

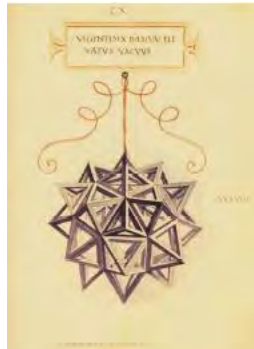
12-TH SCIENTIFIC MEETING
CLASSIFICATION AND DATA ANALYSIS

© 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 Pociecha</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 Cappelozzo, 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 Gallagher, 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, Irimi 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

Preface

This book collects the short papers presented at CLADAG 2019, the 12th Scientific Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society (SIS).

The meeting has been organized by the Department of Economics and Law of the University of Cassino and Southern Lazio, under the auspices of the SIS and the International Federation of Classification Societies (IFCS). CLADAG is a member of the IFCS, a federation of national, regional, and linguistically-based classification societies. It is a non-profit, non-political scientific organization, whose aims are to further classification research.

Every two years, CLADAG organizes a scientific meeting, devoted to the presentation of theoretical and applied papers on classification and related methods of data analysis in the broad sense. This includes advanced methodological research in multivariate statistics, mathematical and statistical investigations, survey papers on the state of the art, real case studies, papers on numerical and algorithmic aspects, applications in special fields of interest, and the interface between classification and data science. The conference aims at encouraging the interchange of ideas in the above-mentioned fields of research, as well as the dissemination of new findings.

CLADAG conferences, initiated in 1997 in Pescara (Italy), were soon considered as an attractive information exchange market and became a most important meeting point for people interested in classification and data analysis. One reason was

certainly the fact that a selection of the presented papers is regularly published in (post-conference) proceedings, typically by Springer Verlag.

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

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

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

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

Cassino, September 11, 2019

Giovanni C. Porzio
Francesca Greselin
Simona Balzano

PENALIZED VS CONSTRAINED MAXIMUM LIKELIHOOD APPROACHES FOR CLUSTERWISE LINEAR REGRESSION MODELING

Roberto Di Mari¹, Stefano Antonio Gattone² and Roberto Rocci^{3,4}

¹ Department of Economics and Business, University of Catania,
(e-mail: roberto.dimari@unict.it)

² Department of Philosophical and Social Sciences, Economics and Quantitative Methods,
University G. d'Annunzio, Chieti-Pescara, (e-mail: gattone@unich.it)

³ Department of Statistical Sciences, University of Rome La Sapienza,

⁴ Department of Economics and Finance, University of Rome Tor Vergata,
(e-mail: roberto.rocci@uniroma2.it)

ABSTRACT: Several approaches exist to avoid singular and spurious solutions in maximum likelihood (ML) estimation of clusterwise linear regression models. We propose to solve the degeneracy problem by using a penalized approach: this is done by adding a penalty term to the log-likelihood function which increasingly penalizes smaller values of the scale parameters and the tuning of the penalty term is done based on the data. Another traditional solution to degeneracy consists in imposing constraints on the variances of the regression error terms (constrained approach). We will compare the penalized approach to the constrained approach in a broad simulation study and an empirical application, providing practical guidelines on which approach to use under different circumstances.

KEYWORDS: clusterwise linear regression, penalized likelihood, scale constraints.

1 Introduction

Let y_1, \dots, y_n be a sample of independent observations drawn from the response random variable Y_i , each observed alongside with a vector of J explanatory variables $\mathbf{x}_1, \dots, \mathbf{x}_n$. Let us assume $Y_i|\mathbf{x}_i$ to be distributed as a finite mixture of linear regression models, that is

$$f(y_i|\mathbf{x}_i; \boldsymbol{\psi}) = \sum_{g=1}^G p_g \phi_g(y_i|\mathbf{x}_i, \sigma_g^2, \boldsymbol{\beta}_g) = \sum_{g=1}^G p_g \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp \left[-\frac{(y_i - \mathbf{x}_i' \boldsymbol{\beta}_g)^2}{2\sigma_g^2} \right], \quad (1)$$

where G is the number of clusters and p_g , $\boldsymbol{\beta}_g$, and σ_g^2 are the mixing proportion, the vector of $J + 1$ regression coefficients that includes an intercept, and the variance term for the g -th cluster. The set of all model parameters is given by $\boldsymbol{\Psi} = \{(p_1, \dots, p_G; \boldsymbol{\beta}_0, \dots, \boldsymbol{\beta}_G; \sigma_1^2, \dots, \sigma_G^2) \in \mathbb{R}^{(G-1)+(J+1)G+G} : p_1 + \dots + p_G = 1, p_g > 0, \sigma_g^2 > 0, \text{ for } g = 1, \dots, G\}$.

The likelihood function can be specified as

$$\mathcal{L}(\boldsymbol{\Psi}) = \prod_{i=1}^n \left\{ \sum_{g=1}^G p_g \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp \left[-\frac{(y_i - \mathbf{x}_i' \boldsymbol{\beta}_g)^2}{2\sigma_g^2} \right] \right\}, \quad (2)$$

which we maximize to estimate $\boldsymbol{\Psi}$ either by means of direct maximization or with the perhaps more popular EM algorithm (Dempster *et al.*, 1977). However, there is a well-known complication in ML estimation of this class of models: the likelihood function of mixtures of (conditional) normals with cluster-specific variances is unbounded (Kiefer & Wolfowitz, 1956; Day, 1969).

A traditional solution to the problem of unboundedness is based on the seminal work of Hathaway (1985) which, in order to have the likelihood function of univariate mixtures of normals bounded, suggested to impose a lower bound to the ratios of the scale parameters in the maximization step. The method is equivariant under linear affine transformations of the data. That is, if the data are linearly transformed, the estimated posterior probabilities do not change and the clustering remains unaltered. Recently, in the multivariate case, Rocci *et al.* (2018) incorporated constraints on the eigenvalues of the component covariances of Gaussian mixtures that are tuned on the data based on a cross-validation strategy. These constraints are built upon Ingrassia (2004)'s reformulation and are an equivariant sufficient condition for Hathaway's constraints. Estimation is done in a familiar ML environment Ingrassia & Rocci (2007), with data-driven selection of the scale balance. Di Mari *et al.* (2017) adapted Rocci *et al.* (2018)'s method to clusterwise linear regression, further investigating its properties.

Another possible approach for handling unboundedness is to modify the log-likelihood function by adding a penalty term, in which smaller values of the scale parameters are increasingly penalized. Representative examples can be found in Chen & Tan (2009); Chen *et al.* (2008); Ciuperca *et al.* (2003).

In this work we review the constrained approach of Di Mari *et al.* (2017) and develop a data-driven equivariant penalized approach for ML estimation. Next, we sketch the bulk of the methodologies.

2 The methodology

2.1 The constrained approach

Di Mari *et al.* (2017) proposed relative constraints on the group conditional variances σ_g^2 of the kind

$$\sqrt{c} \leq \frac{\sigma_g^2}{\bar{\sigma}^2} \leq \frac{1}{\sqrt{c}}, \quad (3)$$

or equivalently

$$\bar{\sigma}^2 \sqrt{c} \leq \sigma_g^2 \leq \bar{\sigma}^2 \frac{1}{\sqrt{c}}. \quad (4)$$

The above constraints are equivariant and have the effect of shrinking the variances to a suitably chosen $\bar{\sigma}^2$, the *target* variance term, and the level of shrinkage is given by the value of c . This constraints are easily implementable within the EM algorithm (Ingrassia, 2004; Ingrassia & Rocci, 2007), which is fully available in closed-form, and the selection of c is based on the data.

2.2 The penalized approach

An alternative to the constrained estimator is the penalized approach, in which a penalty $s_n(\sigma_1^2, \dots, \sigma_G^2)$ is put on the component variances and it is added to the log-likelihood. Under certain conditions on the penalty function, the penalized estimator is know to be consistent (Chen & Tan, 2009). A function s_n that satisfies these conditions is

$$s_n(\sigma_1^2, \dots, \sigma_G^2) = -\lambda \sum_{g=1}^G \left(\frac{\bar{\sigma}^2}{\sigma_g^2} + \log(\sigma_g^2) \right), \quad (5)$$

where $\bar{\sigma}^2$, the *target* variance, can be seen as our *prior* information on the scale structure and λ is the penalizing constant that is selected based on the data. Thus, the penalized log-likelihood can be written as

$$p\ell(\Psi) = \ell(\Psi) + s_n(\sigma_1^2, \dots, \sigma_G^2) \quad (6)$$

and the set of unknown parameters is found by ML with computation done by means of an EM algorithm that is available in closed-form. As well as with the constrained approach, the penalized approach is equivariant with respect to linear transformation in the response.

References

- CHEN, J., & TAN, X. 2009. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, **100**(7), 1367–1383.
- CHEN, J., TAN, X., & ZHANG, R. 2008. Inference for normal mixtures in mean and variance. *Statistica Sinica*, 443–465.
- CIUPERCA, G., RIDOLFI, A., & IDIER, J. 2003. Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, **30**(1), 45–59.
- DAY, N.E. 1969. Estimating the components of a mixture of normal distributions. *Biometrika*, **56**(3), 463–474.
- DEMPSTER, A.P., LAIRD, N.M., & RUBIN, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.
- DI MARI, R., ROCCI, R., & GATTONE, S.A. 2017. Clusterwise linear regression modeling with soft scale constraints. *International Journal of Approximate Reasoning*, **91**, 160–178.
- HATHAWAY, R.J. 1985. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**(2), 795–800.
- INGRASSIA, S. 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications*, **13**(2), 151–166.
- INGRASSIA, S., & ROCCI, R. 2007. Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Computational Statistics & Data Analysis*, **51**(11), 5339–5351.
- KIEFER, J., & WOLFOWITZ, J. 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, 887–906.
- ROCCI, R., GATTONE, S.A., & DI MARI, R. 2018. A data driven equivariant approach to constrained Gaussian mixture modeling. *Advances in Data Analysis and Classification*, **12**(2), 235–260.

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.

