Statistical Methods & Applications (2004) 13: 151–166
DOI: 10.1007/s10260-004-0092-4

SMA

© Springer-Verlag 2004

A likelihood-based constrained algorithm for multivariate normal mixture models

Salvatore Ingrassia[⋆]

Dipartimento di Economia e Statistica, Università della Calabria, 87036 Arcavacata di Rende (CS), Italy
(e-mail: s.ingrassia@unical.it)

Abstract. It is well known that the log-likelihood function for samples coming from normal mixture distributions may present spurious maxima and singularities. For this reason here we reformulate some Hathaway's results and we propose two constrained estimation procedures for multivariate normal mixture modelling according to the likelihood approach. Their perfomances are illustrated on the grounds of some numerical simulations based on the EM algorithm. A comparison between multivariate normal mixtures and the hot-deck approach in missing data imputation is also considered.

Key words: Mixture models, likelihood, EM algorithm, missing data

1. Introduction

Finite mixture distributions play a central role in statistical modelling as they combine much of the flexibility of non parametric models with some analytic properties of parametric models, see e.g. [20,13]. In this last decade such models have attracted the interest of many researchers so that mixtures have found a lot of new and interesting fields of application, see e.g. [10,19,17,1].

Let $f(\cdot;\boldsymbol{\psi})$ be the density function of a normal mixture distribution, where $\boldsymbol{\psi}$ assumes values in some parameter space $\boldsymbol{\Psi}$, and let $\mathcal{L}(\boldsymbol{\psi})$ be the log-likelihood function corresponding to a sample of size N with law $f(\cdot;\boldsymbol{\psi})$. The MLE $\hat{\boldsymbol{\psi}}$ is usually computed by means of suitable optimization procedures which generate a sequence of estimates $\{\boldsymbol{\psi}^{(m)}\}_m$ – starting from some initial guess $\boldsymbol{\psi}^{(0)}$ – so that the corresponding sequence $\{\mathcal{L}(\boldsymbol{\psi}^{(m)})\}_m$ is not decreasing. [**?**]). However,

[⋆] S. Ingrassia carried out the research as part of the project *Metodi Statistici e Reti Neuronali per l'Analisi di Dati Complessi* (PRIN 2000, resp. G. Lunetta).

the convergence towards $\hat{\psi}$ is not guaranteed for two main reasons: *i*) the log-likelihood function $\mathcal{L}(\psi)$ may be unbounded and this may cause the failure of the algorithm; *ii*) the log-likelihood function $\mathcal{L}(\psi)$ presents local maxima, so that the final estimate depends on the initial guess $\psi^{(0)}$.

This problem has been investigated in a sequence of papers, in particular in [5] it is proved that the likelihood function should be maximised in a suitable subset of $\boldsymbol{\Psi}$ and [6,8] present many related numerical results concerning the mixtures of univariate normal distributions; some ideas will be outlined in Sect. 2.

The rest of the paper concerns the multivariate case; some preliminary ideas were summarized in [3]. In Sect. 3 we reformulate some constraints introduced in [5] for the multivariate normal mixture decomposition according to the likelihood approach; in Sect. 4 we illustrate two simple procedures in order to implement such constraints in practical algorithms. Afterwards in Sect. 5 we present some numerical simulations based on the EM algorithm showing that our constrained formulation leads to no failures and to a clearly smaller number of spurious maxima. In Sect. 6 we consider an application in missing data imputation data by means of a mixture of multivariate normal distributions. Finally in Sect. 7 we shall present some concluding remarks and ideas for future work.

2. The univariate case

To begin with, we summarize some results concerning the univariate case. Let $f(x;\psi)$ be the density function of a mixture of k univariate normal components:

$$f(x;\psi) = \alpha_1 p(x;\mu_1,\sigma_1^2) + \cdots + \alpha_k p(x;\mu_k,\sigma_k^2) \tag{1}$$

where $p(x;\mu_j,\sigma_j)$ is the density function of a normal distribution with parameters μ_j,σ_j^2 ($j=1,\ldots,k$), and $\psi \in \boldsymbol{\Psi}$ where:

$$\begin{aligned}\boldsymbol{\Psi} = \{&(\alpha_1,\ldots,\alpha_k,\mu_1,\ldots,\mu_k,\sigma_1,\ldots,\sigma_k) \in \mathbb{R}^{3k} \; : \\ &\alpha_1+\cdots+\alpha_k = 1,\ \alpha_j \geq 0,\ \sigma_j > 0 \text{ for } j=1,\ldots,k\}.\end{aligned}$$

Let $\mathcal{L}(\psi)$ be the log-likelihood function of ψ for a sample $\mathbf{x} = (x_1,\ldots,x_N)$ of size N drawn from $f(x;\psi)$. It is well known that $\mathcal{L}(\psi;\mathbf{x})$ may show some singularities, thus the definition of estimate of maximum likelihood as the absolute maximum of $\mathcal{L}(\psi;\mathbf{x})$ lacks numerical sense. Under reasonable conditions, there exists only one strongly consistent solution of the likelihood equations:

$$\frac{\partial \mathcal{L}(\psi;\mathbf{x})}{\partial \psi} = \mathbf{0}$$

and it is a point of local maximum of the likelihood function, see [11,18]. Herewith the MLE estimate will be such a point of $\boldsymbol{\Psi}$. In [5] it is proved that if the sample $\mathbf{x} = (x_1,\ldots,x_N)$ drawn with law (1) contains at least $k+1$ distinct points, then for $c \in (0,1]$, there exists a point of absolute maximum which is strongly consistent in the subset of $\boldsymbol{\Psi}$ satisfying the constraint:

$$\min_{i\neq j}\left(\frac{\sigma_j^2}{\sigma_i^2}\right) \geq c > 0\,.$$

Numerical experiments about the univariate case are presented in [6,8].

3. The multivariate case

Let $f(\mathbf{x};\boldsymbol{\psi})$ be the density of a mixture of k multinormal distribution:

$$f(\mathbf{x};\boldsymbol{\psi}) = \alpha_1 p(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \cdots + \alpha_k p(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \tag{2}$$

where the α_j $(j = 1,\ldots,k)$ are the mixing weights and $p(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ is the density function of the multivariate normal distribution with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ (which is required to be positive definite, denoted by $\boldsymbol{\Sigma}_j > 0$). 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:

$$\begin{aligned}\boldsymbol{\Psi} = \{&(\alpha_1,\ldots,\alpha_k,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_k) \in \mathbb{R}^{k[1+p+(p^2+p)/2]} :\\ &\alpha_1 + \cdots + \alpha_k = 1, \alpha_j \geq 0,\ \boldsymbol{\Sigma}_j > 0 \quad \text{for } j = 1,\ldots,k\}\,.\end{aligned} \tag{3}$$

It is well known that the matrix $\boldsymbol{\Sigma}_j$ is positive definite if and only if all its eigenvalues are strictly positive. Thus, if we denote by $\lambda_i(\boldsymbol{\Sigma}_j)$ the i-th eigenvalue of the j-th covariance matrix $\boldsymbol{\Sigma}_j$, the parameter space (3) can be rewritten as:

$$\begin{aligned}\boldsymbol{\Psi} = \{&(\alpha_1,\ldots,\alpha_k,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_k) \in \mathbb{R}^{k[1+p+(p^2+p)/2]} :\\ &\alpha_1 + \cdots + \alpha_k = 1, \alpha_j \geq 0,\ \lambda_i(\boldsymbol{\Sigma}_j) > 0 \text{ for } j = 1,\ldots,k,\ i = 1,\ldots,p\}\,.\end{aligned} \tag{4}$$

Assume that we are provided with a sample drawn with law (2) containing at least $k+p$ distinct points and let $c \in (0,1]$. In [5] the following constraint:

$$\min_{1\leq h\neq j\leq k} \lambda(\boldsymbol{\Sigma}_h\boldsymbol{\Sigma}_j^{-1}) \geq c > 0 \tag{5}$$

on the eigenvalues of $\boldsymbol{\Sigma}_h\boldsymbol{\Sigma}_j^{-1}$ is imposed for some positive number c (satisfied by the true parameter), because this leads to a constrained (global) maximum-likelihood formulation as the assumptions of [12] are satisfied.

In practice the constraint (5) presents a drawback: indeed even if the bound (5) can be easily checked, in the present form it is inapplicable in the optimization procedures like the EM algorithm where the estimates of the covariance matrices are iteratively updated. Thus our goal becomes to impose suitable constraints on the eigenvalues of each covariance matrix $\boldsymbol{\Sigma}_j$ $(j = 1,\ldots,k)$ such that the bound (5) is satisfied. For this aim a suitable reformulation of the above constraint is proposed in Proposition 1. We premise the following definition of *matrix norm*, see e.g. [7, 9,16].

Definition 1 (Matrix norm) *Let $\mathcal{M}_p$ be the set of all $(p\times p)$ matrices over $\mathbb{R}$. A function $\|\cdot\| : \mathcal{M}_p \to \mathbb{R}$ is a matrix norm if for all $\mathbf{A},\mathbf{B} \in \mathcal{M}_p$ and $c \in \mathbb{R}$ it satisfies the following five axioms:*

i. $\|\mathbf{A}\| \geq 0$
ii. $\|\mathbf{A}\| = 0$ *if and only if* $\mathbf{A} = \mathbf{0}$,
iii. $\|c\mathbf{A}\| = |c|\cdot\|\mathbf{A}\|$,
iv. $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$,

v. $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$ *(submultiplicative* or *multiplivative property).*

Proposition 1 *Let* $\mathbf{A}, \mathbf{B}$ *two* $p \times p$ *symmetric and positive definite matrices. Then we have:*

$$\lambda_{\max}(\mathbf{AB}^{-1}) \leq \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{B})} \tag{6}$$

$$\lambda_{\min}(\mathbf{AB}^{-1}) \geq \frac{\lambda_{\min}(\mathbf{A})}{\lambda_{\max}(\mathbf{B})} \tag{7}$$

where $\lambda_{\min}(\cdot)$ *and* $\lambda_{\max}(\cdot)$ *are respectively the smallest and the largest eigenvalue of the matrix* $(\cdot)$.

Proof. The proof is based on some results of matrix analysis, see e.g. [7,16] for details. The hypotheses on the matrices $\mathbf{A}$ and $\mathbf{B}$ imply that $\mathbf{AB}^{-1}$ is positive definite (we remark that in general the matrix $\mathbf{AB}^{-1}$ is not symmetric even if $\mathbf{A}$ and $\mathbf{B}$ are) and thus all its eigenvalues are stricly positive. In particular the spectral radius $\rho(\mathbf{AB}^{-1}) = \max_i |\lambda_i(\mathbf{AB}^{-1})|$ of the matrix $\mathbf{AB}^{-1}$ – that is the largest eigenvalue of $\mathbf{AB}^{-1}$ in absolute value – is equal to largest eigenvalue $\lambda_{\max}(\mathbf{AB}^{-1}) := \max_i \lambda_i(\mathbf{AB}^{-1}) = \max_i |\lambda_i(\mathbf{AB}^{-1})|$.

The multiplicative property of matrix norms specializes in our case:

$$\|\mathbf{AB}^{-1}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}^{-1}\| \, . \tag{8}$$

Moreover, since for any matrix norm it results $\rho(\mathbf{AB}^{-1}) \leq \|\mathbf{AB}^{-1}\|$, we have:

$$\rho(\mathbf{AB}^{-1}) = \lambda_{\max}(\mathbf{AB}^{-1}) \leq \|\mathbf{AB}^{-1}\| \, . \tag{9}$$

Let us consider the spectral norm $\|\cdot\|_2$ of a matrix $\mathbf{A}$, which is defined as:

$$\|\mathbf{A}\|_2 := \max\{\sqrt{\lambda} : \ \lambda \text{ is an eigenvalue of } \mathbf{A}'\mathbf{A}\};$$

in particular if $\mathbf{A}$ is symmetric then $\|\mathbf{A}\|_2$ coincides with its spectral radius, that is $\|\mathbf{A}\|_2 = \rho(\mathbf{A})$. Thus we get (above we remarked that in general the matrix $\mathbf{AB}^{-1}$ is not symmetric even if $\mathbf{A}$ and $\mathbf{B}$ are):

$$\begin{aligned} \lambda_{\max}(\mathbf{AB}^{-1}) &\leq \|\mathbf{AB}^{-1}\|_2 \\ \|\mathbf{A}\|_2 &= \lambda_{\max}(\mathbf{A}) \\ \|\mathbf{B}^{-1}\|_2 &= \lambda_{\max}(\mathbf{B}^{-1}), \end{aligned} \tag{10}$$

so that from the relations (8), (9) and (10) we get:

$$\lambda_{\max}(\mathbf{AB}^{-1}) \leq \|\mathbf{AB}^{-1}\|_2 \leq \|\mathbf{A}\|_2 \cdot \|\mathbf{B}^{-1}\|_2 = \lambda_{\max}(\mathbf{A})\lambda_{\max}(\mathbf{B}^{-1}) \, .$$

The proof is completed considering the equality $\lambda_{\min}(\mathbf{A}) = 1/\lambda_{\max}(\mathbf{A}^{-1})$ because the eigenvalues of $\mathbf{A}^{-1}$ are the inverse of those of $\mathbf{A}$ and this yields:

$$\lambda_{\max}(\mathbf{AB}^{-1}) \leq \lambda_{\max}(\mathbf{A})\lambda_{\max}(\mathbf{B}^{-1}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{B})} \, .$$

The other statement (7) results in the same way by applying the inequality (8) to the inverse matrix:

$$\begin{aligned}\lambda_{\min}(\mathbf{A}\mathbf{B}^{-1}) &= \frac{1}{\lambda_{\max}[(\mathbf{A}\mathbf{B}^{-1})^{-1}]} = \frac{1}{\lambda_{\max}[(\mathbf{B}\mathbf{A}^{-1})]} \\ &\geq \frac{1}{\lambda_{\max}(\mathbf{B})\lambda_{\max}(\mathbf{A}^{-1})} = \frac{\lambda_{\min}(\mathbf{A})}{\lambda_{\max}(\mathbf{B})} \,.\end{aligned}$$

This completes the proof. □

In particular since the eigenvalues of $\boldsymbol{\Sigma}_j$ give the variances along the principal axes, then imposing bounds on the eigenvalues of $\boldsymbol{\Sigma}_j$ amounts to imposing bounds on these variances. According to the relation (7), the bound (5) can be satisfied by imposing suitable constraints on the eigenvalues of each covariance matrix $\boldsymbol{\Sigma}_j$ $(1 \leq j \leq k)$.

Let a, b two strictly positive constants such that $a/b \geq c$, where c satisfies the relation (5), and assume that the eigenvalues of the covariance matrices $\boldsymbol{\Sigma}_j$ satisfy the constraints:

$$a \leq \lambda_i(\boldsymbol{\Sigma}_j) \leq b \qquad i = 1, \ldots, p \,, j = 1, \ldots, k \,. \tag{11}$$

Then for any pair of covariance matrices $\boldsymbol{\Sigma}_h, \boldsymbol{\Sigma}_j$, the inequality (7) yields:

$$\lambda_{\min}(\boldsymbol{\Sigma}_h\boldsymbol{\Sigma}_j^{-1}) \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_h)}{\lambda_{\max}(\boldsymbol{\Sigma}_j)} \geq \frac{a}{b} \geq c > 0 \,, \qquad 1 \leq h \neq j \leq k \,. \tag{12}$$

In the following the log-likelihood will be maximised over the subspace $\boldsymbol{\Psi}_{a,b}$ of $\boldsymbol{\Psi}$:

$$\begin{aligned}\boldsymbol{\Psi}_{a,b} = \{(\alpha_1, \ldots, \alpha_k, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k) \in \mathbb{R}^{k[1+p+(p^2+p)/2]} \;: \\ \alpha_1 + \cdots + \alpha_k = 1, \; \alpha_j \geq 0, \; a \leq \lambda_i(\boldsymbol{\Sigma}_j) \leq b, \quad \text{for } j = 1, \ldots, k\} \,,\end{aligned}$$

for suitable positive numbers a, b such that $a/b \geq c$.

4. Two different procedures

In order to implement the constraints (11) in the EM algorithm, we suggest two different strategies based on some results of matrix theory, see e.g. [16].

Procedure 1. For any eigenvalue $\lambda(\mathbf{A})$ of a $p \times p$ matrix $\mathbf{A}$ it results:

$$\lambda(\mathbf{A} + \epsilon\mathbf{I}_p) = \lambda(\mathbf{A}) + \epsilon \tag{13}$$

$$\lambda(\gamma\mathbf{A}) = \gamma \cdot \lambda(\mathbf{A}) \tag{14}$$

where $\mathbf{I}_p$ is the p-dimensional identity matrix and γ is a non-zero real number. Let $\boldsymbol{\Sigma}_j^{(m)}$ be the j-th covariance matrix evaluated at the m-th iteration of the algorithm and denote respectively by $\lambda_{\min}(\boldsymbol{\Sigma}_j^{(m)})$ and $\lambda_{\max}(\boldsymbol{\Sigma}_j^{(m)})$ the smallest and the largest eigenvalue of $\boldsymbol{\Sigma}_j^{(m)}$ ($j = 1, \ldots, k$). Afterwards let us consider:

$$\lambda_*^{(m)} = \min\left\{\lambda_{\min}(\boldsymbol{\Sigma}_1^{(m)}), \ldots, \lambda_{\min}(\boldsymbol{\Sigma}_k^{(m)})\right\}$$
$$\lambda^{*(m)} = \max\left\{\lambda_{\max}(\boldsymbol{\Sigma}_1^{(m)}), \ldots, \lambda_{\max}(\boldsymbol{\Sigma}_k^{(m)})\right\} ,$$

and finally set $\epsilon_*^{(m)} = a - \lambda_*^{(m)}$. If $\lambda_{\min}(\boldsymbol{\Sigma}_j^{(m)}) < a$ for some $j \in \{1, \ldots, k\}$, the matrix $\boldsymbol{\Sigma}_j^{(m)} + \epsilon_*^{(m)}\mathbf{I}_p$ satisfies the lower bound in (11). Indeed, according to the relation (13), we have:

$$\begin{aligned}\lambda_{\min}(\boldsymbol{\Sigma}_j^{(m)} + \epsilon_*^{(m)}\mathbf{I}_p) = \lambda_{\min}(\boldsymbol{\Sigma}_j^{(m)}) + \epsilon_*^{(m)} &= \lambda_{\min}(\boldsymbol{\Sigma}_j^{(m)}) - \lambda_*^{(m)} + a \\ &\geq a \ . \end{aligned} \tag{15}$$

Analogous arguments apply in order to impose the upper bound in (11) according to the relation (14). If $\lambda_{\max}(\boldsymbol{\Sigma}_j^{(m)}) > b - a$ for some $j \in \{1, \ldots, k\}$, then let us set $\gamma = (b-a)/\lambda^{*(m)}$; thus the matrix $\gamma\boldsymbol{\Sigma}_j^{(m)}$ satisfies the constraint:

$$\lambda_{\max}\left(\frac{b-a}{\lambda^{*(m)}}\boldsymbol{\Sigma}_j^{(m)}\right) = \frac{b-a}{\lambda^{*(m)}}\lambda_{\max}(\boldsymbol{\Sigma}_j^{(m)}) \leq b - a < b \ . \tag{16}$$

Here we impose a slightly stronger constraint than necessary, i.e $\lambda_{\max}(\boldsymbol{\Sigma}_j^{(m)}) \leq b - a$ rather than $\lambda_{\max}(\boldsymbol{\Sigma}_j^{(m)}) \leq b$, in order to prevent an eccessive translation of the spectrum of the eigenvalues due to the constraint (15) on the smallest eigenvalue $\lambda_{\min}(\boldsymbol{\Sigma}_j^{(m)})$.

The two constraints (15) and (16) described above can be implemented quite easily in optimization algorithms, like the EM algorithm, performing the following steps at the m-th iteration, after computing the eigenvalues of $\boldsymbol{\Sigma}_j^{(m)}$:

1. If $\lambda_{\max}(\boldsymbol{\Sigma}_j^{(m)}) > b - a$ then set $\gamma = (b-a)/\lambda^{*(m)}$ and afterwards $\boldsymbol{\Sigma}_j^{(m)} \leftarrow \gamma\boldsymbol{\Sigma}_j^{(m)}$;
2. If $\lambda_{\min}(\boldsymbol{\Sigma}_j^{(m)}) < a$ then set $\epsilon_*^{(m)} = a - \lambda_*^{(m)}$ and afterwards $\boldsymbol{\Sigma}_j^{(m)} \leftarrow \boldsymbol{\Sigma}_j^{(m)} + \epsilon_*^{(m)}\mathbf{I}_p$.

Procedure 2. The previous recipe is quite simple but it involves a drawback. Indeed, when the constraints (13) and (14) act, they modify the entire spectrum (that is all eigenvalues) of the covariance matrix. A more elegant strategy will be presented below based on the spectral decomposition theorem and it allows the imposition of a constraint only to some selected eigenvalues of the covariance matrix. It is well known that any symmetric matrix $\mathbf{A}$ can be decomposed as:

$$\mathbf{A} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}' \tag{17}$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues of $\mathbf{A}$, and $\boldsymbol{\Gamma}$ is an orthogonal matrix whose columns are standardized eigenvectors; the symbol $'$ denotes matrix transpose.

The idea is to utilise formula (17) in a constructive way rather than as a decomposition relation, in order to build a covariance matrix having given eigenvalues: if any eigenvalue of $\boldsymbol{\Sigma}_j^{(m)}$ is smaller (greater) than a (b) than it can be replaced by a (b) while the eigenvectors remain unchanged. The procedure can be summarized as follows at the m-th iteration:

1. Compute the diagonal matrix $\boldsymbol{\Lambda}^{(m)} = \operatorname{diag}(\lambda_1^{(m)}, \dots, \lambda_p^{(m)})$ of the eigenvalues and the orthogonal matrix $\boldsymbol{\Gamma}^{(m)}$ whose columns are the standardized eigenvectors of $\boldsymbol{\Sigma}_j^{(m)}$;
2. if $\lambda_i^{(m)} < a$ then set $\lambda_i^{(m)} \leftarrow a$, $i = 1, \dots, p$;
3. if $\lambda_i^{(m)} > b$ then set $\lambda_i^{(m)} \leftarrow b$, $i = 1, \dots, p$;
4. set $\boldsymbol{\Lambda}^{(m)} \leftarrow \operatorname{diag}(\lambda_1^{(m)}, \dots, \lambda_p^{(m)})$;
5. set $\boldsymbol{\Sigma}_j^{(m)} \leftarrow \boldsymbol{\Gamma}^{(m)} \boldsymbol{\Lambda}^{(m)} \boldsymbol{\Gamma}'^{(m)}$.

5. Numerical simulations

The performances of the proposed constraints have been evaluated on the grounds of two different sets of problems involving data modelling by multivariate normal mixtures. The first one concerns some mixture decompositions and will be illustrated in this section; the second one lies in the area of missing data imputation and it will be treated in the next section.

As far as the first applications are concerned, we have considered three mixtures of k components of p-normal distributions for different parameters $\boldsymbol{\psi}$. For each of them we have first generated a sample $\mathbf{X}$ of N data; the parameters have been estimated using the EM algorithm. The point of local maximum corresponding to the consistent estimator $\boldsymbol{\psi}^*$ has been chosen to be the limit of the EM algorithm using $\boldsymbol{\psi}$ as initial estimate.

For each mixture we have generated a set of 100 points as initial estimates $\boldsymbol{\psi}^{(0)}$ selected as follows. The initial weights $(\alpha_1^{(0)}, \dots, \alpha_k^{(0)})$ have been randomly chosen with uniform distribution in the unit interval $[0, 1]$ satisfying the constraint $\sum_i \alpha_i^{(0)} = 1$. The initial mean vectors $\boldsymbol{\mu}_1^{(0)}, \dots, \boldsymbol{\mu}_k^{(0)}$ have been selected as follows. Let $\bar{\mathbf{x}}$ and $\mathbf{S}$ be respectively the sample mean and the sample covariance matrix of the data $\mathbf{X}$; afterwards we randomly generated the $\boldsymbol{\mu}_1^{(0)}, \dots, \boldsymbol{\mu}_k^{(0)}$ independently with multivariate normal distribution as:

$$\boldsymbol{\mu}_1^{(0)}, \dots, \boldsymbol{\mu}_k^{(0)} \overset{\text{i.i.d}}{\sim} N(\bar{\mathbf{x}}, \mathbf{S}) ,$$

see also [17]. The initial guesses $\boldsymbol{\Sigma}_1^{(0)}, \dots, \boldsymbol{\Sigma}_k^{(0)}$ have been chosen as the covariance matrices of k different random subsamples of $\mathbf{X}$.

Finally we run a hundred times the unconstrained EM algorithm and the two constrained EM algorithms implementing respectively the Procedure 1 and the Procedure 2 starting from the initial estimates previously selected. Many different pairs of constraints (a, b) have been taken into account.

The computation stopped when the difference between two consecutive log-likelihood values, say $\mathcal{L}(\boldsymbol{\psi}^{(m)}) - \mathcal{L}(\boldsymbol{\psi}^{(m-1)})$, resulted less than 0.01. Computer programs were written in the R language; the different experiments and the obtained results are described below.

Mixture 1: $k = 3$, $p = 2$, $N = 200$.
The sample was generated according to the following parameters of the mixture:

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

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

The covariance matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ have respectively the following eigenvalues:

$$\boldsymbol{\lambda}_1 = (1, 2)' \qquad \boldsymbol{\lambda}_2 = (0.382, 2.618)' \qquad \boldsymbol{\lambda}_3 = (1, 3)'.$$

Note that the smallest eigenvalue is equal to 0.382 while the largest eigenvalue resulted equal to 3.

First we considered the unconstrained algorithm: we observed failures (that is the algorithm did not generate a bounded sequence of loglikelihood values) in 3% of cases and convergence towards to some spurious maximum in 11% of cases; then the right maximum was attained in the remaining 86% of cases. On the contrary both the constrained algorithms based on Procedures 1 and 2 gave no failures and moreover converged to the right maximum in a larger percentage of cases. Table 1 summarizes the percentage of convergence to the right maximum for the two constrained procedures for some different values of the constraints (a, b) in (11).

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

	a			
b	0.20	0.25	0.30	0.35
4.0	96%	98%	97%	96%
4.5	93%	93%	91%	94%
5.0	94%	93%	90%	93%

Procedure 1

	a			
b	0.20	0.25	0.30	0.35
4.0	98%	97%	98%	99%
4.5	99%	98%	99%	99%
5.0	98%	98%	100%	100%

Procedure 2

The two recipes are practically equivalent even if Procedure 2 worked slightly better than the other recipe.

Mixture 2: $k = 3$, $p = 2$, $N = 200$.
The second mixture has the same parameters as the previous one except a variance in $\boldsymbol{\Sigma}_3$ that now is equal to 0.6 rather than 2. This sample was generated according to the following parameters of the mixture:

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

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & 1 \\ 1 & 0.6 \end{pmatrix}$$

The covariance matrices $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ have respectively the following eigenvalues:

$$\boldsymbol{\lambda}_1 = (1, 2)' \qquad \boldsymbol{\lambda}_2 = (0.382, 2.618)' \qquad \boldsymbol{\lambda}_3 = (0.079, 2.521)' .$$

Note that in this case the smallest eigenvalue is equal to 0.079 while the largest eigenvalue resulted equal to 2.618.

As concerns the unconstrained algorithm, we observed failures in 16% of cases while the algorithm converged to some spurious maximum in 15% of cases; then the right maximum was attained in 69% of cases.

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

	a			
b	0.01	0.02	0.03	0.05
3.5	78%	81%	82%	95%
4.0	85%	93%	93%	95%
4.5	89%	95%	89%	87%

Procedure 1

	a			
b	0.01	0.02	0.03	0.05
3.5	88%	89%	97%	99%
4.0	88%	86%	97%	99%
4.5	87%	87%	99%	98%

Procedure 2

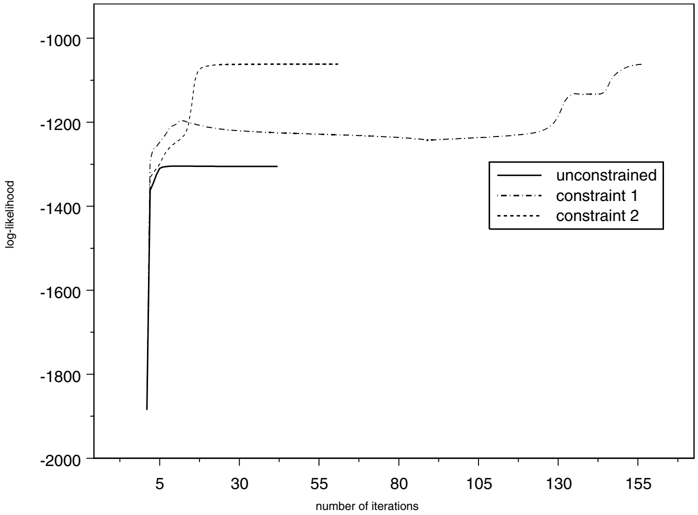

Fig. 1. Comparison among the trajectories of the log-likelihood functions when the unconstrained algorithm converges to a spurious maximum

On the contrary again both the constrained algorithms based on Procedures 1 and 2 gave no failures and moreover converged to the right maximum in a larger percentage of cases. The obtained results have been summarized in Table 2 which gives the percentage of convergence to the right maximum for the two constrained procedures for some different values of (a, b) in (11). Also in this case Procedure 2 worked in general slightly better than the first recipe.

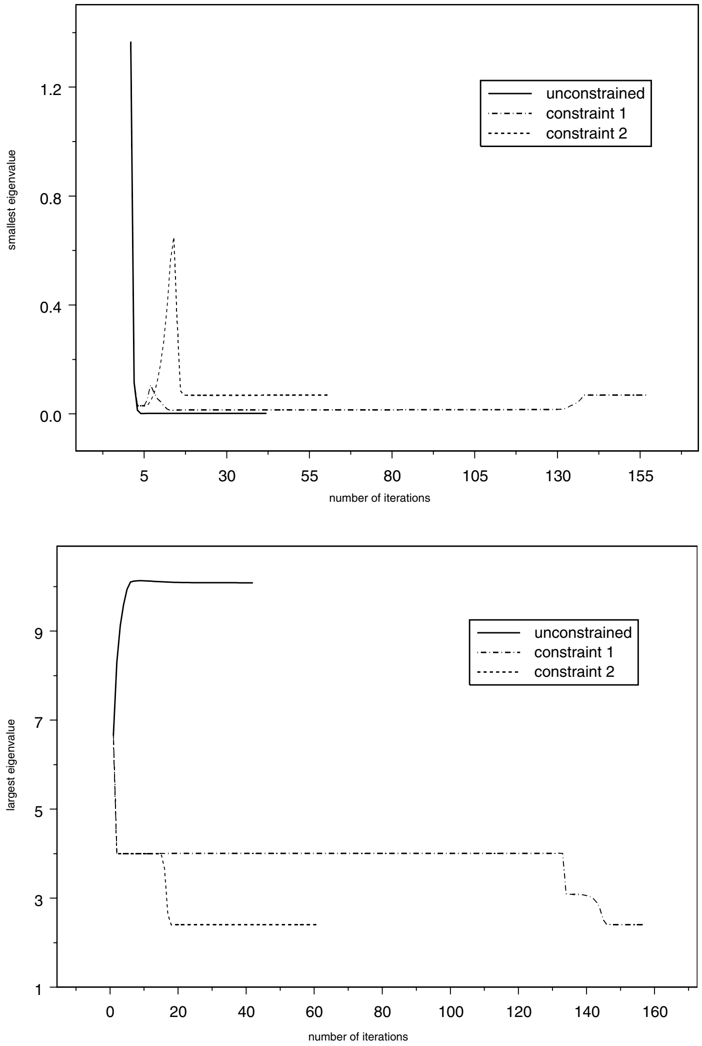

Fig. 2. Comparisons among the trajectories of the smallest eigenvalue (*top*) and of the largest eigenvalue (*bottom*) when the unconstrained algorithm converges to a spurious maximum

Figures 1, 2 and 3 show examples of how the two constrained versions work when the unconstrained converges respectively to a spurious maximum or tends to a singular point for some cases concerning the Mixture 2; Figs. 1 and 2 refer to the same initial guess. The following constraints have been selected: $a = 0.03$ and $b = 4$ (note that here the algorithm based on Procedure 2 required a smaller number of iterations, but this is not true in general).

Mixture 3: $k = 2$, $p = 3$, $N = 300$.
The third mixture is based on 2 three-dimensional multivariate normal distributions.

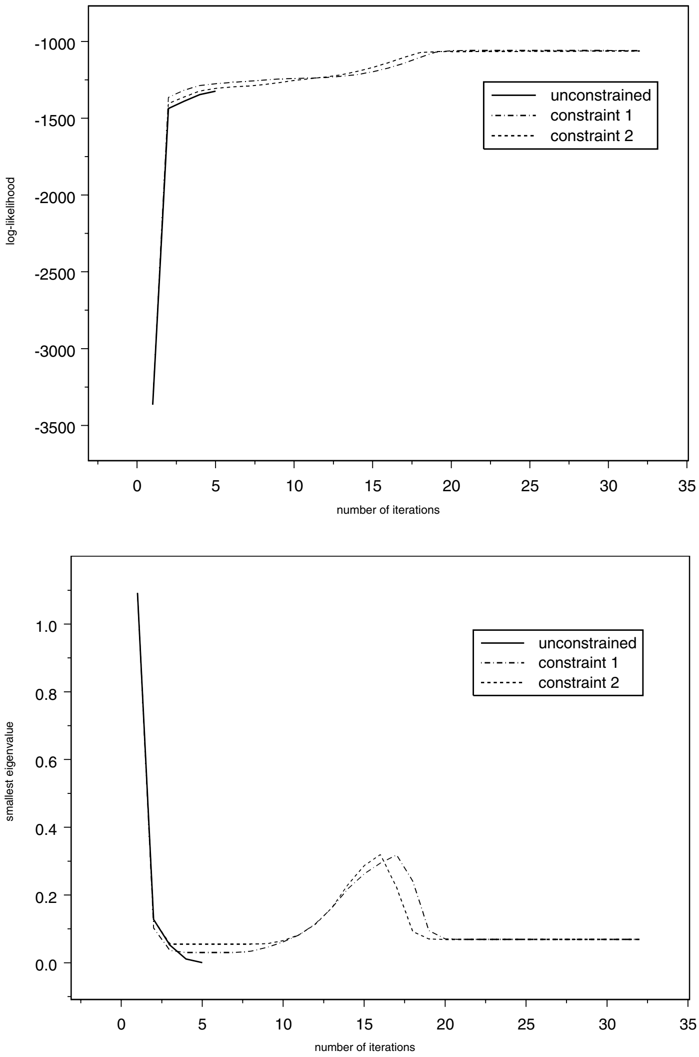

Fig. 3. Example of how the constrained versions of the EM algorithm work when the unconstrained algorithm tends to a singularity: trajectories of the log-likelihood (top) and of the smallest eigenvalue (bottom).

The sample was generated according to the following parameters of the mixture:

$$\boldsymbol{\alpha} = (0.3, 0.7)' \qquad \boldsymbol{\mu}_1 = (0, 0, 0)' \qquad \boldsymbol{\mu}_2 = (5, -2, 3)'$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 4 & -2 & 1 \\ -2 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

The covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ have respectively the following eigenvalues:

$$\boldsymbol{\lambda}_1 = (1, 1, 1)' \qquad \boldsymbol{\lambda}_2 = (0.551, 2.000, 5.449)' .$$

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

	a		
b	0.2	0.3	0.4
6.0	76%	75%	74%
6.5	83%	82%	80%
7.0	80%	81%	72%
8.0	52%	55%	55%

Procedure 1

	a		
b	0.2	0.3	0.4
6.0	73%	73%	69%
6.5	76%	74%	71%
7.0	74%	74%	71%
8.0	72%	70%	70%

Procedure 2

As far as the unconstrained algorithm is concerned, we observed failures in 36% of cases while the algorithm converged to some spurious maximum in 32% of cases; then the right maximum was attained in 32% of cases. Also in this case both the constrained algorithms based on Procedures 1 and 2 gave no failures; Table 3 gives the percentage of convergence to the right maximum for the two constrained procedures for some different values of a, b in (11). Unlike the two previous mixtures, in this case either procedure worked slightly better than the other one depending on the choice of the constraints (a, b).

6. Mixture models vs. hot deck imputation: a case study

An important field of application of multivariate normal mixture models concerns missing data imputation, see e.g. [14]. Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be a sample of size N from a Gaussian mixture with density $f(\mathbf{x}; \psi)$ given in (2). Here we assumed components with the same covariance matrix $\boldsymbol{\Sigma}$, that is $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$ $(j = 1, \ldots, k)$. The maximum likelihood estimation of the parameters has been obtained via the EM algorithm, see [15].

Since the scope is now missing data imputation rather than mixture decomposition, our analysis was carried out along different lines and we utilised the well known Fisher's iris data set for comparison with the results given in [4]. It is well known that this data set contains 150 cases of 4-dimensional observations concerning three species of iris flowers. Moreover since we placed in the context of statistical learning by examples, we considered 90 units (30 for each species) as the learning set (for parameter estimation) and the remaining 60 units (20 for each species) as the test set.

Starting from the complete data set, first of all we have generated six groups of a hundred incomplete data sets eliminating at random some values (the whole data set contains 600 values): 100 data sets with 25 missing values, 100 data sets with 50 missing values, ..., 100 data sets with 150 missing values. In this case the missing at random (MAR) hypotheses are satisfied. We point out that the missing data in the test set have been estimated using the model obtained from the learning set.

We compared the estimates obtained using imputation based on the mixture models (MM) with the ones obtained by means of the *hot-deck* (HD) method (using the Euclidean distance); for this aim we considered the following *standardized mean distances*:

$$\delta_L := \frac{1}{n_L} \sum_{(i,j)\in X_L} \frac{|\hat{x}_{ij} - x_{ij}|}{s_j} \quad \text{and} \quad \delta_T := \frac{1}{n_T} \sum_{(i,j)\in X_T} \frac{|\hat{x}_{ij} - x_{ij}|}{s_j} \tag{18}$$

where X_L (X_T) denotes the learning (test) set containing n_L (n_T) points, $\hat{x}_{ij}$ denotes the estimated ij-th value and x_{ij} the corresponding value in the complete data set, and s_j is the standard deviation of the j-th variable, see also [2].

The initial estimate of the parameters has been obtained via the sample estimate based on the subset of complete data; moreover, in view of our scope only the simplest constraining recipe has been implemented, i.e. Procedure 1. The algorithm was stopped when the difference between two consecutive values of the log-likelihood resulted less than 0.5. Finally, as we assumed the same covariance matrix for the three multivariate normal components, only a lower bound $a = 0.01$ on the smallest eigenvalue must be imposed.

The results obtained using the mixture models have been summarized in Table 4, where for each group of 100 datasets we give: the common total number of missing values, the minimum and the maximum number of missing values for both the learning sets and the test sets, the mean distances $\overline{\delta_L}$ and $\overline{\delta_T}$ computed according to the distances (18).

Table 4. Results for different numbers of missing values in Iris data estimated using MM: distances from the original data

	Learning Set			*Test Set*		
# missing	# missing			# missing		
tot	min	max	$\overline{\delta_L}$	min	max	$\overline{\delta_T}$
25	9	21	0.1214	4	16	0.1016
50	23	37	0.1281	13	27	0.1083
75	37	53	0.1260	22	38	0.1078
100	52	71	0.1238	29	48	0.1119
125	65	86	0.1255	39	60	0.1186
150	79	103	0.1270	47	71	0.1272

Table 5 summarizes analogous quantities when using the HD imputation method on the same datasets.

Table 6 reports the percentages of times (for each group of 100 datasets) in which the MM approach led to a smaller value of the standardized mean distances δ_L and δ_T, respectively for the learning set and the test set, than the HD approach

Table 5. Results for different numbers of missing values estimated in Iris data using HD: distances from the original data

	Learning Set			*Test Set*		
# missing	# missing			# missing		
tot	min	max	$\overline{\delta_L}$	min	max	$\overline{\delta_T}$
25	9	21	0.1444	4	16	0.1526
50	23	37	0.1488	13	27	0.1504
75	37	53	0.1542	22	38	0.1552
100	52	71	0.1570	29	48	0.1588
125	65	86	0.1622	39	60	0.1663
150	79	103	0.1660	47	71	0.1755

Table 6. MM versus HD: percentage of times in which the MM method led to smaller distances δ_L and δ_T than the HD method

# missing tot	δ_L	δ_T
25	74%	87%
50	81%	92%
75	98%	97%
100	99%	97%
125	99%	100%
150	99%	98%

did. The results obtained show that almost always the mixture models have led to better results than the hot-deck method.

7. Discussion and concluding remarks

An important point concerns the choice of the bounds a, b in (11). The results presented in Sect. 5 show that Procedure 1 is more sensible than Procedure 2 with regard to this choice; however the difficulty is when no prior information about a, b is avalaible. For this aim, we tried to implement a criterion based on the eigenvalues of the initial estimates of the covariance matrices $\boldsymbol{\Sigma}_1^{(0)}, \ldots, \boldsymbol{\Sigma}_k^{(0)}$: let $\lambda_{\min}^{(0)}$ and $\lambda_{\max}^{(0)}$ respectively the smallest and the largest eigenvalue of $\boldsymbol{\Sigma}_1^{(0)}, \ldots, \boldsymbol{\Sigma}_k^{(0)}$, and afterwards we set:

$$a := \lambda_{\min}^{(0)}/a' \qquad\qquad b := \lambda_{\max}^{(0)} \cdot b' \,.$$

for suitable positive numbers a', b'; the results were unsatisfactory since we didn't find a unique suitable pair of values (a, b). Perhaps a fruitful strategy is to look for some adaptive procedure for varying the constraints dynamically as the algorithm carries out.

As concerns the application in missing data imputation, the scope was to evaluate mixture models in comparison with the hot-deck approach: in our opinion, on the grounds of the obtained results, the use of mixture models is questionable for small percentages of missing data. We remark that we compared also the hot-deck approach with the usual mean imputation method: the hot-deck method gave always better estimates than the other one both for the learning and the test sets.

A final important question concerns the monotonicity of the constrained EM algorithm. Our simulations seems to suggest that a suitable choice of the bounds (a, b) could preserve the monotonicity when Procedure 2 is taken into account; on the contrary Procedure 1 often destroyed the monotonicity of the EM algorithm but this did not constitute a problem: the trajectories of the likelihood shown in Figs. 1 and 3 are representative of the obtained results in our simulations.

The proposed constraints on the parameter space of the likelihood function of a multinormal mixture distribution, and the practical recipes we implemented, have proved to work quite well in our simulations: problems with singularities do not exist while the number of spurious maxima have been at least reduced. No overwhelming superiority of either procedure can be assessed in general; the two recipes reflect two different constraints on the geometric properties of the covariance matrices and more suggestions could follow from further theoretical studies.

Acknowledgements The author thanks the two anonymous referees for their comments and suggestions which contributed to improve the earlier version of the paper very much; thanks are also due to Filippo Domma for helpful discussions.

References

1. Böhning D, Seidel W (2003) Recent develompments in mixture models. Computational Statistics & Data Analysis 41: 349–357
2. Domma F, Ingrassia S (2001) Mixture models for maximum likelihood estimation from incomplete values. In: Borra S, Rocci M, Vichi M, Schader M (eds) Advances in classification and data analysis. Springer, Berlin Heidelberg New York, pp 201–208
3. Domma F, Ingrassia S (2002) A constrained MLE formulation for multinormal mixture decomposition. In: Atti della XLI Riunione Scientifica della Società Italiana di Statistica, Milano, 5–7 Giugno 2002, vol 2, pp 371–374
4. Ghahramani Z, Jordan M (1997) Learning from incomplete data. In: Greiner R, Petsche T, Hanson SJ (eds) Computational learning theory and natural learning systems. MIT Press, Cambridge, pp 67–85
5. Hathaway RJ (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions. The Annals of Statistics 13: 795–800
6. Hathaway RJ (1986) A constrained EM algorithm for univariate normal mixture. Journal of Statistical Computing and Simulation 23: 211–230
7. Horn RA, Johnson CR (1999) Matrix analysis. Cambridge University Press, New York
8. Ingrassia S (1992) A comparison between the simulated annealing and the EM algorithms in normal mixture decompositions. Statistics and Computing 2: 203–211
9. Isaacsson E, Keller HB (1966) Analysis of numerical methods. Wiley, New York
10. Jordan MI, Jacobs RA (1994) Hierarchical mixtures of experts and the EM algorithm. Neural Computation 3: 79–97
11. Kiefer NM (1978) Discrete parameter variation: efficient estimation of a switching regression model. Econometrica 46: 427–434
12. Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics 27: 888–906
13. Lindsay BG (1995) Mixture models: Theory, geometry and applications. NSF-CBMS Regional Conference Series in Probability and Statistics. IMS-ASA
14. Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York

15. Little RJA, Schluchter MD (1985) Maximum likelihood estimation for mixed continuous and categorial data with missing values. Biometrika 72: 497–512
16. Lütkepohl H (1996) Handbook of matrices. Wiley, Chichester
17. McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
18. Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM Reviews 46: 195–239
19. Sharkey AJC (1999) Combining artificial neural nets. Springer, Berlin Heidelberg New York
20. Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, Chichester