romano, Domenico Vistocco</i>	154
On the robustness of the cosine distribution depth classifier <i>Houyem Demni, Amor Messaoud, Giovanni C. Porzio</i>	158
Network effect on individual scientific performance: a longitudinal study on an Italian scientific community <i>Domenico De Stefano, Giuseppe Giordano, Susanna Zaccarin</i>	162
Penalized vs constrained maximum likelihood approaches for clusterwise linear regression modelling <i>Roberto Di Mari, Stefano Antonio Gattone, Roberto Rocci</i>	166
Local fitting of angular variables observed with error <i>Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor</i>	170
Quantile composite-based path modeling to estimate the conditional quantiles of health indicators <i>Pasquale Dolce, Cristina Davino, Stefania Taralli, Domenico Vistocco</i>	174
AUC-based gradient boosting for imbalanced classification <i>Martina Dossi, Giovanna Menardi</i>	178
How to measure material deprivation? A latent Markov model based approach <i>Francesco Dotto</i>	182
Decomposition of the interval based composite indicators by means of biclustering <i>Carlo Drago</i>	186
Consensus clustering via pivotal methods <i>Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli</i>	190

Robust model-based clustering with mild and gross outliers <i>Alessio Farcomeni, Antonio Punzo</i>	194
Gaussian processes for curve prediction and classification <i>Sara Fontanella, Lara Fontanella, Rosalba Ignaccolo, Luigi Ippoliti, Pasquale Valentini</i>	198
A new proposal for building immigrant integration composite indicator <i>Mario Fordellone, Venera Tomaselli, Maurizio Vichi</i>	199
Biodiversity spatial clustering <i>Francesca Fortuna, Fabrizio Maturo, Tonio Di Battista</i>	203
Skewed distributions or transformations? Incorporating skewness in a cluster analysis <i>Michael Gallaughier, Paul McNicholas, Volodymyr Melnykov, Xuwen Zhu</i>	207
Robust parsimonious clustering models <i>Luis Angel Garcia-Escudero, Agustin Mayo-Isacar, Marco Riani</i>	208
Projection-based uniformity tests for directional data <i>Eduardo García-Portugués, Paula Navarro-Esteban, Juan Antonio Cuesta-Albertos</i>	212
Graph-based clustering of visitors' trajectories at exhibitions <i>Martina Gentilin, Pietro Lovato, Gloria Menegaz, Marco Cristani, Marco Minozzo</i>	214
Symmetry in graph clustering <i>Andreas Geyer-Schulz, Fabian Ball</i>	218
Bayesian networks for the analysis of entrepreneurial microcredit: evidence from Italy <i>Lorenzo Giammei, Paola Vicard</i>	222
The PARAFAC model in the maximum likelihood approach <i>Paolo Giordani, Roberto Rocci, Giuseppe Bove</i>	226
Structure discovering in nonparametric regression by the GRID procedure <i>Francesco Giordano, Soumendra Nath Lahiri, Maria Lucia Parrella</i>	230
A microblog auxiliary part-of-speech tagger based on bayesian networks <i>Silvia Golia, Paola Zola</i>	234
Recent advances in model-based clustering of high dimensional data <i>Isobel Claire Gormley</i>	238
Tree embedded linear mixed models <i>Anna Gottard, Leonardo Grilli, Carla Rampichini, Giulia Vannucci</i>	239

Weighted likelihood estimation of mixtures <i>Luca Greco, Claudio Agostinelli</i>	243
A canonical representation for multiblock methods <i>Mohamed Hanafi</i>	247
An adequacy approach to estimating the number of clusters <i>Christian Hennig</i>	251
Classification with weighted compositions <i>Karel Hron, Julie Rendlova, Peter Filzmoser</i>	255
MacroPCA: an all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers <i>Mia Hubert, Peter J. Rousseeuw, Wannes Van den Bossche</i>	256
Marginal effects for comparing groups in regression models for ordinal outcome when uncertainty is present <i>Maria Iannario, Claudia Tarantola</i>	258
A multi-criteria approach in a financial portfolio selection framework <i>Carmela Iorio, Giuseppe Pandolfo, Roberta Siciliano</i>	262
Clustering of trajectories using adaptive distances and warping <i>Antonio Irpino, Antonio Balzanella</i>	266
Sampling and learning Mallows and generalized Mallows models under the Cayley distance: short paper <i>Ekhine Irurozki, Borja Calvo, Jose A. Lozano</i>	270
The gender parity index for the academic students progress <i>Aglaia Kalamatianou, Adele H. Marshall, Mariangela Zenga</i>	274
Some asymptotic properties of model selection criteria in the latent block model <i>Christine Keribin</i>	278
Invariant concept classes for transcriptome classification <i>Hans Kestler, Robin Szekely, Attila Klimmek, Ludwig Lausser</i>	282
Clustering of ties defined as symbolic data <i>Luka Kronegger</i>	283
Application of data mining in the housing affordability analysis <i>Viera Labudová, Eubica Sipková</i>	284
Cylindrical hidden Markov fields <i>Francesco Lagona</i>	288

Comparing tree kernels performances in argumentative evidence classification <i>Davide Liga</i>	292
Recent advancement in neural network analysis of biomedical big data <i>Pietro Liò, Giovanna Maria Dimitri, Chiara Sopegno</i>	296
Bias reduction for estimating functions and pseudolikelihoods <i>Nicola Lunardon</i>	297
Large scale social and multilayer networks <i>Matteo Magnani</i>	301
Uncertainty in statistical matching by BNs <i>Daniela Marella, Paola Vicard, Vincenzina Vitale</i>	305
Evaluating the recruiters' gender bias in graduate competencies <i>Paolo Mariani, Andrea Marletta</i>	309
Dynamic clustering of network data: a hybrid maximum likelihood approach <i>Maria Francesca Marino, Silvia Pandolfi</i>	313
Stability of joint dimension reduction and clustering <i>Angelos Markos, Michel Van de Velden, Alfonso Iodice D'Enza</i>	317
Hidden Markov models for clustering functional data <i>Andrea Martino, Giuseppina Guatteri, Anna Maria Paganoni</i>	321
Composite likelihood inference for simultaneous clustering and dimensionality reduction of mixed-type longitudinal data <i>Antonello Maruotti, Monia Ranalli, Roberto Rocci</i>	325
Bivariate semi-parametric mixed-effects models for classifying the effects of Italian classes on multiple student achievements <i>Chiara Masci, Francesca Ieva, Tommaso Agasisti, Anna Maria Paganoni</i>	329
Multivariate change-point analysis for climate time series <i>Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio, Carlo Blasi</i>	333
A dynamic stochastic block model for longitudinal networks <i>Catherine Matias, Tabea Rebafka, Fanny Villers</i>	337
Unsupervised fuzzy classification for detecting similar functional objects <i>Fabrizio Mauro, Francesca Fortuna, Tonio Di Battista</i>	339
Mixture modelling with skew-symmetric component distributions <i>Geoffrey McLachlan</i>	343

New developments in applications of pairwise overlap <i>Volodymyr Melnykov, Yana Melnykov, Domenico Perrotta, Marco Riani, Francesca Torti, Yang Wang</i>	344
Modelling unobserved heterogeneity of ranking data with the bayesian mixture of extended Plackett-Luce models <i>Cristina Mollica, Luca Tardella</i>	346
Issues in nonlinear time series modeling of European import volumes <i>Gianluca Morelli, Francesca Torti</i>	350
Gaussian parsimonious clustering models with covariates and a noise component <i>Keefe Murphy, Thomas Brendan Murphy</i>	352
Illumination in depth analysis <i>Stanislav Nagy, Jiří Dvořák</i>	353
Copula-based non-metric unfolding on augmented data matrix <i>Marta Nai Ruscone, Antonio D'Ambrosio</i>	357
A statistical model for software releases complexity prediction <i>Marco Ortu, Giuseppe Destefanis, Roberto Tonelli</i>	361
Comparison of serious diseases mortality in regions of V4 <i>Viera Pacáková, Lucie Kopecká</i>	365
Price and product design strategies for manufacturers of electric vehicle batteries: inferences from latent class analysis <i>Friederike Paetz</i>	369
A Mahalanobis-like distance for cylindrical data <i>Lucio Palazzo, Giovanni C. Porzio, Giuseppe Pandolfo</i>	373
Archetypes, prototypes and other types <i>Francesco Palumbo, Giancarlo Ragozini, Domenico Vistocco</i>	377
Generalizing the skew-t model using copulas <i>Antonio Parisi, Brunero Liseo</i>	381
Contamination and manipulation of trade data: the two faces of customs fraud <i>Domenico Perrotta, Andrea Cerasa, Lucio Barabesi, Mario Menegatti, Andrea Cerioli</i>	385
Bayesian clustering using non-negative matrix factorization <i>Michael Porter, Ketong Wang</i>	389

Exploring gender gap in international mobility flows through a network analysis approach <i>Ilaria Primerano, Marialuisa Restaino</i>	393
Clustering two-mode binary network data with overlapping mixture model and covariates information <i>Saverio Ranciati, Veronica Vinciotti, Ernst C. Wit, Giuliano Galimberti</i>	395
A stochastic blockmodel for network interaction lengths over continuous time <i>Riccardo Rastelli, Michael Fop</i>	399
Computationally efficient inference for latent position network models <i>Riccardo Rastelli, Florian Maire, Nial Friel</i>	403
Clustering of complex data stream based on barycentric coordinates <i>Parisa Rastin, Basarab Matei, Guénaél Cabanes</i>	407
An INDSCAL based mixture model to cluster mixed-type of data <i>Roberto Rocci, Monia Ranalli</i>	411
Topological stochastic neighbor embedding <i>Nicoleta Rogovschi, Nistor Grozavu, Basarab Matei, Younès Bennani, Seiichi Ozawa</i>	415
Functional data analysis for spatial aggregated point patterns in seismic science <i>Elvira Romano, Jonatan González Monsalve, Francisco Javier Rodríguez Cortés, Jorge Mateu</i>	419
ROC curves with binary multivariate data <i>Lidia Sacchetto, Mauro Gasparini</i>	420
Silhouette-based method for portfolio selection <i>Marco Scaglione, Carmela Iorio, Antonio D'Ambrosio</i>	424
Item weighted Kemeny distance for preference data <i>Mariangela Sciandra, Simona Buscemi, Antonella Plaia</i>	428
A fast and efficient modal EM algorithm for Gaussian mixtures <i>Luca Scrucca</i>	432
Probabilistic archetypal analysis <i>Sohan Seth</i>	436
Multilinear tests of association between networks <i>Daniel K. Sewell</i>	438

Use of multi-state models to maximise information in pressure ulcer prevention trials <i>Linda Sharples, Isabelle Smith, Jane Nixon</i>	442
Partial least squares for compositional canonical correlation <i>Violetta Simonacci Massimo Guarino, Michele Gallo</i>	445
Dynamic modelling of price expectations <i>Rosaria Simone, Domenico Piccolo, Marcella Corduas</i>	449
Towards axioms for hierarchical clustering of measures <i>Philipp Thomann, Ingo Steinwart, Nico Schmid</i>	453
Influence of outliers on cluster correspondence analysis <i>Michel Van de Velden, Alfonso Iodice D'Enza, Lisa Schut</i>	454
Earthquake clustering and centrality measures <i>Elisa Varini, Antonella Peresan, Jiancang Zhuang</i>	458
Co-clustering high dimensional temporal sequences summarized by histograms <i>Rosanna Verde, Antonio Irpino, Antonio Balzanella</i>	462
Statistical analysis of item pre-knowledge in educational tests: latent variable modelling and optimal statistical decision <i>Chen Yunxiao, Lu Yan, Iriini Moustaki</i>	466
Evaluation of the web usability of the University of Cagliari portal: an eye tracking study <i>Gianpaolo Zammarchi, Francesco Mola</i>	468
Application of survival analysis to critical illness insurance data <i>David Zapletal, Lucie Kopecka</i>	472

COMPOSITE LIKELIHOOD INFERENCE FOR SIMULTANEOUS CLUSTERING AND DIMENSIONALITY REDUCTION OF MIXED-TYPE LONGITUDINAL DATA

Antonello Maruotti^{1,2}, Monia Ranalli³ 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, 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 $\mathbf{x}_t = [x_1, \dots, x_O]'$ and $\mathbf{y}_t^{\bar{O}} = [y_{O+1}, \dots, y_P]'$ be O ordinal and $\bar{O} = P - O$ continuous variables, respectively, with $t = 1, \dots, T$. The associated categories for each ordinal variable are denoted by $c_i = 1, 2, \dots, C_i$ with $i = 1, 2, \dots, O$. Following the URV approach, the ordinal variables \mathbf{x} are considered as a categorization of a continuous multivariate latent variable $\mathbf{y}_t^O = [y_1, \dots, y_O]'$. We assume that the temporal evolution of these data is driven by a multinomial process in discrete time $\boldsymbol{\xi}_{1:T} = (\boldsymbol{\xi}_t, t = 1, \dots, T)$, where $\boldsymbol{\xi}_t = (\xi_{t1}, \dots, \xi_{tK})$ is a multinomial random variable with K classes. We specifically assume that such process is distributed as a homogeneous Markov chain, whose distribution, say $p(\boldsymbol{\xi}_{1:T}; \mathbf{p})$, 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 $h, k = 1, \dots, K$ 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 $\mathbf{y}^{\bar{O}}$ is explained by the threshold model, $x_i = c_i \Leftrightarrow \gamma_{c_{i-1}}^{(i)} \leq y_i < \gamma_{c_i}^{(i)}$, with $c_i = 1, \dots, C_i$ and where $-\infty = \gamma_0^{(i)} < \gamma_1^{(i)} < \dots < \gamma_{C_i-1}^{(i)} < \gamma_{C_i}^{(i)} = +\infty$ 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 $\mathbf{y}_t = [\mathbf{y}_t^{O'}, \mathbf{y}_t^{\bar{O}}]'$ given a particular point in time, say t and conditioning on $\boldsymbol{\xi}_t$, follows a partially observed multivariate normal, $f(\mathbf{y}_{nt} | \boldsymbol{\xi}_t) = \prod_{k=1}^K \phi_P(\mathbf{y}_{nt} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{\xi_{nkt}}$, 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 $\boldsymbol{\Psi} = \{\mathbf{p}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K, \boldsymbol{\Gamma}\} \in \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the parameter space. For a random i.i.d. sample of size N , $(\mathbf{x}_1, \mathbf{y}_1^{\bar{O}}), \dots, (\mathbf{x}_N, \mathbf{y}_N^{\bar{O}})$, the log-likelihood is

$$\ell(\boldsymbol{\Psi}) = \sum_{n=1}^N \log \left[\sum_{\boldsymbol{\xi}_{1:T}} p(\boldsymbol{\xi}_t, \mathbf{p}) \phi_{\bar{O}}(\mathbf{y}_{nt}^{\bar{O}} | \boldsymbol{\xi}_t, \boldsymbol{\mu}_k^{\bar{O}}, \boldsymbol{\Sigma}_k^{\bar{O}}) \pi_{nt}(\boldsymbol{\mu}_{n;k}^{O|\bar{O}}, \boldsymbol{\Sigma}_k^{O|\bar{O}}, \boldsymbol{\Gamma}, \boldsymbol{\xi}_t) \right], \quad (1)$$

where, with obvious notation

$$\pi_{nt}(\boldsymbol{\mu}_{n;k}^{O|\bar{O}}, \boldsymbol{\Sigma}_k^{O|\bar{O}}, \boldsymbol{\Gamma}, \boldsymbol{\xi}_t) = \int_{\gamma_{c_1-1}^{(1)}}^{\gamma_{c_1}^{(1)}} \dots \int_{\gamma_{c_O-1}^{(O)}}^{\gamma_{c_O}^{(O)}} \phi_O(\mathbf{u}_{nt}; \boldsymbol{\mu}_{n;k}^{O|\bar{O}}, \boldsymbol{\Sigma}_k^{O|\bar{O}}) d\mathbf{u}_{nt},$$

where $\pi_n(\boldsymbol{\mu}_{n;k}^{O|\bar{O}}, \boldsymbol{\Sigma}_k^{O|\bar{O}}, \boldsymbol{\gamma})$ is the conditional joint probability of response pattern $\mathbf{x}_{nt} = (c_1^{(1)}, \dots, c_O^{(O)})$ 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 $E(\tilde{\mathbf{y}}^Q | k) = \boldsymbol{\eta}_k$ and $\text{Cov}(\tilde{\mathbf{y}}^Q | k) = \boldsymbol{\Omega}_k$, 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: $E(\tilde{\mathbf{y}}^{\bar{Q}} | k) = \boldsymbol{\eta}_0$ and $\text{Cov}(\tilde{\mathbf{y}}^{\bar{Q}} | k) = \boldsymbol{\Omega}_0$. 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.

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.

