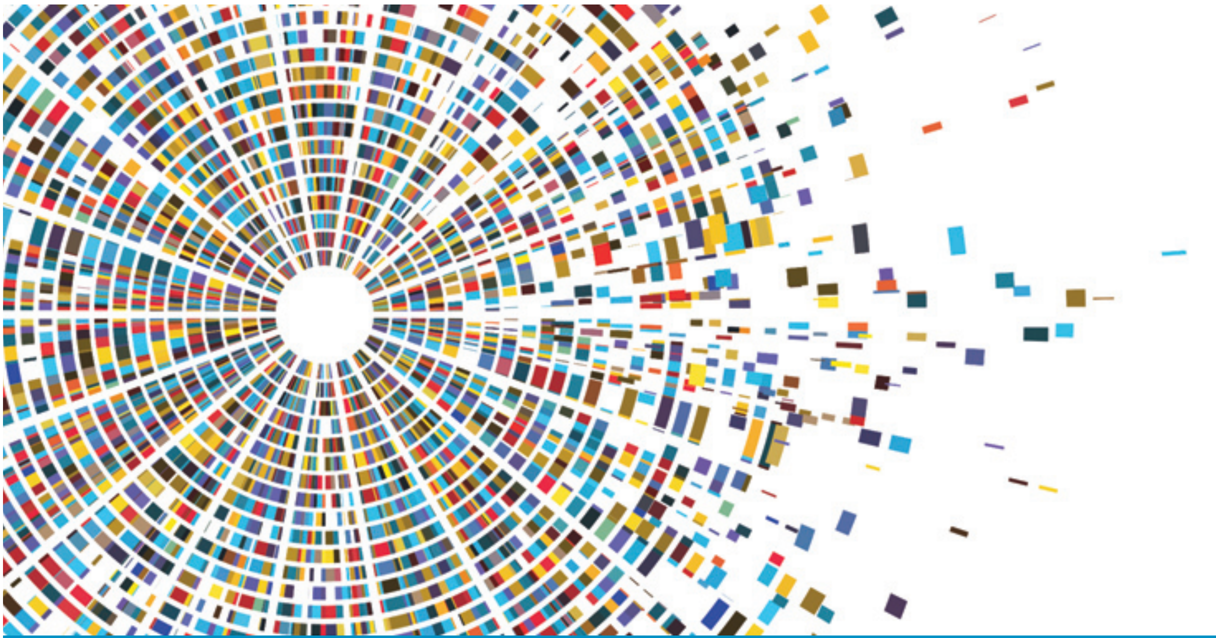

CLADAG 2023

BOOK OF ABSTRACTS AND SHORT PAPERS

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

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

Pearson

SCIENTIFIC PROGRAM COMMITTEE

LOCAL PROGRAM COMMITTEE

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

Front cover: Genome sequence map, chromosome architecture and genetic sequencing chart abstract data,

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

INDEX

Preface XVII

Plenary Session 1

Francesco Bartolucci, Michael Greenacre, Silvia Pandolfi and Fulvia Pennoni
Discrete latent variable models: recent advances and perspectives 3

Gerda Claeskens, Sarah Pirenne, Snigdha Panigrahi and Yiling Huang
Selective inference after variable selection by the randomized group Lasso method 7

Fancesca Greselin
To get the best, tame the beast: robust ML estimation for mixture models 8

Thomas Kneib
Rage against the mean - an introduction to distributional regression 12

Sofia Charlotta Olhede
On graph limits as models for interaction data 13

Invited Papers 15

Alessandro Albano, Mariangela Sciandra and Antonella Plaia
Ensemble method for text classification in medicine with multiple rare classes 17

Alessandro Albano, Mariangela Sciandra and Antonella Plaia
Distance-based aggregation and consensus for preference-approvals 21

Marco Alfò, Dimitris Pavlopoulos and Roberta Varriale
Flexible employment, a machine learning approach 25

Federico Ambrogi and Matteo Di Maso
Clinically useful measures in survival analysis: the restricted mean survival time as an alternative to the hazard ratio 29

Jose Ameijeiras-Alonso
Data-driven smoothing parameter selection for circular data analysis 33

Laura Anderlucci, Silvia Dallari and Angela Montanari
View it differently: finding groups in microbiome data 34

Rabea Aschenbruck, Gero Szepannek and Adalbert F. X. Wilhelm
Random-based initialization for clustering mixed-type data with the k-prototypes algorithm 38

Filippo Ascolani and Valentina Ghidini
Posterior clustering for Dirichlet process mixtures of Gaussians with constant data 42

Vincent Audigier and Ndèye Niang
Multiple imputation for clustering on incomplete data **46**

Alejandra Avalos-Pacheco and Roberta De Vito
Integrative factor models for biomedical applications **50**

Silvia Bacci, Bruno Bertaccini, Carla Galluccio, Leonardo Grilli and Carla Rampichini
Test equating with evolving latent ability **54**

Michela Baccini, Alessandra Mattei, Elena Degli Innocenti, Giulio Biscardi and Aitana Lertxundi
Causal inference on the impact of extreme ambient temperatures on population health **58**

Zsuzsa Bakk
Measurement invariance testing of latent class models using residual statistics and likelihood ratio test **61**

Falco J. Bargagli-Stoffi, Costanza Tortù and Laura Forastiere
Network interference and effect modification **65**

Francesco Barile, Simonón Lunagómez and Bernardo Nipoti
Flexible modelling of heterogeneous populations of networks: a Bayesian nonparametric approach **69**

Mario Beraha and Jim E. Griffin
Normalized latent measure factor models **70**

Silvia Bianconcini and Silvia Cagnone
Estimation issues in multivariate panel data **74**

Alessandro Bitetto and Paola Cerchiello
The nexus between ESG and initial coin offerings: evidence from text analysis **78**

Laura Bocci and Donatella Vicari
A clustering model for three-way asymmetric proximity data **82**

Ilaria Bombelli, Ichcha Manipur and Maria Brigida Ferraro
Cluster analysis for networks using a fuzzy approach **86**

Davide Buttarazzi and Giovanni C. Porzio
Visualizing anomalies in circular data **90**

Andrea Cappozzo, Chiara Masci, Francesca Ieva and Anna Maria Paganoni
Model-based clustering of right-censored lifetime data with frailties and random covariates **91**

Michelle Carey and Catherine Higgins
Clustering imbalanced functional data **95**

Alessandro Casa, Thomas Brendan Murphy and Michael Fop
Partial membership models for high-dimensional spectroscopy data **99**

Fabio Centofanti, Antonio Lepore and Biagio Palumbo
Sparse clustering for functional data **103**

Yunxiao Chen, Motonori Oka and Matthias von Davier
Interpretable and accurate scaling in large-scale assessment: a variable selection approach to latent regression **107**

Katharine M. Clark and Paul D. McNicholas
Clustering three-way data with outliers **111**

Roberto Colombi and Sabrina Giordano
A two-component markov switching regression model **115**

Federica Conte and Paola Paci
The broad phenotype-specific applications of the network-based SWIM tool **119**

Houyem Demni, Pierre Miasnikof, Alexander Y. Shestopaloff, Cristián Bravo and Yuri Lawryshyn
Testing graph clusterability: a density based statistical test for directed graphs **123**

Anna Denkowska, Krystian Szczf ôsny, Joao Paulo Vieito and Stanisław Wanat
Deep neural network in the modeling of the dependence structure in risk aggregation **124**

Marco Di Marzio, Chiara Passamonti and Charles Taylor
Circular regression with measurement errors **128**

Marco Di Zio, Romina Filippini, Gaia Rocchetti and Simona Toti
Classification tree to improve data quality in official statistics **132**

Rosa Fabbricatore and Maria Iannario
Uncertainty and response style in latent trait models to assess emotional intelligence of elite swimmers **136**

Rosa Fabbricatore, Roberto Di Mari, Zsuzsa Bakk, Mark de Rooij and Francesco Palumbo
Three-step rectangular latent Markov modeling based on ML correction **140**

Alessio Farcomeni, Alfonso Russo and Marco Geraci
Mid-quantile regression for discrete panel data **144**

Matteo Farnè
Trimmed factorial k-means **148**

Florian Felice and Christophe Ley
Estimation of team's strength for handball games predictions **152**

Peter Filzmoser and Marcus Mayrhofer
Outlier explanation based on Shapley values for vector- and matrix-valued observations **156**

Lara Fontanella, Emiliano del Gobbo and Alex Cucco
Identification of misogynistic accounts on Twitter through Graph Convolutional Networks **159**

Giacomo Francisci and Anand Vidyashankar
Depth functions for tree-indexed processes **163**

Carla Galluccio, Matteo Magnani, Davide Vega, Giancarlo Ragozini and Alessandra Petrucci
Analysing the effect of different design choices in network-based topic detection **164**

Luis A. García-Escudero, Christian Hennig, Agustín Mayo-Iscar, Gianluca Morelli and Marco Riani
A proposal for the joint automated detection of clusters and anomalies **168**

V. G. Genova, C. Edling, H. Mondani, A. M. Rostami and M. Tumminello
Mobility across crimes: statistically validated networks and temporal pattern recognition **172**

Paolo Giordani, Susanna Levantesi, Andrea Nigri and Virginia Zarulli
A cohort study on the gender gap in mortality through the Tucker3 model **176**

Luca Greco, Giovanna Menardi and Marco Rudelli
Trimmed kernel mean shift **180**

Bettina Grün, Thomas Petzoldt and Helga Wagner
Modeling zone diameter measurements to infer antibiotic susceptibility of bacteria **184**

Julien Jacques and Francesco Amato
Clustering longitudinal ordinal data **185**

Daniyal Kazempour and Peer Kröger
"You call it a manifold, I call it a subspace" - selected examples on the interface between computer science and statistics in the context of clustering and manifold learning **187**

Annika M. T. U. Kestler, Nensi Ikonomi, Silke D. Werle, Julian D. Schwab, Friedhelm Schwenker and Hans A. Kestler
Sparse rule generating fold-change classification for molecular high-throughput profiles **188**

Silvia Komara, Martina Košiková, Erik Šoltés and Tatiana Šoltésová
Comparison of the households' work intensity in Slovakia and Czechia through least squares means analysis based on GLM **192**

Arnost Komárek
Model based clustering procedures for multivariate mixed type longitudinal data **193**

Tomasz Kwarciński, Paweł Ulman
Inequality, populism, and unfairness: a comparison of unfair income inequalities in Poland and Norway 196

Francesco Lagona and Marco Mingione
Segmenting toroidal time series by nonhomogeneous hidden semi-Markov models 197

Roland Langrock and Sina Mews
How to build your latent Markov model: the role of time and space 201

Paweł Lula, Zsuzsanna Géring, Mńagdalena Talaga, Ildikó Dén-Nagy and Réka Tamássy
The comparative analysis of publication activity in Hungary and Poland in the field of economics, finance and business 205

Johan Lyrvall, Roberto Di Mari, Zsuzsa Bakk, Jennifer Oser and Jouni Kuha
An R package for multilevel latent class analysis with covariates 206

R. Neal Mackenzie and Paul D. McNicholas
Longitudinal hidden Markov models: problems and methods 210

Matteo Magnani, Matias Piqueras, Alexandra Segerberg, Davide Vega and Victoria Yantseva
Cluster analysis for the study of online visual communication 214

Ichcha Manipur, Ilaria Granata, Lucia Maddalena and Mario R. Guarracino
Cluster analysis of cancer metabolic network ensembles 218

Carlo Metta, Marco Fantozzi, Andrea Papini, Gianluca Amato, Matteo Bergamaschi, Silvia Giulia Galfrè, Alessandro Marchetti, Michelangelo Vegliò, Maurizio Parton and Francesco Morandin
Improving performance in neural networks by dendrite-activated connection 219

Rodolfo Metulini, Francesco Biancalani and Giorgio Gnecco
The Generalized Shapley measure for ranking players in basketball: applications and future directions 223

Rouven Michels, Timo Adam and Marius Ötting
Tree-based regression within a hidden Markov model framework 227

Boris Mirkin
Scoring distances between equivalence and preference relations 231

Fabio Morea and Domenico De Stefano
Evaluation of the performance of a modularity-based consensus community detection algorithm 234

Vincenzo Nardelli and Niccolò Salvini
Assessing and improving data quality in open spatial data: a case study with ANAC data 238

M. Rosário Oliveira, Diogo Pinheiro and Lina Oliveira
Visualizing interval Fisher Discriminant Analysis results 239

Niels Lundtorp Olsen, Alessia Pini and Simone Vantini
Nonparametric local inference for functional data defined on manifold domains 242

Silvia Pandolfi and Francesco Bartolucci
Case-control variational inference for large scale stochastic block models 246

Francesca Panero
Issues with sparse spatial random graphs 250

Barbara Pawełek and Maria Sadko
Corporate bankruptcy prediction: application of statistical learning methods 254

Daniele Pretolesi, Andrea Vian and Annalisa Barla
Using machine learning and AI in science of science 255

Pascal Préa
Distances, orders and spaces 259

Antonio Punzo, Luca Bagnato and Salvatore Daniele Tomarchio
Model-based clustering via parsimonious mixtures of dimension-wise scaled normal mixtures 263

Monia Ranalli and Roberto Rocci
Model-based simultaneous classification and reduction for three-way ordinal data 264

Jakob Raymaekers and Peter J. Rousseeuw
The cellwise Minimum Covariance Determinant estimator 268

Maurizio Romano and Roberta Siciliano
A new accurate heuristic algorithm to solve the rank aggregation problem with a large number of objects 269

Jorge Rueda, Maria del Mar Rueda, Ramón Ferri and Beatriz Cobo
Using ML techniques for estimation with non-probabilistic survey data 273

Ana Santos, Sónia Dias, Paula Brito and Paula Amaral
Multiclass classification of distributional data 276

Lorenzo Schiavon
Latent Bayesian clustering for topic modelling 280

Michael G. Schimek, Bastian Pfeifer and Marcus D. Bloice
A novel multi-view ensemble clustering framework for cancer subtype discovery 284

Francesco Schirripa Spagnolo, Gaia Bertarelli, Nicola Salvati, Donato Summa, Monica Scannapieco, Stefano Marchetti and Monica Pratesi
Reducing selection bias in non-probability sample by Small Area Estimation **288**

Pedro Duarte Silva, Peter Filzmoser and Paula Brito
Sparse and robust estimators for outlier detection in distributional data **292**

Andrea Sottosanti, Sara Agavni' Castiglioni, Stefania Pirrotta, Enrica Calura and Davide Risso
Clustering genes spatial expression profiles with the aid of external biological knowledge **296**

Arthur Tenenhaus, Michel Tenenhaus and Theo Dijkstra
Structural equation modeling with latent/emergent variables: RGCCAc **300**

Yoshikazu Terada
On some properties of reconstructed trajectories from sparse longitudinal data **301**

Daniel J.W. Touw, Patrick J.F. Groenen, Ines Wilms and Andreas Alfons
Clusterpath Gaussian graphical modeling **302**

Paweł Ulman, Małgorzata Ćwiek and Maria Sadko
Housing poverty in Europe. Multidimensional analysis **305**

Anand Vidyashankar, Fengnan Deng, Giacomo Francisci and Xiaoran Jiang
Efficiency and robustness in supervised learning **306**

Frédéric Vrins
Optimal and robust combination of forecasts via constrained optimization and shrinkage **307**

Gabriel Wallin, Yunxiao Chen and Irini Moustaki
DIF analysis with unknown groups and anchor items **308**

Felix M. Weidner, Mirko Rossini, Joachim Ankerhold and Hans A. Kestler
Constraint-based attractor search in Boolean networks using quantum computing **309**

Michio Yamamoto and Yoshikazu Terada
Clustering for sparsely sampled longitudinal data based on basis expansions **312**

Naoto Yamashita
Two extensions of extended redundancy analysis for exploratory data analysis **313**

Giorgia Zaccaria
Ultrametric Gaussian Mixture models with parsimonious structures **314**

Li-Chun Zhang
Using retail transactions for consumer price index and expenditure statistics **318**

Contributed Papers 323

Giuseppe Alfonzetti, Luca Grassetti and Laura Rizzi
Propensity towards Master's degree: choices of northern students after BAs? 325

Giuseppe Alfonzetti, Luca Grassetti and Laura Rizzi
Classifying northern Italian students in their transition to Master degree 329

Rosa Arboretti, Elena Barzizza, Nicolò Biasetton and Marta Disegna
Customer satisfaction through time: structured time series from sentiment analysis of TripAdvisor data 333

Roberto Ascari and Alice Giampino
A flexible topic model 334

Golnoosh Babaei, Paolo Pagnottoni and Thanh Thuy Do
Explainable machine learning for lending default classification 338

Elena Barzizza, Riccardo Ceccato, Solomon Harrar, Fortunato Pesarin and Luigi Salmaso
A multivariate permutation test for association 342

Michela Battauz
A competing risk analysis of academic careers with students' ability and speed as predictors 343

Andriette Bekker, J.T. Ferreira, J. Pillay and M. Arashi
Bayesian analysis for a graphical t-model 347

Marco Berrettini, Giuliano Galimberti, Thomas Brendan Murphy and Saverio Ranciati
Modelling soccer players field position via mixture of Gaussians with flexible weights 351

Antonella Bianchino, Daniela Fusco, Paola Giordano, Maria Antonietta Liguori, Maria Carmina Palma and Donato Summa
Tourism as support in economic development of inner areas: a multi-sources approach 355

Luisa Bisaglia and Francesco Lisi
SARIMA models with multiple seasonality 358

Stefano Bonnini and Michela Borghesi
Adoption of 4.0 technologies and related obstacles. Application of a multivariate nonparametric test for categorical variables 362

Giuseppe Bove
An application of asymmetric multidimensional scaling to the VQR 2015-2019 data 366

Luca Brusa and Fulvia Pennoni
Improving clustering in temporal networks through an evolutionary algorithm **370**

Andrea Carta
A support vector machine approach to create oblique decision trees for regression **374**

Giulia Cereda, Fabio Corradi and Cecilia Viscardi
Comparing soft classification methods for the rare type match problem **378**

Annalisa Cerquetti
Bayesian Shannon entropy estimation under normalized inverse Gaussian priors via Monte Carlo sampling **382**

Lax Chan and Aldo Goia
Goodness-of-fit test for single functional index model **386**

Silvia Columbu, Nicola Piras and Jeroen K. Vermunt
Multilevel cross-classified latent class models **390**

Giulia Contu, Luca Frigau, Marco Ortu and Sara Pau
Multivariate regression tree to investigate the Italian mortality rates **394**

Luca Coraggio and Pietro Coretto
Empirical analysis of the quadratic scoring for selecting clustering solutions **398**

Marcella Corduas and Domenico Piccolo
Classification of daily streamflow data: a study on regime changes **402**

Noemi Corsini and Giovanna Menardi
Modal clustering for categorical data **406**

Cristina Davino, Tormod Næs, Rosaria Romano and Domenico Vistocco
The use of principal components in quantile regression: a simulation study **410**

Antonio De Falco and Antonio Irpino
An interdisciplinary methodology for socio-economic segregation analysis **414**

Houyem Demni and Simona Balzano
Visualizing classification results: graphical tools for DD-classifiers **418**

Claudia Di Caterina
Detecting the positions of nonconsensus amino acids in HIV patients by marginal likelihood thresholding **419**

Davide Di Cecco, Andrea Tancredi and Tiziana Tuoto
One-inflated Bayesian mixtures for population size estimation **423**

Marta Di Lascio and Roberta Pappadà
Cluster analysis and conditional copula: a joint approach to analyse energy demand **427**

Marta Di Lascio, Fabrizio Durante and Aurora Gatto
Hierarchical percentile clustering to analyse greenhouse gas emissions from agriculture in European Union **431**

Cinzia Di Nuzzo and Salvatore Ingrassia
Maximum likelihood approach to parameter selection in the spectral clustering algorithm **435**

José G. Dias
Finite mixture models: a systematic review **439**

Francesco Dotto, Roberto Di Mari, Alessio Farcomeni and Antonio Punzo
Measurement invariance: a method based on latent Markov models **441**

Niccolò Ducci, Leonardo Grilli and Marta Pittavino
A comparison between the varying-thresholds model and quantile regression **445**

Augusto Fasano, Niccolò Anceschi, Beatrice Franzolini and Giovanni Rebaudo
Efficient computation of predictive probabilities in probit models via expectation propagation **449**

Donata Favaro and Anna Giraldo
How women react to their partners' work instability. The added-worker effect **453**

Carlina C. Feldmann, Sina Mews, Rouven Michels and Roland Langrock
Inference on the state distribution in periodic hidden Markov models **457**

Giuseppe Feo, Francesco Giordano, Marcella Niglio, Sara Milito and Maria Lucia Parrella
Testing clusters of locations in spatial dynamic panel data models **461**

Beatrice Franzolini, Laura Bondi, Augusto Fasano and Giovanni Rebaudo
Bayesian forecasting of multivariate longitudinal zero-inflated counts: an application to civil conflict **465**

Francesco Freni and Giovanna Menardi
Efficient disentangling γ-ray sources from diffuse background in the sky map **469**

Luca Frigau, Giulia Contu, Marco Ortu and Andrea Carta
A method to validate clustering partitions **473**

Flora Fullone, Gianmarco Farina, Enza Compagnone, Mirella Morrone and Gioacchino de Candia
Analysis of the need for working timber starting from Istat industrial production data **477**

Ravi Kumar Gangadharan, Vanessa Petrarca, Maria Chiara Pagliarella and Giovanni C. Porzio
Stratified sampling on data nuggets: a strategy for data reduction **481**

Ewa Genge
Is the subjective financial well-being of Polish families changing with time? An empirical study based on constrained latent Markov models **482**

Sara Geremia, Fabio Morea and Domenico De Stefano
Visualization of proximity and role-based embedding in a regional labour flow network **486**

Massimiliano Giacalone, Vincenzo Dottorini, Giuseppe Oddo, Vito Santarcangelo and Angelo Romano
Method for the quality control and operators training in maintenance activities **490**

Lorenzo Giammei, Flaminia Musella, Fulvia Mecatti and Paola Vicard
Building improved gender equality composite indicators by object-oriented Bayesian networks **494**

Sabrina Giordano, Roberta Varriale and Mariangela Zenga
A comparative study of financial literacy using data from PISA survey **498**

Natalia Golini, Francesca Martella and Antonello Maruotti
On model-based clustering for equitable and sustainable well-being at local level: how many Italies? **499**

Luca Greco, Antonio Lucadamo and Claudio Agostinelli
Model-based clustering for torus data **503**

Giulio Grossi and Emilia Rocco
AutoSynth index: a synthetic indicator for socio-economic development based on autoencoders **507**

Lucia Guastadisegni, Irini Moustaki, Silvia Cagnone and Vassilis Vasdekis
A statistical test to assess the non-normality of the latent variable distribution **511**

Christian Hennig and Keefe Murphy
Quantifying variable importance in cluster analysis **515**

Mia Hubert, Iwein Vranckx, Jakob Raymaekers, Bart De Ketelaere and Peter Rousseeuw
Real-time discriminant analysis in the presence of label and measurement noise **519**

Carmela Iorio, Giuseppe Pandolfo and Antonio D'Ambrosio
A proposal to evaluate the solution of a fuzzy clustering algorithm **520**

Aazm Kheyri, Andriette Bekker and Mohammad Arashi
A fused-type elastic net Gaussian graphical model for paired data **524**

Amir Khorrami Chokami
Complete records over independent FGM sequences **528**

Ursula Laa and Dianne Cook
New tour methods for visualizing high-dimensional data **532**

Michele Lambardi di San Miniato, Michela Battauz, Ruggero Bellio and Paolo Vidoni
Bayesian aggregation of crowd judgments for quantitative fact checking **536**

Salvatore Latora and Luigi Augugliaro
Supervised classification of curves by functional data analysis: an application to neuromarketing data **540**

Gertraud Malsiner-Walli, Bettina Grün and Sylvia Frühwirth-Schnatter
Capturing correlated clusters using mixtures of latent class models **544**

Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore
A three-way "indirect" redundancy analysis **545**

Maria Francesca Marino, Matteo Sani and Monia Lupparelli
Multi-level stochastic blockmodels for multiplex networks **549**

Francesca Martella, Xiaoke Qin, Wangshu Tu and Sanjena Subedi
The multivariate cluster-weighted disjoint factor analyzers model **553**

Raffaele Mattera, Germana Scepi, Pooria Ebrahimi and Fabio Matano
Spatial modelling of pyroclastic cover deposit thickness with remote sensing data and ground measurements: a forecasting combination approach **557**

Fiammetta Menchetti
Granger network on Santa Maria del Fiore Dome **561**

Giuseppe Mignemi, Ioanna Manolopoulou and Antonio Calcagnì
Group's heterogeneity in rating tasks: a Bayesian semi-parametric approach **565**

Dung Ngoc Nguyen and Alberto Roverato
Lattice of Gaussian graphical models for paired data with common undirected structure **569**

Marco Ortu, Giulia Contu and Luca Frigau
Multivariate regression tree topic modeling **573**

Lucio Palazzo, Alfonso Iodice D'Enza, Francesco Palumbo and Domenico Vistocco
Dendrogram slicing through a permutation test approach reconsidered **577**

Roberta Paroli and Luigi Spezia
Markov switching autoregressive models for the analysis of hydrological time series **581**

Davide Passaro, Luca Tardella, Giovanna Jona Lasinio, Tiziana Fragasso, Valeria Raggi and Zaccaria Ricci
A case study of electronic medical records use for predicting kidney injury **585**

Matteo Pedone, Raffaele Argiento and Francesco C. Stingo
Personalized treatment selection model for survival outcomes **589**

Danilo Petti, Marcella Niglio and Marialuisa Restaino
Variable ranking in bivariate copula survival models **593**

Pia Pfeiffer and Peter Filzmoser
Robust penalized multivariate analysis for high-dimensional data **597**

Francesco Porro
Structural zeros in regression models with compositional explanatory variables **600**

Kemmawadee Preedalikit, Daniel Fernández, Ivy Liu, Louise McMillan, Marta Nai Ruscone and Roy Costilla
One-dimensional mixture-based clustering for ordinal responses **604**

Iuliia Promskaia, Adrian O'Hagan and Michael Fop
A compositional stochastic block model for the analysis of the Erasmus programme network **608**

Claudia Rampichini and Maria Brigida Ferraro
A proposal of deep fuzzy clustering by means of the simultaneous approach **609**

Maria Giovanna Ranalli, Fulvia Pennoni, Francesco Bartolucci and Antonietta Mira
When nonresponse makes estimates from a census a small area estimation problem: the case of the survey on Graduates' Employment Status in Italy **613**

Edoardo Redivo and Cinzia Viroli
A supervised classification strategy based on the novel directional distribution depth function **617**

Ilaria Rocco
An application of CART algorithm to administrative data: analysis of youth initial employment trajectories **621**

Dorota Rozmus
Resampling for stability estimation vs. cluster validation via data splitting and subsampling. Which approach is better in detection of clusters in taxonomy? **625**

Annalina Sarra, Adelia Evangelista, Tonio Di Battista, and Sergio Palermi
Functional data analysis approach for identifying redundancy in air quality monitoring stations **627**

Luca Scaffidi Domianello
Student mobility in higher education: a destination-specific local analysis **631**

Rosaria Simone
Residuals diagnostics for model-based trees for ordered rating responses **635**

Alexa Sochaniwsky and Paul D. McNicholas
Hidden Markov models for multivariate longitudinal data **639**

Andrzej Sokołowski, Małgorzata Markowska and Maciej Laburda
K-means clustering - new variations **643**

Daniele Spinelli, Salvatore Ingrassia and Giorgio Vittadini
A Stata implementation of cluster weighted models: the CWMGLM package **644**

Salvatore D. Tomarchio, Antonio Punzo and Antonello Maruotti
Matrix-variate hidden Markov regressions **648**

Cristian Usala, Isabella Sulis and Mariano Porcu
Inequalities at entrance, labour market conditions and university dropout: first evidence from Italy **652**

Rosanna Verde, Gianmarco Borrata and Antonio Balzanella
A clustering method for distributional data based on a LDQ transformation **656**

Helga Wagner and Roman Pfeiler
Shrinkage of time-varying effects in panel data models **657**

Carlo Zaccardi, Pasquale Valentini and Luigi Ippoliti
A Bayesian spatio-temporal regression approach for confounding adjustment **661**

Gianpaolo Zammarchi
Linear random forest to predict energy consumption **665**

Preface

This book collects the abstracts and short papers presented at CLADAG 2023, the 14th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS). The meeting has been organized by the Department of Economics and Statistics of the University of Salerno, under the auspices of the University of Salerno, the SIS and the International Federation of Classification Societies (IFCS).

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

The Scientific Committee of CLADAG 2023 conceived the Keynote Sessions to provide a fresh perspective on the state of the art of knowledge and research in the field. The scientific program of CLADAG 2023 is particularly rich. All in all, it comprises 5 Keynote Lectures, 31 Invited Sessions promoted by the members of the Scientific Program Committee, and 27 Contributed Sessions. We thank all the session organizers for inviting renowned speakers, coming from many different countries. We are greatly indebted to the referees, for the time spent in a careful review of the abstracts and short papers collected in this book. Special thanks are finally due to the members of the Local Organizing Committee and all the people who collaborated for CLADAG 2023. Last but not least, we thank all the authors and participants, without whom the conference would not have been possible.

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

Salerno, September 2023

A Three-Way "indirect" Redundancy Analysis

Laura Marcis [1], Maria Chiara Pagliarella[1] and Renato Salvatore [1]

[1] Department of Economics and Law, University of Cassino and Southern Lazio, (e-mail: laura.marcis@unicas.it, mc.pagliarella@unicas.it, rsalvatore@unicas.it)

Abstract: This work introduces a composite Three-Way application of the High Order Singular Value Decomposition. Two of the three component data matrices are processed by a standard Redundancy Analysis. The remaining "external" data matrix is related to the others in a heterogeneous system of relations, that can be well suited to tensor analysis. The external data are set to be linked with the first matrix, while with the second matrix the relations are explained only through multivariate linear regression. An application introduces the method, based on the official data from the Italian Equitable and Sustainable Well-being indicators.

Keywords: Tucker decomposition, high order singular value decomposition, redundancy analysis.

1 Introduction and background

Tensor decomposition (Kolda & Bader, 2009) has the main objective of reducing complex information detected by higher dimensional arrays of data. From a pure statistical perspective, there are two important exploitations of the tensor analysis: the Candecomp/Parafac decomposition and the Tucker decomposition. They play the role of the extension to tensor objects of the principal component analysis (PCa), recognized as an explorative way to approach multidimensional information (Kroonenberg, 2008). In the literature, the most popular tensor decompositions are "Canonical Decomposition" and the "High Order SVD" (HOSVD, De Lathauwer *et al.*, 2000). The HOSVD decomposes an N-mode tensor, as a multidimensional array, in a core reduced-order tensor, multiplied by component matrices alongside each of the N modes. Three-way PCa was the first extension of the PCa to a three-way data set, giving the first useful employment of tensor analysis to explorative statistical analysis. In standard PCa, the components that come from the SVD that summarize individuals are uniquely related to the components that summarize variables. In a three-way PCa the components that summarize entities in each of the modes are related with the remaining two. Redundancy Analysis (RDA, Legendre and Legendre, 2012) was originally introduced in order to capture the effect onto a reduced space $\widehat{\mathbf{Y}}_X = \mathbf{X}\widehat{\mathbf{B}}$ of the linear dependence by a set of criterion variables $\mathbf{Y}$ from a set of predictors $\mathbf{X}$, where $\widehat{\mathbf{B}}$ is the matrix of the ordinary least squares multivariate regression estimates. RDA provides a constrained analysis of the whole linear relations between the two sets of variables, and an unconstrained analysis given by the set of multivariate regression residuals. It can be considered as an extension of multivariate regression because models the effects of the explanatory variables on a response matrix. Partial RDA (pRDA) explores the effects of the predictors in $\mathbf{X}$ on the $\mathbf{Y}$ variables, given the covariates of some additional exploratory variables in a matrix $\mathbf{Z}$. It is a standard RDA performed taking into account the $\mathbf{X}$ variables as predictors on $\mathbf{Y} - \widehat{\mathbf{Y}}_Z$, with the "effect" by $\mathbf{Z}$ removed. Nevertheless, the relations between the variables $\mathbf{Y}$ and $\mathbf{Z}$ may be quite several. While remaining the same the role of the predictors $\mathbf{X}$ on $\mathbf{Y}$, a third set of variables $\mathbf{Z}$ may be related and depend on $\mathbf{Y}$, by an existing but not well defined dependence. Thus, applying multivariate regression may result hardly appropriate. Variables in $\mathbf{Z}$ in some cases can not be modeled on $\mathbf{Y}$ as predictors in a multivariate regression, while $\mathbf{X}$ predict $\mathbf{Y}$ and, indirectly through $\mathbf{Y}$, the variables in $\mathbf{Z}$. Residuals $\mathbf{Y} - \widehat{\mathbf{Y}}_X$ may take in account the role of $\mathbf{X}$ in the "indirect" explanation of $\mathbf{Z}$. This is somewhat different from pRDA, because $\mathbf{Y}$ is not regressed on $\mathbf{Z}$, as the external set of covariates from which we remove the effect on $\mathbf{Y}$, and also $\mathbf{Z}$ is not related with $\mathbf{Y}$ through linear regression. Given a 3rd-order tensor $\mathcal{X} \in \mathbb{R}^{I\times J\times K}$, the Tucker decomposition through the HOSVD decomposes the tensor $\mathcal{X}$ into a core tensor $\mathcal{G}$ and factor matrices along each mode, as follows:

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}$$

with the correspondent elementwise expression $x_{ijk} = \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{t=1}^{T} g_{rst}a_{ir}b_{js}c_{kt}$, with $i = 1,...,I, j = 1,...,J, k = 1,...,K$. The factor matrices are columnwise orthonormal, $\mathbf{A} = [\mathbf{a}_1,...,\mathbf{a}_R], \mathbf{B} = [\mathbf{b}_1,...,\mathbf{b}_S], \mathbf{C} = [\mathbf{c}_1,...,\mathbf{c}_T]$, with $r = 1,...,R, s = 1,...,S, t = 1,...,T$. The matricized forms, one per mode, of the 3-way tensor $\mathcal{X}$ are:

$$\begin{aligned}
\mathbf{X}_{(1)} &\approx \mathbf{A}(\mathbf{C}\odot\mathbf{B})' = \mathbf{A}\mathbf{G}_{(1)}(\mathbf{C}\otimes\mathbf{B})',\\
\mathbf{X}_{(2)} &\approx \mathbf{B}(\mathbf{C}\odot\mathbf{A})' = \mathbf{B}\mathbf{G}_{(2)}(\mathbf{C}\otimes\mathbf{A})',\\
\mathbf{X}_{(3)} &\approx \mathbf{C}(\mathbf{B}\odot\mathbf{A})' = \mathbf{C}\mathbf{G}_{(3)}(\mathbf{B}\otimes\mathbf{A})',
\end{aligned}$$

with the symbols $\odot$ and $\otimes$ that are the Khatri-Rao and Kronecker products, respectively. If $r_R(\mathcal{X})$ is the rank of the tensor $\mathcal{X}$ alongside one of the modes,

Table 1. *Description of the variables used for the application*

Variables	Description
S8	Age-standardised mortality rate for dementia and nervous system diseases
IF3	People having completed tertiary education (30-34 years old)
L12	Share of employed persons who feel satisfied with their work
REL4	Social participation
POL5	Trust in other institutions like the police and the fire brigade
SIC1	Homicide rate
BS3	Positive judgement for future perspectives
PATR9	Presence of Historic Parks/Gardens and other Urban Parks recognised of significant public interest
AMB9	Satisfaction for the environment - air, water, noise
INN1	Percentage of R&D expenditure on GDP
Q2	Children who benefited of early childhood services
BE1	Per capita adjusted disposable income
LBE1	Logarithm of Per capita adjusted disposable income

the HOSVD may uses Alternating Least Squares, in order to find:

$$\min_{\mathcal{G},\mathbf{A},\mathbf{B},\mathbf{C}} \left\|\mathcal{X} - \mathcal{G}\times_1\mathbf{A}\times_2\mathbf{B}\times_3\mathbf{C}\right\|.$$

Making the substitutions $\mathbf{A} = \mathbf{Y}$, $\mathbf{B} = \mathbf{Y} - \widehat{\mathbf{Y}}_X$, $\mathbf{C} = \mathbf{Z}$, with $I = J = K = n$, $R = S = r(\mathbf{Y}) = r(\mathbf{Y} - \widehat{\mathbf{Y}}_X)$, and $T = r(\mathbf{Z})$, we achieve the desired result, by finding a Three-Way version of the ”indirect” RDA, with the proper data matrices. Like in the standard RDA, the data in **Y**, **X**, and **Z** have to be preprocessed by centering and standardazing their column vectors. This is requested before the application of the RDA of **Y** on **X**.

2 Application study

The Equitable and Sustainable Well-being indicators (BES) are designed to define the economic policies which largely act on some fundamental aspects of the quality of life. Table 2 reports the description of these indicators. We use the latter as the predictor variable in the RDA that gives the constrained analysis in the subspace of $\widehat{\mathbf{Y}}_X$. Table 2 reports the correlation matrix between the column vectors of **Y**, $\mathbf{Y}^*$, and **Z**. Correlations in bold are significant. It is interesting to remark that in some cases the variables in **Z** are correlated with the columns of **Y**, while they are generally poorly related with the RDA residuals vectors (given by the unconstrained RDA). In particular, the evidence is that even if **Z** may be regressed on **Y**, for some variables the regression on **X** results inappropriate. One of the important cases is shown by the variable AMB9. This variable (Satisfaction for the environment - air, water, noise) is permanently correlated with the variable BS3 (Positive judgement for future

Table 2. *Correlations - Matrices* **Y**, **Y***, *and* **Z**

Variable	$Y1_{BS3}$	$Y2_{INN1}$	$Y3_{IF3}$	$Y4_{Q2}$	$Y5_{L12}$	$Y6_{S8}$
$Z1_{AMB9}$	**0,4029**	−0,0239	**0,4570**	**0,6852**	**0,8090**	**0,6926**
$Z2_{POL5}$	0,1906	**0,3629**	**0,2594**	**0,6395**	**0,6330**	**0,5973**
$Z3_{PATR9}$	0,1800	**0,3759**	0,0426	0,0353	0,0146	**0,2420**
$Z4_{REL4}$	**0,5133**	**0,2601**	**0,4413**	**0,7026**	**0,8380**	**0,6507**
$Z5_{SIC1}$	**−0,2215**	−0,1150	**−0,4665**	**−0,5397**	**−0,5925**	**−0,6343**
Variable	$Y1^\star_{BS3}$	$Y2^\star_{INN1}$	$Y3^\star_{IF3}$	$Y4^\star_{Q2}$	$Y5^\star_{L12}$	$Y6^\star_{S8}$
$Z1_{AMB9}$	**0,4605**	−0,1075	**0,2848**	0,1294	0,0423	−0,0119
$Z2_{POL5}$	0,0042	−0,1972	−0,0523	0,0662	−0,0624	−0,0755
$Z3_{PATR9}$	−0,1311	0,2081	**−0,2749**	**0,2794**	0,0053	0,1774
$Z4_{REL4}$	**0,3595**	−0,0025	−0,0056	0,0993	−0,1227	−0,1229
$Z5_{SIC1}$	−0,2029	−0,0184	**−0,3021**	−0,1787	−0,0291	−0,0234

perspectives), whatever is $\mathbf{y}$ or $\mathbf{y}^* = \mathbf{y} - \widehat{\mathbf{y}}_X$ (with $corr(y, y^*) = 0.7293$). We have a moderate correlation between the variable BS3 and the correspondent RDA residuals, and a moderate explanation of this variable is given by the BE1 (Per capita adjusted disposable income). Then, a tentative conclusion is that the "Satisfaction for the environment" (a **Z** variable) does not depend on the "Disposable income" (the RDA predictor **X**). An opposite case occurs when we try to assess the same AMB9 variable, versus L12 (Share of employed persons who feel satisfied with their work). Even we have that $corr(y, y^*) = -0.2395$, AMB9 has the greatest correlation with the observed L12 (y), which reduces to be not significant in terms of L12 RDA residuals (y^*). Thus, even the "Share of employed persons who feel satisfied with their work" depends on the "Disposable income", and the "Satisfaction for the environment" can be explained by the relation with "People that feel satisfied with their work", the "Satisfaction for the environment" depends on the "Disposable income" through its relation with the "People that feel satisfied with their work".

References

DE LATHAUWER, LIEVEN, DE MOOR, BART, & VANDEWALLE, JOOS. 2000. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, **21**(4), 1253–1278.

KOLDA, TAMARA G, & BADER, BRETT W. 2009. Tensor decompositions and applications. *SIAM review*, **51**(3), 455–500.

KROONENBERG, PIETER M. 2008. *Applied multiway data analysis*. John Wiley & Sons.

LEGENDRE, PIERRE, & LEGENDRE, LOUIS. 2012. *Numerical ecology*. Elsevier.