Rapporto di ricerca n. 233

Francesca GRESELIN – Salvatore INGRASSIA

Constrained EM Algorithms for Gaussian Mixtures of Factor Analyzers

Settembre 2012

Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali
Università degli Studi di Milano Bicocca
Segreteria di redazione: Andrea Bertolini
Via Bicocca degli Arcimboldi, 8 - 20126 Milano - Italia
Tel +39/02/64483102/3 - Fax +39/02/64483105

Noname manuscript No.
(will be inserted by the editor)

Constrained EM Algorithms for Gaussian Mixtures of Factor Analyzers

Francesca Greselin · Salvatore Ingrassia

Received: date / Accepted: date

Abstract Mixtures of factor analyzers are becoming more and more popular in the area of model based clustering of high-dimensional data. In data modeling, according to the likelihood approach, it is well known that the loglikelihood function may present spurious maxima and singularities and this is due to specific patterns of the estimated covariance structure. To reduce such drawbacks, in this paper we introduce and implement a procedure for the parameter 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. We then analyze and measure its performance, compared to the usual non-constrained approach, via some simulations and applications to real data sets.

Keywords Factor Analyzers Modeling, Mixture Models, Model-Based Clustering, Parsimonious gaussian models.

1 Introduction and motivation

Finite mixture distributions have been receiving a growing interest in statistical modeling. Their central role is mainly due to their double nature: they combine the flexibility of non-parametric models with the strong and useful mathematical properties of parametric models. According to this approach, when we know that a sample of observations has been

Francesca Greselin
Dipartimento di Metodi Quantitativi per l'Economia e le Scienze Aziendali
Università di Milano-Bicocca
Via Bicocca degliArcimboldi 8 - 20126 Milano (Italy). E-mail: francesca.greselin@unimib.it

Salvatore Ingrassia
Dipartimento di Economia e Impresa
Università di Catania
Corso Italia 55, - Catania (Italy). E-mail: s.ingrassia@unict.it

drawn from different populations, we assume a specific distributional form in each of the underlying populations. The purpose is to decompose the sample into its mixture components (Peel and McLachlan, 2000), which, for quantitative data, are usually modeled as a multivariate Gaussian distribution, and to estimate parameters. The assumption of underlying normality, besides the elegant analytic properties, allows also to employ the EM algorithm for the ML estimation of the parameters. On the other side, when considering a large number of observed variables, Gaussian mixture models can provide an over-parameterized solution as, besides the mixing weights, it is required to estimate the mean vector and the covariance matrix for each component (Peel and McLachlan, 2000). As a consequence, we observe at the same time an undue load of computationally intensive procedures for the estimation.

This is the reason why a number of strategies have been introduced in the literature to avoid over-parameterized solutions. Among the various proposal, some authors developed methodologies for variable selection (see, f.i., Liu *et al.* (2003) and Hoff (2005) in the Bayesian framework, Pan and Shen (2007) and Raftery and Dean (2006) in the frequentist one). They further motivate their approach from the observation that the presence of non-informative variables can be strongly misleading for some clustering methods. With the same purpose of parsimony, but a completely different approach, Banfield and Raftery (1993) devised a methodology to identify common patterns among the component-covariance matrices, which arose a great attention in the literature. Along a slightly different line of thinking, Ghahramani and Hilton (1997) and McLachlan *et al.* (2003) proposed to employ latent variables to perform dimensional reduction in each component, starting from the consideration that in many phenomena some few unobserved features could be explained by the many observed ones.

In this paper we address mixtures of factor analyzers by assuming 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, by facing some of its issues and giving practical recipes to overcome them. It is well known that the EM algorithm generates a sequence of estimates, starting from an initial guess, so that the corresponding sequence of the log-likelihood values is not decreasing. However, the convergence toward the MLE is not guaranteed, because the log-likelihood is unbounded and presents local maxima, so that the final estimate crucially depends on the initial guess. This issue has been investigated by many authors, starting from the seminal paper of Redner and Walker (1984). Along the lines of (Ingrassia, 2004), here we will 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. We then analyze and compare its performance, compared to the usual non-constrained approach.

We have organized the rest of the paper as follows. In Section 2 we summarize main ideas about Gaussian Mixtures of Factor Analyzer model; in Section 3 we provide fairly extensive notes concerning the likelihood function and the AECM algorithm. Some well known considerations (Hathaway, 1985) related to spurious maximizers and singularities in the EM algorithm are recalled in Section 4, which motivate our proposal to introduce constraints on factor analyzers. Further, we give a detailed methodology to implement such constraints into the EM algorithm. In Section 6 we show and discuss the improved performance of our procedure, on the ground of some numerical results based on both simulated and real data. Section 7 contains concluding notes and provides ideas for future research.

2 The Gaussian Mixture of Factor analyzers

Within the Gaussian Mixture (GM) model-based approach to density estimation and clustering, the density of the d-dimensional random variable $\mathbf{Y}$ of interest is modelled as a mixture of a number, say G, of multivariate normal densities in some unknown proportions $\pi_1, \dots \pi_G$. That is, each data point is taken to be a realization of the mixture probability density function,

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

where $\phi_d(\mathbf{x};\mu,\Sigma)$ denotes the d-variate normal density function with mean μ and covariance matrix Σ. Here the vector $\theta_{GM}(d,G)$ of unknown parameters consists of the $(G-1)$ mixing proportions π_g, the $G \times d$ elements of the component means μ_g, and the $\frac{1}{2}Gd(d+1)$ distinct elements of the component-covariance matrices Σ_g. Therefore, the G-component normal mixture model (1) with unrestricted component-covariance matrices is a highly parametrized model. We crucially need some method for parsimonious parametrization of the matrices Σ_g, because they requires $O(d^2)$ parameters. Among the various proposals for dimensionality reduction, we are interested here in considering Mixtures of Gaussian Factor Analyzers (MGFA), which allows to explain data by explicitly modeling correlations between variables in multivariate observations. We postulate a finite mixture of linear sub-models for the distribution of the full observation vector $\mathbf{Y}$, given the (unobservable) factors $\mathbf{U}$. That is we can provide a local dimensionality reduction method by assuming that the distribution of the observation $\mathbf{Y}_i$ can be given as

$$\mathbf{X}_i = \mu_g + \Lambda_g \mathbf{U}_{ig} + \mathbf{e}_{ig} \quad \text{with probability} \quad \pi_g (g = 1,\dots,G) \quad \text{for } i = 1,\dots,n, \tag{2}$$

where Λ_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 Ψ_g is a $d \times d$ diagonal matrix $(g = 1,\dots,G)$. We suppose that $q < d$, which means that q unobservable factors are jointly explaining the d observable features of the statistical units. Under these assumptions, the mixture of factor analyzers model is given by (1), where the g-th component-covariance matrix Σ_g has the form

$$\Sigma_g = \Lambda_g \Lambda_g' + \Psi_g \quad (g = 1,\dots,G). \tag{3}$$

The parameter vector $\theta_{GMFA}(d,q,G)$ now consists of the elements of the component means μ_g, the Λ_g, and the Ψ_g, along with the mixing proportions π_g $(g = 1,\dots,G-1)$, on putting $\pi_G = 1 - \sum_{i=1}^{G-1} \pi_g$.

Comparing the two approaches and willing now to measure the gained parsimony when we use mixtures of factor analyzers, with respect to the more usual gaussian mixtures, we have to choose values of q such that $q < \frac{1}{2}(d-1)$. This is the only requirement for parsimony. Further, denoting by $|\theta_{CovGM}(d,G)|$ and $|\theta_{CovGMFA}(d,q,G)|$, the number of the estimated parameters for the covariance matrices in the GM and MFA models, respectively, we can express the relative reduction $RR(d,q,G) = RR(d,q)$ given by

$$\begin{aligned} RR(d,q) &= \frac{|\theta_{CovGM}(d,G)| - |\theta_{CovGMFA}(d,q,G)|}{|\theta_{CovGM}(d,G)|} \\ &= \frac{\frac{1}{2}d(d+1)G - d(q+1)G}{\frac{1}{2}d(d+1)G} \\ &= \frac{d-2q-1}{d+1}. \end{aligned}$$

In Tables 2 we report the relative reduction, in term of lower number of the estimated parameters for the covariance matrices in the GMFA models, with respect to the GM models. The relative reduction represents the extent to which the factor model offers a simpler interpretation for the behaviour of **x** than the alternative assumption given by the gaussian mixture model.

Table 1 Relative reduction $RR(d,q)$ for different values of d and q (where '-' means: 'no reduction')

$d \vert q$	1	2	3	4	5	6	7
1	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-
4	0.20	-	-	-	-	-	-
5	0.33	-	-	-	-	-	-
6	0.43	0.14	-	-	-	-	-
7	0.50	0.25	-	-	-	-	-
8	0.56	0.33	0.11	-	-	-	-
9	0.60	0.40	0.20	-	-	-	-
10	0.64	0.45	0.27	0.09	-	-	-
11	0.67	0.50	0.33	0.17	-	-	-
12	0.69	0.54	0.38	0.23	0.08	-	-
13	0.71	0.57	0.43	0.29	0.14	-	-
14	0.73	0.60	0.47	0.33	0.20	0.07	-
15	0.75	0.63	0.50	0.38	0.25	0.13	-

3 The likelihood function and the EM algorithm for GMFA

In this section we summarize the main steps of the EM algorithm for mixtures of Factor analyzers, see e.g. McLachlan and Peel (2000) for details.

Let $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$ be a sample of size n from density (1), and let $\mathbf{x}_i$ $(i = 1,\ldots,n)$ denotes the realization of $\mathbf{X}_i$ in (2). For given data $\mathbf{X}$, parameters in (1) can be estimated according to the likelihood approach via the EM algorithm, where the likelihood function is given by:

$$\begin{aligned} L(\theta;\mathbf{Y}) &= \prod_{i=1}^{n}\left\{\sum_{g=1}^{G}\phi_d(\mathbf{x}_i;\mu_g,\Sigma_g)\,\pi_g\right\} \\ &= \prod_{i=1}^{n}\left\{\sum_{g=1}^{G}\phi_d(\mathbf{x}_i;\mu_g,\Lambda_g,\Psi_g)\,\pi_g\right\}, \end{aligned}$$

where we set $\Sigma_g = \Lambda_g\Lambda_g' + \Psi_g$ $(g = 1,\ldots,G)$. Consider the augmented data $\{(\mathbf{x}_i,\mathbf{u}_{ig},\mathbf{z}_i),\ i = 1,\ldots,n\}$, where $\mathbf{z}_i = (z_{i1},\ldots,z_{ig})'$, with $z_{ig} = 1$ if $\mathbf{x}_i$ comes from the g-th population and $z_{ig} = 0$ otherwise. Then, the complete-data likelihood function can be written in the form:

$$L_c(\theta;\mathbf{Y}) = \prod_{i=1}^{n}\prod_{g=1}^{G}\left[\phi_d\left(\mathbf{x}_i|\mathbf{u}_i;\mu_g,\Lambda_g,\Psi_g\right)\phi_q(\mathbf{u}_{ig})\pi_g\right]^{z_{ig}}. \tag{4}$$

In particular, due to the factor structure of the model, see Meng and van Dyk (1997), we have to consider the alternating expectation-conditional maximization (AECM) algorithm. Such a procedure is an extension of the EM algorithm that uses different specifications of missing data at each stage. The idea is to partition $\theta = (\theta_1',\theta_2')'$ in such a way that $L(\theta;\mathbf{Y})$ is easy to maximize for θ_1 given θ_2 and vice versa. Then, we can iterate between these two conditional maximizations until convergence. In this case $\theta_1 = \{\pi_g,\mu_g\, g = 1,\ldots,G\}$ where the missing data are the unobserved group labels $\mathbf{Z} = (\mathbf{z}_1',\ldots,\mathbf{z}_n')$, and the second part of the parameters vector is given by $\theta_2 = \{(\Lambda_g,\Psi_g),\ g = 1,\ldots,G\}$ where the missing data are the group labels $\mathbf{Z}$ and the unobserved latent factors $\mathbf{U} = (\mathbf{U}_{11},\ldots,\mathbf{U}_{nG})$. In this case, the application of the AECM algorithm consists of two cycles, and there is one E-step and one CM-step alternatively considering θ_1 and θ_2 in each pair of cycles.

First Cycle. Here $\theta_1 = \{\pi_g,\mu_g,\ g = 1,\ldots,G\}$ where the missing data are the unobserved group labels $\mathbf{Z} = (\mathbf{z}_1',\ldots,\mathbf{z}_n')$. The complete data likelihood is

$$L_{c1}(\theta_1) = \prod_{i=1}^{n}\prod_{g=1}^{G}\left[\phi_d\left(\mathbf{x}_i;\mu_g,\Sigma_g\right)\pi_g\right]^{z_{ig}} \tag{5}$$

The E-step on the first cycle on the $(k+1)$-th iteration requires the calculation of $Q_1(\theta_1;\theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}\{\mathscr{L}_c(\theta_1)|\mathbf{Y}\}$ which is the expected complete-data log-likelihood given the data $\mathbf{Y}$ and using the current estimate $\theta^{(k)}$ for θ. In practice it requires calculating $\mathbb{E}_{\theta^{(k)}}\{Z_{ig}|\mathbf{Y}\}$ and usual computations show that this step is achieved by replacing each z_{ig} by its current conditional expectation given the observed data $\mathbf{x}_i$, that is we replace z_{ig} by $z_{ig}^{(k+1/2)}$, where

$$z_{ig}^{(k+1)} = \frac{\phi_d\left(\mathbf{x}_i|\mu_g^{(k)},\Lambda_g^{(k)},\Psi_g^{(k)}\right)\pi_g^{(k)}}{\sum_{j=1}^{G}\phi_d\left(\mathbf{x}_i|\mu_j^{(k)},\Lambda_j^{(k)},\Psi_j^{(k)}\right)\pi_j^{(k)}}, \tag{6}$$

On the M-step, the maximization of this complete-data log-likelihood yields

$$\begin{aligned}
\pi_g^{(k+1)} &= \frac{\sum_{i=1}^n z_{ig}^{(k+1)}}{n} \\
\mu_g^{(k+1)} &= \frac{1}{n_g}\sum_{i=1}^n z_{ig}^{(k+1)}\mathbf{x}_i
\end{aligned}$$

where $n_g^{(k+1)} = \sum_{i=1}^n z_{ig}^{(k+1)}$. According to notation in McLachlan and Peel (2000), we set $\theta^{(k+1/2)} = (\theta_1^{(k+1)'}, \theta_2^{(k)'})'$.

Second Cycle. Here $\theta_2 = \{\Sigma_g, g = 1,\dots,G\} = \{(\Lambda_g, \Psi_g), g = 1,\dots,G\}$ where the missing data are the unobserved group labels $\mathbf{Z}$ and the latent factors $\mathbf{U}$. Therefore, the complete data likelihood is

$$\begin{aligned}
L_{c2}(\theta_2) &= \prod_{i=1}^n\prod_{g=1}^G \left[\phi_d\left(\mathbf{x}_i|\mathbf{u}_{ig};\mu_g^{(k+1)},\Sigma_g\right)\phi_q\left(\mathbf{u}_{ig}\right)\pi_g^{(k+1)}\right]^{z_{ig}} \\
&= \prod_{i=1}^n\prod_{g=1}^G \left[\phi_d\left(\mathbf{x}_i|\mathbf{u}_{ig};\mu_g^{(k+1)},\Lambda_g,\Psi_g\right)\phi_q\left(\mathbf{u}_{ig}\right)\pi_g^{(k+1)}\right]^{z_{ig}}, \qquad (7)
\end{aligned}$$

where

$$\begin{aligned}
\phi_d\left(\mathbf{x}_i|\mathbf{u}_{ig};\mu_g^{(k+1)},\Lambda_g,\Psi_g\right) &= \frac{1}{|2\pi\boldsymbol{\Psi}_g|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\mu_g^{(k+1)}-\Lambda_g\mathbf{u}_{ig})'\boldsymbol{\Psi}_g^{-1}(\mathbf{x}_i-\mu_g^{(k+1)}-\Lambda_g\mathbf{u}_{ig})\right\}. \\
\phi_q(\mathbf{u}_{ig}) &= \frac{1}{(2\pi)^{q/2}}\exp\left\{-\frac{1}{2}\mathbf{u}_{ig}'\mathbf{u}_{ig}\right\}.
\end{aligned}$$

Now the complete data log-likelihood is given by

$$\begin{aligned}
\mathscr{L}_{c2}(\theta_2) = &-\frac{nd}{2}\ln 2\pi + \sum_{g=1}^G n_g\ln\pi_g + \frac{1}{2}\sum_{i=1}^n\sum_{g=1}^G z_{ig}\ln|\boldsymbol{\Psi}_g^{-1}| \\
&-\frac{1}{2}\sum_{i=1}^n\sum_{g=1}^G z_{ig}\mathrm{tr}\left\{(\mathbf{x}_i-\mu_g^{(k+1)}-\Lambda_g\mathbf{u}_{ig})(\mathbf{x}_i-\mu_g^{(k+1)}-\Lambda_g\mathbf{u}_{ig})'\boldsymbol{\Psi}_g^{-1}\right\}. \qquad (8)
\end{aligned}$$

Some algebras lead to the following estimate of $\{(\Lambda_g, \Psi_g), g = 1,\dots,G\}$:

$$\begin{aligned}
\hat{\Lambda}_g &= \mathbf{S}_g^{(k+1)}\gamma_g^{(k)'}[\Theta_g^{(k)}]^{-1} \\
\hat{\Psi}_g &= \mathrm{diag}\left\{\mathbf{S}_g^{(k+1)} - \hat{\Lambda}_g\gamma_g^{(k)}\mathbf{S}_g^{(k+1)}\right\}.
\end{aligned}$$

where we set

$$\begin{aligned}
\mathbf{S}_g^{(k+1)} &= (1/n_g^{(k+1)})\sum_{i=1}^n z_{ig}^{(k+1)}(\mathbf{x}_i-\mu_g^{(k+1)})(\mathbf{x}_i-\mu_g^{(k+1)})' \\
\gamma_g^{(k)} &= \Lambda_g^{(k)'}(\Lambda_g^{(k)}\Lambda_g^{(k)'}+\boldsymbol{\Psi}_g^{(k)})^{-1} \\
\Theta_{ig}^{(k)} &= \mathbf{I}_q - \gamma_g^{(k)}\Lambda_g^{(k)} + \gamma_g^{(k)}(\mathbf{x}_i-\mu_g)(\mathbf{x}_i-\mu_g)'\gamma_g^{(k)'}.
\end{aligned}$$

Hence the maximum likelihood estimates $\hat{\Lambda}_g$ and $\hat{\Psi}_g$ for Λ and Ψ can be obtained by alternatively computing the update estimates Λ_g^+ and Ψ_g^+, by

$$\Lambda_g^+ = \mathbf{S}_g^{(k+1)} \gamma_g^{(k)\prime} [\Theta_g^{(k)}]^{-1} \qquad \text{and} \qquad \Psi_g^+ = \operatorname{diag}\left\{\mathbf{S}_g^{(k+1)} - \Lambda_g^+ \gamma_g^{(k)} \mathbf{S}_g^{(k+1)}\right\}, \tag{9}$$

and, from the latter, computing the update estimates γ_g^+ and Θ_g^+ by

$$\gamma_g^+ = \Lambda_g^{'}(\Lambda_g \Lambda_g^{'} + \Psi_g)^{-1} \qquad \text{and} \qquad \Theta_g^+ = \mathbf{I}_q - \gamma_g \Lambda_g + \gamma_g \mathbf{S}_g^{(k+1)} \gamma_g^{'}, \tag{10}$$

iterating these two steps until convergence on $\hat{\Lambda}_g$ and $\hat{\Psi}_g$, so giving $\Lambda_g^{(k+1)}$ and $\Psi_g^{(k+1)}$.

In summary, the procedure can be described as follows. For a given initial random clustering $\mathbf{z}^{(0)}$, on the $(k+1)-th$ iteration, the algorithm carries out the following steps, for $g = 1,\dots,G$:

1. Compute $z_{ig}^{(k+1)}$ and consequently obtain $\pi_g^{(k+1)}$ and $\mu_g^{(k+1)}$ and also $n_g^{(k+1)}$ and $\mathbf{S}_g^{(k+1)}$;
2. Set a starting value for Λ_g and Ψ_g from $\mathbf{S}_g^{(k+1)}$;
3. Repeat the following steps, until convergence on $\hat{\Lambda}_g$ and $\hat{\Psi}_g$:
   (a) Compute γ_g^+ and Θ_g^+ from (10);
   (b) Set $\gamma_g \leftarrow \gamma_g^+$ and $\Theta_g \leftarrow \Theta_g^+$;
   (c) Compute $\Lambda_g^+ \leftarrow \mathbf{S}_g^{(k+1)} \gamma_g^{\prime} (\Theta_g^{-1})$ and $\Psi_g^+ \leftarrow \operatorname{diag}\left\{\mathbf{S}_g^{(k+1)} - \Lambda_g^+ \gamma_g \mathbf{S}_g^{(k+1)}\right\}$;
   (d) Set $\Lambda_g \leftarrow \Lambda_g^+$ and $\Psi_g \leftarrow \Psi_g^+$;

To completely describe the algorithm, here we give more details on how to specify the starting values for Λ_g and Ψ_g from $\mathbf{S}_g^{(k+1)}$, as it is needed in Step 2.

Starting from the eigen-decomposition of $\mathbf{S}_g^{(k+1)}$, say $\mathbf{S}_g^{(k+1)} = \mathbf{A}_g \mathbf{B}_g \mathbf{A}_g^{\prime}$, computed on the base of $z_{ig}^{(k+1)}$, the main idea is that Λ_g has to synthesize the ”more important” relations between the d observed features. Then, looking at the equality $\Sigma_g = \Gamma_g \Gamma_g^{\prime} + \Psi_g$, the initial values of Λ_g were set as

$$\lambda_{ij} = \sqrt{d_j} a_{ij} \tag{11}$$

where b_j is the jth largest eigenvalue of $\mathbf{S}_g^{(k+1)}$ and a_{ij} is the ith element of the corresponding eigenvector $\mathbf{a}_j$ (the jth column in A_g), for $i \in \{1,2,\dots,p\}$ and $j \in \{1,2,\dots,q\}$. Finally the Ψ_g matrices can be initialized by the position $\Psi_g = \operatorname{diag}\{\mathbf{S}_g^{(k+1)} - \Lambda_g \Lambda_g^{\prime}\}$.

4 Likelihood maximization in constrained parametric spaces

Properties of maximum likelihood estimation for normal mixture models have been deeply investigated. It is well known that $\mathscr{L}(\theta)$ is unbounded on Θ and may present many local maxima. Day (1969) was perhaps the first noting that any small number of sample points, grouped sufficiently close together, can give raise to spurious maximizers, corresponding to parameters points with greatly differing component standard deviation. To overcome this issue and to prevent $\mathscr{L}(\theta)$ from singularities, Hathaway (1985) proposed a constrained maximum likelihood formulation for mixture of univariate normal distributions, suggesting a natural extension to the multivariate case. Let $c \in (0,1]$, then the following constraints

$$\min_{1\leq h\neq j\leq k} \lambda(\Sigma_h\Sigma_j^{-1}) \geq c \tag{12}$$

on the eigenvalues λ of $\Sigma_h\Sigma_j^{-1}$ leads to properly defined, scale-equivariant, consistent ML-estimators for the mixture-of-normal case, see Hennig (2004). It is easy to show that a sufficient condition for (12) is

$$a \leq \lambda_{ig} \leq b, \qquad i = 1,\ldots,d; \qquad g = 1,\ldots,G \tag{13}$$

where λ_{ig} denotes the ith eigenvalue of Σ_g i.e. $\lambda_{ig} = \lambda_i(\Sigma_g)$, and for $a, b \in \mathbb{R}^+$ such that $a/b \geq c$, see Ingrassia (2004). Differently from (12), condition (13) can be easily implemented in any optimization algorithm. Let us consider the constrained parameter space Θ_c of Θ:

$$\begin{aligned}\Theta_c = \{&(\pi_1,\ldots,\pi_G,\mu_1,\ldots,\mu_G,\Sigma_1,\ldots,\Sigma_G) \in \mathbb{R}^{k[1+d+(d^2+d)/2]} : \\ &\pi_g \geq 0,\, \pi_1+\cdots+\pi_G = 1,\, a \leq \lambda_{ig} \leq b, \quad g = 1,\ldots,G \quad i = 1,\ldots,d\}.\end{aligned} \tag{14}$$

Due to the structure of the covariance matrix Σ_g given in (3), the bound (13) yields

$$\lambda_{\min}(\Lambda_g\Lambda_g' + \Psi_g) \geq a \qquad \text{and} \qquad \lambda_{\max}(\Lambda_g\Lambda_g' + \Psi_g) \leq b, \qquad g = 1,\ldots,G \tag{15}$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and the largest eigenvalue of $(\cdot)$ respectively. Since $\Lambda_g\Lambda_g'$ and Ψ_g are symmetric and positive definite, then it results:

$$\lambda_{\min}(\Lambda_g\Lambda_g' + \Psi_g) \geq \lambda_{\min}(\Psi_g) \tag{16}$$

$$\lambda_{\max}(\Lambda_g\Lambda_g' + \Psi_g) \leq \lambda_{\max}(\Lambda_g\Lambda_g') + \lambda_{\max}(\Psi_g), \tag{17}$$

for $g = 1,\ldots,G$, see Lütkepohl (1996). Since Ψ_g is a diagonal matrix, then

$$\lambda_{\min}(\Psi_g) = \min_i \psi_{ig}, \tag{18}$$

$$\lambda_{\max}(\Psi_g) = \max_i \psi_{ig}, \tag{19}$$

where ψ_{ig} denotes the i-th entry along the diagonal of the matrix Ψ_g.

Concerning the square $d \times d$ matrix $\Lambda_g\Lambda_g'$ ($g = 1,\ldots,G$), we can get its eigenvalue decomposition, i.e. we can find Λ_g and Γ_g such that

$$\Lambda_g\Lambda_g' = \Gamma_g\Delta_g\Gamma_g' \tag{20}$$

where Γ_g is the orthonormal matrix whose rows are the eigenvectors of $\Lambda_g\Lambda_g'$ and $\Delta_g = \text{diag}(\delta_{1g},\ldots,\delta_{dg})$ is the diagonal matrix of the eigenvalues of $\Lambda_g\Lambda_g'$, sorted in non increasing order, i.e. $\delta_{1g} \geq \delta_{2g} \geq \ldots \geq \delta_{dg}$.

Now, on the $d \times q$ rectangular matrix Λ_g we can apply the singular value decomposition, so giving $\Lambda_g = \mathbf{U}_g\mathbf{D}_g\mathbf{V}_g'$, where $\mathbf{U}_g$ is a $d \times d$ unitary matrix (i.e., such that $\mathbf{U}_g'\mathbf{U}_g = \mathbf{I}_d$) and $\mathbf{D}_g$ is a $d \times q$ rectangular diagonal matrix with q nonnegative real numbers on the diagonal, known as *singular values*, and $\mathbf{V}_g$ is a $q \times q$ unitary matrix. The d columns of $\mathbf{U}$ and the q columns of $\mathbf{V}$ are called the left singular vectors and right singular values of Λ_g, respectively. Now we have that

$$\Lambda_g \Lambda_g' = (\mathbf{U}_g \mathbf{D}_g \mathbf{V}_g')(\mathbf{V}_g \mathbf{D}_g' \mathbf{U}_g') = \mathbf{U}_g \mathbf{D}_g \mathbf{I}_q \mathbf{D}_g' \mathbf{U}_g' = \mathbf{U}_g \mathbf{D}_g \mathbf{D}_g' \mathbf{U}_g' \tag{21}$$

and equating (20) and (21) we get $\Gamma_g = \mathbf{U}_g$ and $\Delta_g = \mathbf{D}_g \mathbf{D}_g'$, that is

$$\operatorname{diag}(\delta_{1g}, \ldots, \delta_{qg}) = \operatorname{diag}(d_{1g}^2, \ldots, d_{qg}^2)\,. \tag{22}$$

with $d_{1g} \geq d_{2g} \geq \cdots \geq d_{qg} \geq 0$. Thus it results

$$\lambda_{\max}(\Lambda_g \Lambda_g') = d_{1g}^2. \tag{23}$$

Results (18), (19) and (23) express some relationships between the eigenvalues of $\boldsymbol{\Sigma}_g$ and the singular values of Λ_g and the diagonal elements of $\boldsymbol{\Psi}_g$. In particular, the bounds on the eigenvalues of $\Lambda_g \Lambda_g' + \boldsymbol{\Psi}_g$ in (15) are satisfied when

$$\min_i \psi_{ig} \geq a \tag{24}$$

$$d_{1g}^2 + \max_i \psi_{ig} \leq b. \tag{25}$$

5 Constraints on the covariance matrix for factor analyzers

The reformulation of the update of the covariance matrices $\boldsymbol{\Sigma}_g$ (for $g = 1, \ldots, G$) presented above, suggests how to modify the EM algorithm in such a way that the eigenvalues of the covariances are confined into suitable ranges. To this aim we have to implement the constraints (24) and (25).

As for the bound on the smallest eigenvalue of $\boldsymbol{\Sigma}_g$, based on (16), on the $(k+1)$th iteration, after the procedure described at the end of Section 3, we carry out the following conditional assignment

- if $\psi_{ig}^{(k+1)} < a$ then set $\psi_{ig}^{(k+1)} \leftarrow a$,

so that condition (24) is satisfied, for a given real value $a > 0$.

Concerning the upper bound on the largest eigenvalue of $\boldsymbol{\Sigma}$, we first recall that $\boldsymbol{\Psi}_g$ is the covariance matrix of the errors $\mathbf{e}_{ig}$ of the G linear submodels in (2), i.e. we have assumed that the $\mathbf{e}_{ig}$ are independently $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$ distributed, with $\boldsymbol{\Psi}_g = \operatorname{diag}\{\psi_{1g}, \ldots, \psi_{dg}\}$. Thus we can assume that the largest eigenvalue of $\boldsymbol{\Psi}_g$, that is $\max_i \psi_{ig}$ is small compared to the eigenvalues $d_{1g}^2, \ldots, d_{qg}^2$ of $\Lambda_g \Lambda_g'$. Hence, in order to satisfy the condition (25), on the $(k+1)$th iteration, after the procedure described at the end of Section 3, for a given real value $b > 0$, we proceed as follows:

1. Decompose Λ_g according to the singular value decomposition as $\Lambda_g = \mathbf{U}_g \mathbf{D}_g \mathbf{V}_g'$;
2. Compute the squared singular values $(d_1^2, \ldots, d_q^2)$ of Λ;

3. Create a copy $\mathbf{D}_g^*$ of $\mathbf{D}_g$;
4. For $i = 1$ to d, if $d_{ig}^2 > b - \max_i \psi_{ig}^{(k+1)}$, then set $d_i \leftarrow \sqrt{b - \max_i \psi_{ig}^{(k+1)}}$ into $\mathbf{D}_g^*$ at the corresponding place, otherwise go to step 5;
5. Compute $\Lambda_g^* = \mathbf{U}_g \mathbf{D}_g^* \mathbf{V}_g'$;
6. Stop.

It is important to remark that the resulting EM algorithm is monotone, once the initial guess, say Σ_g^0 satisfies the constraints. Further, as shown in the case of gaussian mixtures in Ingrassia and Rocci (2007), the maximization of the complete loglikelihood is guaranteed. Form the other side, it is apparent that he above recipes require some a priori information on the covariance structure of the mixture, throughout the bounds a and b. A weaker constraint could be imposed directly on the ratio a/b in

6 Numerical studies

The aim of this section is to show the improvement we attain by using the constrained EM algorithm, in terms of its ability to provide the ”right” solution, when compared to the unconstrained approach. This analysis will be carried out by firstly considering two synthetic samples and then some real data-sets.

6.1 Simulations

We have considered three mixtures of G components of d-variate normal distributions, for different values of the parameters vector θ. To generate a synthetic dataset, the Λ_g and Ψ_g matrices have to be chosen for $g = 1, \ldots, G$, so that multivariate normal variates with covariances given by $\Sigma_g = \Lambda_g \Lambda_g' + \Psi_g$ can be easily generated.

The point of local maximum corresponding to the consistent estimator θ^* (i.e. the ”right” solution), has been chosen to be the limit of the EM algorithm using θ as initial estimate, i.e. considering the crisp z_{ig} which assign each observation to the mixture component from which it has been generated, for $i = 1, \ldots, N$ and $g = 1, \ldots, G$.

We run a hundred times the unconstrained EM algorithm, followed by the constrained EM algorithm, both were fed with the same starting values. We generated a set of 100 different random initial clusterings to initialize the algorithm at each run. To this aim, we draw each time a set of random starting values for the z_{ig} from the multinomial distribution, with the same given values of the parameters $(\alpha_0, \alpha_1, \ldots, \alpha_G)$ and G we employed before, when generating the synthetic sample. In this way, when the algorithms evaluate the first cycle of the EM, the randomly generated $\hat{z}_{ig}$ were taken as the starting group membership labels. The initial values for the elements of Λ_g and Ψ_g were obtained as described at the end of Section 3 from the eigen-decomposition of $\mathbf{S}_g$, and the algorithms run until convergence or reached the maximum number of iterations (failure).

We decided to employ the convergence criterion based on the Aitken acceleration procedure (Aitken, 1926), to estimate the asymptotic maximum of the log-likelihood at each iteration of the EM algorithm. Based on this estimate, a decision can be made regarding whether or not the algorithm has reached convergence; that is, whether or not the log-likelihood is sufficiently close to its estimated asymptotic value. The Aitken acceleration at iteration k is given by

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}},$$

where $l^{(k+1)}$, $l^{(k)}$, and $l^{(k-1)}$ are the log-likelihood values from iterations $k+1$, k, and $k-1$, respectively. Then, the asymptotic estimate of the log-likelihood at iteration $k+1$ is given by

$$l_{\infty}^{(k+1)} = l^{(k)} + \frac{1}{1-a^{(k)}}\left(l^{(k+1)} - l^{(k)}\right),$$

see Böhning *et al.* (1994). In our analyses, the algorithms stopped when $l_{\infty}^{(k+1)} - l^{(k)} < \varepsilon$, with $\varepsilon = 0.01$.

Computer programs were written in the R language; the different experiments and the obtained results are described below.

Mixture 1: $G = 3$, $d = 6$, $q = 2$, $N = 150$.
The sample was generated with weights $\alpha = (0.3, 0.4, 0.3)'$ according to the following parameters:

$$\begin{aligned}
\mu_1 &= (0,0,0,0,0,0)' \\
\mu_2 &= (5,5,5,5,5,5)' \\
\mu_3 &= (10,10,10,10,10,10)'
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Psi}_1 &= \operatorname{diag}(0.1,0.1,0.1,0.1,0.1,0.1) \\
\boldsymbol{\Psi}_2 &= \operatorname{diag}(0.4,0.4,0.4,0.4,0.4,0.4) \\
\boldsymbol{\Psi}_3 &= \operatorname{diag}(0.2,0.2,0.2,0.2,0.2,0.2)
\end{aligned}$$

$$\Lambda_1 = \begin{pmatrix} 0.50 & 1.00 \\ 1.00 & 0.45 \\ 0.05 & -0.50 \\ -0.60 & 0.50 \\ 0.50 & 0.10 \\ 1.00 & -0.15 \end{pmatrix} \qquad \Lambda_2 = \begin{pmatrix} 0.10 & 0.20 \\ 0.20 & 0.50 \\ 1.00 & -1.00 \\ -0.20 & 0.50 \\ 1.00 & 0.70 \\ 1.20 & -0.30 \end{pmatrix} \qquad \Lambda_3 = \begin{pmatrix} 0.10 & 0.20 \\ 0.20 & 0.00 \\ 1.00 & 0.00 \\ -0.20 & 0.00 \\ 1.00 & 0.00 \\ 0.00 & -1.30 \end{pmatrix}.$$

The covariance matrices $\Sigma_g = \Lambda_g\Lambda_g' + \Psi_g$ $(g = 1,2,3)$ have respectively the following eigenvalues:

$$\lambda(\Sigma_1) = (3.17, 1.63, 0.10, 0.10, 0.10, 0.10)'$$
$$\lambda(\Sigma_2) = (4.18, 2.27, 0.40, 0.40, 0.40, 0.40)'$$
$$\lambda(\Sigma_3) = (2.29, 1.93, 0.20, 0.20, 0.20, 0.20)',$$

in particular the largest eigenvalue is equal to 4.18.

First we run the uncontrained algorithm: the right solution has been attained in 30% of cases. Afterwards, to compare how the choice of the values a and b could influence the performance of the constrained EM, we run the constrained algorithm for different values of the upper bound b on the largest eigenvalue, , while maintaining $a = 0.0001$, see Table 3. In Figure 1 we plot the classified data on the factor spaces given by $\hat{\Lambda}_1, \hat{\Lambda}_2$ and $\hat{\Lambda}_3$. under the right classification, while in Figure 2 we give the classification obtained according to a spurious maximum of the likelihood function.

Table 2 Mixture 1: Percentage of convergence to the right maximum of the constrained EM algorithms for some pairs (a,b).

		b		
6	10	15	20	25
100%	100%	98%	97%	87%

Mixture 2: $G = 4$, $d = 7$, $q = 2$, $N = 100$.
The sample was generated with weights $\alpha = (0.2, 0.3, 0.35, 0.15)'$ according to the following parameters:

$$\mu_1 = (0,0,0,0,0,0,0)'$$
$$\mu_2 = (5,5,5,5,5,5,5)'$$
$$\mu_3 = (10,10,10,10,10,10,10,)'$$
$$\mu_4 = (15,15,15,15,15,15,15)'$$

$$\Psi_1 = \text{diag}(0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2)$$
$$\Psi_2 = \text{diag}(0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25)$$
$$\Psi_3 = \text{diag}(0.15, 0.15, 0.15, 0.15, 0.15, 0.15, 0.15)$$
$$\Psi_4 = \text{diag}(0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$$

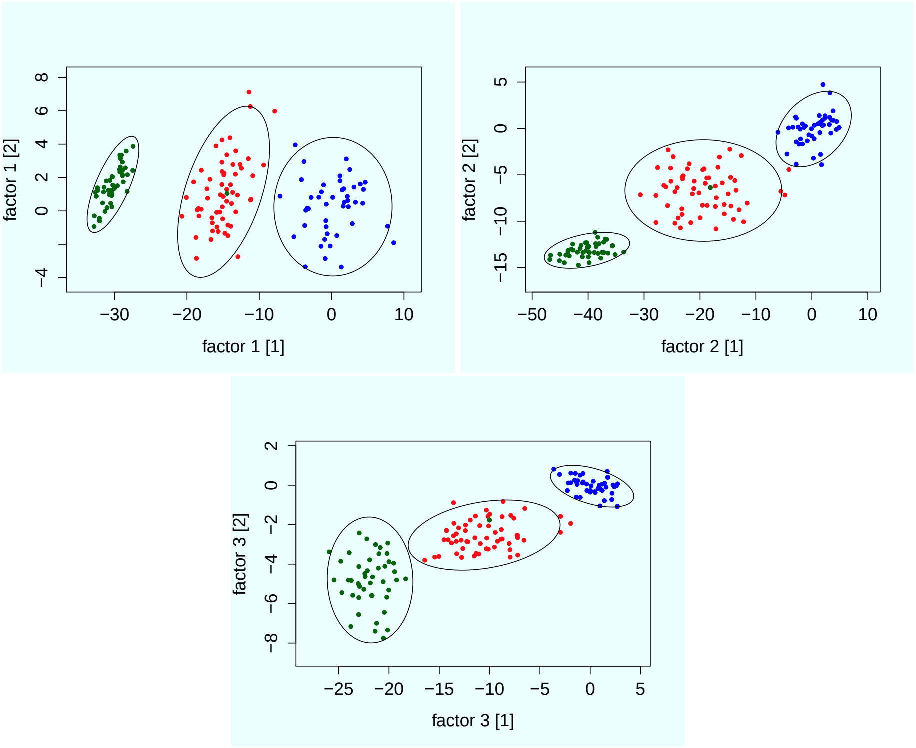

Fig. 1 Mixture 1: plot of the classified data on the factor spaces, under the ”right” solution given by the algorithm

$$\Lambda_1 = \begin{pmatrix} 0.30 & 0.60 \\ 0.60 & 0.27 \\ 0.03 & -0.30 \\ -0.36 & 0.30 \\ 0.30 & 0.06 \\ 0.60 & -0.09 \\ -0.63 & 1.50 \end{pmatrix} \quad \Lambda_2 = \begin{pmatrix} 0.08 & 0.16 \\ 0.16 & 0.40 \\ 0.80 & -0.80 \\ -0.16 & 0.40 \\ 0.80 & 0.56 \\ 0.96 & -0.24 \\ 1.60 & -0.24 \end{pmatrix} \quad \Lambda_3 = \begin{pmatrix} 0.07 & 0.14 \\ 0.14 & 0.00 \\ 0.70 & 0.00 \\ -0.14 & 0.00 \\ 0.70 & 0.00 \\ 0.00 & -0.91 \\ 0.70 & -0.70 \end{pmatrix} \quad \Lambda_4 = \begin{pmatrix} 0.04 & 0.08 \\ 0.08 & 0.00 \\ 0.40 & 0.00 \\ -0.08 & 0.00 \\ 0.40 & 0.00 \\ 0.00 & -0.52 \\ -0.40 & 0.80 \end{pmatrix}.$$

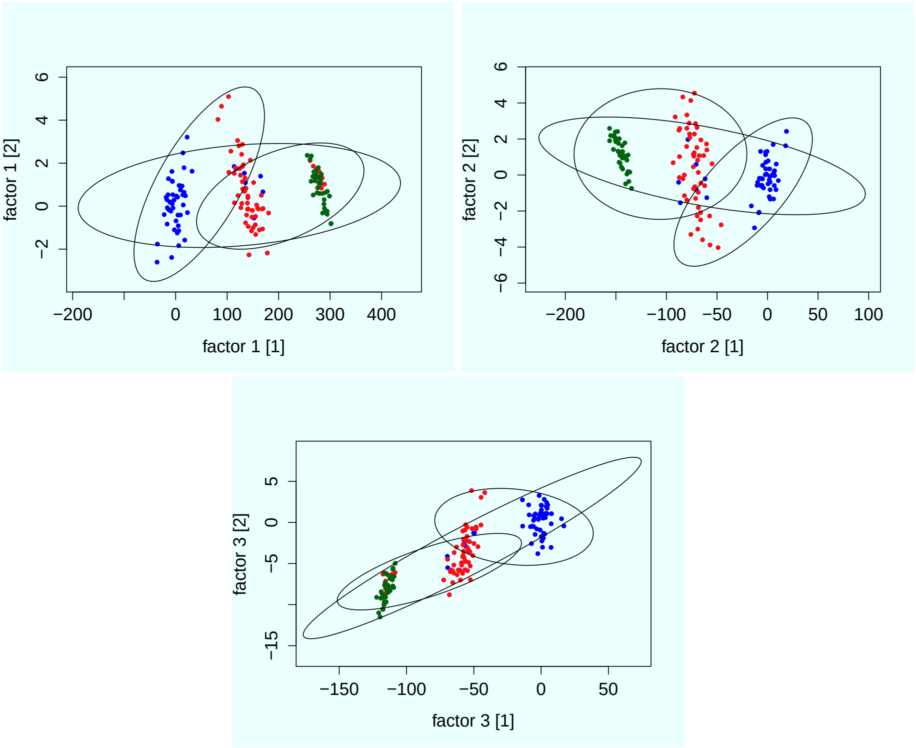

Fig. 2 Mixture 1: plot of the classified data on the factor spaces, giving an example of the wrong classification, which is obtained when the algorithm converges to a spurious maximum of the loglikelihood

The covariance matrices $\Sigma_g = \Lambda_g \Lambda_g' + \Psi_g$ $(g = 1,2,3)$ have respectively the following eigenvalues:

$$\lambda(\Sigma_1) = (4.10, 1.14, 0.33, 0.21, 0.15, 0.09, 0.04)'$$
$$\lambda(\Sigma_2) = (7.62, 1.18, 0.34, 0.20, 0.18, 0.12, 0.05)'$$
$$\lambda(\Sigma_3) = (3.36, 1.36, 0.24, 0.17, 0.14, 0.10, 0.09)'$$
$$\lambda(\Sigma_4) = (2.08, 0.48, 0.11, 0.09, 0.07, 0.06, 0.02)',$$

in particular the largest eigenvalue is equal to 7.62.

First we run the uncontrained algorithm: the right solution has been never attained on the hundred run. Afterwards, to compare how the choice of the values a and b could influence the performance of the constrained EM, we run the constrained algorithm for different values of the upper bound b on the largest eigenvalue, while maintaining $a = 0.0001$, see Table **??**. In Figure 3 we plot the classified data on the factor spaces given by $\hat{\Lambda}_1, \hat{\Lambda}_2, \hat{\Lambda}_3$ and $\hat{\Lambda}_4$. under the right classification, while in Figure 4 we give the classification obtained according to a spurious maximum of the likelihood function.

eigenval

Table 3 Mixture 2: Percentage of convergence to the right maximum of the constrained EM algorithms for some pairs (a,b).

		b		
6	10	15	20	25
71%	62%	48%	29%	29%

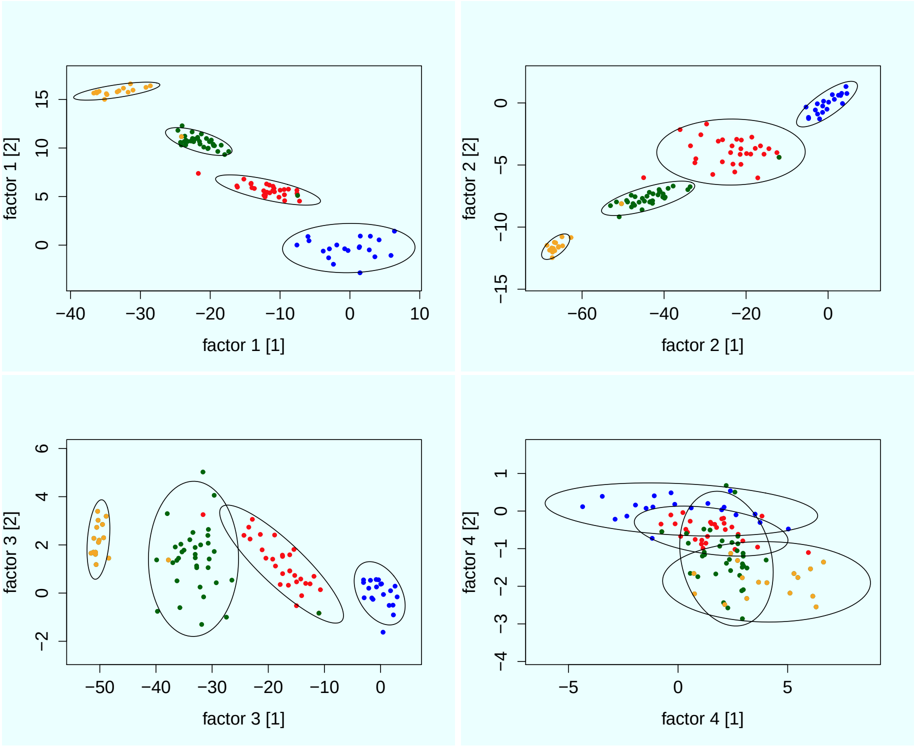

Fig. 3 Mixture 2: plot of the classified data on the factor spaces, under the ”right” solution given by the algorithm

7 Concluding remarks

In this paper we have considered mixtures of factor analyzers, with the purpose of considerably reducing the number of parameters to be estimated with respect to the mixtures of gaussian models. Latent variables can indeed be employed to perform dimensional reduction in each component, starting from the consideration that in many phenomena some few unobserved features could be explained by the many observed ones. For both family of models, the gaussian mixtures and the mixtures of factor analyzers, however, the loglikelihood function may present spurious maxima and singularities and this is due to specific patterns of the estimated covariance structure. It is known, from the literature, that a con-

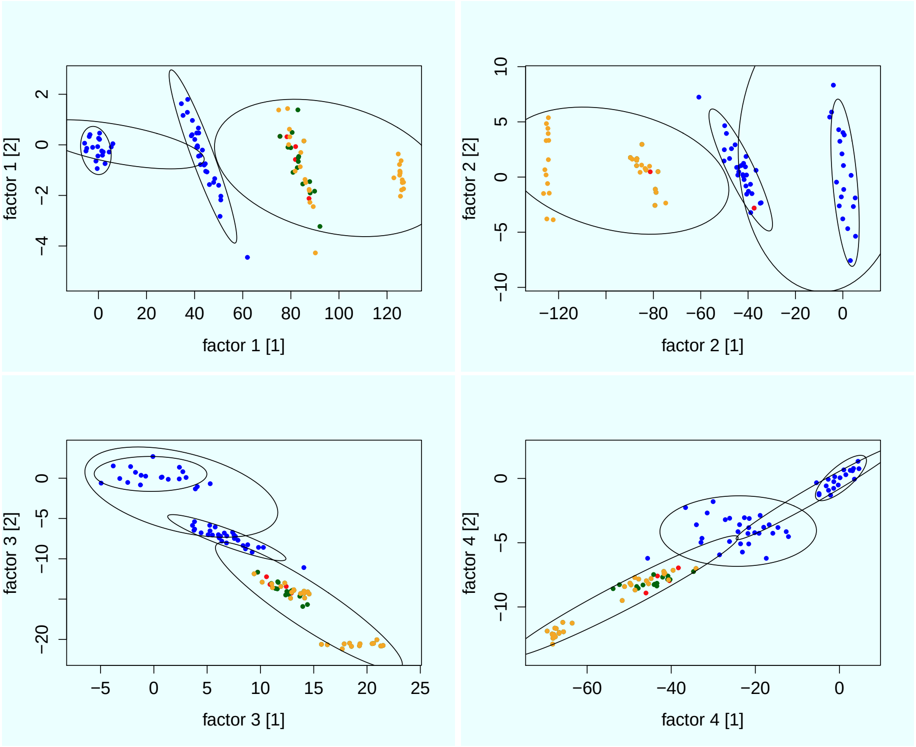

Fig. 4 Mixture 2: plot of the classified data on the factor spaces, giving an example of the wrong classification, which is obtained when the algorithm converges to a spurious maximum of the loglikelihood

strained formulation of the EM algorithm have shown to be able to considerably reduce such drawbacks for gaussian mixtures. Motivated by these considerations, in this paper we introduced a constrained approach for gaussian mixtures of factor analyzers. In particular, we implemented a methodology to maximize the likelihood function in a constrained parameter space, having no singularities and a reduced number of spurious local maxima. The performance of the newly introduced estimation approach has been shown and compared to the usual non-constrained one. The results shows that the problematic convergence of the EM, even more critical when dealing with mixture of gaussian factor analyzers, can be greatly improved. This improvement has been observed both via some numerical simulations on synthetic samples and via applications to real data sets.

Appendix: formulas for evaluating the inverse and the determinant of the matrix $\Sigma = \Lambda\Lambda' + \Psi$

The following notes are given to show that the formulas in Mardia's and in McLachlan's books (see Mardia *et al.* (2003) and McLachlan and Peel (2000)), for the inverse and the determinant of Σ are equivalent. We employed these formulas when writing the code for the EM algorithm, because the latter allows to avoid inverting any non-diagonal $p \times p$ matrix and the former offers a computational shortcut.

Let us begin with the formula of the inverse of $\Sigma = \Lambda\Lambda' + \Psi$.
Starting from Mardia's formula (Section A.2.4 Inverse, page 458, formula A.2.4f):

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA$$

and setting $A = \Psi$, $B = \Lambda$, $C = I$ e $D = \Lambda'$ then we obtain formula (8.11) in (McLachlan and Peel, 2000)

$$(\Psi + \Lambda I \Lambda')^{-1} = \Psi^{-1} - \Psi^{-1}\Lambda(I^{-1} + \Lambda'\Psi^{-1}\Lambda)^{-1}\Lambda'\Psi.$$

The formula is also called the ”Woodbury identity”. The AECM algorithm requires the inversion of the $p \times p$ covariance matrices $\Sigma_1, \ldots, \Sigma_G$ at each iteration; this becomes increasingly computationally expensive as the number of variables p gets larger. One of the main computational advantages of using the Woodbury identity, is that it can be used to avoid the inversion of any non-diagonal $p \times p$ matrices.

Now, focusing on the determinant of $\Sigma = \Lambda\Lambda' + \Psi$, starting from Mardia *et al.* (2003) (Section A.2.3 Determinants and cofactors, page 457, formula A.2.3k), we have

$$\begin{aligned}\det(A+BC) &= \det(A^{-1})\det(I_d + A^{-1}BC) \\ &= \det(A^{-1})\det(I_q + CA^{-1}B).\end{aligned}$$

Now, taking into account the first and the third term in the chain of equalities and setting $A = \Psi$, $B = \Lambda, C = \Lambda'$, we get

$$\det(\Lambda\Lambda' + \Psi) = \det(\Psi)\det(I - \Lambda'\Psi^{-1}\Lambda).$$

This result seems quite different from McLachlan's formula, which involves a ratio between determinants instead of a product. Indeed, it is easy to show that

$$I - \Lambda'\Psi^{-1}\Lambda$$

is the inverse of

$$I - \Lambda'(\Lambda\Lambda' + \Psi)^{-1}\Lambda.$$

Now, using the relation

$$\det(A^{-1}) = 1/\det(A)$$

we obtain the desired equality between Mardia's and McLachlan's formulas.

References

Aitken, A. (1926). On Bernoulli's numerical solution of algebraic equations. In *Proceedings of the Royal Society of Edinburgh*, volume 46, pages 289–305.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**(3), 803–821.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**(2), 373–388.

Ghahramani, Z. and Hilton, G. (1997). The EM algorithm for mixture of factor analyzers. *Techical Report CRG-TR-96-1*.

Hathaway, R. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, **13**(2), 795–800.

Hoff, P. (2005). Subset clustering of binary sequences, with an application to genomic abnormality data. *Biometrics*, **61**, 1027–1036.

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

Ingrassia, S. and Rocci, R. (2007). Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Computational Statistics & Data Analysis*, **51**, 5339–5351.

Liu, J., Zhang, J., Palumbo, M., and Lawrence, C. (2003). Bayesian clustering with variable and transformation selection (with discussion). *Bayesian Statistics*, **7**, 249–275.

Lütkepohl, H. (1996). *Handbook of matrices*. John Wiley & Sons, Chichester.

Mardia, K., Kent, J., and Bibby, J. (2003). *Multivariate analysis*. Academic Press, London, UK.

McLachlan, G., Peel, D., and Bean, R. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, **41**, 379–388.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, New York.

Meng, X. and van Dyk, D. (1997). The EM algorithm – an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**(3), 511–567.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection). *Journal of machine learning research*, **8**, 1145–1164.

Peel, D. and McLachlan, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

Raftery, A. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, **101**(473), 168–178.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, **26**(2), pp. 195–239.