Studies in Classification, Data Analysis, and Knowledge Organization

Titles in the Series

E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy (Eds.)
New Approaches in Classification and Data Analysis. 1994 (out of print)

W. Gaul and D. Pfeifer (Eds.)
From Data to Knowledge. 1995

H.-H. Bock and W. Polasek (Eds.)
Data Analysis and Information Systems. 1996

E. Diday, Y. Lechevallier, and O. Opitz (Eds.)
Ordinal and Symbolic Data Analysis. 1996

R. Klar and O. Opitz (Eds.)
Classification and Knowledge Organization. 1997

C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, and Y. Baba (Eds.)
Data Science, Classification, and Related Methods. 1998

I. Balderjahn, R. Mathar, and M. Schader (Eds.)
Classification, Data Analysis, and Data Highways. 1998

A. Rizzi, M. Vichi, and H.-H. Bock (Eds.)
Advances in Data Science and Classification. 1998

M. Vichi and O. Opitz (Eds.)
Classification and Data Analysis. 1999

W. Gaul and H. Locarek-Junge (Eds.)
Classification in the Information Age. 1999

H.-H. Bock and E. Diday (Eds.)
Analysis of Symbolic Data. 2000

H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, and M. Schader (Eds.)
Data Analysis, Classification, and Related Methods. 2000

W. Gaul, O. Opitz, and M. Schader (Eds.)
Data Analysis. 2000

R. Decker and W. Gaul (Eds.)
Classification and Information Processing at the Turn of the Millenium. 2000

S. Borra, R. Rocci, M. Vichi, and M. Schader (Eds.)
Advances in Classification and Data Analysis. 2001

W. Gaul and G. Ritter (Eds.)
Classification, Automation, and New Media. 2002

K. Jajuga, A. Sokołowski, and H.-H. Bock (Eds.)
Classification, Clustering and Data Analysis. 2002

M. Schwaiger and O. Opitz (Eds.)
Exploratory Data Analysis in Empirical Research. 2003

M. Schader, W. Gaul, and M. Vichi (Eds.)
Between Data Science and Applied Data Analysis. 2003

H.-H. Bock, M. Chiodi, and A. Mineo (Eds.)
Advances in Multivariate Data Analysis. 2004

D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul (Eds.)
Classification, Clustering, and Data Mining Applications. 2004

D. Baier and K.-D. Wernecke (Eds.)
Innovations in Classification, Data Science, and Information Systems. 2005

M. Vichi, P. Monari, S. Mignani, and A. Montanari (Eds.)
New Developments in Classification and Data Analysis. 2005

D. Baier, R. Decker, and L. Schmidt-Thieme (Eds.)
Data Analysis and Decision Support. 2005

C. Weihs and W. Gaul (Eds.)
Classification – the Ubiquitous Challenge. 2005

M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul (Eds.)
From Data and Information Analysis to Knowledge Engineering. 2006

V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna (Eds.)
Data Science and Classification. 2006

Sergio Zani · Andrea Cerioli
Marco Riani · Maurizio Vichi
Editors

Data Analysis, Classification and the Forward Search

Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma, June 6–8, 2005

With 118 Figures and 50 Tables

Springer

Prof. Sergio Zani
Department of Economics
Section of Statistics
and Computing
University of Parma
Via Kennedy 6
43100 Parma, Italy
sergio.zani@unipr.it

Prof. Marco Riani
Department of Economics
Section of Statistics
and Computing
University of Parma
Via Kennedy 6
43100 Parma, Italy
mriani@unipr.it

Prof. Andrea Cerioli
Department of Economics
Section of Statistics
and Computing
University of Parma
Via Kennedy 6
43100 Parma, Italy
andrea.cerioli@unipr.it

Prof. Maurizio Vichi
Department of Statistics,
Probability and Applied
Statistics
University of Rome
"La Sapienza"
Piazzale Aldo Moro 5
00185 Roma, Italy
maurizio.vichi@uniroma1.it

ISSN 1431-8814
ISBN 10 3-540-35977-X Springer Berlin Heidelberg New York
ISBN 13 978-3-540-35977-7 Springer Berlin Heidelberg New York

Springer-Verlag is a part of Springer Science+Business Media

springer.com

Cover: Erich Kirchner, Heidelberg
Production: LE-TEX, Jelonek, Schmidt & Vöckler GbR, Leipzig

SPIN 11789703 Printed on acid-free paper – 43/3100 – 5 4 3 2 1 0

Preface

This volume contains revised versions of selected papers presented at the biennial meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, which was held in Parma, June 6-8, 2005. Sergio Zani chaired the Scientific Programme Committee and Andrea Cerioli chaired the Local Organizing Committee.

The scientific programme of the conference included 127 papers, 42 in specialized sessions, 68 in contributed paper sessions and 17 in poster sessions. Moreover, it was possible to recruit five notable and internationally renowned invited speakers (including the 2004–2005 President of the International Federation of Classification Societies) for plenary talks on their current research work. Among the specialized sessions, two were organized by Wolfgang Gaul with five talks by members of the GfKl (German Classification Society), and one by Jacqueline J. Meulman (Dutch/Flemish Classification Society). Thus, the conference provided a large number of scientists and experts from home and abroad with an attractive forum for discussion and mutual exchange of knowledge. The topics of all plenary and specialized sessions were chosen to fit, in the broadest possible sense, the mission of CLADAG, the aim of which is "to further methodological, computational and applied research within the fields of Classification, Data Analysis and Multivariate Statistics".

A peer-review refereeing process led to the selection of 46 extended papers, which are contained in this book. The more methodologically oriented papers focus on developments in clustering and discrimination, multidimensional data analysis, data mining, and robust statistics with a special emphasis on the novel Forward Search approach. Many papers also provide significant contributions in a wide range of fields of application. Customer satisfaction and service evaluation are two examples of such emerging fields. This suggested the presentation of the 46 selected papers in six parts as follows:

1. Clustering and Discrimination
2. Multidimensional Data Analysis and Multivariate Statistics
3. Robust Methods and the Forward Search
4. Data Mining Methods and Software
5. Multivariate Methods for Customer Satisfaction and Service Evaluation
6. Multivariate Methods in Applied Science

We wish to express our gratitude to the other members of the Scientific Programme Committee

B. Chiandotto, N.C. Lauro, P. Monari, A. Montanari, C. Provasi, G. Vittadini and to the specialized session organizers

F. Camillo, M. Chiodi, W. Gaul, S. Ingrassia, J.J. Meulman

for their ability to attract interesting contributions, and to the authors, whose enthusiastic participation made the meeting possible. We would also like to extend our thanks to the chairpersons and discussants of the sessions for their stimulating comments and suggestions. We are very grateful to the referees for their careful reviews of all submitted papers and for the time spent in this professional activity.

We gratefully acknowledge the University of Parma and its Department of Economics for financial support and hospitality. We are also indebted to *Istat - Istituto Nazionale di Statistica* and *SAS* for their support.

We thank all the members of the Local Organizing Committee

A. Corbellini, G. Gozzi, L. Grossi, F. Laurini, M.A. Milioli, G. Morelli, I. Morlini

for their excellent work in managing the organization of the CLADAG-2005 conference. Special thanks go to Prof. Isabella Morlini, for her skilful accomplishment of the duties of Scientific Secretary of CLADAG-2005, and to Dr. Fabrizio Laurini for his assistance in producing this volume.

Finally, we would like to thank Dr. Martina Bihn of Springer-Verlag, Heidelberg, for her support and dedication to the production of this volume.

Parma and Rome,
June 2006

Sergio Zani
Andrea Cerioli
Marco Riani
Maurizio Vichi

Contents

Part I. Clustering and Discrimination

Genetic Algorithms-based Approaches for Clustering Time Series ... 3
Roberto Baragona, Salvatore Vitrano

On the Choice of the Kernel Function in Kernel Discriminant Analysis Using Information Complexity ... 11
Hamparsum Bozdogan, Furio Camillo, Caterina Liberati

Growing Clustering Algorithms in Market Segmentation: Defining Target Groups and Related Marketing Communication ... 23
Reinhold Decker, Sören W. Scholz, Ralf Wagner

Graphical Representation of Functional Clusters and MDS Configurations ... 31
Masahiro Mizuta

Estimation of the Structural Mean of a Sample of Curves by Dynamic Time Warping ... 39
Isabella Morlini, Sergio Zani

Sequential Decisional Discriminant Analysis ... 49
Rafik Abdesselam

Regularized Sliced Inverse Regression with Applications in Classification ... 59
Luca Scrucca

Part II. Multidimensional Data Analysis and Multivariate Statistics

Approaches to Asymmetric Multidimensional Scaling with External Information ... 69
Giuseppe Bove

Variable Architecture Bayesian Neural Networks: Model Selection Based on EMC ... 77
Silvia Bozza, Pietro Mantovan

Missing Data in Optimal Scaling ... 85
Pier Alda Ferrari, Paola Annoni

Simple Component Analysis Based on RV Coefficient ... 93
Michele Gallo, Pietro Amenta, Luigi D'Ambra

Baum-Eagon Inequality in Probabilistic Labeling Problems ... 103
Crescenzio Gallo, Giancarlo de Stasio

Monotone Constrained EM Algorithms for Multinormal Mixture Models ... 111
Salvatore Ingrassia, Roberto Rocci

Visualizing Dependence of Bootstrap Confidence Intervals for Methods Yielding Spatial Configurations ... 119
Henk A.L. Kiers, Patrick J.F. Groenen

Automatic Discount Selection for Exponential Family State-Space Models ... 127
Andrea Pastore

A Generalization of the Polychoric Correlation Coefficient ... 135
Annarita Roscino, Alessio Pollice

The Effects of MEP Distributed Random Effects on Variance Component Estimation in Multilevel Models ... 143
Nadia Solaro, Pier Alda Ferrari

Calibration Confidence Regions Using Empirical Likelihood ... 153
Diego Zappa

Part III. Robust Methods and the Forward Search

Random Start Forward Searches with Envelopes for Detecting Clusters in Multivariate Data ... 163
Anthony Atkinson, Marco Riani, Andrea Cerioli

Robust Transformation of Proportions Using the Forward Search ... 173
Matilde Bini, Bruno Bertaccini

The Forward Search Method Applied to Geodetic Transformations ... 181
Alessandro Carosio, Marco Piras, Dante Salvini

An R Package for the Forward Analysis of Multivariate Data ... 189
Aldo Corbellini, Kjell Konis

A Forward Search Method for Robust Generalised Procrustes Analysis 199
Fabio Crosilla, Alberto Beinat

A Projection Method for Robust Estimation and Clustering in Large Data Sets 209
Daniel Peña, Francisco J. Prieto

Robust Multivariate Calibration 217
Silvia Salini

Part IV. Data Mining Methods and Software

Procrustes Techniques for Text Mining 227
Simona Balbi, Michelangelo Misuraca

Building Recommendations from Random Walks on Library OPAC Usage Data 235
Markus Franke, Andreas Geyer-Schulz, Andreas Neumann

A Software Tool via Web for the Statistical Data Analysis: R-php 247
Angelo M. Mineo, Alfredo Pontillo

Evolutionary Algorithms for Classification and Regression Trees 255
Francesco Mola, Raffaele Miele

Variable Selection Using Random Forests 263
Marco Sandri, Paola Zuccolotto

Boosted Incremental Tree-based Imputation of Missing Data . 271
Roberta Siciliano, Massimo Aria, Antonio D'Ambrosio

Sensitivity of Attributes on the Performance of Attribute-Aware Collaborative Filtering 279
Karen H. L. Tso, Lars Schmidt-Thieme

Part V. Multivariate Methods for Customer Satisfaction and Service Evaluation

Customer Satisfaction Evaluation: An Approach Based on Simultaneous Diagonalization 289
Pietro Amenta, Biagio Simonetti

Analyzing Evaluation Data: Modelling and Testing for Homogeneity 299
Angela D'Elia, Domenico Piccolo

Archetypal Analysis for Data Driven Benchmarking 309
Giovanni C. Porzio, Giancarlo Ragozini, Domenico Vistocco

Determinants of Secondary School Dropping Out: a Structural Equation Model 319
Giancarlo Ragozini, Maria Prosperina Vitale

Testing Procedures for Multilevel Models with Administrative Data 329
Giorgio Vittadini, Maurizio Sanarico, Paolo Berta

Multidimensional Versus Unidimensional Models for Ability Testing 339
Stefania Mignani, Paola Monari, Silvia Cagnone, Roberto Ricci

Part VI. Multivariate Methods in Applied Science

Economics

A Spatial Mixed Model for Sectorial Labour Market Data 349
Marco Alfó, Paolo Postiglione

The Impact of the New Labour Force Survey on the Employed Classification 359
Claudio Ceccarelli, Antonio R. Discenza, Silvia Loriga

Using CATPCA to Evaluate Market Regulation 369
Giuseppe Coco, Massimo Alfonso Russo

Credit Risk Management Through Robust Generalized Linear Models 377
Luigi Grossi, Tiziano Bellini

Classification of Financial Returns According to Thresholds Exceedances 387
Fabrizio Laurini

Environmental and Medical Sciences

Nonparametric Clustering of Seismic Events 397
Giada Adelfio, Marcello Chiodi, Luciana De Luca, Dario Luzio

A Non-Homogeneous Poisson Based Model for Daily Rainfall Data 405
Alberto Lombardo, Antonina Pirrotta

A Comparison of Data Mining Methods and Logistic Regression to Determine Factors Associated with Death Following Injury 417
Kay Penny, Thomas Chesney

Author Index 425

Monotone Constrained EM Algorithms for Multinormal Mixture Models

Salvatore Ingrassia[1] and Roberto Rocci[2]

[1] Dipartimento di Economia e Metodi Quantitativi,
Università di Catania, Italy
s.ingrassia@unict.it
[2] Dipartimento SEFEMEQ,
Università di Tor Vergata, Italy
roberto.rocci@uniroma2.it

Abstract. We investigate the spectral decomposition of the covariance matrices of a multivariate normal mixture distribution in order to construct constrained EM algorithms which guarantee the monotonicity property. Furthermore we propose different set of constraints which can be simply implemented. These procedures have been tested on the ground of many numerical experiments.

1 Introduction

Let $f(\mathbf{x};\boldsymbol{\psi})$ be the density of a mixture of k multinormal distribution $f(\mathbf{x};\boldsymbol{\psi}) = \alpha_1 p(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \cdots + \alpha_k p(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ where the α_j are the mixing weights and $p(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ is the density function of a $q-$multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. Finally, we set $\boldsymbol{\psi} = \{(\alpha_j,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j),\ j=1,\ldots,k\} \in \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the parameter space. It is well known that the log-likelihood function $\mathcal{L}(\boldsymbol{\psi})$ coming from a sample of N i.i.d. observations with law $f(\mathbf{x};\boldsymbol{\psi})$ is unbounded and presents many local spurious maxima, see McLachlan and Peel (2000); however under suitable hypotheses in Hathaway (1985) a constrained (global) maximum-likelihood formulation has been proposed which presents no singularities and a smaller number of spurious maxima by imposing the following constraint satisfied by the true set of parameters

$$\min_{1\le h\ne j\le k} \lambda_{\min}(\boldsymbol{\Sigma}_h\boldsymbol{\Sigma}_j^{-1}) \ge c, \qquad c\in(0,1] \tag{1}$$

where $\lambda_{\min}(\mathbf{A})$ is the smallest eigenvalue of $\mathbf{A}$. Such constraints are difficult to apply in algorithms like the EM, where the estimates of the covariance matrices are iteratively updated; for this aim they have been reformulated as

$$a \le \lambda_i(\boldsymbol{\Sigma}_j) \le b \qquad i=1,\ldots,q, \quad j=1,\ldots,k \tag{2}$$

where $\lambda_i(\mathbf{A})$ is the i^{th} eigenvalue of $\mathbf{A}$, in non increasing order, and a and b are positive numbers such that $a/b=c$, see Ingrassia (2004). In this way a set of stronger constraints are obtained; in fact, the inequalities

$$\lambda_{\min}(\boldsymbol{\Sigma}_h \boldsymbol{\Sigma}_j^{-1}) \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_h)}{\lambda_{\max}(\boldsymbol{\Sigma}_j)} \geq \frac{a}{b} = c\,, \qquad 1 \leq h \neq j \leq k \tag{3}$$

show that (2) implies (1).

In this paper we analyze the eigenvalue and eigenvector structure of the covariance matrices $\boldsymbol{\Sigma}_j$ in order to construct constrained EM algorithms which guarantee the monotonicity property of the unconstrained version. We also propose a new set of simple constraints which are weaker than (2).

2 An Algebraic Analysis of the Covariance Matrices

The EM algorithm generates a sequence of estimates $\{\boldsymbol{\psi}^{(m)}\}_m$, where $\boldsymbol{\psi}^{(0)}$ denotes the initial guess and $\boldsymbol{\psi}^{(m)} \in \boldsymbol{\Psi}$ for $m \in \mathbb{N}$, so that the corresponding sequence $\{\mathcal{L}(\boldsymbol{\psi}^{(m)})\}_m$, is not decreasing. The theory of the EM algorithm assures that $\mathcal{L}(\boldsymbol{\psi}^{(m+1)}) \geq \mathcal{L}(\boldsymbol{\psi}^{(m)})$. The E-step, on the $(m+1)$ iteration computes the quantities

$$u_{nj}^{(m+1)} = \frac{\alpha_j^{(m)} p(\mathbf{x}_n; \boldsymbol{\mu}_j^{(m)}, \boldsymbol{\Sigma}_j^{(m)})}{\sum_{h=1}^k \alpha_h^{(m)} p(\mathbf{x}_n; \boldsymbol{\mu}_h^{(m)}, \boldsymbol{\Sigma}_h^{(m)})} \qquad n = 1, \ldots, N, \quad j = 1, \ldots, k\,.$$

The M-step on the $(m+1)$ iteration requires the global maximization of the complete log-likelihood

$$Q(\boldsymbol{\psi}) = \sum_{j=1}^k \sum_{n=1}^N u_{nj}^{(m+1)} (\ln \alpha_j) + \sum_{j=1}^k q_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{4}$$

with respect to $\boldsymbol{\psi}$ over the parameter space $\boldsymbol{\Psi}$ to give the update estimate $\boldsymbol{\psi}^{(m+1)}$. To achieve this global maximization, let us first study the three separate maximizations with respect to $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_k]'$, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ $(j = 1, 2, \ldots, k)$.

1) Maximization with respect to $\boldsymbol{\alpha}$. It can be easily shown that the complete log-likelihood (4) obtains a maximum with respect to $\boldsymbol{\alpha}$ by setting

$$\alpha_j = \frac{1}{N} u_{\cdot j}^{(m+1)} = \frac{1}{N} \sum_{n=1}^N u_{nj}^{(m+1)} \qquad j = 1, \ldots, k\,,$$

2) Maximization with respect to $\boldsymbol{\mu}_j$. In this case, the maximization of (4) can be split into k independent maximizations of the terms $q_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$

$(j = 1, \ldots, k)$. Thus, let us write

$$q_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \sum_{n=1}^{N} u_{nj}^{(m+1)} \ln p(\mathbf{x}_n; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$
$$= \frac{1}{2} \sum_{n=1}^{N} u_{nj}^{(m+1)} \left[-q \ln(2\pi) - \ln|\boldsymbol{\Sigma}_j| - (\mathbf{x}_n - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_j)\right] \quad (5)$$

>From (5) easily follows that (4) obtains a maximum with respect to $\boldsymbol{\mu}_j$ when

$$\boldsymbol{\mu}_j = \frac{1}{u_{\cdot j}^{(m+1)}} \sum_{n=1}^{N} u_{nj}^{(m+1)} \mathbf{x}_n \qquad j = 1, \ldots, k\,;$$

3) Maximization with respect to $\boldsymbol{\Sigma}_j$ which is the most relevant for our scope. Again the maximization of $Q(\boldsymbol{\psi})$ can be split into k independent maximizations of the terms $q_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ $(j = 1, \ldots, k)$. By noting that

$$(\mathbf{x}_n - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_j) = \mathrm{tr}((\mathbf{x}_n - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_j))$$
$$= \mathrm{tr}(\boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_j)(\mathbf{x}_n - \boldsymbol{\mu}_j)'),$$

the function (5) can be also written as

$$q_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) =$$
$$= \frac{1}{2} \sum_{n=1}^{N} u_{nj}^{(m+1)} \left[-q \ln(2\pi) - \ln|\boldsymbol{\Sigma}_j| - \mathrm{tr}(\boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_j)(\mathbf{x}_n - \boldsymbol{\mu}_j)')\right].$$

After some algebras we get

$$q_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \gamma_j^{(m+1)} - \frac{1}{2} u_{\cdot j}^{(m+1)} \left[\ln|\boldsymbol{\Sigma}_j| + \mathrm{tr}\left(\boldsymbol{\Sigma}_j^{-1} \mathbf{S}_j\right)\right] \quad (6)$$

where for simplicity we set

$$\gamma_j = -\frac{q}{2} \ln(2\pi) u_{\cdot j}^{(m+1)}, \mathbf{S}_j = \frac{1}{u_{\cdot j}^{(m+1)}} \sum_{n=1}^{N} u_{nj}^{(m+1)} (\mathbf{x}_n - \boldsymbol{\mu}_j)(\mathbf{x}_n - \boldsymbol{\mu}_j)'.$$

The relation (6) shows that the maximization of $q_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with respect to $\boldsymbol{\Sigma}_j$ amounts to the minimization of $\ln|\boldsymbol{\Sigma}_j| + \mathrm{tr}\left(\boldsymbol{\Sigma}_j^{-1} \mathbf{S}_j\right)$.

Let $\boldsymbol{\Sigma}_j = \boldsymbol{\Gamma}_j \boldsymbol{\Lambda}_j \boldsymbol{\Gamma}_j'$ be the spectral decomposition of $\boldsymbol{\Sigma}_j$, where $\boldsymbol{\Lambda}_j = \mathrm{diag}(\lambda_{1j}, \ldots, \lambda_{qj})$ is the diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}_j$ in non decreasing order, and $\boldsymbol{\Gamma}_j$ is an orthogonal matrix whose columns are the standardized eigenvectors of $\boldsymbol{\Sigma}_j$. It is well known that (see for example: Theobald, 1975)

$$\mathrm{tr}\left(\boldsymbol{\Sigma}_j^{-1} \mathbf{S}_j\right) \geq \mathrm{tr}\left(\boldsymbol{\Lambda}_j^{-1} \mathbf{L}_j\right) = \sum_{i=1}^{q} \lambda_{ij}^{-1} l_{ij} \quad (7)$$

where $\mathbf{L}_j = \mathrm{diag}(l_{1j}, \ldots, l_{qj})$ is the diagonal matrix of the eigenvalues, in non decreasing order, of $\mathbf{S}_j$. In particular, the equality in (7) holds if and only if $\boldsymbol{\Sigma}_j$ and $\mathbf{S}_j$ have the same eigenvectors which are ordered with respect both $\lambda_{1j}, \ldots, \lambda_{qj}$ and $l_{1j}, \ldots, l_{qj}$. This implies that the minimum can be reached if and only if $\boldsymbol{\Sigma}_j$ has the same eigenvectors of $\mathbf{S}_j$. Under this condition, since $\ln|\boldsymbol{\Sigma}_j| = \sum_i \ln \lambda_{ij}$, the minimization of $\ln|\boldsymbol{\Sigma}_j| + \mathrm{tr}\left(\boldsymbol{\Sigma}_j^{-1}\mathbf{S}_j\right)$ with respect to $\boldsymbol{\Sigma}_j$ amounts to the minimization of

$$\sum_{i=1}^{q} \ln \lambda_{ij} + \sum_{i=1}^{q} \lambda_{ij}^{-1} l_{ij} = \sum_{i=1}^{q} \left(\ln \lambda_{ij} + \lambda_{ij}^{-1} l_{ij}\right) \tag{8}$$

with respect to $\lambda_{1j}, \ldots, \lambda_{qj}$, which is equivalent to q independent minimizations of $\ln \lambda_{ij} + \lambda_{ij}^{-1} l_{ij}$ with respect to $\lambda_{1j}, \ldots, \lambda_{qj}$, which give $\lambda_{ij} = l_{ij}$. In conclusion, the optimal $\boldsymbol{\Sigma}_j$ is obtained first by setting its eigenvectors equal to the ones of $\mathbf{S}_j$ and then doing the same with the eigenvalues. This can be simply done by setting $\boldsymbol{\Sigma}_j = \mathbf{S}_j$.

On the basis of the previous results, it should be noted that only the maximization with respect to $\boldsymbol{\Sigma}_j$ depends on the current values of the other parameters. It follows that the M-step can be done by maximizing:

1. $Q(\boldsymbol{\alpha}, \boldsymbol{\mu}_1^{(m)}, \ldots, \boldsymbol{\mu}_k^{(m)}, \boldsymbol{\Sigma}_1^{(m)}, \ldots, \boldsymbol{\Sigma}_k^{(m)})$ with respect to $\boldsymbol{\alpha}$ to get
$$\alpha_j^{(m+1)} = \frac{1}{N} u_{\cdot j}^{(m)} \quad ;$$
2. $Q(\boldsymbol{\alpha}^{(m+1)}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1^{(m)}, \ldots, \boldsymbol{\Sigma}_k^{(m)})$ with respect to $\boldsymbol{\mu}_j$ $(j = 1, \ldots, k)$ to get
$$\boldsymbol{\mu}_j^{(m+1)} = \frac{1}{u_{\cdot j}^{(m+1)}} \sum_{n=1}^{N} u_{nj}^{(m+1)} \mathbf{x}_n \quad ;$$
3. $Q(\boldsymbol{\alpha}^{(m+1)}, \boldsymbol{\mu}_1^{(m+1)}, \ldots, \boldsymbol{\mu}_k^{(m+1)}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k)$ with respect to $\boldsymbol{\Sigma}_j$ $(j = 1, \ldots, k)$ to get
$$\boldsymbol{\Sigma}_j^{(m+1)} = \mathbf{S}_j^{(m+1)} = \frac{1}{u_{\cdot j}^{(m+1)}} \sum_{n=1}^{N} u_{nj}^{(m+1)} (\mathbf{x}_n - \boldsymbol{\mu}_j^{(m+1)})(\mathbf{x}_n - \boldsymbol{\mu}_j^{(m+1)})' \quad .$$

The third step can be regarded as obtained according to the following three substeps:

i) set $\boldsymbol{\Gamma}_j^{(m+1)}$ equal to the orthonormal matrix whose columns are standardized eigenvectors of $\mathbf{S}_j^{(m+1)}$
ii) set $\boldsymbol{\Lambda}_j^{(m+1)} \leftarrow \mathrm{diag}(l_{1j}^{(m+1)}, \ldots, l_{qj}^{(m+1)})$;
iii) compute $\boldsymbol{\Sigma}_j^{(m+1)} \leftarrow \boldsymbol{\Gamma}_j^{(m+1)} \boldsymbol{\Lambda}_j^{(m+1)} \boldsymbol{\Gamma}_j^{(m+1)\prime}$.

This split into three substeps is not convenient in the ordinary EM algorithm. However, in the next section it will be shown how this help to formulate monotone algorithms for the constrained case.

3 Constrained Monotone EM Algorithms

The reformulation of the update of the covariance matrices $\boldsymbol{\Sigma}_j$ $(j = 1, \dots, k)$ presented in the previous section suggests some ideas for the the construction of EM algorithms such that the constraints (1) are satisfied while the monotonicity is preserved.

Approach A. The simplest approach is the following:

i) if $\lambda_{\min}(\mathbf{S}_j^{(m+1)})/\lambda_{\max}(\mathbf{S}_j^{(m+1)}) \geq c$ then set $\boldsymbol{\Sigma}_j^{(m+1)} \leftarrow \mathbf{S}_j^{(m+1)}$ otherwise set $\boldsymbol{\Sigma}_j^{(m+1)} \leftarrow \mathbf{S}_j^{(m)}$.

Approach B. A more refined strategy is:

i) set $\boldsymbol{\Gamma}_j^{(m+1)}$ equal to the orthogonal matrix whose columns are standardized eigenvectors of $\mathbf{S}_j^{(m+1)}$;
ii) if $\lambda_{\min}(\mathbf{S}_j^{(m+1)})/\lambda_{\max}(\mathbf{S}_j^{(m+1)}) \geq c$ then
set $\boldsymbol{\Lambda}_j^{(m+1)} \leftarrow \mathrm{diag}(l_{1j}^{(m+1)}, \dots, l_{qj}^{(m+1)})$ otherwise set $\boldsymbol{\Lambda}_j^{(m+1)} \leftarrow \boldsymbol{\Lambda}_j^{(m)}$;
iii) compute the covariance matrix by $\boldsymbol{\Sigma}_j^{(m+1)} \leftarrow \boldsymbol{\Gamma}_j^{(m+1)} \boldsymbol{\Lambda}_j^{(m+1)} \boldsymbol{\Gamma}_j^{(m+1)'}$.

Approach C. Another approach consists in imposing the constraints to the eigenvalues, that is to find an update of $\boldsymbol{\Sigma}_j$ which maximizes (4) under the constraints (3), see Ingrassia (2004). According to the results given in the previous section, the optimal update is a symmetric matrix having the same eigenvectors as $\mathbf{S}_j^{(m+1)}$, and eigenvalues minimizing (8) under (2). It is easy to show that this can be achieved by setting

$$\lambda_{ij} = \begin{cases} a & \text{if } l_{ij}^{(m+1)} < a \\ l_{ij}^{(m+1)} & \text{if } a \leq l_{ij}^{(m+1)} \leq b \\ b & \text{if } l_{ij}^{(m+1)} > b \,. \end{cases}$$

We can summarize this strategy as follows:

i) set $\boldsymbol{\Gamma}_j^{(m+1)}$ equal to the orthogonal matrix whose columns are standardized eigenvectors of $\mathbf{S}_j^{(m+1)}$;
ii) afterward set

$$\lambda_{ij}^{(m+1)} \leftarrow \min\left(b, \max\left(a, l_{ij}^{(m+1)}\right)\right) \,. \tag{9}$$

iii) update the covariance matrix as $\boldsymbol{\Sigma}_j^{(m+1)} \leftarrow \boldsymbol{\Gamma}_j^{(m+1)} \boldsymbol{\Lambda}_j^{(m+1)} \boldsymbol{\Gamma}_j^{(m+1)'}$.

In this way the monotonicity is guaranteed once the initial guess $\boldsymbol{\Sigma}_j^{(0)}$ satisfies the constraints.

Approach D. A different kind of constraints on eigenvalues is here proposed by introducing a suitable parameterization for the covariance matrices of the mixture components. Let us rewrite $\boldsymbol{\Sigma}_j = \eta^2 \boldsymbol{\Omega}_j$ $(j = 1, \ldots, k)$, where the matrices $\boldsymbol{\Omega}_j$ are such that

$$\lambda_i(\boldsymbol{\Omega}_j) \leq \frac{1}{c}, \qquad \text{and} \qquad \min_{ij} \lambda_i(\boldsymbol{\Omega}_j) = 1 \tag{10}$$

for $i = 1, \ldots, q$ and $j = 1, \ldots, k$. They are weaker than (2), indeed if constraints (2) are satisfied and we set $\eta^2 = \min_{ij} \lambda_i(\boldsymbol{\Sigma}_j)$ and $\boldsymbol{\Omega}_j = \boldsymbol{\Sigma}_j/\eta^2$ then, by noting that $\lambda_i(\boldsymbol{\Sigma}_j) = \lambda_i(\boldsymbol{\Omega}_j)\eta^2$, we get (10) since

$$\lambda_i(\boldsymbol{\Sigma}_j) \leq b \Rightarrow \lambda_i(\boldsymbol{\Omega}_j) \leq \frac{b}{\eta^2} \leq \frac{b}{a} = \frac{1}{c} \quad \text{and} \quad \min_{ij} \lambda_i(\boldsymbol{\Omega}_j) = 1.$$

However they are stronger than (1), indeed if the constraints (10) hold then

$$\lambda_{\min}(\boldsymbol{\Sigma}_h \boldsymbol{\Sigma}_j^{-1}) \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_h)}{\lambda_{\max}(\boldsymbol{\Sigma}_j)} = \frac{\lambda_{\min}(\boldsymbol{\Omega}_h)}{\lambda_{\max}(\boldsymbol{\Omega}_j)} \geq \frac{1}{1/c} = c, \qquad 1 \leq h \neq j \leq k$$

In order to implement in the EM algorithm the new set of constraints, only the last step must be changed. The update of η^2 and $\boldsymbol{\Omega}_j$ must maximize the complete log-likelihood, i.e., they have to maximize the function

$$\sum_{j=1}^{k} q_j(\boldsymbol{\mu}_j, \eta^2, \boldsymbol{\Omega}_j) = \gamma_j - \frac{1}{2} \sum_{j=1}^{k} u_{\cdot j}^{(m+1)} \left[\ln |\eta^2 \boldsymbol{\Omega}_j| + \operatorname{tr}\left(\eta^{-2} \boldsymbol{\Omega}_j^{-1} \mathbf{S}_j\right) \right].$$

It can be shown that the maximum with respect to η^2 is achieved by setting

$$\eta^2 = \frac{1}{Nq} \sum_{j=1}^{k} u_{\cdot j}^{(m+1)} \left[\operatorname{tr}\left(\boldsymbol{\Omega}_j^{-1} \mathbf{S}_j\right) \right], \tag{11}$$

while, on the basis of the results shown in this section and in the previous one, it can be easily shown that the maximum with respect to $\boldsymbol{\Omega}_j$ is obtained by setting its eigenvectors equal to the eigenvectors of $\mathbf{S}_j$ and the eigenvalues

$$\lambda_i(\boldsymbol{\Omega}_j) = \min\left(\frac{1}{c}, \max\left(1, \frac{l_{ij}}{\eta^2}\right)\right). \tag{12}$$

We can summarize this fourth strategy as follows:

i) set $\boldsymbol{\Gamma}_j^{(m+1)}$ equal to the orthogonal matrix whose columns are standardized eigenvectors of $\mathbf{S}_j^{(m+1)}$;

ii) update η^2 as

$$(\eta^{(m+1)})^2 = \frac{1}{Nq} \sum_{j=1}^{k} u_{\cdot j}^{(m+1)} \left[\operatorname{tr}\left((\boldsymbol{\Omega}_j^{(m)})^{-1} \mathbf{S}_j^{(m+1)} \right) \right];$$

iii) set

$$\lambda_{ij}^{(m+1)} \leftarrow (\eta^{(m+1)})^2 \min\left(\frac{1}{c}, \max\left(1, \frac{l_{ij}^{(m+1)}}{(\eta^{(m+1)})^2}\right)\right) ;$$

iv) update the covariance matrix as $\boldsymbol{\Sigma}_j^{(m+1)} \leftarrow \boldsymbol{\Gamma}_j^{(m+1)} \boldsymbol{\Lambda}_j^{(m+1)} \boldsymbol{\Gamma}_j^{(m+1)'}$.

It is important to note that in (12) the maximizer depends on the current value of η^2, while in (11) the maximizer depends on the current values of $\boldsymbol{\Omega}_j$ $(j = 1, \ldots, k)$. It follows that the sequential implementation of the above four steps leads to an increment of the complete log-likelihood but does not necessarily maximize it with respect to η^2 and $\boldsymbol{\Omega}_j$ $(j = 1, \ldots, k)$. This implies that the resulting algorithm is of the class ECM (Expectation Conditional Maximization) (see e.g. McLachlan & Krishnan, 1997) rather than EM. It is also important to note that the proposed algorithm does not necessarily gives a solution satisfying the constraint $\min_{ij} \lambda_i(\boldsymbol{\Omega}_j) = 1$ in (10); in this case, a correct solution can be obtained by setting

$$\lambda_i(\boldsymbol{\Omega}_j) \longleftarrow \frac{\lambda_i(\boldsymbol{\Omega}_j)}{\min_{ij} \lambda_i(\boldsymbol{\Omega}_j)} \quad \text{and} \quad \eta^2 \longleftarrow \eta^2 \min_{ij} \lambda_i(\boldsymbol{\Omega}_j) \tag{13}$$

and thus a new solution is obtained that satisfy the complete set of constraints by giving the same value of the log-likelihood. Also in this case the monotonicity is guaranteed once the initial guess $\boldsymbol{\Sigma}_j^{(0)}$ satisfies the constraints. Finally, it should be noted that strategies A and B do not necessarily maximize the complete log-likelihood at each iteration.

4 Numerical Results and Concluding Remarks

In this section we present some numerical results in order to evaluate the performance of approaches C and D, corresponding to the constraints (2) and (10). Further experiments have been carried out and they are presented in Ingrassia and Rocci (2006). We considered samples of size $N = 200$ generated from a mixture of three bi-variate normal distributions ($k = 3$ and $q = 2$) having the parameters $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3)$ where

$$\boldsymbol{\alpha} = (0.3, 0.4, 0.3)' \qquad \boldsymbol{\mu}_1 = (0,3)' \qquad \boldsymbol{\mu}_2 = (1,5)' \qquad \boldsymbol{\mu}_3 = (-3,8)'$$
$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

and the eigenvalues of the covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ are respectively: $\boldsymbol{\lambda}_1 = (1,2)'$, $\boldsymbol{\lambda}_2 = (0.382, 2.618)'$ and $\boldsymbol{\lambda}_3 = (1,3)'$. We generated 200 samples from this mixture. For each sample, we run the constrained EM algorithms following approaches C and D, starting from a set of points randomly chosen; the computation stopped when the difference between two consecutive log-likelihood values resulted less than 0.0001. The results, displayed in table 1, have been summarized by considering the mean of the sum of squared differences between the true parameters and the corresponding estimates (SSE), and the average number of iterations (# iter). On the same

Table 1. Mean values of the sum of squared errors of estimation and mean values for the number of iterations for constrained EM algorithms C and D

.

Strategy C			Strategy D		
a, b	SSE	# iter	c	SSE	# iter
0.38, 3	1.50	75	0.38/3	2.08	111
0.20, 4	2.02	79	0.20/4	4.42	129
0.10, 8	3.67	95	0.10/8	6.05	103
0.01,80	6.11	99	0.01/80	6.11	100
1.14, 9	3.89	161			

datasets we run also the unconstrained algorithm obtaining an average number of iterations equal to 99 and an average SSE equal to 6.11.
We can not draw general conclusions from this limited simulation study. However, we note that the two constrained algorithms outperforms always the unconstrained one. They are equivalent only when $(a, b) = (0.01, 80)$ because in this case the constraints are not active. We note also that the performances of the algorithms decreases when the constraints are less tight. The same consideration applies if we compare the two constrained algorithms when $a/b = c$: approach C is always better than D because it is the most constrained. Only in the last setting, D is better than C (note that $1.14/9 = 0.38/3$). This is due to the fact that the constraints are wrong for C. In conclusion, it seems that the choice between the two approaches depends on the information available on the eigenvalues: use C if the location is known, use D if only the ratio between the highest and the lowest is known.

References

HATHAWAY, R. J. (1985): A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics, 13, 795–800.*

INGRASSIA, S. (2004): A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods & Applications, 13, 151–166.*

INGRASSIA, S. and ROCCI, R. (2006): On the geometry of the EM algorithm for multinormal mixture models and constrained monotone versions. Technical report.

McLACHLAN, G. J. and KRISHNAN, T. (1997): *The EM Algorithm and Extensions.* John Wiley & Sons, New York.

McLACHLAN, G. J. and PEEL, D. (2000): *Finite Mixture Models.* John Wiley & Sons, New York.

THEOBALD, C. M. (1975): An inequality with applications to multivariate analysis. *Biometrika, 62, 461–466.*