



**UNIMORE**  
UNIVERSITÀ DEGLI STUDI DI  
MODENA E REGGIO EMILIA



# CLADAG 2013

9<sup>th</sup> SCIENTIFIC MEETING OF THE CLASSIFICATION  
AND DATA ANALYSIS GROUP OF THE  
ITALIAN STATISTICAL SOCIETY

**September 18 - 20, 2013**

University of Modena and Reggio Emilia  
San Geminiano Complex - Modena, Italy

# Book of Abstracts

Editors: Tommaso Minerva, Isabella Morlini, Francesco Palumbo

CLEUP

ISBN: 9788867871179

Patronage



Dipartimento di  
Comunicazione ed Economia

Dipartimento di  
Scienze Fisiche, Informatiche e Matematiche

<b>Tommaso Agasisti, Patrizia Falzetti</b> <i>Socioeconomic sorting and test scores: an empirical analysis in the Italian junior secondary schools</i>	<b>pag. 2</b>
<b>Dario Albarello, Vera D'Amico</b> <i>Empirical testing of probabilistic seismic hazard models</i>	<b>pag. 9</b>
<b>Federico Andreis, Pier Alda Ferrari</b> <i>A proposal for the multidimensional extension of CUB models</i>	<b>pag. 15</b>
<b>Morten Arendt Rasmussen, Evrim Acar</b> <i>Data fusion in the framework of coupled matrix tensor factorization with common, partially common and unique factors</i>	<b>pag. 19</b>
<b>Luigi Augugliaro, Angelo M. Mineo</b> <i>Estimation of Sparse Generalized Linear Models: the dglars package</i>	<b>pag. 20</b>
<b>Antonio Balzanella, Lidia Rivoli, Elvira Romano</b> <i>A comparison between two tools for data stream summarization</i>	<b>pag. 24</b>
<b>Lucio Barabesi, Giancarlo Diana, Pier Francesco Perri</b> <i>Gini Index Estimation in Randomized Response Surveys</i>	<b>pag. 28</b>
<b>Francesco Bartolucci, Federico Belotti, Franco Peracchi</b> <i>A test for time-invariant individual effects in generalized linear models for panel data</i>	<b>pag. 32</b>
<b>Erich Battistin, Carlos Lamarche, Enrico Rettore</b> <i>Identification of the distribution of the causal effect of an intervention using a generalised factor model</i>	<b>pag. 36</b>
<b>Matilde Bini, Lucio Masserini</b> <i>Internal effectiveness of educational offer and students' satisfaction: a SEM approach</i>	<b>pag. 37</b>
<b>Matilde Bini, Leopoldo Nascia, Alessandro Zeli</b> <i>Groups heterogeneity and sectors concentration: a structural equation modeling for micro level analysis of firms</i>	<b>pag. 41</b>
<b>Giuseppe Boari, Marta Nai Ruscone</b> <i>Use of Relevant Principal Components to Define a Simplified Multivariate Test Procedure of Optimal Clustering</i>	<b>pag. 45</b>
<b>Giuseppe Boari, Gabriele Cantaluppi, Angelo Zanella</b> <i>Some Distance Proposals for Cluster Analysis in Presence of Ordinal Variables</i>	<b>pag. 49</b>

<b>Laura Bocci, Donatella Vicari</b> <i>A general model for INDCLUS with external information</i>	<b>pag. 53</b>
<b>Paola Bongini, Paolo Trivellato, Mariangela Zenga</b> <i>The financial literacy and the undergraduates</i>	<b>pag. 57</b>
<b>Riccardo Bramante, Marta Nai Ruscone, Pasquale Spani</b> <i>Credit risk measurement and ethical issue: some evidences from the italian banks</i>	<b>pag. 61</b>
<b>Pierpaolo Brutti, Lucio Ceccarelli, Fulvio De Santis, Stefania Gubbiotti</b> <i>On the Stylometric Authorship of Ovid's Double Heroides: An Ensemble Clustering Approach</i>	<b>pag. 65</b>
<b>Silvia Caligaris, Fulvia Mecatti and Patrizia Farina</b> <i>Causal Inference in Gender Discrimination in China: Nutrition, Health, Care</i>	<b>pag. 69</b>
<b>Giorgio Calzolari, Antonino Di Pino</b> <i>Self-Selection and Direct Estimation of Across-Regime Correlation Parameter</i>	<b>pag. 73</b>
<b>Maria Gabriella Campolo, Antonino Di Pino, Ester Lucia Rizzi</b> <i>Modern Vs. Traditional: A cluster-based specification of gender and familistic attitudes and their influence on the division of labour of Italian couples</i>	<b>pag. 77</b>
<b>Gabriele Cantaluppi, Marco Passarotti</b> <i>Clustering the Four Gospels in the Greek, Latin, Gothic and Old Church Slavonic Translations</i>	<b>pag. 81</b>
<b>Carmela Cappelli, Francesca Di Iorio</b> <i>Regression Trees for change point analysis: methods, applications and recent developments</i>	<b>pag. 85</b>
<b>Roberto Casarin and Marco Tronzano and Domenico Sartore</b> <i>Bayesian Stochastic Correlation Models</i>	<b>pag. 89</b>
<b>Rosalia Castellano, Gennaro Punzo, Antonella Rocca</b> <i>Evaluating the selection effect in labour markets with a low female participation</i>	<b>pag. 93</b>
<b>Paola Cerchiello, Paolo Giudici</b> <i>A statistical based H index for the evaluation of e-markets</i>	<b>pag. 97</b>
<b>Annalisa Cerquetti</b> <i>Bayesian nonparametric estimation of global disclosure risk</i>	<b>pag. 101</b>
<b>Enrico Ciavolino, Roberto Savona</b> <i>The Forecasting side of Sovereign Risk: a Generalized Cross Entropy Approach</i>	<b>pag. 105</b>

<b>Nicoletta Cibella, Tiziana Tuoto, Luca Valentino</b> <i>What data tell you that models can't say</i>	<b>pag. 109</b>
<b>Roberto Colombi, Sabrina Giordano</b> <i>Multiple Hidden Markov Models for Categorical Time Series</i>	<b>pag. 114</b>
<b>Pier Luigi Conti, Daniela Marella</b> <i>Asymptotics in survey sampling for high entropy sampling designs</i>	<b>pag. 118</b>
<b>Claudio Conversano, Massimo Cannas, Francesco Mola</b> <i>On the Use of Recursive Partitioning in Casual Inference: A Proposal</i>	<b>pag. 122</b>
<b>Franca Crippa, Marcella Mazzoleni, Mariangela Zenga</b> <i>Keeping the pace with higher education. A fuzzy states gender study</i>	<b>pag. 128</b>
<b>F. Cugnata, C. Guglielmetti and S. Salini</b> <i>CUB model to validate FACIT TS-PS measurement instrument</i>	<b>pag. 133</b>
<b>Rosario D'Agata, Venera Tomaselli</b> <i>Multilevel Approach in Meta-Analysis of Pre-Election Poll Accuracy</i>	<b>pag. 137</b>
<b>Alfonso Iodice D'Enza and Angelos Markos</b> <i>Low-dimensional tracking of association structures in categorical data</i>	<b>pag. 141</b>
<b>Giulio D'Epifani</b> <i>Self-censored Categorical Responses A device for recovering latent behaviors</i>	<b>pag. 145</b>
<b>Pierpaolo D'Urso, Marta Disegna, Riccardo Massari</b> <i>Tourism Market Segmentation with Imprecise Information</i>	<b>pag. 150</b>
<b>Utkarsh J. Dang, Salvatore Ingrassia, Paul D. McNicholas and Ryan Browne</b> <i>Cluster-weighted models for multivariate response and extensions</i>	<b>pag. 154</b>
<b>Cristina Davino, Domenico Vistocco</b> <i>Unsupervised Classification through Quantile Regression</i>	<b>pag. 158</b>
<b>F. Marta L. Di Lascio, Simone Giannerini</b> <i>A copula-based approach to discover inter-cluster dependence relationships</i>	<b>pag. 162</b>
<b>Josè G. Dias, Sofia B. Ramos</b> <i>Hierarchical market structure of Euro area regime dynamics</i>	<b>pag. 166</b>
<b>Drago Carlo, Balzanella Antonio</b> <i>Consensus Community Detection: a Nonmetric MDS Approach</i>	<b>pag. 170</b>
<b>Fabrizio Durante, Roberta Pappad`a and Nicola Torelli</b> <i>Clustering financial time series by measures of tail dependence</i>	<b>pag. 174</b>

<b>Marco Enea, Antonella Plaia</b> <i>Influence diagnostics for generalized linear mixed models: a gradient-like statistic</i>	<b>pag. 178</b>
<b>Enrico Fabrizi, Maria R. Ferrante, Carlo Trivisano</b> <i>Joint estimation of poverty and inequality parameters in small areas</i>	<b>pag. 182</b>
<b>Giorgio Fagiolo, Andrea Roventini</b> <i>Macroeconomic Policy in DSGE and Agent-Based Models</i>	<b>pag. 187</b>
<b>Salvatore Fasola, Mariangela Sciandra</b> <i>New Flexible Probability Distributions for Ranking Data</i>	<b>pag. 191</b>
<b>Maria Brigida Ferraro, Paolo Giordani</b> <i>A new fuzzy clustering algorithm with entropy regularization</i>	<b>pag. 195</b>
<b>Camilla Ferretti, Piero Ganugi, Renato Pieri</b> <i>Mobility measures for the dairy farms in Lombardy</i>	<b>pag. 199</b>
<b>Silvia Figini, Marika Vezzoli</b> <i>Model averaging and ensemble methods for risk corporate estimation</i>	<b>pag. 203</b>
<b>Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, Agustín Mayo-Iscar</b> <i>New proposals for clustering based on trimming and restrictions</i>	<b>pag. 207</b>
<b>Andreas Geyer-Schulz, Fabian Ball</b> <i>Formal Diagnostics for Graph Clustering: The Role of Graph Automorphisms</i>	<b>pag. 211</b>
<b>Massimiliano Giacalone, Angela Alibrandi</b> <i>An overview on multiple regression models based on permutation tests</i>	<b>pag. 215</b>
<b>Francesca Giambona, Mariano Porcu</b> <i>The determinants of Italian students' reading scores: a Quantile Regression analysis</i>	<b>pag. 219</b>
<b>Paolo Giordani, Henk A.L. Kiers, Maria Antonietta Del Ferraro</b> <i>The R Package ThreeWay</i>	<b>pag. 223</b>
<b>Giuseppe Giordano, Ilaria Primerano</b> <i>Co-occurrence Network from Semantic Differential Data</i>	<b>pag. 227</b>
<b>Paolo Giudici</b> <i>Financial risk data analysis</i>	<b>pag. 231</b>
<b>Silvia Golia, Anna Simonetto</b> <i>A Comparison between SEM and Rasch model: the polytomous case</i>	<b>pag. 237</b>
<b>Anna Gottard</b> <i>Some considerations on VCUB models</i>	<b>pag. 241</b>

<b>Francesca Greselin, Salvatore Ingrassia</b> <i>Data driven EM constraints for mixtures of factor analyzers</i>	<b>pag. 245</b>
<b>Leonardo Grilli, Carla Rampichini, Roberta Varriale</b> <i>Predicting students' academic performance: a challenging issue in statistical modelling</i>	<b>pag. 249</b>
<b>Luigi Grossi, Fany Nan</b> <i>Robust estimation of regime switching models</i>	<b>pag. 255</b>
<b>Kristian Hovde Liland</b> <i>Variable selection in sequential multi-block analysis</i>	<b>pag. 259</b>
<b>Maria Iannario</b> <i>Robustness issues for a class of models for ordinal data</i>	<b>pag. 260</b>
<b>Maria Iannario, Domenico Piccolo</b> <i>A class of ordinal data models in R</i>	<b>pag. 264</b>
<b>Salvatore Ingrassia, Antonio Punzo</b> <i>Parsimony in Mixtures with Random Covariates</i>	<b>pag. 268</b>
<b>Hiroshi Inoue</b> <i>International Relations Based on the Voting Behavior in General Assembly</i>	<b>pag. 272</b>
<b>Carmela Iorio, Massimo Aria, Antonio D'Ambrosio</b> <i>Visual model representation and selection for classification and regression trees</i>	<b>pag. 276</b>
<b>Monia Lupparelli, Luca La Rocca, Alberto Roverato</b> <i>Log-Mean Linear Parameterizations for Smooth Independence Models</i>	<b>pag. 284</b>
<b>Marica Manisera, Paola Zuccolotto</b> <i>Nonlinear CUB models</i>	<b>pag. 288</b>
<b>Marica Manisera, Marika Vezzoli</b> <i>Finding number of groups using a penalized internal cluster quality index</i>	<b>pag. 292</b>
<b>Daniela Marella, Paola Vicard</b> <i>Object-Oriented Bayesian Network to deal with measurement error in household surveys</i>	<b>pag. 296</b>
<b>Angelos Markos, Alfonso Iodice D'Enza, Michel Van de Velden</b> <i>Beyond tandem analysis: joint dimension reduction and clustering in R</i>	<b>pag. 300</b>
<b>F. Martella and M. Alfò</b> <i>A biclustering approach for discrete outcomes</i>	<b>pag. 304</b>

<b>Mariagiulia Matteucci, Stefania Mignani, Roberto Ricci</b> <i>A Multidimensional IRT approach to analyze learning achievement of Italian students</i>	<b>pag. 309</b>
<b>Sabina Mazza</b> <i>Extending the Forward Search to the Combination of Multiple Classifiers: A Proposal</i>	<b>pag. 314</b>
<b>Fulvia Mecatti, M. Giovanna Ranalli</b> <i>Plug-in Bootstrap for Sample Survey Data</i>	<b>pag. 318</b>
<b>Alessandra Menafoglio, Matilde Dalla Rosa and Piercesare Secchi</b> <i>A BLU Predictor for Spatially Dependent Functional Data of a Hilbert Space</i>	<b>pag. 322</b>
<b>Maria Adele Milioli, Lara Berzieri, Sergio Zani</b> <i>Comparing fuzzy and multidimensional methods to evaluate well-being at regional level</i>	<b>pag. 326</b>
<b>Michelangelo Misuraca, Maria Spano</b> <i>Comparing text clustering algorithms from a multivariate perspective</i>	<b>pag. 331</b>
<b>Cristina Mollica, Luca Tardella</b> <i>Mixture models for ranked data classification</i>	<b>pag. 335</b>
<b>Isabella Morlini, Stefano Orlandini</b> <i>Cluster analysis of three-way atmospheric data</i>	<b>pag. 339</b>
<b>Roberto Nardecchia, Roberto Sanzo, Margherita Velucchi, Alessandro Zeli</b> <i>Productivity transition probabilities: A microlevel data analysis for Italian manufacturing sectors (1998-2007)</i>	<b>pag. 345</b>
<b>Andrea Neri, Giuseppe Ilardi</b> <i>Interviewers, co-operation and data accuracy: is there a link?</i>	<b>pag. 349</b>
<b>Akinori Okada, Satoru Yokoyama</b> <i>Nonhierarchical Asymmetric Cluster Analysis</i>	<b>pag. 353</b>
<b>Marco Perone Pacifico</b> <i>SuRF: Subspace Ridge Finder</i>	<b>pag. 357</b>
<b>Andrea Pagano, Francesca Torti, Jessica Cariboni, Domenico Perrotta</b> <i>Robust clustering of EU banking data</i>	<b>pag. 361</b>
<b>Giuseppe Pandolfo, Giovanni C. Porzio</b> <i>On depth functions for directional data</i>	<b>pag. 365</b>
<b>Andrea Pastore, Stefano F. Tonellato</b> <i>A generalised Silhouette-width measure</i>	<b>pag. 369</b>

<b>Fulvia Pennoni, Giorgio Vittadini</b> <i>Hospital efficiency under two competing panel data models</i>	<b>pag. 373</b>
<b>Alessia Pini, Simone Vantini</b> <i>The Interval-Wise Control of the Family-Wise Error Rate for Testing Functional Data</i>	<b>pag. 377</b>
<b>Mariano Porcu, Isabella Sulis</b> <i>Detecting differences between primary schools in mathematics and reading achievement by using schools added-value measures of performance</i>	<b>pag. 381</b>
<b>Antonio Punzo, Paul D. McNicholas, Katherine Morris, Ryan P. Browne</b> <i>Outlier Detection via Contaminated Mixture Distributions</i>	<b>pag. 387</b>
<b>Emanuela Raffinetti, Pier Alda Ferrari</b> <i>New perspectives for the RDI index in social research fields</i>	<b>pag. 392</b>
<b>Monia Ranalli, Roberto Rocci</b> <i>Mixture models for ordinal data: a pairwise likelihood approach</i>	<b>pag. 396</b>
<b>Marco Riani, Andrea Cerioli, Gianluca Morelli</b> <i>Issues in robust clustering</i>	<b>pag. 400</b>
<b>Stéphane Robin</b> <i>Deciphering and modeling heterogeneity in interaction networks</i>	<b>pag. 404</b>
<b>Rosaria Romano, Francesco Palumbo</b> <i>Partial Possibilistic Regression Path Modeling</i>	<b>pag. 409</b>
<b>Renata Rotondi</b> <i>Classification of composite seismogenic sources through probabilistic score indices</i>	<b>pag. 413</b>
<b>Gabriella Schoier</b> <i>On Wild Bootstrap and M Unit Root Test</i>	<b>pag. 417</b>
<b>Luca Scrucca</b> <i>On the implementation of a parallel algorithm for variable selection in model-based clustering</i>	<b>pag. 421</b>
<b>Paolo Sestito</b> <i>The Role of Learning Measurement in the Governance of an Education System: an Overview of the Issues</i>	<b>pag. 425</b>
<b>John Shawe-Taylor, Blaz Zlicar</b> <i>Novelty Detection with Support Vector Machines</i>	<b>pag. 430</b>
<b>Nadia Solaro</b> <i>Multidimensional scaling with incomplete distance matrices: an insight into the problem</i>	<b>pag. 431</b>



<b>Luigi Spezia, Cecilia Pinto</b> <i>Markov switching models for high-frequency time series: flapper skate's depth profile as a case study</i>	<b>pag. 435</b>
<b>Ralf Stecking, Klaus B. Schebesch</b> <i>Data Privacy in Credit Scoring: Evaluating SVM Approaches Based on Microaggregated Data</i>	<b>pag. 439</b>
<b>Isabella Sulis, Francesca Giambona, Nicola Tedesco</b> <i>Analyzing university students' careers using Multi-State Models</i>	<b>pag. 443</b>
<b>Luca Tardella, Danilo Alunni Fegatelli</b> <i>BBRecap for Bayesian Behavioural Capture-Recapture Modeling</i>	<b>pag. 447</b>
<b>Cristina Tortora, Paul D. McNicholas, Ryan P. Browne</b> <i>Mixtures of generalized hyperbolic factor analyzers</i>	<b>pag. 451</b>
<b>Giovanni Trovato</b> <i>Testing for endogeneity and country heterogeneity</i>	<b>pag. 455</b>
<b>Joaquin Vanschoren and Mikio L. Braun, Cheng Soon Ong</b> <i>Open science in machine learning</i>	<b>pag. 461</b>
<b>Valerio Veglio</b> <i>Logistic Regression and Decision Tree: Performance Comparisons in Estimating Customers' Risk of Churn</i>	<b>pag. 465</b>
<b>Maurizio Vichi</b> <i>Robust Two-mode clustering</i>	<b>pag. 469</b>
<b>Vincenzina Vitale</b> <i>Hierarchical Graphical Models and Item Response Theory</i>	<b>pag. 470</b>
<b>Sara Viviani</b> <i>Extending the JM library</i>	<b>pag. 474</b>
<b>Adalbert F.X. Wilhelm</b> <i>Visualisations of Classification Tree Models: An Evaluative Comparison</i>	<b>pag. 478</b>

# Data driven EM constraints for mixtures of factor analyzers

Francesca Greselin and Salvatore Ingrassia

**Abstract** Mixtures of factor analyzers are becoming more and more popular in the area of model based clustering of high-dimensional data. In this paper we implement a data-driven methodology to maximize the likelihood function in a constrained parameter space, to overcome the well known issue of singularities and to reduce spurious maxima in the EM algorithm. Simulation results and applications to real data show that the problematic convergence of the EM, even more critical when dealing with factor analyzers, can be greatly improved.

**Key words:** Mixture of Factor Analyzers, Model-Based Clustering, Constrained EM algorithm.

## 1 Introduction and motivation

Finite mixture distributions, dating back to the seminal work of Newcomb and Pearson, have been receiving a growing interest in statistical modeling all along the last century. Along the lines of Ghahramani and Hilton (1997) we assume that the data have been generated by a linear factor model with latent variables modeled as Gaussian mixtures. Our purpose is to improve the performances of the EM algorithm, giving practical recipes to overcome some of its issues. Following Ingrassia (2004), in this paper we introduce and implement a procedure for the parameters estimation of mixtures of factor analyzers, which maximizes the likelihood function in a constrained parameter space, having no singularities and a reduced number of spurious local maxima. Within the Gaussian Mixture (GM) model-based approach to density estimation and clustering, the density of the  $d$ -dimensional random variable  $\mathbf{X}$  of interest is modeled as a mixture of a number, say  $G$ , of multivariate normal densities in some unknown proportions  $\pi_1, \dots, \pi_G$ ,

---

Francesca Greselin  
Department of Statistics and Quantitative Methods  
Milano-Bicocca University  
Via Bicocca degli Arcimboldi 8 - 20126 Milano (Italy), e-mail: francesca.greselin@unimib.it

Salvatore Ingrassia  
Department of Economics and Business  
University of Catania  
Corso Italia 55 - Catania (Italy), e-mail: s.ingrassia@unict.it

$$f(\mathbf{x}; \theta) = \sum_{g=1}^G \pi_g \phi_d(\mathbf{x}; \mu_g, \Sigma_g)$$

where  $\phi_d(\mathbf{x}; \mu, \Sigma)$  denotes the  $d$ -variate normal density function with mean  $\mu$  and covariance matrix  $\Sigma$ . Then, we postulate a finite mixture of linear sub-models  $\mathbf{X}_i = \mu_g + \Lambda_g \mathbf{U}_{ig} + \mathbf{e}_{ig}$  with probability  $\pi_g$  ( $g = 1, \dots, G$ ) for  $i = 1, \dots, n$ , for the distribution of the full observation vector  $\mathbf{X}$ , given the (unobservable) latent factors  $\mathbf{U}$ , where  $\Lambda_g$  is a  $d \times q$  matrix of *factor loadings*, the *factors*  $\mathbf{U}_{1g}, \dots, \mathbf{U}_{ng}$  are  $\mathcal{N}(\mathbf{0}, \mathbf{I}_q)$  distributed independently of the *errors*  $\mathbf{e}_{ig}$ , which are independently  $\mathcal{N}(\mathbf{0}, \Psi_g)$  distributed, and  $\Psi_g$  is a  $d \times d$  diagonal matrix ( $g = 1, \dots, G$ ). We suppose that  $q < d$ , which means that  $q$  latent factors are jointly explaining the  $d$  observable features of the statistical units. Under these assumptions,  $\Sigma_g = \Lambda_g \Lambda_g' + \Psi_g$  ( $g = 1, \dots, G$ ). The parameter vector  $\theta_{MGFA}(d, q, G)$  now consists of the elements of the component means  $\mu_g$ , the  $\Lambda_g$ , and the  $\Psi_g$ , along with the mixing proportions  $\pi_g$  ( $g = 1, \dots, G-1$ ).

## 2 The likelihood function and the EM algorithm for MGFA

In this section we summarize the main steps of the EM algorithm for mixtures of Factor analyzers, see e.g. McLachlan *et al.* (2003) for details. Let  $\mathbf{x}_i$  ( $i = 1, \dots, n$ ) denotes the realization of  $\mathbf{X}_i$ . Then, the complete-data likelihood function for a sample  $\tilde{\mathbf{X}}$  of size  $n$  can be written as

$$L_c(\theta; \tilde{\mathbf{X}}) = \prod_{i=1}^n \prod_{g=1}^G [\phi_d(\mathbf{x}_i | \mathbf{u}_i; \mu_g, \Lambda_g, \Psi_g) \phi_q(\mathbf{u}_{ig}) \pi_g]^{z_{ig}}. \quad (1)$$

Due to the factor structure of the model, we consider the alternating expectation-conditional maximization (AECM) algorithm, consisting of the iteration of two conditional maximizations, until convergence. There is one E-step and one CM-step, alternatively i) considering  $\theta_1 = \{\pi_g, \mu_g, g = 1, \dots, G\}$  where the missing data are the unobserved group labels  $\mathbf{Z} = (\mathbf{z}'_1, \dots, \mathbf{z}'_n)$  and ii) considering  $\theta_2 = \{\Lambda_g, \Psi_g, g = 1, \dots, G\}$  where the missing data are the group labels  $\mathbf{Z}$  and the unobserved latent factors  $\mathbf{U} = (\mathbf{U}_{11}, \dots, \mathbf{U}_{nG})$ . In the First Cycle, after updating the  $z_{ig}^{(k+1)}$  in the E-step, the M-step provides new values for  $\pi_g^{(k+1)}, \mu_g^{(k+1)}, n_g^{(k+1)}$ . In the Second Cycle, after writing the complete data log-likelihood, some algebras lead to the following estimate of  $\{\Lambda_g, \Psi_g, g = 1, \dots, G\}$

$$\hat{\Lambda}_g = \mathbf{S}_g^{(k+1)} \gamma_g^{(k)'} [\Theta_g^{(k)}]^{-1} \quad \hat{\Psi}_g = \text{diag} \left\{ \mathbf{S}_g^{(k+1)} - \hat{\Lambda}_g \gamma_g^{(k)} \mathbf{S}_g^{(k+1)} \right\},$$

where

$$\mathbf{S}_g^{(k+1)} = (1/n_g^{(k+1)}) \sum_{i=1}^n z_{ig}^{(k+1)} (\mathbf{x}_i - \mu_g^{(k+1)}) (\mathbf{x}_i - \mu_g^{(k+1)})'$$

$$\gamma_g^{(k)} = \Lambda_g^{(k)'} (\Lambda_g^{(k)} \Lambda_g^{(k)'} + \Psi_g^{(k)})^{-1} \quad \text{and} \quad \Theta_g^{(k)} = \mathbf{I}_q - \gamma_g^{(k)} \Lambda_g^{(k)} + \gamma_g^{(k)} \mathbf{S}_g^{(k+1)} \gamma_g^{(k)'}$$

### 3 Likelihood maximization in constrained parametric spaces

Along the lines of Ingrassia (2004) let us consider the constrained parameter space

$$\Theta_c = \{(\pi_1, \dots, \pi_G, \mu_1, \dots, \mu_G, \Sigma_1, \dots, \Sigma_G) \in \mathbb{R}^{k[1+d+(d^2+d)/2]} : \pi_g \geq 0, \pi_1 + \dots + \pi_G = 1, a \leq \lambda_{ig} \leq b, \quad g = 1, \dots, G \quad i = 1, \dots, d\}. \quad (2)$$

Applying the eigenvalue decomposition to the square  $d \times d$  matrix  $\Lambda_g \Lambda_g'$  we can find  $\Gamma_g$  and  $\Delta_g$  such that  $\Lambda_g \Lambda_g' = \Gamma_g \Delta_g \Gamma_g'$  where  $\Gamma_g$  is the orthonormal matrix whose rows are the eigenvectors of  $\Lambda_g \Lambda_g'$  and  $\Delta_g = \text{diag}(\delta_{1g}, \dots, \delta_{dg})$  is the sorted diagonal matrix of the eigenvalues of  $\Lambda_g \Lambda_g'$ , i.e.  $\delta_{1g} \geq \dots \geq \delta_{qg} \geq 0$ , and  $\delta_{(q+1)g} = \dots = \delta_{dg} = 0$ . Further, applying now the singular value decomposition to  $\Lambda_g$ , we get  $\Lambda_g = \mathbf{U}_g \mathbf{D}_g \mathbf{V}_g'$ . This yields  $\Lambda_g \Lambda_g' = (\mathbf{U}_g \mathbf{D}_g \mathbf{V}_g') (\mathbf{V}_g \mathbf{D}_g' \mathbf{U}_g') = \mathbf{U}_g \mathbf{D}_g \mathbf{D}_g' \mathbf{U}_g'$  hence  $\text{diag}(\delta_{1g}, \dots, \delta_{qg}) = \text{diag}(d_{1g}^2, \dots, d_{qg}^2)$ . We can now modify the EM algorithm in such a way that the eigenvalues of the covariances  $\Sigma_g$  (for  $g = 1, \dots, G$ ) are confined into suitable ranges. To this aim we exploit the following inequalities

$$\begin{aligned} \lambda_{\min}(\Lambda_g \Lambda_g' + \Psi_g) &\geq \lambda_{\min}(\Lambda_g \Lambda_g') + \lambda_{\min}(\Psi_g) \geq a \\ \lambda_{\max}(\Lambda_g \Lambda_g' + \Psi_g) &\leq \lambda_{\max}(\Lambda_g \Lambda_g') + \lambda_{\max}(\Psi_g) \leq b \end{aligned}$$

which enforce (2) when imposing the following constraints

$$d_{ig}^2 + \psi_{ig} \geq a \quad i = 1, \dots, d \quad (3)$$

$$d_{ig} \leq \sqrt{b - \psi_{ig}} \quad i = 1, \dots, q \quad (4)$$

$$\psi_{ig} \leq b \quad i = q + 1, \dots, d \quad (5)$$

for  $g = 1, \dots, G$ , where  $\psi_{ig}$  denotes the  $i$ -th diagonal entry of  $\Psi_g$ . In particular, we note that condition (3) reduces to  $\psi_{ig} \geq a$  for  $i = (q + 1), \dots, d$ .

It is important to remark that the resulting EM algorithm is monotone, once the initial guess, say  $\Sigma_g^0$ , satisfies the constraints. Further, as shown in the case of gaussian mixtures in Ingrassia and Rocci (2007), the maximization of the complete log-likelihood is guaranteed. On another note, a data driven method to gauge the bounds  $a$  and  $b$  is needed.

### 4 Numerical studies

A brief numerical study is presented here, to compare the performance of the constrained vs unconstrained EM algorithm. More simulations have been performed, also with real datasets are available in (see Greselin and Ingrassia, 2013). A sample of  $N = 150$  data has been generated with weights  $\alpha = (0.3, 0.4, 0.3)'$  with parameters such that  $\max_{i,g} \lambda_i(\Sigma_g) = 4.18$ . We run 100 times both the unconstrained and the constrained AECM algorithms (for different values of the constraints  $a, b$ ) using a common random initial clusterings. The unconstrained algorithm attains the right solution in 24% of cases; summary statistics about the misclassification error, over

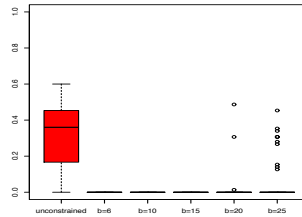
the 100 runs, are reported in Table 1. To compare how  $a$  and  $b$  influences the performance of the constrained EM, different pairs of values has been considered, and Table 2 shows the more interesting cases. Further results are reported in Figure 1, where the boxplots of the distribution of the misclassification errors show the poor performance of the unconstrained algorithm compared to its constrained version. For all values of the upper bound  $b$ , the third quartile of the misclassification error is steadily equal to 0. Indeed, for  $b = 6, 10$  and  $15$  we had no misclassification error, while we observed very low and rare misclassification errors only for  $b = 20$  and  $b = 25$  (respectively 3 and 11 not null values, over 100 runs). Moreover, the robustness of the results with respect to the choice of the upper constraint is apparent. A data driven method to select the bounds can be derived from the observed results, by running the constrained EM for increasing values of the upper bound, till a decrease in the final likelihood. The value of  $b$  before the decrease, observed over a series of run, will be chosen as upper value for the constrained estimation.

**Table 1** Summary statistics for the Misclassif Error over 100 runs of the unconstrained EM alg

min	$Q_1$	$Q_2$	$Q_3$	max
0%	17%	36%	45.3%	60%

**Table 2** Percentage of convergence to the right maximum of the constrained EM algorithm for  $a = 0.01$  and some values of the upper constraint  $b$

$b : +\infty$	6	10	15	20	25
24%	100%	100%	100%	97%	89%



**Fig. 1** Boxplots of the misclassification error. From left to right, boxplots refer to the unconstrained algorithm, then to the constrained algorithm, for  $a = 0.01$  and  $b = 6, 10, 15, 20, 25$ .

## References

- Ghahramani, Z. and Hilton, G. (1997). The EM algorithm for mixture of factor analyzers. *Tech. Rep. CRG-TR-96-1*.
- Greselin, F. and Ingrassia, S. (2013) Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers, <http://arxiv.org/abs/1301.1505>.
- Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Stat. Meth. & Appl.*, **13**, 151–166.
- Ingrassia, S. and Rocci, R. (2007). Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Comp. Stat. & Data Anal.*, **51**, 5339–5351.
- McLachlan, G., Peel, D., and Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Comp. Stat. and Data Anal.*, **41**, 379–388.