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

Eigenvalues and constraints in mixture modeling: geometric and computational issues

Luis-Angel García-Escudero · Alfonso Gordaliza · Francesca Greselin · Salvatore Ingrassia · Agustín Mayo-Iscar

Received: date / Accepted: date

Abstract This paper presents a review about the usage of eigenvalues restrictions for constrained parameter estimation in mixtures of elliptical distributions according to the likelihood approach. These restrictions serve a twofold purpose: to avoid convergence to degenerate solutions and to reduce the onset of non interesting (spurious) maximizers, related to complex likelihood surfaces. The paper shows how the constraints may play a key role in the theory of Euclidean data clustering. The aim here is to provide a reasoned review of the constraints and their applications, along the contributions of many authors, spanning the literature of the last thirty years.

Keywords Mixture Model · EM algorithm · Eigenvalues · Model-based clustering

1 Introduction

Finite mixture distributions play a central role in statistical modelling, as they combine much of the flexibility of non parametric models with nice analytical properties of parametric models, see e.g. Titterington *et al.* (1985), Lindsay (1995), McLachlan and Peel (2000). In the last decades such models have

Luis-Angel García-Escudero, Alfonso Gordaliza, Agustín Mayo-Iscar
Department of Statistics and Operations Research andIMUVA
University of Valladolid, Valladolid, Spain
E-mail: [lagarcia, alfonsog, agustinm]@eio.uva.es

Francesca Greselin
Department of Statistics and Quantitative Methods
Milano-Bicocca University, Milan, Italy
E-mail: francesca.greselin@unimib.it

Salvatore Ingrassia
Department of Economics and Business, University of Catania
Corso Italia 55, 95128 Catania, Italy
E-mail: s.ingrassia@unict.it second address

attracted the interest of many researchers and found a number of new and interesting fields of application. For parameter estimation in mixture models several approaches may be considered, as the ones exposed in McLachlan and Krishnan (2008a). The maximum likelihood (ML) framework is among the most commonly used approaches to mixture parameter estimation, and it is the approach we consider here. In the case of Gaussian mixtures, it is well-known that the likelihood function increases without bound if one of the mixture means coincides with a sample observation and if the corresponding variance tends to zero or, in the multivariate situation, if the variance matrix tends to a singular matrix.

This paper presents a review about the usage of eigenvalues restrictions for constrained estimation, that serves a twofold purpose: to avoid convergence to degenerate solutions and to reduce the onset of non interesting (spurious) maximizers, related to complex likelihood surfaces. From the seminal paper of Hathaway (1985), we will see how the constraints may play a key role in the theory of Euclidean data clustering. The aim here is to provide a reasoned review of the constraints and their applications, along the contributions of many authors, spanning the literature of the last thirty years. Applications of the constraints in robustness, jointly with trimming techniques, requires an extensive discussion per se, hence it will be the argument of a further paper, see García-Escudero *et al.* (2017). The plan of the paper is the following. Maximum likelihood estimation for mixture models is briefly recalled in Section 2, along with conditions assuring the existence and consistency of the estimator, in Section 3 different constrained formulations of maximum-likelihood estimation are presented, in Section 4 degeneracy of the maximum likelihood estimation is investigated. The last part of the paper is devoted to the role of eigenvalues in parsimonious models: Gaussian parsimonious clustering models are summarized in Section 5 while in Section 6 mixture of factor analyzers are presented. Finally, conclusions are given in Section 7.

2 Maximum likelihood estimation of parameters in mixture models

Let $\mathbf{X}$ be a random vector defined on a heterogeneous population Ω with values in a d-dimensional Euclidean space. Here, we consider mixtures of elliptical distributions, having density function

$$p(\mathbf{x};\boldsymbol{\psi}) = \pi_1 f(\mathbf{x};\boldsymbol{\theta}_1) + \cdots + \pi_G f(\mathbf{x};\boldsymbol{\theta}_G) \qquad \mathbf{x} \in \mathbb{R}^d \tag{1}$$

where $\pi_1, \ldots, \pi_G$ are the *mixing weights* and

$$f(\mathbf{x};\boldsymbol{\theta}_g) = \eta_g |\boldsymbol{\Sigma}_g|^{-1/2} h\{(\mathbf{x}-\boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}-\boldsymbol{\mu}_g)\} \tag{2}$$

denotes the density of the elliptical distribution (for more details see, e.g., Fang and Anderson, 1990) where $\boldsymbol{\mu}_g \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_g$ are positive definite matrices in $\mathbb{R}^d \times \mathbb{R}^d$, h is a strictly positive, continuous function on $\mathbb{R}$, symmetrical about 0 and monotonically decreasing on $[0,\infty)$, η_g is a positive constant depending on

the dimension d of the Euclidean space. Moreover, $\boldsymbol{\theta}_g = \{\nu_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}$ belongs to the same parameter space $\boldsymbol{\Theta}$, i.e. $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G \in \boldsymbol{\Theta}$. Finally, we denote by K the number of parameters of $\boldsymbol{\psi}$.

The most popular case of (1) is given by *Gaussian mixtures*, i.e., mixtures with multivariate Gaussian components:

$$p(\mathbf{x}; \boldsymbol{\psi}) = \pi_1 \phi(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \cdots + \pi_G \phi(\mathbf{x}; \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_G)\,, \tag{3}$$

where $\phi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ is the density function of the multivariate normal distribution with parameters $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. Finally, we set $\boldsymbol{\psi} = \{(\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g),\ g = 1, \ldots, G\} \in \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is the parameter space:

$$\begin{aligned} \boldsymbol{\Psi} = \{(\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G) \in \mathbb{R}^{G[1+d+(d^2+d)/2]} \ : \\ \pi_1 + \cdots + \pi_G = 1, \pi_g > 0,\ |\boldsymbol{\Sigma}_g| > 0 \quad \text{for } g = 1, \ldots, G\}\,. \end{aligned} \tag{4}$$

Thus, $K = (G-1) + G[1 + d + (d^2 + d)/2]$.

A more general family is given by mixtures of multivariate t distributions, with densities in (2) taking the form

$$f(\mathbf{x}; \boldsymbol{\theta}_g) = \frac{\Gamma(\frac{\nu_g + d}{2})|\boldsymbol{\Sigma}_g|^{-1/2}}{(\pi\nu_g)^{d/2}\Gamma(\frac{\nu_g}{2})\{1 + d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)/\nu_g\}^{(\nu_g+d)/2}}\,, \tag{5}$$

with location parameter $\boldsymbol{\mu}_g$, a positive definite inner product matrix $\boldsymbol{\Sigma}_g$, degrees of freedom ν_g, and where

$$d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = (\mathbf{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x} - \boldsymbol{\mu}_g) \tag{6}$$

denotes the Mahalanobis distance between $\mathbf{x}$ and $\boldsymbol{\mu}_g$, with respect to the matrix $\boldsymbol{\Sigma}_g$. t distributions are progressively becoming popular in multivariate statistics, providing more realistic tails for real-world data with respect to the alternative Gaussian models. They are a robust alternative able to cope with moderate outliers, as the Mahalanobis distance in the denominator of (5) downweight the contribution of data mildly deviating from the assumed model. For the sake of simplicity, throughout this paper we will generally consider Gaussian mixtures defined in (3). The extension to mixtures of t distributions, and to more general mixtures with elliptical components is usually straightforward.

For a given sample $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of size N drawn from (1), let us consider the log-likelihood function of $\boldsymbol{\psi}$,

$$\mathcal{L}(\boldsymbol{\psi}) = \sum_{n=1}^{N} \log \left(\sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right) \tag{7}$$

and denote by

$$\hat{\boldsymbol{\psi}} = \arg_{\psi \in \Psi} \max \mathcal{L}(\boldsymbol{\psi}) \tag{8}$$

the maximum likelihood estimator (MLE) of $\boldsymbol{\psi}$.

We can introduce the population counterpart of (7). Let $P = P(\mathbf{X})$ be the probability measure in $\mathbb{R}^d$ induced by the random variable $\mathbf{X}$ describing the

heterogeneous population and let $\mathbb{E}_P()$ denote the expectation with respect to P. The aim is to maximize

$$\mathcal{L}^*(\boldsymbol{\psi}) = \mathbb{E}_P \left[\log \left(\sum_{g=1}^{G} \pi_g \phi(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g \right) \right] \tag{9}$$

in terms of parameters $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, given in (4). If P_N stands for the empirical measure, $P_N = (1/N) \sum_{n=1}^{N} \delta_{\{\mathbf{x}_n\}}$, where $\delta_{\{\mathbf{x}_n\}}$ denotes the Dirac function with probability 1 in $\mathbf{x}_n$ and 0 elsewhere, we recover the original problem (7) by replacing P by P_N.

The maximum likelihood estimate is usually attained through the EM algorithm, that is an iterative procedure for maximizing a likelihood function, in the context of partial information, see e.g. Dempster *et al.* (1977); McLachlan and Krishnan (2008b) for details. The algorithm generates a sequence of estimates $\{\boldsymbol{\psi}^{(r)}\}_r$ – starting from some initial guess $\boldsymbol{\psi}^{(0)}$ – so that the corresponding sequence $\{\mathcal{L}(\boldsymbol{\psi}^{(r)})\}_r$ is not decreasing. In particular, for multivariate Gaussian mixtures, on the $(r+1)$th iteration the EM algorithm computes the quantities

$$\tau_{ng}^{+} = \frac{\pi_g^{-} \phi(\mathbf{x}_n; \boldsymbol{\mu}_g^{-}, \boldsymbol{\Sigma}_g^{-})}{\sum_{j=1}^{G} \pi_j^{-} \phi(\mathbf{x}_n; \boldsymbol{\mu}_j^{-}, \boldsymbol{\Sigma}_j^{-})}$$

where the superscript $^{-}$ denotes the estimate on the previous rth iteration. In the M-step the parameters $\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g$ are updated according to the following rule

$$\boldsymbol{\mu}_g^{+} = \frac{\sum_{n=1}^{N} \tau_{ng}^{+} \mathbf{x}_n}{\sum_{n=1}^{N} \tau_{ng}^{+}}$$

$$\boldsymbol{\Sigma}_g^{+} = \frac{\sum_{n=1}^{N} \tau_{ng}^{+} (\mathbf{x}_n - \boldsymbol{\mu}_g^{+})(\mathbf{x}_n - \boldsymbol{\mu}_g^{+})'}{\sum_{n=1}^{N} \tau_{ng}^{+}}$$

Under some regularity assumptions for the likelihood function, in Boyles (1983) it is proved that the sequence $\{\boldsymbol{\psi}^{(r)}\}_r$ converges to a compact set of local maxima of the likelihood function; moreover, this limit set may not be a singleton.

We have already said that, in general, the method of maximum likelihood in the case of mixtures of elliptical distributions (2) leads to an ill-posed optimization problem. The *observed data* log-likelihood of Gaussian mixtures $\mathcal{L}(\boldsymbol{\psi})$ is unbounded (Day, 1969), as it can be easily seen by posing $\mu_1 = x_1$ and $\sigma_1 \to 0$ (or $\boldsymbol{\mu}_1 = \mathbf{x}_1$ and $|\boldsymbol{\Sigma}_1| \to 0$ in the multivariate case). Thus, the definition of estimate of maximum likelihood as the absolute maximum of $\mathcal{L}(\boldsymbol{\psi})$ lacks mathematical sense. In particular, the unboundedness of $\mathcal{L}(\boldsymbol{\psi})$ causes the failure of optimization algorithms, like the EM, and the occurrence of *degenerate* components.

Formally, let $\overline{\boldsymbol{\Psi}}$ be the closure of the parameter space $\boldsymbol{\Psi}$ and consider the set

$$\left\{ \boldsymbol{\psi} \in \overline{\boldsymbol{\Psi}} \, : \, \exists \, j_0 \in \{1, \ldots, G\} \text{ and } n \in \mathbb{N} \text{ such that } \boldsymbol{\mu}_{j_0} = \mathbf{x}_n, |\boldsymbol{\Sigma}_{j_0}| = 0 \right\}. \tag{10}$$

The component corresponding to j_0 in (10) will be referred to as a *degenerate* component hereinafter, and the corresponding solution provided by the EM algorithm is a *degenerate solution*, see Biernacki and Chrétien (2003) and Ingrassia and Rocci (2011).

Secondly, the likelihood function may present local spurious maxima which occur as a consequence of a fitted component having a very small variance or generalized variance (i.e., the determinant of the covariance matrix) compared to the others, see e.g. Day (1969). Such a component usually corresponds to a cluster containing few data points either relatively close together or almost lying in a lower-dimensional subspace, in the case of multivariate data.

In this paper, the focus will be to compare methods to maximize the likelihood function in some constrained subset of the parameter space, through the EM algorithm.

Alternative strategies can be adopted to deal with the two aforementioned issues, and we will briefly give some references, without entering into details. One option is to employ stochastic algorithms for global optimization, like the simulated annealing (see, e.g., van Laarhoven and Aarts, 1988). With this approach, on the one hand, the procedure converges with uniform distribution, at a slow rate, to the maximizing points, and on the other, the optimal tuning of the annealing parameter requires additional effort in time. Overall, no overwhelming superiority has been demonstrated with respect to the EM algorithm (Ingrassia, 1992).

A second very practical strategy is to monitor the relative size of the fitted mixing proportions (McLachlan and Peel, 2000). In the same context of avoiding degeneracy, but with a Bayesian point of view, Ciuperca *et al.* (2003) introduced a bounded penalized likelihood, that does not degenerate in any point of the closure of the parameter space and, therefore, assures the existence of the penalized maximum likelihood estimator. They also provide statistical asymptotic properties of the penalized MLE, namely strong consistency, asymptotic efficiency, and rate of convergence.

Fraley and Raftery (2007) proposed a Bayesian regularisation, and suggested replacing the MLE by the maximum a posteriori estimate. They employed the normal inverse gamma conjugate priors for the conditional mean and the variance, a uniform prior distribution for the vector of component proportions, and derived the prior hyperparameters, assumed to be the same for all components. Their approach has also been implemented in MCLUST as described in Fraley *et al.* (2012).

Resuming now our main stream, the following examples show how singularities and spurious maximizers can undermine the estimation:

– *Simdata 1:* A sample $\mathcal{S}$ of size $N = 200$ has been generated from a three components mixture of bivariate normal ($G = 3$ and $d = 2$) having the

following parameters:

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

Figure 1 shows the sample data (panel a) and two obtained classifications, the first derived from the MLE $\hat{\boldsymbol{\psi}}$ (panel b), and the second from a local maximum $\boldsymbol{\psi}^*$ (panel c). Further, panel d) plots the log-likelihood along thesegment joining $\hat{\boldsymbol{\psi}}$ with $\boldsymbol{\psi}^*$, to reveal a local mode w.r.t. the absolute maximum in the log-likelihood surface.

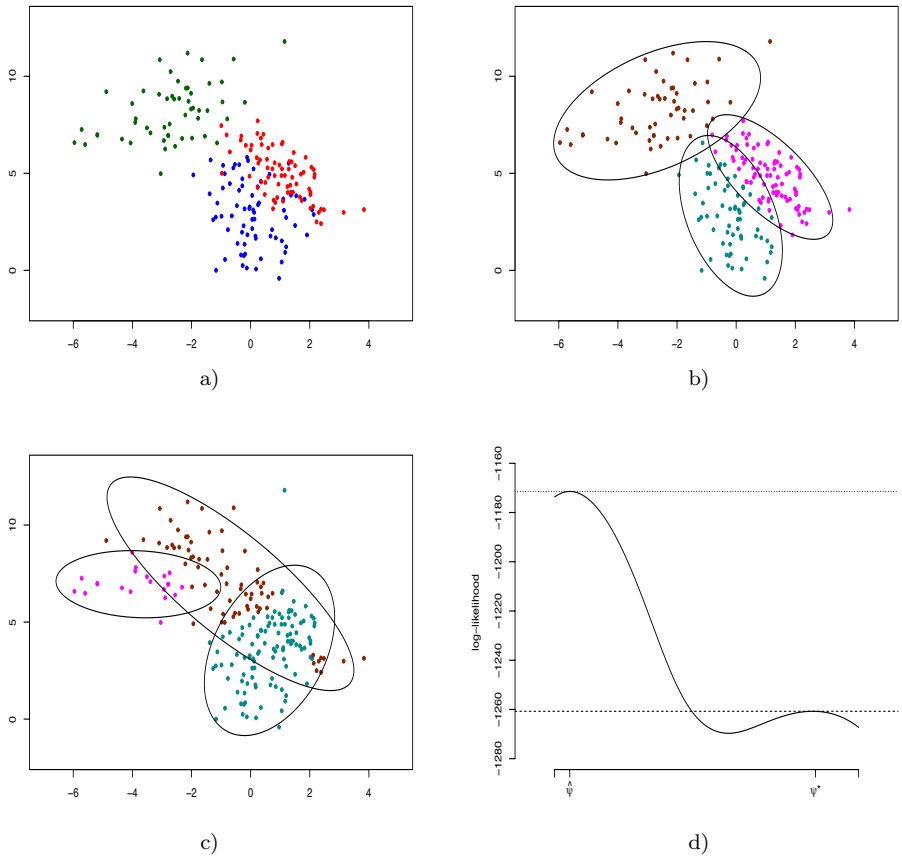


Fig. 1 *Simdata 1:* a) original data; b) classification based on the MLE $\hat{\boldsymbol{\psi}}$; c) classification based on a local maximum $\boldsymbol{\psi}^*$; d) plot of the log-likelihood function along the direction $\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*$, where we have $\mathcal{L}(\hat{\boldsymbol{\psi}}) = -1172.37$ and $\mathcal{L}(\boldsymbol{\psi}^*) = -1261.90$.

– *Simdata 2:* A second example is provided by considering a smaller subset (of size 100) drawn from the same set of parameters we considered before. Figure 2 shows the original data in panel a), a first classification based on the MLE $\hat{\boldsymbol{\psi}}$ in panel b), a second classification based on a local maximum $\boldsymbol{\psi}^*$ in panel c), and finally a third classification based on a spurious maximizer (in panel d), where one component is overfitted to a set composed by 4 data points).

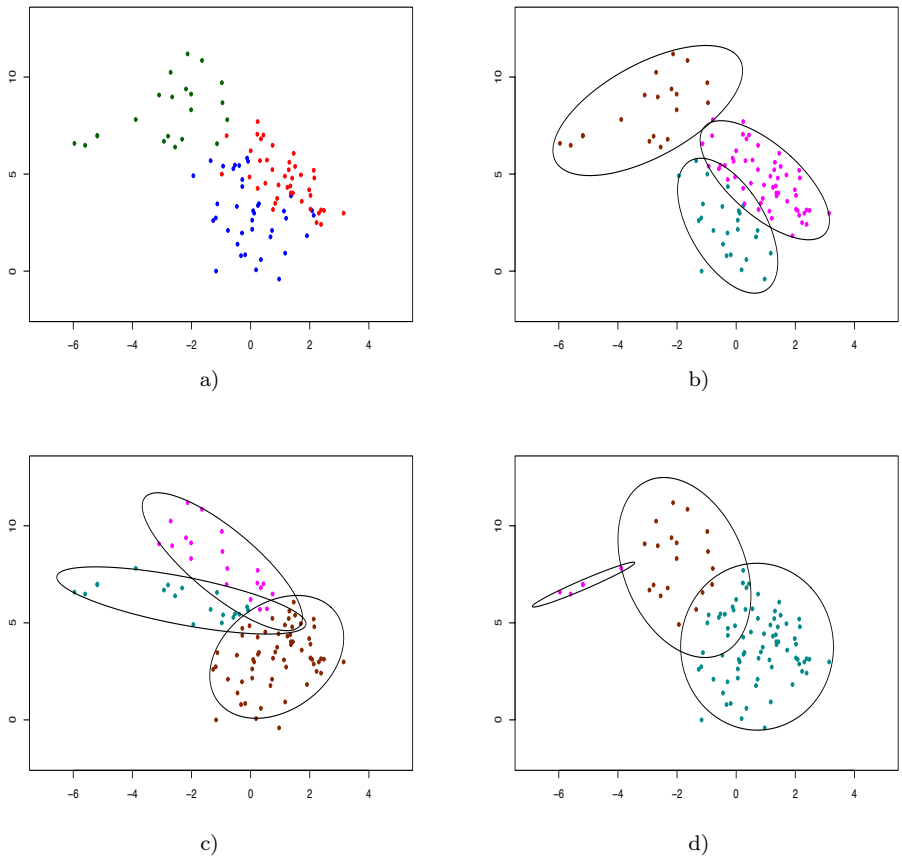


Fig. 2 *Simdata 2:* a) original data; b) classification based on the MLE $\hat{\boldsymbol{\psi}}$; c) classification based on a local maximum $\boldsymbol{\psi}^*$; d) classification based on a spurious solution.

Issues related to spurious maximizers have been investigated in detail in Section 3 of McLachlan and Peel (2000) where it is pointed out that "these solutions often have a high likelihood, but are of little practical use or real world interpretation". Many times, they may even have a higher likelihood than the

one for the "true" mixture parameters. McLachlan and Peel (2000) offers a 11 pages analysis, applied to different synthetic and benchmark datasets, to show how the issue arises. In particular, the well-known Iris data set is analyzed to evaluate whether the "virginica" species should be split into two subspecies or not. To provide hints for distinguishing spurious from useful solutions, 15 possible local maxima were considered, together with different quantities summarizing clusters as the size of the smallest cluster, the determinants of the scatter matrices, the value of the smallest eigenvalue, and the intercomponent mean distances.

The likelihood estimate $\boldsymbol{\psi}$ must be one of the roots of the equation

$$\frac{\partial \mathcal{L}(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \mathbf{0}\,. \tag{11}$$

In a quite general framework (i.e. not restricted to mixture distributions), Cramér (1946) showed that a unique consistent root exists for the equation (11) in the univariate case under certain conditions. The proof has been extended to the multivariate case by Chanda (1954), with a corrected version of Theorem 2 there provided by Tarone and Gruenhage (1975, 1979). See also Kiefer (1978) and Redner and Walker (1984).

Theorem 1 (Redner and Walker, 1984) *Let $p(\mathbf{x};\boldsymbol{\psi})$ be a probability density function depending on some parameter $\boldsymbol{\psi}$ and assume we are provided with a sample $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ drawn from $p(\mathbf{x};\boldsymbol{\psi})$. Let $\boldsymbol{\psi}_0$ be the true value of the parameter $\boldsymbol{\psi}$ and exists at some point in the parameter space $\boldsymbol{\Psi}$. Then, given the existence of, and certain boundedness conditions on, derivatives of the mixture density $p(\mathbf{x};\boldsymbol{\psi})$, of orders up to 3, there exists a unique consistent estimator $\boldsymbol{\psi}_N$ corresponding to a solution of the likelihood equation* (11). *Further, $\sqrt{N}(\boldsymbol{\psi}_N - \boldsymbol{\psi}_0)$ is asymptotically normally distributed with mean zero and covariance $I^{-1}(\boldsymbol{\psi}_0)$, where $I(\boldsymbol{\psi}_0)$ is the Fisher information matrix.*

In the following, by maximum likelihood estimate (MLE), we shall mean such a point of $\boldsymbol{\Psi}$ and it will be denoted by $\hat{\boldsymbol{\psi}}$.

3 Constrained formulations of maximum-likelihood estimation

The MLE $\hat{\boldsymbol{\psi}}$ is usually computed by means of suitable optimization procedures which generate a sequence of estimates $\{\boldsymbol{\psi}^{(r)}\}_r$ – starting from some initial guess $\boldsymbol{\psi}^{(0)}$ – so that the corresponding sequence $\{\mathcal{L}(\boldsymbol{\psi}^{(r)})\}_r$ is not decreasing. To this end, the EM algorithm is usually implemented for parameter estimation in mixture modeling, see e.g. McLachlan and Krishnan (2008b). However, the convergence towards $\hat{\boldsymbol{\psi}}$ is not guaranteed because of the numerical issues associated to the maximization of the log-likelihood function $\mathcal{L}(\boldsymbol{\psi})$, described above. In particular, i) singularities may cause the failure of the algorithm; ii) spurious maximizers may appear along the estimation, and provide a mathematical solution lacking statistical meaning. Further, the final estimate depends on the initial guess $\boldsymbol{\psi}^{(0)}$ (Boyles, 1983).

3.1 Hathaway's approach

The idea of a constrained estimation became popular due to Hathaway (1985), in the framework of Gaussian mixtures. In the *univariate case*, the Gaussian mixture has density

$$p(x;\boldsymbol{\psi}) = \pi_1\phi(x;\mu_1,\sigma_1^2) + \cdots + \pi_G\phi(x;\boldsymbol{\mu}_G,\sigma_G^2)\,, \tag{12}$$

where $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ and $\boldsymbol{\Psi}$ is the parameter space

$$\begin{aligned}\boldsymbol{\Psi} = \{&(\pi_1,\ldots,\pi_G,\mu_1,\ldots,\mu_G,\sigma_1,\ldots,\sigma_G) \in \mathbb{R}^{3G} \ : \\ &\pi_1 + \cdots + \pi_G = 1,\ \pi_g > 0,\ \sigma_g > 0 \text{ for } g = 1,\ldots,G\}.\end{aligned} \tag{13}$$

For some $c > 0$, let $\boldsymbol{\Psi}_c$ be the subset of $\boldsymbol{\Psi}$ such that $\sigma_g^2 \geq c\sigma_j^2$ for $g \neq j$, i.e., such that

$$\min_{g\neq j}\frac{\sigma_g^2}{\sigma_j^2} \geq c \ > \ 0\,. \tag{14}$$

Hathaway (1985) pointed out that the first mention of constraints like (14) is found in Dennis (1981) who, in turn, gives credit to Beale and Thompson (oral communication). For this reason, Gallegos and Ritter (2009) called them the Hathaway-Dennis-Beale-Thompson (HDBT) constraints.

The following result states that the constraint (14) yields an optimization problem having a global solution in a constrained parameter space with no singularities and at least with a smaller number of local maxima.

Theorem 2 (Hathaway, 1985) *Let $\mathcal{X} = \{x_1,\ldots,x_N\}$ be a sample drawn with law* (12) *containing at least $G+1$ distinct points. Then for $c \in (0,1]$ there exists a constrained global maximizer of $\mathcal{L}(\boldsymbol{\psi})$ over the set $\boldsymbol{\Psi}_c$ defined by* (14).

Moreover, also strong consistency of the constrained estimator has been proven in Hathaway (1985), by applying existing maximum-likelihood theory due to Kiefer and Wolfowitz (1956). From a practical point of view, Hathaway (1996) provides an algorithm for building a consistent estimator under a slightly different kind of constraints:

$$\frac{\sigma_g^2}{\sigma_{g+1}^2} \geq c \quad \text{for all} \quad g = 1,\ldots,G-1 \quad \text{and} \quad \frac{\sigma_G^2}{\sigma_1^2} \geq c \ > \ 0\,.$$

Strong consistency is shown in Hathaway (1985) for the univariate case by applying existing maximum-likelihood theory due to Kiefer and Wolfowitz (1956), see Theorems 3.1 and 3.2.

Consider now the multivariate Gaussian mixture

$$p(\mathbf{x};\boldsymbol{\psi}) = \pi_1\phi(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \cdots + \pi_G(\mathbf{x};\boldsymbol{\mu}_G,\boldsymbol{\Sigma}_G) \tag{15}$$

where $\boldsymbol{\psi} \in \boldsymbol{\Psi}$ with

$$\begin{aligned}\boldsymbol{\Psi} = \{&(\pi_1,\ldots,\pi_G,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_G,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_G) \in \mathbb{R}^{G[1+d+(d^2+d)/2]} \ : \\ &\pi_1 + \cdots + \pi_G = 1, \pi_g > 0,\ |\boldsymbol{\Sigma}_g| > 0 \quad \text{for } g = 1,\ldots,G\}.\end{aligned} \tag{16}$$

To generalize results in the *multivariate case*, Hathaway (1985) states only the following sentence in the concluding remarks: "For a mixture of G d-variate normals, constraining all characteristic roots of $\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_j^{-1}$ $(1 \leq g \neq j \leq G)$ to be greater than or equal to some minimum value $c > 0$ (satisfied by the true parameter) lead to a constrained (global) maximum-likelihood formulation". Being just a brief statement, this sentence motivated research summarized in Section 3.3.

3.2 Constraints on the determinants of the covariance matrices

The natural multivariate generalization of (14) appears to be a constraint on the ratio of the component generalized variances, i.e. of the determinants of the covariance matrices, which are required not to be too disparate:

$$\min_{g\neq j}\frac{|\boldsymbol{\Sigma}_g|}{|\boldsymbol{\Sigma}_j|} = \min_{g\neq j}|\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_j^{-1}| \geq c\,, \tag{17}$$

for some $c > 0$. The maximization of the loglikelihood function (7) under the constraint (17) is discussed in McLachlan and Peel (2000, Section 3.9.1).

Recalling that the volume is proportional to the square root of the determinant, we see that this type of constraint limits the relative volumes of the aforementioned equidensity ellipsoids, but not the cluster shapes. The use of this constraint is particularly advisable when affine equivariance is required. Constraints (17) have been implemented in the R package *tclust*, see Fritz *et al.* (2012).

Finally, we recall that in the framework of the trimming approach to robust modeling, Gallegos (2002) implicitly assumed the stronger condition $|\boldsymbol{\Sigma}_1| = |\boldsymbol{\Sigma}_2| = \cdots = |\boldsymbol{\Sigma}_G|$.

3.3 Constraints on the eigenvalues of the covariance matrices

As we recalled in Section 3.1, Hathaway (1985) proposed the constraint

$$\min_{g\neq j}\,\lambda(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_j^{-1}) \geq c\; >\; 0\,, \tag{18}$$

stating that it leads to a constrained (global) maximum-likelihood formulation, without any further development. The constant c in (18) is still referred to as the HDBT constant.

It is worth to note that, while this bound can be easily checked, as far as we know, it cannot be directly implemented in optimization procedures like the EM algorithm, where the estimates are iteratively updated. To this end, different approaches have been pursued.

The next proposition allows to reformulate (18) in terms of stronger (and algorithmically feasible) constraints on the eigenvalues of each covariance matrix $\boldsymbol{\Sigma}_g$.

Proposition 3 (Ingrassia, 2004) *Let* $\mathbf{A}, \mathbf{B}$ *two* $d \times d$ *symmetric and positive definite matrices. Then we have:*

$$\lambda_{\max}(\mathbf{A}\mathbf{B}^{-1}) \leq \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{B})} \tag{19}$$

$$\lambda_{\min}(\mathbf{A}\mathbf{B}^{-1}) \geq \frac{\lambda_{\min}(\mathbf{A})}{\lambda_{\max}(\mathbf{B})} \tag{20}$$

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

The proof is based on the properties of the spectral norm $\|\mathbf{A}\|_2$ of a matrix $\mathbf{A}$. Denote by $\lambda_{ig} = \lambda_i(\boldsymbol{\Sigma}_g)$ the ith eigenvalue of the gth covariance matrix. For given $a, b > 0$, such that $a/b \geq c$, where c satisfies the relation (18), assume that the eigenvalues of the covariance matrices $\boldsymbol{\Sigma}_g$ satisfy the constraints:

$$a \leq \lambda_i(\boldsymbol{\Sigma}_g) \leq b \qquad i = 1, \ldots, d \qquad g = 1, \ldots, G. \tag{21}$$

Then for any pair of covariance matrices $\boldsymbol{\Sigma}_g, \boldsymbol{\Sigma}_j$, the inequality (20) yields:

$$\lambda_{\min}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_j^{-1}) \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_g)}{\lambda_{\max}(\boldsymbol{\Sigma}_j)} \geq \frac{a}{b} \geq c > 0\,, \qquad 1 \leq g \neq j \leq G,$$

assuring that

$$\frac{\lambda^*_{\min}}{\lambda^*_{\max}} \geq c \tag{22}$$

where

$$\lambda^*_{\min} = \min_{g=1,\ldots,G} \min_{i=1,\ldots,d} \lambda_i(\boldsymbol{\Sigma}_g) \tag{23}$$

$$\lambda^*_{\max} = \max_{g=1,\ldots,G} \max_{i=1,\ldots,d} \lambda_i(\boldsymbol{\Sigma}_g). \tag{24}$$

Finally, according to (21), we introduce the following constrained parameter space

$$\boldsymbol{\Psi}_{a,b} = \Big\{ (\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G) \in \mathbb{R}^{G[1+d+(d^2+d)/2]} \; : \; \sum_{g=1}^{G} \pi_g = 1, \pi_g > 0,\; a \leq \lambda_{ig} \leq b,\; g = 1, \ldots, G\,, i = 1, \ldots, d \Big\}. \tag{25}$$

To establish the existence of the ML, we remark that we have to discard the case of components with arbitrarily small variances. However, maxima may be forced by various types of scale constraint. A popular argument is based on the compactness of the parameter space. Indeed, its combination with the continuity of the likelihood function guarantees the existence of a maximum. However, when the natural parameter space of a model is not compact, we may prove that the sequence of ML estimates remains in a compact subset, as in the following theorem.

Theorem 4 *Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a sample drawn from a multivariate Gaussian mixture containing at least $G + d$ distinct points. Then, for any positive real numbers a, b, with $a < b$, there exists a constrained global maximizer of $\mathcal{L}(\boldsymbol{\psi})$ over the set $\boldsymbol{\Psi}_{a,b}$ defined by* (25).

Proof. To maximize $\mathcal{L}(\boldsymbol{\psi})$ means to jointly maximize $|\boldsymbol{\Sigma}_g|^{-1/2}$ and minimize the argument of the exponential, i.e. $(\mathbf{x}_n - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_g)$, for each $g = 1, \ldots, G$. Hence, firstly we will show that, for a given $\boldsymbol{\Sigma}_g$, the mean vector $\boldsymbol{\mu}_g$ has to lie in a compact subset in $\mathbb{R}^d$. Let C be the convex hull of $\mathcal{X}$, i.e. the intersection of all convex sets containing the N points, given by

$$C(\mathcal{X}) = \left\{ \sum_{n=1}^{N} u_n \mathbf{x}_n \mid \sum_{n=1}^{N} u_n = 1 \, , u_n \geq 0 \right\}. \tag{26}$$

Suppose now that $\bar{\boldsymbol{\psi}} \in \boldsymbol{\Psi}_{a,b}$ satisfies $\boldsymbol{\mu}_g \notin C(\mathcal{X})$. Then $\mathcal{L}(\bar{\boldsymbol{\psi}}) \leq \mathcal{L}(\boldsymbol{\psi}^*)$ where $\boldsymbol{\psi}^* \in \boldsymbol{\Psi}_{a,b}$ is obtained from $\bar{\boldsymbol{\psi}}$ by changing the gth mean component to $\boldsymbol{\mu}_g^* = \alpha\boldsymbol{\mu}_g$ for some $\alpha \in (0, 1)$ (i.e., along the line joining $\mathbf{0}$ and $\boldsymbol{\mu}_g$) such that $\boldsymbol{\mu}_g^* \in C(\mathcal{X})$.

Let us set $S = \{\boldsymbol{\psi} \in \boldsymbol{\Psi}_{a,b} \,|\, \boldsymbol{\mu}_g \in C(\mathcal{X}); 0 < a \leq \lambda_i(\boldsymbol{\Sigma}_g) \leq b < +\infty \quad g = 1, \ldots, G\}$. Then, it follows that

$$\sup_{\boldsymbol{\psi} \in \boldsymbol{\Psi}_{a,b}} \mathcal{L}(\boldsymbol{\psi}) = \sup_{\boldsymbol{\psi} \in S} \mathcal{L}(\boldsymbol{\psi}).$$

By the compactness of S and the continuity of $\mathcal{L}(\boldsymbol{\psi})$, there exists a parameter $\hat{\boldsymbol{\psi}} \in \boldsymbol{\Psi}_{a,b}$ satisfying

$$\mathcal{L}(\hat{\boldsymbol{\psi}}) = \sup_{\boldsymbol{\psi} \in \boldsymbol{\Psi}_{a,b}} \mathcal{L}(\boldsymbol{\psi}) = \sup_{\boldsymbol{\psi} \in S} \mathcal{L}(\boldsymbol{\psi})$$

by Weierstrass' theorem. □

The above recipes obviously require some a priori information on the covariance structure of the mixture throughout the bounds a and b in such a way that $\boldsymbol{\Psi}_{a,b}$ contains the maximum likelihood estimate $\hat{\boldsymbol{\psi}}$ introduced at the end of Section 2 and, at least, a reduced number of spurious maximizers.

When we lack such information, this position introduce subjectivity and this is a drawback. Hence, in Ingrassia and Rocci (2007), a weaker constraint is directly imposed on the ratio a/b, setting $a/b \geq c$, and using a suitable parameterization for $\boldsymbol{\Sigma}_g$. Let us rewrite them as $\boldsymbol{\Sigma}_g = \eta^2\boldsymbol{\Omega}_g$ $(g = 1, \ldots, G)$, where $\min_{ig} \lambda_i(\boldsymbol{\Omega}_g) = 1$ and impose the constraints

$$1 \leq \lambda_i(\boldsymbol{\Omega}_g) \leq \frac{1}{c}, \tag{27}$$

for $i = 1, \ldots, d$ and $g = 1, \ldots, G$. The constraints (27) are weaker than (21), in fact if (21) are satisfied and we set $\eta^2 = \min_{ig} \lambda_i(\boldsymbol{\Sigma}_g)$ and $\boldsymbol{\Omega}_g = \frac{\boldsymbol{\Sigma}_j}{\eta^2}$ then, by noting that $\lambda_i(\boldsymbol{\Omega}_g) = \eta^{-2}\lambda_i(\boldsymbol{\Sigma}_g)$, we obtain

$$1 \leq \lambda_i(\boldsymbol{\Omega}_g) \leq \frac{b}{a} \leq \frac{1}{c} \, .$$

Constraints (27), in turn, are stronger than (18). In fact, if the former are satisfied then

$$\lambda_{\min}(\boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_j^{-1}) \geq \frac{\lambda_{\min}(\boldsymbol{\Sigma}_g)}{\lambda_{\max}(\boldsymbol{\Sigma}_j)} = \frac{\lambda_{\min}(\boldsymbol{\Omega}_g)}{\lambda_{\max}(\boldsymbol{\Omega}_j)} \geq c, \qquad 1 \leq g \neq j \leq G\,.$$

Other similar constraints are proposed in Ingrassia and Rocci (2007).

The constraints (21) (as well as the constraints (27) and others of the same kind) can be implemented quite easily in the EM algorithm, using the spectral decomposition theorem. It is well known that any symmetric matrix $\mathbf{A}$ can be decomposed as:

$$\mathbf{A} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}' \tag{28}$$

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. Based on the formula (28), at the r-th iteration of the EM algorithm we can build an estimate $\boldsymbol{\Sigma}_g^{(r)}$ of the covariance matrix $\boldsymbol{\Sigma}_g$ such that the eigenvalues $\lambda_{ig}^{(r)} = \lambda_i(\boldsymbol{\Sigma}_g^{(r)})$ satisfy the constraints (21), by setting:

$$\lambda_{ig}^{(r)} = \min\left(b, \max\left(a, l_{ig}^{(r)}\right)\right) \tag{29}$$

where $l_{ig}^{(r)}$ is the update of $\lambda_i(\boldsymbol{\Sigma}_g)$ computed in the unconstrained M-step of the EM algorithm. The behaviour of (29) is illustrated in Figure 3. These constraints have been implemented in Ingrassia (2004), Ingrassia and Rocci (2007) and in Greselin and Ingrassia (2010), both for mixtures of multivariate Gaussian distributions and mixtures of multivariate t distributions. We remark that the approach (29) is not scale invariant.

An important issue concerns the *monotonicity* of the constrained EM algorithms described above. To this end, the following results hold.

Theorem 5 (Ingrassia and Rocci, 2007) *Let $\mathcal{L}(\boldsymbol{\psi})$ be the loglikelihood function for a mixture of elliptical distributions* (2)*, given a sample $\mathcal{X}$ of size N. Denote by $\{\boldsymbol{\psi}^{(r)}\}_r$ the sequence of the estimates generated by the EM algorithm, where $\boldsymbol{\psi}^{(r)} \in \boldsymbol{\Psi}_{a,b}$, for $r \geq 1$. Then, the resulting sequence of the loglikelihood values $\{\mathcal{L}(\boldsymbol{\psi}^{(r)})\}_r$ is not decreasing, once the initial guess $\boldsymbol{\psi}^{(0)} \in \boldsymbol{\Psi}_{a,b}$.*

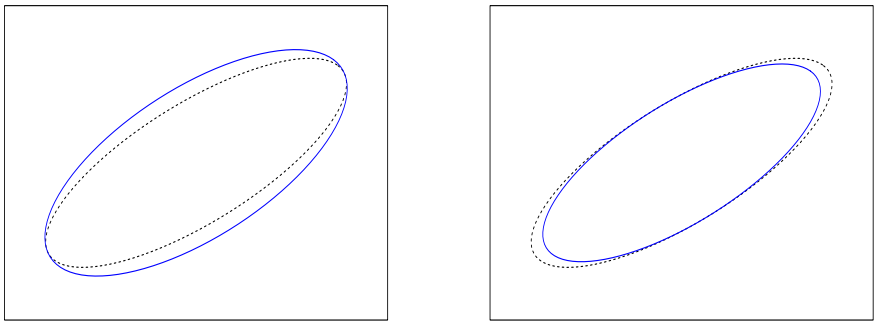
The proof relies on the following inequality, due to Theobald (1975, 1976). Let $\mathbf{A}, \mathbf{B}$ be two real symmetric $d \times d$ matrices and let $\boldsymbol{\Lambda}_A, \boldsymbol{\Lambda}_B$ be the corresponding diagonal matrices of the eigenvalues. Then, it results $\mathrm{tr}(\mathbf{A}\mathbf{B}^{-1}) \geq \mathrm{tr}(\boldsymbol{\Lambda}_A \boldsymbol{\Lambda}_B^{-1})$.

We remark that, even if the constraint (29) leads to a monotone EM algorithm, the choice of a, b is quite critical.

3.4 Equivariant constraints based on eigenvalues of the covariance matrices

Very recently, Rocci *et al.* (2017) proposed a generalization of the constraint (21) that enforces the equivariance with respect to linear affine transformation of the data. Let $\mathbf{A}$ be a $d \times d$ non singular matrix and $\mathbf{b} \in \mathbb{R}^d$ and consider

$$\mathbf{x}^* = \mathbf{A}\mathbf{x} + \mathbf{b}. \tag{30}$$



a) b)

Fig. 3 Results of enforcing constraints (29) on the eigenvalues of the covariance matrix at r-th step: a) If $l_{ig}^{(r)} < a$ (the covariance ellipse in dashed black) then the lowest eigenvalue should be forced to be equal to constant a (the covariance ellipse in blue). On the other hand, in b), if $l_{ig}^{(r)} > b$ (covariance ellipse in dashed black), then the greatest eigenvalue should be shrunk to constant b (the covariance ellipse in blue).

We easily see that

$$\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\mathbf{A}|\phi(\mathbf{x}^*; \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) \tag{31}$$

where $\boldsymbol{\mu}^* = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ and $\boldsymbol{\Sigma}^* = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$. If we consider Gaussian mixtures, it can be easily proven that the relation between the loglikelihood of the original data (7) and the loglikelihood of the transformed data is given by

$$\mathcal{L}(\boldsymbol{\psi}) = N \log |\mathbf{A}| + \mathcal{L}(\boldsymbol{\psi}^*) \tag{32}$$

where $\mathcal{L}(\boldsymbol{\psi}^*) = \sum_{n=1}^{N} \log \left(\sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_n^*; \boldsymbol{\mu}_g^*, \boldsymbol{\Sigma}_g^*) \right)$ is the loglikelihood based on the transformed data. Moreover, it can be proved that the classification of units, based on the maximum posterior probabilities, is invariant under this group of linear affine transformations.

It can be easily shown that constraints of kind (21) do not preserve the equivariance and this implies that the clustering depends on the choice of the matrix $\mathbf{A}$ in (30). In particular, constraints of kind (21) are sensitive to change in the units of data measurement. To overcome this drawback, Rocci *et al.* (2017) recently proposed to generalize the constraint (21) by considering

$$a \leq \lambda_i(\boldsymbol{\Sigma}_g \boldsymbol{\Xi}^{-1}) \leq b \tag{33}$$

where $\boldsymbol{\Xi}$ is a symmetric positive definite matrix representing the prior information about the covariance structure. In particular, the constraint (33) reduces to (21) when $\boldsymbol{\Xi} = \mathbf{I}$, i.e. the identity matrix. It can be proven that

this constraint implies (18), indeed we have

$$\lambda_{\min}(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_j^{-1}) \geq \lambda_{\min}(\boldsymbol{\Sigma}_g\boldsymbol{\Xi}^{-1})\lambda_{\min}(\boldsymbol{\Xi}\boldsymbol{\Sigma}_j^{-1}) = \frac{\lambda_{\min}(\boldsymbol{\Sigma}_g\boldsymbol{\Xi}^{-1})}{\lambda_{\max}(\boldsymbol{\Xi}\boldsymbol{\Sigma}_j^{-1})} \geq \frac{a}{b} \geq c.$$

The constraint (33) is shown to lead to an affine equivariant maximum likelihood function. Rocci *et al.* (2017) discuss also some data-driven choices for $\boldsymbol{\Xi}$ and c and propose an algorithm for maximizing (7) under the constraint (33).

3.5 Constraints on the ratio between maximum and minimum eigenvalues of the covariance matrices

A fully developed approach based on controlling the ratio between the maximum and the minimum eigenvalues of the groups scatter matrices has been proposed in García-Escudero *et al.* (2015). It is based on the robust classification framework previously developed by the same authors in García-Escudero *et al.* (2008), where a proportion of contaminating data was also discarded, to guarantee the robustness of the estimation.

The crucial feature introduced in Fritz *et al.* (2012) is that the optimum of the eigenvalues in the constrained space is obtained in closed form at each step of the EM algorithm. All previous attempts only offered approximations for it. For instance, García-Escudero *et al.* (2008) was based on the Dykstra (1983) algorithm. Ingrassia and Rocci (2007) and Greselin and Ingrassia (2010) were just based on truncating the scatter matrices eigenvalues to assure monotonicity of the likelihood throughout (29). They implemented the constraint

$$\frac{\lambda^*_{\max}}{\lambda^*_{\min}} \leq c' \tag{34}$$

where $c' \geq 1$ is a fixed constant, and $\lambda^*_{\min}$ and $\lambda^*_{\max}$ have been defined in (23) and (24), respectively. Note that the constraint (34) is equivalent to the one introduced in (22), with $c' = 1/c$.

Let $\boldsymbol{\Psi}^*_c$ be the set of mixture parameters obeying that eigenvalues ratio constraint for constant $c' = 1/c \geq 1$.

We take the opportunity of discussing in more depth this type of constraints here. They simultaneously control differences between groups and departure for sphericity. Note that the relative length of the equidensity ellipsoids axes, based on $\phi(\cdot; \mu_g, \Sigma_g)$, is forced to be smaller than $\sqrt{c}$, see Figure 4. The smaller c, the more similarly scattered and spherical the mixture components are. For instance, for $c = 1$ these ellipsoids reduce to balls with the same radius, so extending k-means, in the sense of allowing different component weights.

García-Escudero *et al.* (2015) gives results guaranteeing the existence of both the empirical and population problem solutions, as well as the consistency of the empirical solution to the population one. These two results requires only mild assumptions on the underlying distribution P. In particular, it is only required P to have a finite second moment, i.e. $\mathbb{E}_P[\|\cdot\|^2] < \infty$ and to avoid

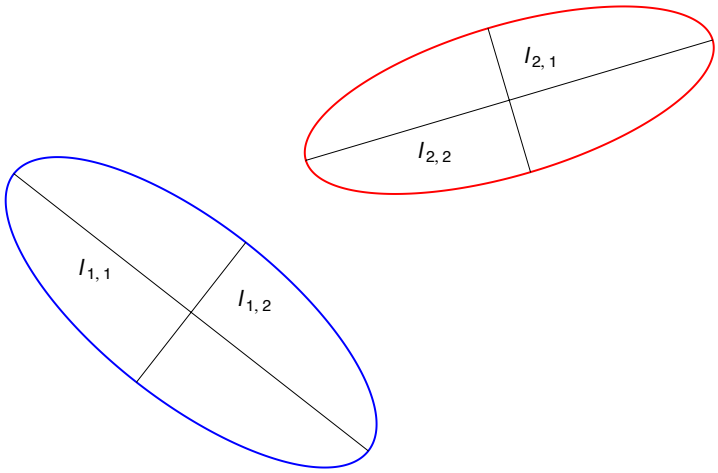


Fig. 4 If $\{l_{h,g}\}$ denotes the length of the semi-axes of the equidensity ellipsoids based on the normal density $\phi(\cdot;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ for $h=1,2$ and $g=1,2$, we set $\max\{l_{h,g}\}/\min\{l_{h,g}\}\leq\sqrt{c'}$ by (34).

that P is completely inappropriate for a mixture fitting approach by requiring that the distribution P is not concentrated on G points.

Proposition 6 (García-Escudero *et al.*, 2015) *If P is not concentrated on G points and $\mathbb{E}_P[\|\cdot\|^2]<\infty$, then there exists some $\boldsymbol{\psi}\in\boldsymbol{\Psi}_c^*$ such that the maximum of* (9) *under the constraint* (34) *is achieved.*

The following consistency result also holds under similar assumptions.

Proposition 7 (García-Escudero *et al.*, 2015) *Let us assume that P is not concentrated on G points and $\mathbb{E}_P[\|\cdot\|^2]<\infty$, and let $\boldsymbol{\psi}_0$ be the unique maximum of* (9) *under the constraint* (34)*. If $\boldsymbol{\psi}_N\in\boldsymbol{\Psi}_c^*$ denotes a sample version of the estimator based on the empirical measure P_N, then $\boldsymbol{\psi}_N\to\boldsymbol{\psi}_0$ almost surely, as $N\to\infty$.*

The consistency result presented in García-Escudero *et al.* (2008) needed an absolutely continuous distribution P with strictly positive density function (in the boundary of the set including the non-trimmed part of the distribution). This condition was needed due to the "trimming" approach considered by the TCLUST methodology. On the other hand, Propositions 6 and 7 do not longer need this assumption, it instead requires the finite second order moments hypothesis to control the tails of the mixture components.

Due to the employed constraints, this approach is rotation and translation equivariant, but not affine equivariant. Though the method becomes closer to affine equivariance when considering large values for c', it is always recommended to standardize the variables when very different measurement scales

are involved. Results in Hennig (2004), Ingrassia (2004), Ingrassia and Rocci (2007) and Greselin and Ingrassia (2010) may be seen as first steps toward the theoretical results presented in García-Escudero *et al.* (2015).

The constraints (34) have been firstly implemented in a Classification EM algorithm by solving several complex optimization problems at each iteration of the algorithm, through Dykstra's algorithm (Dykstra, 1983), to minimize a multivariate function on Gd parameters under $Gd(Gd-1)/2$ linear constraints. This original problem is computationally expensive, even for moderately high values of G or p. After introducing an efficient algorithm for solving the constrained maximization of the M step (Fritz *et al.*, 2013), besides the singular-value decompositions of the covariance matrices, only the evaluation of a univariate function $2Gd+1$ times is needed in the new M-step, with affordable computing time with respect to standard EM algorithms. More recently, the versions for constrained mixture estimation (García-Escudero *et al.*, 2015) and robust mixture modeling (García-Escudero *et al.*, 2014) have also been developed by the same authors.

It is important to note that the estimator in García-Escudero *et al.* (2015) is well defined as a maximum of the likelihood in the constrained space. The proposed algorithm tries to find this maximum by applying the constrained EM algorithm with multiple random initializations. Proposition 7 shows that the solution of the sample problem converges to the solution of the population one. The consistency result in Redner and Walker (1984) states that there is a sequence of local maxima of the likelihood converging to the optimum. Unfortunately, it does not provide a constructive way to choose the right optimal local maximum, for a fixed sample problem of size N, to obtain the consistent sequence.

We have seen that constraining the ratio between the scatter matrices eigenvalues to be smaller than a fixed in advance constant c' leads to an approach with nice theoretical properties (existence and consistency results) and a feasible algorithm for its practical implementation. Now we still may wonder on how to properly select the c' constant. In García-Escudero *et al.* (2015) it is argued that usually the researcher has some information about the populations underlying the dataset at hand, and this could help in setting a reasonable value for c'. If this were not the case, the suggestion is to choose small or moderate value of c', just to avoid degeneracies of the target function and to reduce the occurrence of non-interesting spurious solutions. However, more useful information can be obtained from a careful analysis of the fitted mixtures when moving parameter c in a controlled way, as shown in García-Escudero *et al.* (2015). A few essentially different solutions arise when considering "sensible" values of c', leading to a very reduced list of candidate mixture fits to be carefully investigated, to distinguish legitimate local maximizers from uninteresting spurious solutions.

3.6 Affine equivariant constraints on covariance matrices based on Löwner partial ordering

A different kind of constraints on the eigenvalues of the covariance matrices has been considered in Gallegos and Ritter (2009), resorting to the Löwner matrix ordering, in the framework of robust clustering. This approach is affine equivariant, as it can be easily seen.

Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric matrices of equal size. We say that $\mathbf{A}$ is less than or equal to $\mathbf{B}$ w.r.t. Löwner's (or semi-definite) order, and we denote it by $\mathbf{A} \preceq \mathbf{B}$, if $\mathbf{B} - \mathbf{A}$ is positive semi-definite. Analogously, if $\mathbf{B} - \mathbf{A}$ is positive definite, then we will write $\mathbf{A} \prec \mathbf{B}$. Puntanen *et al.* (2011) remark that the Löwner matrix ordering is a surprisingly strong and useful property. In particular, we recall that if $\mathbf{A} \preceq \mathbf{B}$, then it results $\lambda_i(\mathbf{A}) \leq \lambda_i(\mathbf{B})$, implying that $\text{trace}(\mathbf{A}) \leq \text{trace}(\mathbf{B})$ and $\det(\mathbf{A}) \leq \det(\mathbf{B})$.

To be more specific, Gallegos and Ritter (2009) constrained the scatter matrices to satisfy

$$c\boldsymbol{\Sigma}_j \preceq \boldsymbol{\Sigma}_g \quad \text{for every} \quad j \neq g = 1, \ldots, G, \tag{35}$$

where $c \geq 0$ is the HDBT constant, introduced in (18). It can be proved that the constraint (35) can be equivalently written as

$$\lambda(\boldsymbol{\Sigma}_j^{-1/2} \boldsymbol{\Sigma}_g \boldsymbol{\Sigma}_j^{-1/2}) \leq c \quad \text{for every} \quad g, j = 1, \ldots, G. \tag{36}$$

Differently from the previous cases, the constraint (35) depends also on the orientation of the covariance matrices. To see it, let $\mathbf{A}$ be a positive definite matrix and $\mathbf{B} = \mathbf{V}\mathbf{A}\mathbf{V}'$ be its rotation, under some orthonormal matrix $\mathbf{V}$. Then, in general, $\mathbf{B} - \mathbf{A}$ is not positive definite, i.e. $\mathbf{A} \not\prec \mathbf{B}$. In Figure 5 the matrix $\mathbf{A} = \text{diag}(1, 4)$ and some rotations of $\mathbf{A}$ are considered according to the rotation matrix

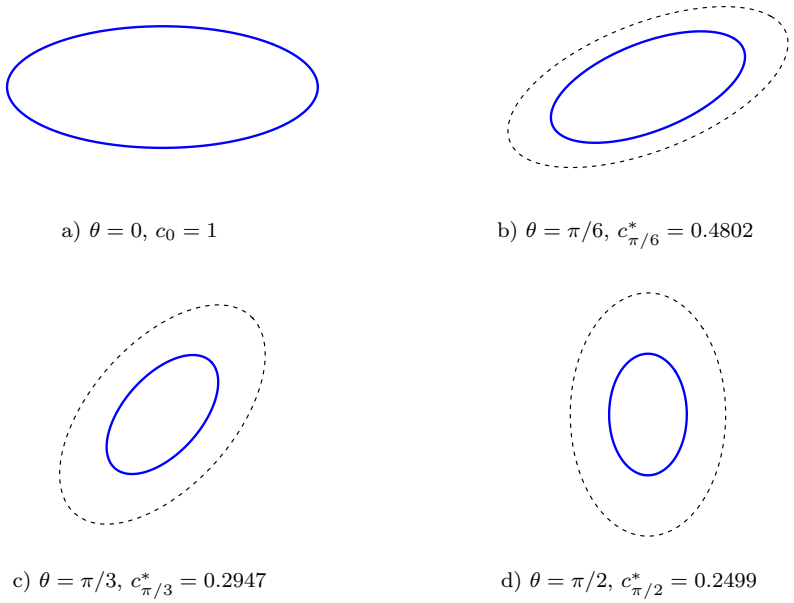
$$\mathbf{V}_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

For instance, let

$$\mathbf{B}_{i\pi/6} = \mathbf{V}_{i\pi/6} \mathbf{A} \mathbf{V}'_{i\pi/6},$$

for $i = 1, 2, 3$. Different suitable values of constant c are required in order to ensure that $c\mathbf{B}_{i\pi/6} \preceq \mathbf{A}$. In other words, the ordering depends not only on the eigenvalues but also on the relative rotation between the ellipsoids of equidensity. The largest values of c such that $c\mathbf{B}_{i\pi/6} \preceq \mathbf{A}$ are $c^*_{\pi/6} = 0.4802$, $c^*_{\pi/3} = 0.2947$ and $c^*_{\pi/2} = 0.2499$, respectively. See Figure 5 for details.

Apart from the peculiarities of the Löwner order discussed so far, a specific algorithm was not given in Gallegos and Ritter (2009) to solve the EM problems under constraint (35) for a fixed value of the constant c. Instead, the authors proposed to obtain all local maxima of the (trimmed) likelihood and derived from them the required values of c.



a) $\theta = 0$, $c_0 = 1$ b) $\theta = \pi/6$, $c^*_{\pi/6} = 0.4802$

c) $\theta = \pi/3$, $c^*_{\pi/3} = 0.2947$ d) $\theta = \pi/2$, $c^*_{\pi/2} = 0.2499$

Fig. 5 The Löwner order is related to the relative size *and* the rotation between ellipsoids of equidensity. The blue ellipse in panel a) is comparable with the blue ellipses in panels b), c) and d). The relation $c\mathbf{B}_\theta \preceq \mathbf{A}$ holds for $c \leq c^*_\theta$ where c^*_θ are the indicated thresholds.

4 On the degeneracy in the maximum likelihood estimation

Convergence properties of the EM algorithm nearby stationary point has been investigated in many papers. See e.g. Boyles (1983), Wu (1983), Meng (1994) and Nettleton (1999). Here we consider convergence properties of the EM towards a solution containing degenerate components, as introduced in Section 2. Let $\{\boldsymbol{\psi}^{(r)}\}_r$ be a sequence of the estimates of $\boldsymbol{\psi}$ provided by the EM and let $\{\mathcal{L}(\boldsymbol{\psi}^{(r)})\}_r$ be the corresponding sequence of the loglikelihood values. Throughout this section, for simplicity, the superscript "$-$" will denote the estimation at iteration r and the superscript "$+$" denotes the estimation at iteration $r + 1$.

The behaviour of the EM algorithm near degenerate components has been first investigated in Biernacki and Chrétien (2003) for mixtures of univariate Gaussian distributions. In particular, they prove the existence of a domain of attraction leading the EM algorithm to degeneracy.

Consider the sequence of the estimates $\{\boldsymbol{\psi}^{(r)}\}_r$ provided by the EM algorithm in the estimation of the parameters of a mixture of univariate Gaussian

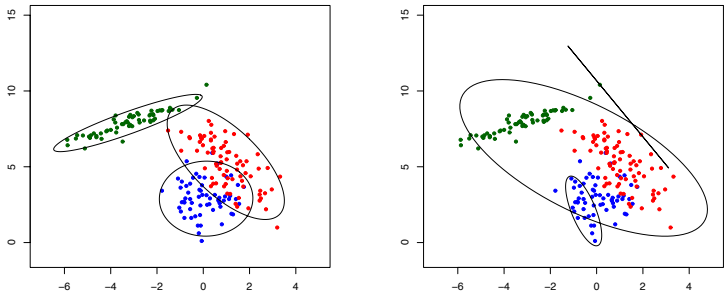


Fig. 6 Example of degenerate components

distributions (12), where $\boldsymbol{\psi}^{(r)} \in \boldsymbol{\Psi}$, with $\boldsymbol{\Psi}$ defined in (13). Let us set

$$f_{ng}^{(r)} = \pi_g^{(r)} \phi(x_n; \mu_g^{(r)}, {\sigma_g^{(r)}}^2), \tag{37}$$

and assume that, at the current iteration of the EM algorithm, the component g_0 $(1 \leq g_0 \leq G)$ is close to degeneracy at the unit x_{n0} $(1 \leq n_0 \leq N)$. Such a situation is equivalent to a high density $f_{n_0 g_0}^{(r)}$ of component g_0 at x_{n_0} and to small densities $f_{n g_0}^{(r)}$ at other units x_n $(n \neq n_0)$ (this occurs with probability one, assuming all individuals to be different with probability one). Afterwards, set the vector

$$\mathbf{v}_0 = \left(1/f_{n_0 g_0}, \{f_{n g_0}\}_{n \neq n_0}\right). \tag{38}$$

For a degenerate component, the Euclidean norm $\|\mathbf{v}_0\|$ is small. Finally, denote by $\mathbf{v}_0^{(r)}$ the value of $\mathbf{v}_0$ evaluated at step r of the EM algorithm. An example of degenerate component is given in Figure 6.

In this framework, Biernacki and Chrétien (2003) prove two results that we summarize below, see Theorems 8 and 9.

Theorem 8 (Biernacki and Chrétien, 2003) *There exists $\varepsilon > 0$ such that if $\|\mathbf{v}_0\| \leq \varepsilon$ then $\|\mathbf{v}_0^{(r)}\| = o\|\mathbf{v}_0\|$ with probability one.*

The proof follows from the Taylor expansions for parameters $\pi_{g_0}^+, \mu_{g_0}^+$ and $\sigma_{g_0}^{2+}$. This results states that, if $\|\mathbf{v}_0\|$ is small enough, than the EM mapping is contracting and, therefore, EM is convergent and its fixed point is degenerated

The second results concerns the speed towards degeneracy.

Theorem 9 (Biernacki and Chrétien, 2003) *There exists $\varepsilon > 0$, $\alpha > 0$ and $\beta > 0$ such that if $\|\mathbf{v}_0\| \leq \varepsilon$ then, with probability one*

$$\sigma_{g_0}^{2+} \leq \alpha \frac{\exp(-\beta/\sigma_{g_0}^{2-})}{\sigma_{g_0}^{2-}} \tag{39}$$

The proof follows again from the Taylor expansions. In particular, this results establishes that the variance of a degenerated component tends to zero with an exponential rate. Since the likelihood tends to infinity as fast as the inverse of the standard deviation, the divergence of the likelihood is exponential too.

These results have been extended to the multivariate case in Ingrassia and Rocci (2011). In this case, a degenerate solutions occurs at some subset of $\mathcal{X}$ containing $q \leq d$ points. Thus, let $\mathcal{D}$ be a subset of $\{1, 2, \ldots, N\}$ containing $q \leq d$ units and consider the vector

$$\mathbf{v}_0 = \left(\{1/f_{ng_0}\}_{n \in \mathcal{D}}, \{f_{ng_0}\}_{n \notin \mathcal{D}} \right) \tag{40}$$

which generalizes (38) in the multivariate setting. The following result extends Theorem 9 to the multivariate case.

Theorem 10 (Ingrassia and Rocci, 2011) *Let g_0 be a degenerate component of the mixture (3)), with $1 \leq g_0 \leq G$. Let $\boldsymbol{\Sigma}_{g_0}^{(m+1)}$ be the estimate of the covariance matrix $\boldsymbol{\Sigma}_{j_0}$ at iteration $m+1 \in \mathbb{N}$. There exist $\varepsilon > 0$ and $\beta > 0$ such that if $\|\mathbf{v}_0\| \leq \varepsilon$ then*

$$\lambda_{\min}(\boldsymbol{\Sigma}_{g_0}^{+}) < \frac{\delta}{[\lambda_{\min}(\boldsymbol{\Sigma}_{g_0}^{-})]^{d/2}} \exp\left\{ -\frac{\beta}{4\lambda_{\min}(\boldsymbol{\Sigma}_{g_0}^{-})} \right\} + \mathrm{o}||\mathbf{v}_0||. \tag{41}$$

The proof follows again from arguments based on the Taylor expansion.

The relation (41) suggests that, near degeneracy, the smallest eigenvalue decreases very fast. Based on this result, in Ingrassia and Rocci (2011) is conjectured that such bad behavior should be prevented by bounding the eigenvalues variations between two consecutive iterations. Thus, the idea is to control the speed of variation of both the smallest and the largest eigenvalues at each iteration and the following constraints have been proposed

$$\lambda_{\min}(\boldsymbol{\Sigma}_j^{-})/\vartheta_a \leq \lambda_{ij}^{+} \leq \vartheta_b \lambda_{\max}(\boldsymbol{\Sigma}_j^{-}), \tag{42}$$

with $\vartheta_a, \vartheta_b > 1$ and thus from (42) we get

$$\lambda_{ij}^{(m+1)} = \min\left(\vartheta_b \lambda_{\max}(\boldsymbol{\Sigma}_j^{(m)}), \max\left(\lambda_{\min}(\boldsymbol{\Sigma}_j^{(m)})/\vartheta_a, l_{ij}^{(m+1)} \right) \right). \tag{43}$$

We refer to such constraints as "dynamic constraints" to point out that the bound on the eigenvalue at the current iteration depends on the value of the eigenvalue computed at the previous step of the algorithm. On the contrary, we shall refer to the type of constraints like in (21) as "static" because the interval remains fixed during the whole computation.

A monotone algorithm implementing (42) can be easily derived by using the constrained EM algorithm of Ingrassia and Rocci (2007) previously described. We remark that this implementation does not lead to an EM algorithm, because in the "M-step" the complete log-likelihood function is not necessarily maximized. However, by noting that at iteration $m+1$ the complete log-likelihood is increased by every update of λ_{ij} lying in the interval

$$[\min(\lambda_{ij}^{m}, l_{ij}^{(m+1)}), \max(\lambda_{ij}^{m}, l_{ij}^{(m+1)})],$$

in the M-step the complete log-likelihood is always increased. This kind of algorithm, where the complete log-likelihood is increased instead of maximized, has been referred to as *generalized* EM in Dempster *et al.* (1977).

We point out that such constraints have a different background with respect to (21), (27) and others proposed in Ingrassia and Rocci (2007). The latter are based on a constrained formulation of the likelihood function for mixture models; the constraints proposed here are in some sense of an algorithmic type being based on the convergence properties of the EM algorithm.

Our results highlighted that, in some way, the convergence of the EM algorithm to some spurious maximum is also due to the properties of the algorithm itself. Indeed, since in the near degeneracy case the covariance matrices converge at exponential rate toward singularity, this implies that in such cases such a covariance matrix could model some spurious small group of data quite quickly and this amounts to an increase in the probability of the algorithm to get stuck into some spurious maximum.

In general, dynamic constraints performed always at least as good as the unconstrained EM algorithm and good performances were attained when both bounds on the variation of the eigenvalues were implemented.

5 Gaussian parsimonious clustering models

In literature, intermediate component covariance matrices lying between homoscedasticity and heteroscedasticity, have been proposed by Banfield and Raftery (1993) and Celeux and Govaert (1995). They proposed a general framework for geometric cross-cluster constraints in multivariate normal mixtures by parameterizing covariance matrices through eigenvalue decomposition in the form

$$\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g' \qquad g = 1, \ldots, G \tag{44}$$

where $\mathbf{D}_g$ is the orthogonal matrix of the eigenvectors of $\boldsymbol{\Sigma}_g$, describing the scatter *orientation*, $\mathbf{A}_g$ is a scaled ($|A_g| = 1$) diagonal matrix whose elements are proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$ in decreasing order, giving the *shape,* and λ_g is an associated constant of proportionality, related to the *volume* of the clusters, which is proportional to $\lambda_g^{d/2} = |\boldsymbol{\Sigma}_g|^{1/2}$. The idea is to treat $\lambda_g, \mathbf{D}_g$ and $\mathbf{A}_g$ as independent sets of parameters and either constrain them to be the same for each cluster or allow them to vary among clusters, see Table 1. In a quite different context, similar ideas have been introduced in Greselin *et al.* (2011).

Hence, we allow the volumes, the shapes and the orientations of clusters to vary or to be equal between clusters. Variations on assumptions on the parameters $\lambda_g, \mathbf{D}_g$ and $\mathbf{A}_g$ ($g = 1, \ldots, G$) lead to fourteen general models of interest, plus two models for the univariate case. For instance, we can assume different volumes and keep the shapes and orientations equal by requiring that $\mathbf{A}_g = \mathbf{A}$ ($\mathbf{A}$ unknown) and $\mathbf{D}_g = \mathbf{D}$ ($\mathbf{D}$ unknown) for $g = 1, \ldots, G$. We denote this model $[\lambda_g \mathbf{D}\mathbf{A}\mathbf{D}']$. With this convention, writing (for instance)

Model ID	*Model*	*Distribution*	*Volume*	*Shape*	*Orientation*	*# parameters*
E		univariate	equal			1
V		univariate	variable			G
EII	$[\lambda\mathbf{I}]$	spherical	equal	equal	NA	$\alpha+1$
VII	$[\lambda_g\mathbf{I}]$	spherical	variable	equal	NA	$\alpha+d$
EII	$[\lambda\mathbf{A}]$	diagonal	equal	equal	coordinate axes	$\alpha+d$
VEI	$[\lambda_g\mathbf{A}]$	diagonal	variable	equal	coordinate axes	$\alpha+d+G-1$
EVI	$[\lambda\mathbf{A}_g]$	diagonal	equal	variable	coordinate axes	$\alpha+dG-G+1$
VVI	$[\lambda_g\mathbf{A}_g]$	diagonal	variable	variable	coordinate axes	$\alpha+dG$
EEE	$[\lambda\mathbf{DAD}']$	ellipsoidal	equal	equal	equal	$\alpha+\beta$
VEE	$[\lambda_g\mathbf{DAD}']$	ellipsoidal	variable	equal	equal	$\alpha+\beta+G-1$
EVE	$[\lambda\mathbf{DA}_g\mathbf{D}']$	ellipsoidal	equal	variable	equal	$\alpha+\beta+(G-1)(d-1)$
VVE	$[\lambda_g\mathbf{DA}_g\mathbf{D}']$	ellipsoidal	variable	variable	equal	$\alpha+\beta+(G-1)d$
EEV	$[\lambda\mathbf{D}_g\mathbf{AD}_g']$	ellipsoidal	equal	equal	variable	$\alpha+G\beta-(G-1)d$
VEV	$[\lambda_g\mathbf{D}_g\mathbf{AD}_g']$	ellipsoidal	variable	equal	variable	$\alpha+G\beta-(G-1)(d-1)$
EVV	$[\lambda\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g']$	ellipsoidal	equal	variable	variable	$\alpha+G\beta-(G-1)$
VVV	$[\lambda_g\mathbf{D}_g\mathbf{A}_g\mathbf{D}_g']$	ellipsoidal	variable	variable	variable	$\alpha+G\beta$

Table 1 Parameterizations of the covariance matrix $\boldsymbol{\Sigma}_g$. We have $\alpha = Gd$ in the restricted case (equal weights, $\pi_g = 1/G$) and $\alpha = Gd + G - 1$ in the unrestricted case. β denotes the number of parameters of each covariance matrix, i.e. $\beta = d(d+1)/2$.

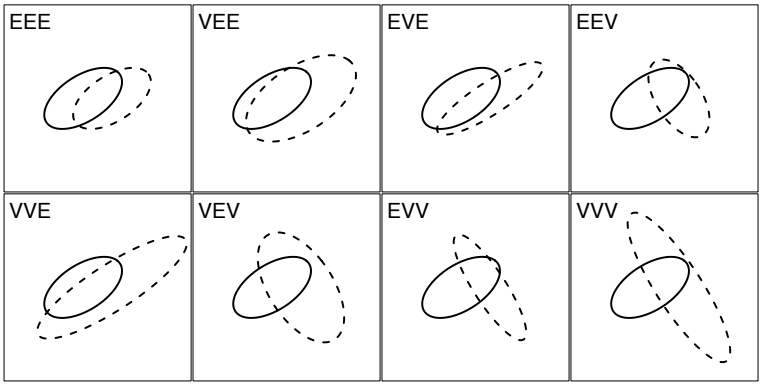


Fig. 7 Example of covariance matrices having different patterns according to Table 1.

$[\lambda\mathbf{D}_g\mathbf{AD}_g']$ means that we consider the mixture model with equal volumes λ, equal shapes $\mathbf{A}$ and different orientations $\mathbf{D}_g$. All these models can be estimated by the MCLUST software (Fraley and Raftery, 1999, 2003, 2006; Fraley *et al.*, 2012) where they are designated by a three-letter symbol indicating volume, shape and orientation, respectively. The letter E indicates cross-cluster equality, while V denotes freedom to vary across clusters, and the letter I designates a spherical shape or an axis-aligned orientation. Patterned covariance matrices are listed in Table 1 and some of them are showed in Figure 7.

This approach has been extended to multivariate t-distributions by Andrews *et al.* (2011).

We see that all models with equal volume and equal shape offer a good alternative when looking at covariance constraints to avoid singularities and reduce spurious solutions. For the other cases, "NA" values could appear in MCLUST results, due to failure in the EM computations caused by singularity and/or shrinking components, meaning that a particular model cannot be estimated. Hence a constraint on the eigenvalues is still needed, and the function `emControl` has been added to MCLUST.

6 Mixtures of factor analyzers

The general Gaussian mixture model (3) is a highly parameterized model, with a total of $(G-1)+G[d+d(d+1)/2]$ parameters. Looking for parsimony, it is of interest to develop some methods for reducing the covariance matrices parametrization $\boldsymbol{\Sigma}_g$, requiring $Gd(d+1)/2$ parameters. To this purpose, Ghahramani and Hinton (1997) and Tipping and Bishop (1999) proposed Gaussian Mixtures of Factor Analyzers (MFA), able to explain multivariate observations, by explicitly modeling correlations between variables.

This approach postulates a finite mixture of linear sub-models for the distribution of the full observation vector $\mathbf{X}$, given the (unobservable) factors $\mathbf{U}$. Hence, it provides local dimensionality reduction by assuming that the distribution of the observation $\mathbf{X}_n$ is given by

$$\mathbf{X}_n = \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{U}_{ng} + \mathbf{e}_{ng} \qquad g = 1, \ldots, G,\ n = 1, \ldots, N, \tag{45}$$

with probability π_g, where $\boldsymbol{\Lambda}_g$ is a $d \times q$ matrix of *factor loadings*, the *factors* $\mathbf{U}_{1g}, \ldots, \mathbf{U}_{Ng}$ are $\mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ distributed independently of the *errors* $\mathbf{e}_{ng}$, which are $\mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_g)$ distributed, and $\boldsymbol{\Psi}_g$ is a $d \times d$ diagonal matrix ($g = 1, \ldots, G$). The diagonality of $\boldsymbol{\Psi}_g$ is one of the key assumptions of factor analysis: the observed variables are independent given the factors.

Note that the factor variables $\mathbf{U}_{ng}$ model correlations between the elements of $\mathbf{X}_n$, while the $\mathbf{e}_{ng}$ variables account for independent noise for $\mathbf{X}_n$. 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 (3), where the g-th component-covariance matrix $\boldsymbol{\Sigma}_g$ has the specific form

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g \quad (g = 1, \ldots, G). \tag{46}$$

The parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}_{MFA}(d, q, G)$ now consists of the elements of the component means $\boldsymbol{\mu}_g$, the $\boldsymbol{\Lambda}_g$, and the $\boldsymbol{\Psi}_g$, along with the mixing proportions π_g ($g = 1, \ldots, G-1$), on putting $\pi_G = 1 - \sum_{g=1}^{G-1} \pi_g$. Note that, in the case of $q > 1$, there is an infinity of choices for $\boldsymbol{\Lambda}_g$, since model (45) is still satisfied if we replace $\boldsymbol{\Lambda}_g$ by $\boldsymbol{\Lambda}_g \mathbf{H}'$, where $\mathbf{H}$ is any orthogonal matrix of order q. As $q(q-1)/2$ constraints are needed for $\boldsymbol{\Lambda}_g$ to be uniquely defined, the number of free parameters for each component of the mixture, is $dq + d - \frac{1}{2}q(q-1)$.

Model ID	*Loading matrix* $\mathbf{\Lambda}_g$	*Error variance* $\mathbf{\Psi}_g$	*Isotropic* $\mathbf{\Psi}_g = \psi_g \mathbf{I}$	*# parameters*
CCC	Constrained	Constrained	Constrained	$[pq - q(q-1)/2] + 1$
CCU	Constrained	Constrained	Unconstrained	$[pq - q(q-1)/2] + p$
CUC	Constrained	Unconstrained	Constrained	$[pq - q(q-1)/2] + G$
CUU	Constrained	Unconstrained	Unconstrained	$[pq - q(q-1)/2] + Gp$
UCC	Unconstrained	Constrained	Constrained	$G[pq - q(q-1)/2] + 1$
UCU	Unconstrained	Constrained	Unconstrained	$G[pq - q(q-1)/2] + p$
UUC	Unconstrained	Unconstrained	Constrained	$G[pq - q(q-1)/2] + G$
UUU	Unconstrained	Unconstrained	Unconstrained	$G[pq - q(q-1)/2] + Gp$

Table 2 Parsimonious covariance structures introduced for mixtures of factor analyzers.

Parameters in mixtures of factor analyzers are usually estimated according to the likelihood approach, based on the AECM algorithm (Meng and van Dyk, 1997). AECM is a variant of the EM procedure, where two E and conditional M steps are alternated, acting on a partition of the parameter space, particularly suitable for ML for Gaussian factors.

McNicholas and Murphy (2008) extended the idea of patterned covariance matrices to mixture of factor analyzers by considering constraints across groups on the $\mathbf{\Lambda}_g$ and $\mathbf{\Psi}_g$ matrices and on whether or not $\mathbf{\Psi}_g = \psi_g \mathbf{I}_p$. The full range of possible constraints provides a class of eight *parsimonious Gaussian mixture of factor analyzer* models, which are given in Table 2. These models provide a unified modeling framework which includes the mixtures of probabilistic principal component analyzers and mixtures of factor of analyzers models as special cases. This approach has been extended in different directions: Andrews and McNicholas (2011) generalized this family of models to multivariate t distributions while Subedi *et al.* (2013) and Subedi *et al.* (2015) introduced their version for cluster-weighted models.

To discuss these models from their ability to deal with singularity and spurious solutions along the ML estimation, we observe that the error matrices are either "unconstrained isotropic" $\mathbf{\Psi}_g = \psi_g \mathbf{I}$ with different ψ_g for each component, or "constrained isotropic", say $\mathbf{\Psi}_g = \mathbf{\Psi} = \psi \mathbf{I}$. This parameterization choice leads the covariance matrices $\mathbf{\Sigma}_g = \mathbf{\Lambda}_g' \mathbf{\Lambda}_g + \mathbf{\Psi}_g$ far from singularities.

In a different approach, Greselin and Ingrassia (2015) constraints of type (21) have been proposed for mixtures of factor analyzers. Due to the structure of the covariance matrix $\mathbf{\Sigma}_g$ given in (46), bound in (21) yields

$$a \leq \lambda(\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g) \leq b, \qquad g = 1, \ldots, G\,. \tag{47}$$

Concerning the square $d \times d$ matrix $\mathbf{\Lambda}_g \mathbf{\Lambda}_g'$ $(g = 1, \ldots, G)$, we can get its eigenvalue decomposition, i.e. we can find $\mathbf{\Lambda}_g$ and $\mathbf{\Gamma}_g$ such that

$$\mathbf{\Lambda}_g \mathbf{\Lambda}_g' = \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}_g' \tag{48}$$

where $\mathbf{\Gamma}_g$ is the orthonormal matrix whose columns are the eigenvectors of $\mathbf{\Lambda}_g \mathbf{\Lambda}_g'$ and $\mathbf{\Delta}_g = \mathrm{diag}(\delta_{1g}, \ldots, \delta_{dg})$ is the diagonal matrix of the eigenvalues

of $\mathbf{\Lambda}_g\mathbf{\Lambda}_g'$, sorted in non increasing order, i.e. $\delta_{1g} \geq \delta_{2g} \geq \ldots \geq \delta_{qg} \geq 0$, and $\delta_{(q+1)g} = \cdots = \delta_{dg} = 0$.

Now, let us consider the singular value decomposition of the $d \times q$ rectangular matrix $\mathbf{\Lambda}_g$, so giving $\mathbf{\Lambda}_g = \mathbf{U}_g\mathbf{D}_g\mathbf{V}_g'$, 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 vectors* of $\mathbf{\Lambda}_g$, respectively. Now we have that

$$\mathbf{\Lambda}_g\mathbf{\Lambda}_g' = (\mathbf{U}_g\mathbf{D}_g\mathbf{V}_g')(\mathbf{V}_g\mathbf{D}_g'\mathbf{U}_g') = \mathbf{U}_g\mathbf{D}_g\mathbf{I}_q\mathbf{D}_g'\mathbf{U}_g' = \mathbf{U}_g\mathbf{D}_g\mathbf{D}_g'\mathbf{U}_g' \tag{49}$$

and equating (48) and (49) we get $\mathbf{\Gamma}_g = \mathbf{U}_g$ and $\mathbf{\Delta}_g = \mathbf{D}_g\mathbf{D}_g'$, that is

$$\operatorname{diag}(\delta_{1g}, \ldots, \delta_{qg}) = \operatorname{diag}(d_{1g}^2, \ldots, d_{qg}^2)\,., \tag{50}$$

with $d_{1g} \geq d_{2g} \geq \cdots \geq d_{qg} \geq 0$. In particular, it is known that only the first q values of $\mathbf{D}_g$ are non negative, and the remaining $d - q$ terms are null.

Denoting now by ψ_{ig} the i-th eigenvalue of $\mathbf{\Psi}_g$, then constraint (47) is satisfied when

$$d_{ig}^2 + \psi_{ig} \geq a \qquad i = 1, \ldots, d \tag{51}$$

$$d_{ig} \leq \sqrt{b - \psi_{ig}} \qquad i = 1, \ldots, q$$

$$\psi_{ig} \leq b \qquad i = q+1, \ldots, d \tag{52}$$

for $g = 1, \ldots, G$. In particular, we remark that condition (51) reduces to $\psi_{ig} \geq a$ for $i = (q+1), \ldots, d$.

The two-fold (eigenvalue and singular value) decomposition of the $\mathbf{\Lambda}_g$ presented above, suggests how to modify the EM algorithm in such a way that the eigenvalues of the covariances $\boldsymbol{\Sigma}_g$ (for $g = 1, \ldots, G$) are confined into suitable ranges. Details are given in Greselin and Ingrassia (2015). Finally, we observe that only constraints on $\mathbf{\Psi}_g$ are needed to discard singularities and to reduce spurious maximizers, as it has been done in a robust approach for estimating Mixtures of Gaussian factors in García-Escudero *et al.* (2016).

7 Concluding remarks

In the maximum likelihood approach for model based clustering and classification, based on mixtures of elliptical components, we have recalled here the need of considering a constrained parameter space for the covariance matrices, to yield a well posed optimization problem.

We have reviewed several different approaches for setting constraints to the eigenvalues of the covariance matrices, as well as the algorithms needed for their exact or approximate implementation. We also discussed the historical path, starting from the popular k-means methods, that implicitly assumes the strongest constraint of equal spherical covariance matrices, till the most

recent contributions in the literature that allows for milder conditions, and arriving to dynamic and/or affine equivariant constraints. We developed a detailed comparison of the advantages of each proposal, also in view of obtaining a sound theoretical framework assuring existence and consistency to the obtained estimator.

Usually, the ML estimation is performed by using the EM algorithm. The latter is a powerful iterative process, leading deterministically to a specific solution, depending from the initial step. This makes the choice of the starting points a very delicate matter. Many efforts have been made in the literature to devise smart initialization methods, mainly to avoid convergence toward singularities or spurious solutions. We have seen that the constrained approach with a reasonable number of random initializations, on the other hand, yields to a reduced set of meaningful solutions. The researcher can devise among them the most convincing one, or the more interesting from the point of view of the obtained clustering, in view of his knowledge of the field of application. We argue that the clustering of a dataset should not be based solely on a single solution of the likelihood equation, but rather on the various solutions considered collectively and analyzed with care.

Finally, we discussed along the paper that the constrained approach in mixture modelling, beyond allowing a proper mathematical setting of the optimization problem, at the same time provides stability to the obtained solutions.

As we stated in the introduction, the role of constrained estimation within robust statistical methods needs a longer discussion and will be the object of a further paper.

References

Andrews, J. L. and McNicholas, P. D. (2011). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing*, **21**, 361–373.

Andrews, J. L., McNicholas, P. D., and Subedi, S. (2011). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics & Data Analysis*, **55**, 520–529.

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

Biernacki, C. and Chrétien, S. (2003). Degeneracy in the maximum likelihood estimation of univariate gaussian mixtures with the EM. *Statistics & Probability Letters*, **61**, 373–382.

Boyles, R. A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society B*, **45**, 47–50.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**, 781–793.

Chanda, K. C. (1954). A note on the consistency and maxima of the roots of likelihood equations. *Biometrika*, **41**, 56–61.

Ciuperca, G., Ridolfi, A., and Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, **30**, 45–59.

Cramér, H. (1946). *Mathematical methods of Statistics*. Princeton University Press, Princeton, New Jersey.

Day, N. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, **56**(3), 463–474.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38.

Dennis, J. E. (1981). Algorithms for non linear fitting. In *Proceedings of the NATO Advanced Research Symposium*, Cambridge, England. Cambridge University.

Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, **78**(384), 837–842.

Fang, K. and Anderson, T. (1990). *Statistical inference in elliptically contoured and related distributions*. Alberton, New York.

Fraley, C. and Raftery, A. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, **24**(2), 155–181.

Fraley, C. and Raftery, A. E. (1999). Mclust: Software for model-based cluster analysis. *Journal of classification*, **16**(2), 297–306.

Fraley, C. and Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust. *Journal of Classification*, **20**(2), 263–286.

Fraley, C. and Raftery, A. E. (2006). Mclust version 3: an r package for normal mixture modeling and model-based clustering. Technical report, DTIC Document.

Fraley, C., Raftery, A., Murphy, T., and Scrucca, L. (2012). mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. *University of Washington: Seattle*.

Fritz, H., García-Escudero, L. A., and Mayo-Iscar, A. (2012). tclust: An r package for a trimming approach to cluster analysis. *Journal of Statistical Software*, **47**(12).

Fritz, H., García-Escudero, L. A., and Mayo-Iscar, A. (2013). A fast algorithm for robust constrained clustering. *Comput. Stat. Data Anal.*, **61**, 124–136.

Gallegos, M. and Ritter, G. (2009). Trimmed ML estimation of contaminated mixture. *Sankhya (Ser. A)*, **71**, 164–220.

Gallegos, M. T. (2002). Maximum likelihood clustering with outliers. In *Classification, Clustering, and Data Analysis*, pages 247–255. Springer.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, **36**(3), 1324–1345.

García-Escudero, L. A., Gordaliza, A., and Mayo-Iscar, A. (2014). A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification*, **8**(1), 27–43.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2015). Avoiding spurious local maximizers in mixture modelling. *Statistics and Computing*, **25**, 619–633.

García-Escudero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S., and Mayo-Iscar, A. (2016). The joint role of trimming and constraints in robust estimation for mixtures of Gaussian factor analyzers. *Computational Statistics & Data Analysis*, (99), 131–147.

García-Escudero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S., and Mayo-Iscar, A. (2017). Eigenvalues in robust approaches to mixture modeling: a review. Technical report, in preparation.

Ghahramani, Z. and Hinton, G. (1997). The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University of Toronto.

Greselin, F. and Ingrassia, S. (2010). Constrained monotone em algorithms for mixtures of multivariate t-distributions. *Statistics and Computing*, **20**, 9–22.

Greselin, F. and Ingrassia, S. (2015). Maximum likelihood estimation in constrained parameter spaces for mixtures of factor analyzers. *Statistics and Computing*, **25**(2), 215–226.

Greselin, F., Ingrassia, S., and Punzo, A. (2011). Assessing the pattern of covariance matrices via an augmentation multiple testing procedure. *Statistical Methods & Applications*, **20**, 141–170.

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

Hathaway, R. J. (1996). A constrained EM algorithm for univariate normal mixtures. *Journal of Statistical Computation and Simulation*, **23**, 211–230.

Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *The Annals of Statistics*, **32**, 1313–1340.

Ingrassia, S. (1992). A comparison between the simulated annealing and the EM algorithms in normal mixture decompositions. *Statistics and Computing*, **2**, 203–211.

Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods & Applications*, **13**(4), 151–166.

Ingrassia, S. and Rocci, R. (2007). Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Computational Statistics & Data Analysis*, **51**(11), 5339–5351.

Ingrassia, S. and Rocci, R. (2011). Degeneracy of the EM algorithm for the MLE of multivariate Gaussian mixtures and dynamic constraints. *Computational Statistics & Data Analysis*, **55**(4), 1715–1725.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, **27**(4), 887–906.

Kiefer, N. M. (1978). Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrica*, **46**(2), 427–434.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics. IMS-ASA.

MacQueen, J. *et al.* (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

McLachlan, G. and Krishnan, T. (2008a). *The EM Algorithm and Extensions (2nd edition)*, volume 589. Wiley Series in Probability and Statistics.

McLachlan, G. J. and Krishnan, T. (2008b). *The EM Algorithm and its Extensions*. John Wiley & Sons, 2nd edition, New York.

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

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**, 285–296.

Meng, X.-L. (1994). On the rate of convergence of the ecm algorithm. *Annals of Statistics*, **22**(1), 326–339.

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

Nettleton, D. (1999). Convergence properties of the EM algorithm in constrained spaces. *The Canadian Journal of Statistics*, **27**(3), 639–644.

Puntanen, S., Styan, G. P., and Isotalo, J. (2011). *Matrix Tricks for Linear Statistical Models*. Springer, Berlin.

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

Rocci, R., Gattone, S. A., and Di Mari, R. (2017). A data driven equivariant approach to constrained Gaussian mixture modeling. *Advances in Data Analysis and Classification*, (forthcoming).

Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. D. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, **7**, 5–40.

Subedi, S., Punzo, A., Ingrassia, S., and McNicholas, P. D. (2015). Cluster-weighted t-factor analyzers for robust model-based clustering and dimension reduction. *Statistical Methods & Applications*, forthcoming.

Tarone, R. D. and Gruenhage, G. (1975). A note on the uniqueness of the roots of the likelihood equations for vector-valued parameters. *Journal of the American Statistical Association*, **70**, 903–904.

Tarone, R. D. and Gruenhage, G. (1979). Corrigenda: A note on the uniqueness of the roots of the likelihood equations for vector-valued parameters. *Journal of the American Statistical Association*, **74**, 744.

Theobald, C. (1975). An inequality with application to multivariate analysis. *Biometrika*, **62**(2), 461–466.

Theobald, C. (1976). Corrections and amendments: An inequality with application to multivariate analysis. *Biometrika*, **63**(3), 685.

Tipping, M. and Bishop, C. M. (1999). Mixtures of probabilistic principal component mixtures of probabilistic principal component analysers. *Neural Computation*, **11**(2), 443–482.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.

van Laarhoven, P. J. M. and Aarts, E. H. L. (1988). *Simulated Annealing: Theory and Practice*. D. Reidel, Dordecht.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, **58**(301), 236–244.

Wu, C. F. J. (1983). On convergence properties of the EM algorithm. *Annals of Statistics*, **11**, 95–103.