



C LADAG 2023



BOOK OF ABSTRACTS AND SHORT PAPERS
14th Scientific Meeting of the Classification and Data Analysis Group
Salerno, September 11-13, 2023

edited by

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



Pearson



SCIENTIFIC PROGRAM COMMITTEE

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

LOCAL PROGRAM COMMITTEE

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

CLADAG 2023 BOOK OF ABSTRACTS AND SHORT PAPERS:

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

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

© 2023

Published by Pearson Education Resources, Italia

www.pearson.it

ISBN: 9788891935632

INDEX

Preface	XVII
Plenary Session	1
<i>Francesco Bartolucci, Michael Greenacre, Silvia Pandolfi and Fulvia Pennoni</i>	
Discrete latent variable models: recent advances and perspectives	3
<i>Gerda Claeskens, Sarah Pirenne, Snigdha Panigrahi and Yiling Huang</i>	
Selective inference after variable selection by the randomized group Lasso method	7
<i>Fancesca Greselin</i>	
To get the best, tame the beast: robust ML estimation for mixture models	8
<i>Thomas Kneib</i>	
Rage against the mean - an introduction to distributional regression	12
<i>Sofia Charlotta Olhede</i>	
On graph limits as models for interaction data	13
Invited Papers	15
<i>Alessandro Albano, Mariangela Sciandra and Antonella Plaia</i>	
Ensemble method for text classification in medicine with multiple rare classes	17
<i>Alessandro Albano, Mariangela Sciandra and Antonella Plaia</i>	
Distance-based aggregation and consensus for preference-approvals	21
<i>Marco Alfò, Dimitris Pavlopoulos and Roberta Varriale</i>	
Flexible employment, a machine learning approach	25
<i>Federico Ambrogi and Matteo Di Maso</i>	
Clinically useful measures in survival analysis: the restricted mean survival time as an alternative to the hazard ratio	29
<i>Jose Ameijeiras-Alonso</i>	
Data-driven smoothing parameter selection for circular data analysis	33
<i>Laura Anderlucci, Silvia Dallari and Angela Montanari</i>	
View it differently: finding groups in microbiome data	34
<i>Rabea Aschenbruck, Gero Szepannek and Adalbert F. X. Wilhelm</i>	
Random-based initialization for clustering mixed-type data with the k-prototypes algorithm	38
<i>Filippo Ascolani and Valentina Ghidini</i>	
Posterior clustering for Dirichlet process mixtures of Gaussians with constant data	42

<i>Vincent Audigier and Ndèye Niang</i>	
Multiple imputation for clustering on incomplete data	46
<i>Alejandra Avalos-Pacheco and Roberta De Vito</i>	
Integrative factor models for biomedical applications	50
<i>Silvia Bacci, Bruno Bertaccini, Carla Galluccio, Leonardo Grilli and Carla Rampichini</i>	
Test equating with evolving latent ability	54
<i>Michela Baccini, Alessandra Mattei, Elena Degli Innocenti, Giulio Biscardi and Aitana Lertxundi</i>	
Causal inference on the impact of extreme ambient temperatures on population health	58
<i>Zsuzsa Bakk</i>	
Measurement invariance testing of latent class models using residual statistics and likelihood ratio test	61
<i>Falco J. Bargagli-Stoffi, Costanza Tortù and Laura Forastiere</i>	
Network interference and effect modification	65
<i>Francesco Barile, Simonón Lunagómez and Bernardo Nipoti</i>	
Flexible modelling of heterogeneous populations of networks: a Bayesian nonparametric approach	69
<i>Mario Beraha and Jim E. Griffin</i>	
Normalized latent measure factor models	70
<i>Silvia Bianconcini and Silvia Cagnone</i>	
Estimation issues in multivariate panel data	74
<i>Alessandro Bitetto and Paola Cerchiello</i>	
The nexus between ESG and initial coin offerings: evidence from text analysis	78
<i>Laura Bocci and Donatella Vicari</i>	
A clustering model for three-way asymmetric proximity data	82
<i>Ilaria Bombelli, Ichcha Manipur and Maria Brigida Ferraro</i>	
Cluster analysis for networks using a fuzzy approach	86
<i>Davide Buttarazzi and Giovanni C. Porzio</i>	
Visualizing anomalies in circular data	90
<i>Andrea Cappozzo, Chiara Masci, Francesca Ieva and Anna Maria Paganoni</i>	
Model-based clustering of right-censored lifetime data with frailties and random covariates	91
<i>Michelle Carey and Catherine Higgins</i>	
Clustering imbalanced functional data	95
<i>Alessandro Casa, Thomas Brendan Murphy and Michael Fop</i>	
Partial membership models for high-dimensional spectroscopy data	99

<i>Fabio Centofanti, Antonio Lepore and Biagio Palumbo</i> Sparse clustering for functional data	103
<i>Yunxiao Chen, Motonori Oka and Matthias von Davier</i> Interpretable and accurate scaling in large-scale assessment: a variable selection approach to latent regression	107
<i>Katharine M. Clark and Paul D. McNicholas</i> Clustering three-way data with outliers	111
<i>Roberto Colombi and Sabrina Giordano</i> A two-component markov switching regression model	115
<i>Federica Conte and Paola Paci</i> The broad phenotype-specific applications of the network-based SWIM tool	119
<i>Houyem Demni, Pierre Miasnikof, Alexander Y. Shestopaloff, Cristián Bravo and Yuri Lawryshyn</i> Testing graph clusterability: a density based statistical test for directed graphs	123
<i>Anna Denkowska, Krystian Szczfôсны, Joao Paulo Vieito and Stanisław Wanat</i> Deep neural network in the modeling of the dependence structure in risk aggregation	124
<i>Marco Di Marzio, Chiara Passamonti and Charles Taylor</i> Circular regression with measurement errors	128
<i>Marco Di Zio, Romina Filippini, Gaia Rocchetti and Simona Toti</i> Classification tree to improve data quality in official statistics	132
<i>Rosa Fabbriatore and Maria Iannario</i> Uncertainty and response style in latent trait models to assess emotional intelligence of elite swimmers	136
<i>Rosa Fabbriatore, Roberto Di Mari, Zsuzsa Bakk, Mark de Rooij and Francesco Palumbo</i> Three-step rectangular latent Markov modeling based on ML correction	140
<i>Alessio Farcomeni, Alfonso Russo and Marco Geraci</i> Mid-quantile regression for discrete panel data	144
<i>Matteo Farnè</i> Trimmed factorial k-means	148
<i>Florian Felice and Christophe Ley</i> Estimation of team's strength for handball games predictions	152
<i>Peter Filzmoser and Marcus Mayrhofer</i> Outlier explanation based on Shapley values for vector- and matrix-valued observations	156

<i>Lara Fontanella, Emiliano del Gobbo and Alex Cucco</i>	
Identification of misogynistic accounts on Twitter through Graph Convolutional Networks	159
<i>Giacomo Francisci and Anand Vidyashankar</i>	
Depth functions for tree-indexed processes	163
<i>Carla Galluccio, Matteo Magnani, Davide Vega, Giancarlo Ragozini and Alessandra Petrucci</i>	
Analysing the effect of different design choices in network-based topic detection	164
<i>Luis A. García-Escudero, Christian Hennig, Agustín Mayo-Iscar, Gianluca Morelli and Marco Riani</i>	
A proposal for the joint automated detection of clusters and anomalies	168
<i>V. G. Genova, C. Edling, H. Mondani, A. M. Rostami and M. Tumminello</i>	
Mobility across crimes: statistically validated networks and temporal pattern recognition	172
<i>Paolo Giordani, Susanna Levantesi, Andrea Nigri and Virginia Zarulli</i>	
A cohort study on the gender gap in mortality through the Tucker3 model	176
<i>Luca Greco, Giovanna Menardi and Marco Rudelli</i>	
Trimmed kernel mean shift	180
<i>Bettina Grün, Thomas Petzoldt and Helga Wagner</i>	
Modeling zone diameter measurements to infer antibiotic susceptibility of bacteria	184
<i>Julien Jacques and Francesco Amato</i>	
Clustering longitudinal ordinal data	185
<i>Daniyal Kazempour and Peer Kröger</i>	
“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
<i>Annika M. T. U. Kestler, Nensi Ikonomi, Silke D. Werle, Julian D. Schwab, Friedhelm Schwenker and Hans A. Kestler</i>	
Sparse rule generating fold-change classification for molecular high-throughput profiles	188
<i>Silvia Komara, Martina Košíková, Erik Šoltés and Tatiana Šoltésová</i>	
Comparison of the households’ work intensity in Slovakia and Czechia through least squares means analysis based on GLM	192
<i>Arnost Komárek</i>	
Model based clustering procedures for multivariate mixed type longitudinal data	193

<i>Tomasz Kwarciński, Paweł Ulman</i>	
Inequality, populism, and unfairness: a comparison of unfair income inequalities in Poland and Norway	196
<i>Francesco Lagona and Marco Mingione</i>	
Segmenting toroidal time series by nonhomogeneous hidden semi-Markov models	197
<i>Roland Langrock and Sina Mews</i>	
How to build your latent Markov model: the role of time and space	201
<i>Paweł Lula, Zsuzsanna Géring, Mária Magdalena Talaga, Ildikó Dén-Nagy and Réka Tamássy</i>	
The comparative analysis of publication activity in Hungary and Poland in the field of economics, finance and business	205
<i>Johan Lyrvall, Roberto Di Mari, Zsuzsa Bakk, Jennifer Oser and Jouni Kuha</i>	
An R package for multilevel latent class analysis with covariates	206
<i>R. Neal Mackenzie and Paul D. McNicholas</i>	
Longitudinal hidden Markov models: problems and methods	210
<i>Matteo Magnani, Matias Piqueras, Alexandra Segerberg, Davide Vega and Victoria Yantseva</i>	
Cluster analysis for the study of online visual communication	214
<i>Ichcha Manipur, Ilaria Granata, Lucia Maddalena and Mario R. Guarracino</i>	
Cluster analysis of cancer metabolic network ensembles	218
<i>Carlo Metta, Marco Fantozzi, Andrea Papini, Gianluca Amato, Matteo Bergamaschi, Silvia Giulia Galfrè, Alessandro Marchetti, Michelangelo Vegliò, Maurizio Parton and Francesco Morandin</i>	
Improving performance in neural networks by dendrite-activated connection	219
<i>Rodolfo Metulini, Francesco Biancalani and Giorgio Gnecco</i>	
The Generalized Shapley measure for ranking players in basketball: applications and future directions	223
<i>Rouven Michels, Timo Adam and Marius Ötting</i>	
Tree-based regression within a hidden Markov model framework	227
<i>Boris Mirkin</i>	
Scoring distances between equivalence and preference relations	231
<i>Fabio Morea and Domenico De Stefano</i>	
Evaluation of the performance of a modularity-based consensus community detection algorithm	234

<i>Vincenzo Nardelli and Niccolò Salvini</i>	
Assessing and improving data quality in open spatial data: a case study with ANAC data	238
<i>M. Rosário Oliveira, Diogo Pinheiro and Lina Oliveira</i>	
Visualizing interval Fisher Discriminant Analysis results	239
<i>Niels Lundtorp Olsen, Alessia Pini and Simone Vantini</i>	
Nonparametric local inference for functional data defined on manifold domains	242
<i>Silvia Pandolfi and Francesco Bartolucci</i>	
Case-control variational inference for large scale stochastic block models	246
<i>Francesca Panero</i>	
Issues with sparse spatial random graphs	250
<i>Barbara Pawelek and Maria Sadko</i>	
Corporate bankruptcy prediction: application of statistical learning methods	254
<i>Daniele Pretolesi, Andrea Vian and Annalisa Barla</i>	
Using machine learning and AI in science of science	255
<i>Pascal Pr�ea</i>	
Distances, orders and spaces	259
<i>Antonio Punzo, Luca Bagnato and Salvatore Daniele Tomarchio</i>	
Model-based clustering via parsimonious mixtures of dimension-wise scaled normal mixtures	263
<i>Monia Ranalli and Roberto Rocci</i>	
Model-based simultaneous classification and reduction for three-way ordinal data	264
<i>Jakob Raymaekers and Peter J. Rousseeuw</i>	
The cellwise Minimum Covariance Determinant estimator	268
<i>Maurizio Romano and Roberta Siciliano</i>	
A new accurate heuristic algorithm to solve the rank aggregation problem with a large number of objects	269
<i>Jorge Rueda, Maria del Mar Rueda, Ram�on Ferri and Beatriz Cobo</i>	
Using ML techniques for estimation with non-probabilistic survey data	273
<i>Ana Santos, S�onia Dias, Paula Brito and Paula Amaral</i>	
Multiclass classification of distributional data	276
<i>Lorenzo Schiavon</i>	
Latent Bayesian clustering for topic modelling	280
<i>Michael G. Schimek, Bastian Pfeifer and Marcus D. Bloice</i>	
A novel multi-view ensemble clustering framework for cancer subtype discovery	284

<i>Francesco Schirripa Spagnolo, Gaia Bertarelli, Nicola Salvati, Donato Summa, Monica Scannapieco, Stefano Marchetti and Monica Pratesi</i>	
Reducing selection bias in non-probability sample by Small Area Estimation	288
<i>Pedro Duarte Silva, Peter Filzmoser and Paula Brito</i>	
Sparse and robust estimators for outlier detection in distributional data	292
<i>Andrea Sottosanti, Sara Agavni' Castiglioni, Stefania Pirrotta, Enrica Calura and Davide Risso</i>	
Clustering genes spatial expression profiles with the aid of external biological knowledge	296
<i>Arthur Tenenhaus, Michel Tenenhaus and Theo Dijkstra</i>	
Structural equation modeling with latent/emergent variables: RGCCAc	300
<i>Yoshikazu Terada</i>	
On some properties of reconstructed trajectories from sparse longitudinal data	301
<i>Daniel J.W. Touw, Patrick J.F. Groenen, Ines Wilms and Andreas Alfons</i>	
Clusterpath Gaussian graphical modeling	302
<i>Pawel Ulman, Małgorzata Ćwiek and Maria Sadko</i>	
Housing poverty in Europe. Multidimensional analysis	305
<i>Anand Vidyashankar, Fengnan Deng, Giacomo Francisci and Xiaoran Jiang</i>	
Efficiency and robustness in supervised learning	306
<i>Frédéric Vrins</i>	
Optimal and robust combination of forecasts via constrained optimization and shrinkage	307
<i>Gabriel Wallin, Yunxiao Chen and Irimi Moustaki</i>	
DIF analysis with unknown groups and anchor items	308
<i>Felix M. Weidner, Mirko Rossini, Joachim Ankerhold and Hans A. Kestler</i>	
Constraint-based attractor search in Boolean networks using quantum computing	309
<i>Michio Yamamoto and Yoshikazu Terada</i>	
Clustering for sparsely sampled longitudinal data based on basis expansions	312
<i>Naoto Yamashita</i>	
Two extensions of extended redundancy analysis for exploratory data analysis	313
<i>Giorgia Zaccaria</i>	
Ultrametric Gaussian Mixture models with parsimonious structures	314
<i>Li-Chun Zhang</i>	
Using retail transactions for consumer price index and expenditure statistics	318

Contributed Papers	323
<i>Giuseppe Alfonzetti, Luca Grassetto and Laura Rizzi</i>	
Propensity towards Master's degree: choices of northern students after BAs?	325
<i>Giuseppe Alfonzetti, Luca Grassetto and Laura Rizzi</i>	
Classifying northern Italian students in their transition to Master degree	329
<i>Rosa Arboretti, Elena Barzizza, Nicolò Biasetton and Marta Disegna</i>	
Customer satisfaction through time: structured time series from sentiment analysis of TripAdvisor data	333
<i>Roberto Ascari and Alice Giampino</i>	
A flexible topic model	334
<i>Golnoosh Babaei, Paolo Pagnottoni and Thanh Thuy Do</i>	
Explainable machine learning for lending default classification	338
<i>Elena Barzizza, Riccardo Ceccato, Solomon Harrar, Fortunato Pesarin and Luigi Salmaso</i>	
A multivariate permutation test for association	342
<i>Michela Battauz</i>	
A competing risk analysis of academic careers with students' ability and speed as predictors	343
<i>Andriette Bekker, J.T. Ferreira, J. Pillay and M. Arashi</i>	
Bayesian analysis for a graphical t-model	347
<i>Marco Berrettini, Giuliano Galimberti, Thomas Brendan Murphy and Saverio Ranciati</i>	
Modelling soccer players field position via mixture of Gaussians with flexible weights	351
<i>Antonella Bianchino, Daniela Fusco, Paola Giordano, Maria Antonietta Liguori, Maria Carmina Palma and Donato Summa</i>	
Tourism as support in economic development of inner areas: a multi-sources approach	355
<i>Luisa Bisaglia and Francesco Lisi</i>	
SARIMA models with multiple seasonality	358
<i>Stefano Bonnini and Michela Borghesi</i>	
Adoption of 4.0 technologies and related obstacles. Application of a multivariate nonparametric test for categorical variables	362
<i>Giuseppe Bove</i>	
An application of asymmetric multidimensional scaling to the VQR 2015-2019 data	366

<i>Luca Brusa and Fulvia Pennoni</i>	
Improving clustering in temporal networks through an evolutionary algorithm	370
<i>Andrea Carta</i>	
A support vector machine approach to create oblique decision trees for regression	374
<i>Giulia Cereda, Fabio Corradi and Cecilia Viscardi</i>	
Comparing soft classification methods for the rare type match problem	378
<i>Annalisa Cerquetti</i>	
Bayesian Shannon entropy estimation under normalized inverse Gaussian priors via Monte Carlo sampling	382
<i>Lax Chan and Aldo Goia</i>	
Goodness-of-fit test for single functional index model	386
<i>Silvia Columbu, Nicola Piras and Jeroen K. Vermunt</i>	
Multilevel cross-classified latent class models	390
<i>Giulia Contu, Luca Frigau, Marco Ortu and Sara Pau</i>	
Multivariate regression tree to investigate the Italian mortality rates	394
<i>Luca Coraggio and Pietro Coretto</i>	
Empirical analysis of the quadratic scoring for selecting clustering solutions	398
<i>Marcella Corduas and Domenico Piccolo</i>	
Classification of daily streamflow data: a study on regime changes	402
<i>Noemi Corsini and Giovanna Menardi</i>	
Modal clustering for categorical data	406
<i>Cristina Davino, Tormod Næs, Rosaria Romano and Domenico Vistocco</i>	
The use of principal components in quantile regression: a simulation study	410
<i>Antonio De Falco and Antonio Irpino</i>	
An interdisciplinary methodology for socio-economic segregation analysis	414
<i>Houyem Demni and Simona Balzano</i>	
Visualizing classification results: graphical tools for DD-classifiers	418
<i>Claudia Di Caterina</i>	
Detecting the positions of nonconsensus amino acids in HIV patients by marginal likelihood thresholding	419
<i>Davide Di Cecco, Andrea Tancredi and Tiziana Tuoto</i>	
One-inflated Bayesian mixtures for population size estimation	423
<i>Marta Di Lascio and Roberta Pappadà</i>	
Cluster analysis and conditional copula: a joint approach to analyse energy demand	427

<i>Marta Di Lascio, Fabrizio Durante and Aurora Gatto</i>	
Hierarchical percentile clustering to analyse greenhouse gas emissions from agriculture in European Union	431
<i>Cinzia Di Nuzzo and Salvatore Ingrassia</i>	
Maximum likelihood approach to parameter selection in the spectral clustering algorithm	435
<i>José G. Dias</i>	
Finite mixture models: a systematic review	439
<i>Francesco Dotto, Roberto Di Mari, Alessio Farcomeni and Antonio Punzo</i>	
Measurement invariance: a method based on latent Markov models	441
<i>Niccolò Ducci, Leonardo Grilli and Marta Pittavino</i>	
A comparison between the varying-thresholds model and quantile regression	445
<i>Augusto Fasano, Niccolò Anceschi, Beatrice Franzolini and Giovanni Rebaudo</i>	
Efficient computation of predictive probabilities in probit models via expectation propagation	449
<i>Donata Favaro and Anna Giraldo</i>	
How women react to their partners' work instability. The added-worker effect	453
<i>Carlina C. Feldmann, Sina Mews, Rouven Michels and Roland Langrock</i>	
Inference on the state distribution in periodic hidden Markov models	457
<i>Giuseppe Feo, Francesco Giordano, Marcella Niglio, Sara Milito and Maria Lucia Parrella</i>	
Testing clusters of locations in spatial dynamic panel data models	461
<i>Beatrice Franzolini, Laura Bondi, Augusto Fasano and Giovanni Rebaudo</i>	
Bayesian forecasting of multivariate longitudinal zero-inflated counts: an application to civil conflict	465
<i>Francesco Freni and Giovanna Menardi</i>	
Efficient disentangling γ-ray sources from diffuse background in the sky map	469
<i>Luca Frigau, Giulia Contu, Marco Ortu and Andrea Carta</i>	
A method to validate clustering partitions	473
<i>Flora Fullone, Gianmarco Farina, Enza Compagnone, Mirella Morrone and Gioacchino de Candia</i>	
Analysis of the need for working timber starting from Istat industrial production data	477
<i>Ravi Kumar Gangadharan, Vanessa Petrarca, Maria Chiara Pagliarella and Giovanni C. Porzio</i>	
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, Irimi 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 Khorrani Chokami

Complete records over independent FGM sequences 528

<i>Ursula Laa and Dianne Cook</i>	
New tour methods for visualizing high-dimensional data	532
<i>Michele Lambardi di San Miniato, Michela Battauz, Ruggero Bellio and Paolo Vidoni</i>	
Bayesian aggregation of crowd judgments for quantitative fact checking	536
<i>Salvatore Latora and Luigi Augugliaro</i>	
Supervised classification of curves by functional data analysis: an application to neuromarketing data	540
<i>Gertraud Malsiner-Walli, Bettina Grün and Sylvia Frühwirth-Schnatter</i>	
Capturing correlated clusters using mixtures of latent class models	544
<i>Laura Marcis, Maria Chiara Pagliarella and Renato Salvatore</i>	
A three-way “indirect” redundancy analysis	545
<i>Maria Francesca Marino, Matteo Sani and Monia Lupporelli</i>	
Multi-level stochastic blockmodels for multiplex networks	549
<i>Francesca Martella, Xiaoke Qin, Wangshu Tu and Sanjena Subedi</i>	
The multivariate cluster-weighted disjoint factor analyzers model	553
<i>Raffaele Mattera, Germana Scepi, Pooria Ebrahimi and Fabio Matano</i>	
Spatial modelling of pyroclastic cover deposit thickness with remote sensing data and ground measurements: a forecasting combination approach	557
<i>Fiammetta Menchetti</i>	
Granger network on Santa Maria del Fiore Dome	561
<i>Giuseppe Mignemi, Ioanna Manolopoulou and Antonio Calcagni</i>	
Group’s heterogeneity in rating tasks: a Bayesian semi-parametric approach	565
<i>Dung Ngoc Nguyen and Alberto Roverato</i>	
Lattice of Gaussian graphical models for paired data with common undirected structure	569
<i>Marco Ortu, Giulia Contu and Luca Frigau</i>	
Multivariate regression tree topic modeling	573
<i>Lucio Palazzo, Alfonso Iodice D’Enza, Francesco Palumbo and Domenico Vistocco</i>	
Dendrogram slicing through a permutation test approach reconsidered	577
<i>Roberta Paroli and Luigi Spezia</i>	
Markov switching autoregressive models for the analysis of hydrological time series	581
<i>Davide Passaro, Luca Tardella, Giovanna Jona Lasinio, Tiziana Fragasso, Valeria Raggi and Zaccaria Ricci</i>	
A case study of electronic medical records use for predicting kidney injury	585

<i>Matteo Pedone, Raffaele Argiento and Francesco C. Stingo</i>	
Personalized treatment selection model for survival outcomes	589
<i>Danilo Petti, Marcella Niglio and Marialuisa Restaino</i>	
Variable ranking in bivariate copula survival models	593
<i>Pia Pfeiffer and Peter Filzmoser</i>	
Robust penalized multivariate analysis for high-dimensional data	597
<i>Francesco Porro</i>	
Structural zeros in regression models with compositional explanatory variables	600
<i>Kemmawadee Preedalikit, Daniel Fernández, Ivy Liu, Louise McMillan, Marta Nai Ruscone and Roy Costilla</i>	
One-dimensional mixture-based clustering for ordinal responses	604
<i>Iuliia Promskaia, Adrian O'Hagan and Michael Fop</i>	
A compositional stochastic block model for the analysis of the Erasmus programme network	608
<i>Claudia Rampichini and Maria Brigida Ferraro</i>	
A proposal of deep fuzzy clustering by means of the simultaneous approach	609
<i>Maria Giovanna Ranalli, Fulvia Pennoni, Francesco Bartolucci and Antonietta Mira</i>	
When nonresponse makes estimates from a census a small area estimation problem: the case of the survey on Graduates' Employment Status in Italy	613
<i>Edoardo Redivo and Cinzia Viroli</i>	
A supervised classification strategy based on the novel directional distribution depth function	617
<i>Ilaria Rocco</i>	
An application of CART algorithm to administrative data: analysis of youth initial employment trajectories	621
<i>Dorota Rozmus</i>	
Resampling for stability estimation vs. cluster validation via data splitting and subsampling. Which approach is better in detection of clusters in taxonomy?	625
<i>Annalina Sarra, Adelia Evangelista, Tonio Di Battista, and Sergio Palermi</i>	
Functional data analysis approach for identifying redundancy in air quality monitoring stations	627
<i>Luca Scaffidi Domianello</i>	
Student mobility in higher education: a destination-specific local analysis	631
<i>Rosaria Simone</i>	
Residuals diagnostics for model-based trees for ordered rating responses	635

<i>Alexa Sochaniwsky and Paul D. McNicholas</i>	
Hidden Markov models for multivariate longitudinal data	639
<i>Andrzej Sokółowski, Małgorzata Markowska and Maciej Laburda</i>	
K-means clustering - new variations	643
<i>Daniele Spinelli, Salvatore Ingrassia and Giorgio Vittadini</i>	
A Stata implementation of cluster weighted models: the CWMGLM package	644
<i>Salvatore D. Tomarchio, Antonio Punzo and Antonello Maruotti</i>	
Matrix-variate hidden Markov regressions	648
<i>Cristian Usala, Isabella Sulis and Mariano Porcu</i>	
Inequalities at entrance, labour market conditions and university dropout: first evidence from Italy	652
<i>Rosanna Verde, Gianmarco Borrata and Antonio Balzanella</i>	
A clustering method for distributional data based on a LDQ transformation	656
<i>Helga Wagner and Roman Pfeiler</i>	
Shrinkage of time-varying effects in panel data models	657
<i>Carlo Zaccardi, Pasquale Valentini and Luigi Ippoliti</i>	
A Bayesian spatio-temporal regression approach for confounding adjustment	661
<i>Gianpaolo Zammarchi</i>	
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

EMPIRICAL ANALYSIS OF THE QUADRATIC SCORING FOR SELECTING CLUSTERING SOLUTIONS

Luca Coraggio¹, Pietro Coretto²

¹ Department of Economics and Statistics, University of Naples Federico II, (e-mail: luca.coraggio@unina.it)

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

ABSTRACT: Selecting an optimal clustering solutions is a difficult problem, and there exist many data-driven validation strategies to perform this task. In this paper, we focus on a recent proposal, the BQH and BQS criteria, based on quadratic discriminant scores and bootstrap resampling. We provide more insight on these criteria, comparing them with a likelihood-based alternative and using different resampling schemes.

KEYWORDS: cluster validation, mixture models, model-based clustering, resampling methods

1 Quadratic scoring, likelihood-based scoring, and resampling

Selecting an optimal clustering solution is not an easy task (von Luxburg *et al.*, 2012). Recently, in Coraggio & Coretto, 2023, we proposed a novel validation index aimed at selecting clustering solutions in cases where clusters can be expected to have elliptic-symmetric shapes, or to be separable by quadratic boundaries.

Let \mathbb{X}_n indicate sample data, and $\mathcal{G}^{(m)} = \{G_k^{(m)}, k = 1, \dots, K_m\}$ be a clustering solution, obtained running clustering method $m \in \mathcal{M}$. We assume that $\mathcal{G}^{(m)}$ can be meaningfully described by K_m triplets $\boldsymbol{\theta}^{(m)} = \{\boldsymbol{\theta}_k^{(m)}, k = 1, \dots, K_m\}$, each collecting unique elements of (i) π_k , the expected fraction of points belonging to the k -th group; (ii) $\boldsymbol{\mu}_k \in \mathbb{R}^p$, the k -th cluster's center; (iii) $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p \times p}$ a positive definite scatter matrix. For a point \mathbf{x} and a triplet $\boldsymbol{\theta}_k$, we define the quadratic score (inspired to Quadratic Discriminant Analysis; e.g., see Hastie *et al.*, 2009) of point \mathbf{x} for the k -th cluster as

$$\text{qs}(x, \boldsymbol{\theta}_k) = \log(\pi_k) - \frac{1}{2} \log(\det(\boldsymbol{\Sigma}_k)) - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k); \quad (1)$$

it can be seen as a measure of how well point \mathbf{x} is accommodated into cluster k . The hard (QH) and (QS) smooth scores are based on (1), and are essentially

Algorithm 1 Bootstrap likelihood-based scoring

input: observed sample \mathbb{X}_n (with ecdf \mathbb{F}_n), $\alpha \in (0, 1)$; clustering method $m \in \mathcal{M}$; integers $B > 0$

output: bootstrap likelihood-based scoring for method m : $\tilde{L}_n^{(m)}$.

(to ease notation, dependence on m is dropped and reintroduced in step 3)

for $b \in \{1, \dots, B\}$ do

(step 1.1) $\mathbb{X}_n^{(b)} \leftarrow$ non-parametric bootstrap resample from \mathbb{X}_n (sample of size n from \mathbb{F}_n)

(step 1.2) $\hat{\boldsymbol{\theta}}_n^{(b)} \leftarrow$ triplets of parameters from clustering solution m fitted on $\mathbb{X}_n^{(b)}$

(step 1.3) $S_n^{(b)} \leftarrow l(\hat{\boldsymbol{\theta}}_n^{(b)}; \mathbb{X}_n)$ (score solution on \mathbb{X}_n)

end for

(step 2) $\tilde{W}_n \leftarrow \frac{1}{B} \sum_{b=1}^B S_n^{(b)}$ $R_n^{(b)} \leftarrow \sqrt{n} (S_n^{(b)} - \tilde{W}_n)$

(step 3) Compute $(\alpha/2)$ -level and $(1 - \alpha/2)$ -level empirical quantiles:

$$\tilde{L}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B} \sum_{b=1}^B \mathbb{I} \{ R_n^{s(b)} \leq t \} \geq \frac{\alpha}{2} \right\}; \quad \tilde{U}_n^{(m)} \leftarrow \inf_t \left\{ t : \frac{1}{B} \sum_{b=1}^B \mathbb{I} \{ R_n^{s(b)} \leq t \} \geq 1 - \frac{\alpha}{2} \right\}$$

weighted averages of the quadratic score (see Coraggio & Coretto, 2023 for details). The quadratic score (1) is strongly connected to likelihood theory, and it is easy to show that it is proportional to the Gaussian density function. Thus, as a natural alternative to the scoring criteria we use the following likelihood function

$$l(\boldsymbol{\theta}^{(m)}; \mathbb{X}_n) = \frac{1}{n} \sum_{\mathbf{x} \in \mathbb{X}_n} \log \left(\sum_{k=1}^{K^{(m)}} \pi_k^{(m)} \phi(\mathbf{x}, \boldsymbol{\theta}_k^{(m)}) \right), \quad (2)$$

where $\phi(\mathbf{x}, \boldsymbol{\theta}_k^{(m)})$ is the density function of a multi-variate Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$.

Choosing the solution that maximizes (2) may give poor results: since the sample data \mathbb{X}_n is used both to estimate $\boldsymbol{\theta}^{(m)}$ and for scoring, overly-complex solutions may be selected due to overoptimism in the evaluation process. Thus, we use the same resampling scheme used for the BQH and BQS scores, proposed in Coraggio & Coretto, 2023, that is to estimate clustering solutions on non-parametric bootstrap resamples (Efron, 1979) from \mathbb{X}_n , while using the full data to evaluate the score. The procedure is reviewed in Algorithm 1 for the likelihood-based scoring criterion.

2 Empirical analysis

The experimental analysis is a scaled-down version of that in Coraggio & Coretto, 2023, using the Pentagon5, T510D and Uniform simulated data sets.

Table 1: Selected solution by selection criteria (left-most column). Each sub-table shows results from a data set: the first column shows the selected solution, and the second column reports its ARI, computed against true classes.

Criterion	(b) Pentagon5		(c) T510D		(d) Uniform	
	Selected m	ARI	Selected m	ARI	Selected m	ARI
QH	M, K=3, VVV	0.86	O, K=10, $\gamma=10^4$	0.51	O, K=10, $\gamma=10^3$	0
QS	M, K=3, VVV	0.86	O, K=10, $\gamma=10^4$	0.51	M, K=8, VVV	0
LK	O, K=10, $\gamma=10^3$	0.44	O, K=10, $\gamma=10^4$	0.51	O, K=10, $\gamma=10^3$	0
CVQH	M, K=3, EVE	0.86	O, K=6, $\gamma=1$	0.73	O, K=5, $\gamma=10^2$	0
CVQS	M, K=3, EVE	0.86	O, K=5, $\gamma=1$	0.97	M, K=1, EEI	1
CVLK	M, K=5, EVI	0.86	O, K=8, $\gamma=1$	0.60	M, K=7, VEE	0
BQH	M, K=3, EVE	0.86	O, K=8, $\gamma=5$	0.57	O, K=9, $\gamma=10^4$	0
BQS	M, K=3, EVE	0.86	O, K=5, $\gamma=5$	0.98	M, K=1, EEI	1
BLK	O, K=5, $\gamma=1$	0.85	O, K=8, $\gamma=5$	0.57	M, K=10, VVI	0

Since likelihood-based scoring is only justified for model-based clustering, \mathcal{M} includes: (i) 140 Gaussian mixture models with covariance matrices restrictions (Banfield & Raftery, 1993), implemented with the Mclust (M) software (Scrucca *et al.*, 2016; setting $K = 1, \dots, 10$, and 14 covariance models); (ii) 180 Gaussian mixture models with eigen-ratio constraints (ERC; Ingrassia, 2004), implemented with Otrimle (O) software (Coretto & Hennig, 2017, Coretto & Hennig, 2021; setting $K \in \{1, \dots, 10\}$, ERC $\gamma \in \{1, 5, 10, 10^2, 10^3, 10^4\}$, and 3 initialization methods). The criteria compared to select optimal solutions are as follows. QH, QS, and LK: clustering solutions are estimated and scored using the full data, \mathbb{X}_n ; CVQH, CVQS, CVLK: clustering solutions are estimated on a “train set” and scored on a non-overlapping “test set”, using a 10-fold cross-validation scheme, as in Smyth, 2000. BQH, BQS, BLK: clustering solution are estimated and scored according to Algorithm 1, selecting the method m maximizing $\tilde{L}_n^{(m)}$. For each criterion, the selected solutions are evaluated against the true class labels, reporting the achieved Adjusted Rand Index (ARI, Hubert & Arabie, 1985).

Results are presented in Table 1. The comparison gives a better understanding on the mechanism that lies behind the effectiveness of the BQH and BQS criteria. First, notice that all criteria where solutions are estimated and scored on the full data (QH, QS, LK) always select overly-complex solutions. The extra penalization of the smooth score on overlapping clusters is key to select better solutions in more complicated settings (T510D and Uniform). Finally, the bootstrap scheme improves on the cross-validation. Overall, both the quadratic scores, QH and QS, and the resampling scheme in Algorithm 1 seem

equally important to consistently achieve good results.

3 Conclusion

In this paper, we run an empirical comparison of the BQH and BQS procedures from Coraggio & Coretto, 2023 with a likelihood-based alternative, using different resampling schemes. Our experiments provide new insights on the criteria, showing that both the bootstrap resampling scheme and the quadratic scores contribute equally to the procedure: (i) the penalization for clusters' overlap from the quadratic scores allows achieving better results in cases where clusters are not well separated; (ii) the bootstrap resampling scheme allows to effectively take into account clustering methods' variability, better than cross-validation would (likely better suited for prediction settings).

References

- BANFIELD, JEFFREY D., & RAFTERY, ADRIAN E. 1993. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, **49**(3), 803.
- CORAGGIO, LUCA, & CORETTO, PIETRO. 2023. Selecting the number of clusters, clustering models, and algorithms. A unifying approach based on the quadratic discriminant score. *Journal of Multivariate Analysis*, **196**(July), 105181.
- CORETTO, PIETRO, & HENNIG, CHRISTIAN. 2017. Consistency, Breakdown Robustness, and Algorithms for Robust Improper Maximum Likelihood Clustering. *Journal of Machine Learning Research*, **18**(142), 1–39.
- CORETTO, PIETRO, & HENNIG, CHRISTIAN. 2021. *otrimle: Robust Model-Based Clustering*. R package version 2.0.
- EFRON, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**(1).
- HASTIE, TREVOR, TIBSHIRANI, ROBERT, & FRIEDMAN, JEROME. 2009. *The Elements of Statistical Learning*. 2 edn. Springer Series in Statistics (SSS). Springer New York.
- HUBERT, LAWRENCE, & ARABIE, PHIPPS. 1985. Comparing partitions. *Journal of Classification*, **2**(1), 193–218.
- INGRASSIA, SALVATORE. 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods & Applications*, **13**(2).
- SCRUCCA, LUCA, FOP, MICHAEL, MURPHY, T. BRENDAN, & RAFTERY, ADRIAN E. 2016. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, **8**(1), 289–317.
- SMYTH, PADHRAIC. 2000. *Statistics and Computing*, **10**(1), 63–72.
- VON LUXBURG, ULRIKE, WILLIAMSON, ROBERT C., & GUYON, ISABELLE. 2012. Clustering: Science or Art? Proceedings of Machine Learning Research, vol. 27. Bellevue, Washington, USA: PMLR.