Model-based clustering with determinant-and-shape constraint

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

Model-Based Clustering with Determinant-and-Shape Constraints

Luis Angel García-Escudero · Agustín Mayo-Iscar · Marco Riani

Received: date / Accepted: date

Abstract Model-based approaches to cluster analysis and mixture modeling often involve maximizing classification and mixture likelihoods. Without appropriate constrains on the scatter matrices of the components, these maximizations result in ill-posed problems. Moreover, without constrains, non-interesting or "spurious" clusters are often detected by the EM and CEM algorithms traditionally used for the maximization of the likelihood criteria. Considering an upper bound on the maximal ratio between the determinants of the scatter matrices seems to be a sensible way to overcome these problems by affine equivariant constraints. Unfortunately, problems still arise without also controlling the elements of the "shape" matrices. A new methodology is proposed that allows both control of the scatter matrices determinants and also the shape matrices elements. Some theoretical justification is given. A fast algorithm is proposed for this doubly constrained maximization. The methodology is also extended to robust model-based clustering problems.

Keywords Clustering · Constraints · Mixture modeling · Robustness.

This research is partially supported by Spanish Ministerio de Economía y Competitividad, grant MTM2017-86061-C2-1-P, and by Consejería de Educación de la Junta de Castilla y León and FEDER, grant VA005P17 and VA002G18. This research benefits from the HPC (High Performance Computing) facility of the University of Parma, Italy. M.R. gratefully acknowledges support from the CRoNoS project, reference CRoNoS COST Action IC1408 and the European Union's Horizon 2020 Research and Innovation Program for its financial support of the PrimeFish project, Grant Agreement No. 635761.

L.A. García-Escudero
Department of Statistics and Operational Research and IMUVA, University of Valladolid
E-mail: lagarcia@eio.uva.es

A. Mayo-Iscar
Department of Statistics and Operational Research and IMUVA, University of Valladolid
E-mail: agustinm@eio.uva.es

M. Riani
Department of Economics and Management and Interdepartmental Centre of Robust Statistics, University of Parma E-mail: mriani@unipr.it

1 Introduction

Given a sample of observations $\{x_1, ..., x_n\}$ in $\mathbb{R}^p$, a widely used method in unsupervised learning is to assume multivariate normal components and to adopt a maximum likelihood approach for clustering purposes. With this idea in mind, well-known classification and mixture likelihood approaches can be followed.

In this work, we use $\phi(\cdot; \mu, \Sigma)$ to denote the probability density function of a p-variate normal distribution with mean μ and covariance matrix Σ.

In the *classification likelihood* approach we search for a partition $\{H_1, ..., H_k\}$ of the indices $\{1, \cdots, n\}$, centres $\mu_1, \cdots, \mu_k$ in $\mathbb{R}^p$, symmetric positive semidefinite $p \times p$ scatter matrices $\Sigma_1, \cdots, \Sigma_k$ and positive weights $\pi_1, \cdots, \pi_k$ with $\sum_{j=1}^{k} \pi_j = 1$, which maximize

$$\sum_{j=1}^{k} \sum_{i \in H_j} \log \left(\pi_j \phi(x_i; \mu_j, \Sigma_j) \right). \tag{1}$$

On the other hand, in the *mixture likelihood* approach, we seek the maximization of

$$\sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j \phi(x_i; \mu_j, \Sigma_j) \right), \tag{2}$$

with similar notation and conditions on the parameters as above. In this second approach, a partition into k groups can be also obtained, from the fitted mixture model, by assigning each observation to the cluster-component with the highest posterior probability.

Unfortunately, it is well-known that the maximization of "log-likelihoods" like (1) and (2) without constraints on the Σ_j matrices is a mathematically ill-posed problem (Kiefer and Wolfowitz 1956; Day 1969). To see this unboundedness issue, we can just take $\mu_1 = x_1$, $\pi_1 > 0$ and $|\Sigma_1| \to 0$ making (2) to diverge to infinity or (1) also to diverge with $H_1 = \{1\}$.

This lack of boundedness can be solved by just focusing on local maxima of the likelihood target functions. However, many local maxima are often found and it is difficult to know which are the most interesting ones. See McLachlan and Peel (2000) for a detailed discussion of this issue. In fact, non-interesting local maxima denoted as "spurious" solutions, which consist of a few, almost collinear, observations, are often detected by the Classification EM algorithm (CEM), traditionally applied when maximizing (1), and by the EM algorithm, traditionally applied when maximizing (2). A recent review of approaches for dealing with this lack of boundedness and for reducing the detection of spurious solutions can be found in García-Escudero et al. (2018).

Affine equivariance is an interesting property which is often required by a clustering method. The property means that the clustering results remain unchanged after applying affine transformations of the input variables. For instance, this is the case when changing variable measurement scales. n that direction, the work by Biernacki and Lourme (2014) nicely explains how scale changes can affect different model-based clustering approaches and their visualization. It is important to note that affine equivariance is not always the most convenient property in specific clustering applications, as Hennig and Liao (2013) show in a social stratification problem.

At first sight, the use of constraints on the relative sizes of the determinant of the Σ_j matrices may be seen as a simple and useful way to overcome these degeneracy issues and to apply affine equivariant constraints. These constraints were already suggested by McLachlan and Peel (2000) (Section 3.9.1) and lie behind the EVV (equal volume, variable shape and orientation) parametrization within the well-known Gaussian parsimonious models family (Celeux and Govaert 1992; Banfield and Raftery 1993). In Section 2, we will see how only constraints on the determinants do not fully avoid the detection of degenerate (spurious) solutions. We will also see how additional constraints on the elements of the "shape" matrices are useful to minimize these degeneracy issues. Moreover, different clustering approaches can be defined depending on the strength of these two, determinant and shape, types of constraints. Regarding affine equivariance, we are virtually affine equivariant if only very mild constraints on the shape elements are posed.

At a first sight, the doubly constrained likelihood maximization seems to be a difficult task. In Section 3, we will show that just a minor modification of the traditional CEM and EM algorithms is needed. In fact, that modification is based on the same "optimal truncation" procedure applied for eigenvalue ratio constraints described in Fritz et al. (2013). A justification of the proposed algorithm will be given in Section 3.3.

A robust extension of the methodology, which improves the performance if a certain fraction of contaminating observations appears in our data set, will be introduced in Section 4. Robustness follows from applying a trimming approach. A simulation study will be given in Section 5 and a real data application in Section 6. Some conclusions and further research lines are outlined in Section 7.

2 Determinant-and-shape constraints

It is known that if one or more $|\Sigma_j|$, but not all, tend to 0 then (1) and (2) can go to $+\infty$. This fact creates a lack of boundedness problem when maximizing these target functions. Therefore, we could consider the maximization of (1) and (2) but under

Determinant constraints: we force

$$\frac{\max_{j=1,\ldots,k} |\Sigma_j|}{\min_{j=1,\ldots,k} |\Sigma_j|} \leq c_1, \tag{3}$$

for a given fixed constant $c_1 \geq 1$.

Notice that (3) implies that if any of the determinants $|\Sigma_j|$ goes to 0 then all the other determinants also have to go to 0 and this solution is not interesting (because, in that case, the log-likelihoods (1) and (2) go to $-\infty$). It is also trivial to see that this type of constraints results in an affine equivariant clustering procedure.

The particular case $c_1 = 1$ forces all the determinants of the scatter matrices to be equal, i.e. $|\Sigma_1| = ... = |\Sigma_k|$. This case corresponds to the approach in McLachlan and Peel (2000) and to the EVV (equal volume, variable shape and orientation) parametrization within the Gaussian parsimonious family. When considering $1 < c_1 < \infty$, we relax the exact "equal determinant" assumption without leaving determinants completely free.

However, even when all the $|\Sigma_j|$ determinants are kept away from 0, degeneracy troubles still arise, because some eigenvalues of the Σ_j matrices may still go to 0. To illustrate this, let us consider the well-known decomposition for the Σ_j scatter matrices as

$$\Sigma_j = \lambda_j \Omega_j \Gamma_j \Omega_j',$$

where Ω_j is an orthogonal matrix of eigenvectors, Γ_j is a diagonal matrix with $|\Gamma_j| = 1$ and with elements $\{\gamma_{j1}, ..., \gamma_{jp}\}$ in its diagonal (proportional to the eigenvalues of the Σ_j matrix) and $|\Sigma_j| = \lambda_j^p$. These Γ_j matrices are commonly known as "shape" matrices, because they determine the "shape" of the fitted cluster components. To see that degeneracy issues may still happen, even with controlled determinant sizes, let us fix $p = 2$ and take $\mu_1 = x_1$, $\pi_1 > 0$, $\lambda_1 = 1$, $\gamma_{11} = C$ and $\gamma_{12} = 1/C$. The remaining Σ_j matrices, $j = 2, ..., k$, are arbitrarily chosen but satisfying $|\Sigma_2| = \cdots = |\Sigma_k| = 1$. Note that the smallest eigenvalue of Σ_1 converges to 0 when $C \uparrow \infty$ and, then, one of the fitted components can be made arbitrarily close to a degenerate normal component.

To overcome the above explained source for degeneracy, we may consider, besides (3), an additional type of constraint which controls the elements of the "shape" matrices as:

Shape constraints: consider the following k constraints:

$$\frac{\max_{l=1,...,p} \gamma_{jl}}{\min_{l=1,...,p} \gamma_{jl}} \leq c_2, \text{ for } j = 1, ..., k, \tag{4}$$

where $c_2 \geq 1$.

Notice that (4) imposes k independent sets of constraints, one for each shape matrix, and nothing relates the shape matrix elements of one component to the other components.

The combination of different c_1 and c_2 values, with $1 \leq c_1 < \infty$ and $1 \leq c_2 < \infty$, entails different clustering approaches throughout their associated constrained maximizations.

If we consider a very large c_2 value (e.g., $c_2 = 10^{10}$) and a small/moderate value for c_1 we are virtually affine equivariant. That choice would constitute just mild constraints on the scatter matrices "condition numbers" (ratios between the largest and smallest eigenvalues). As will be seen in the simulation study, this type of constraint is a kind of convenient "computational precision" protection especially when dimension increases.

2.1 Illustrative example

As a first simple illustrative example, we show the result of applying the proposed methodology in the classification likelihood case with $k = 2$, $c_1 = 1$ and $c_2 = 10^{10}$. The data set in Figure 1,(a) is drawn from two spherical bivariate normal components with the same scatter (and, consequently, $|\Sigma_1|/|\Sigma_2| = 1$). An affine transformation is applied to this data set which notably affects the measurement scales. The new transformed data set in Figure 1,(b) can be seen as randomly drawn from two non-spherical normal bivariate distributions where the Σ_1 and

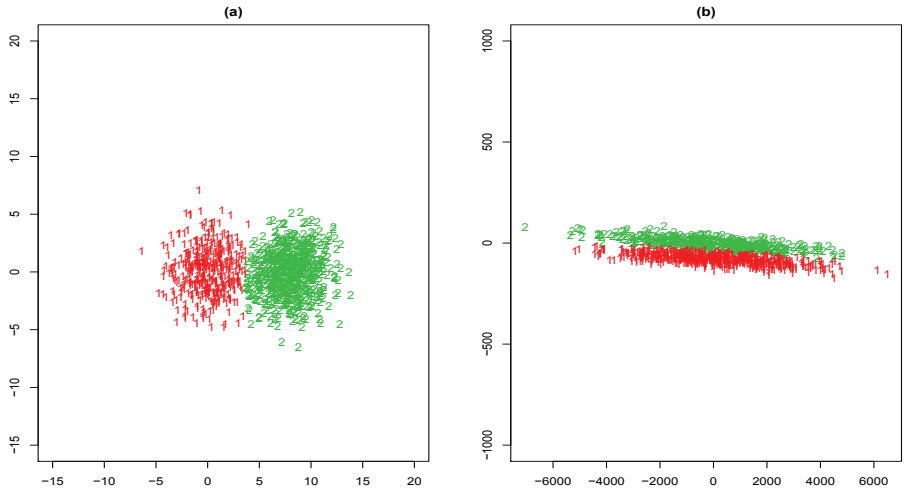

Fig. 1 Results of applying the proposed methodology with $c_1 = 1$ and $c_2 = 10^{10}$. (a) Original data set (b) The same data set after applying an affine transformation.

Σ_2 still satisfy $|\Sigma_1|/|\Sigma_2| = 1$. The result of applying the same determinant-and-shape constrained approach, with the same c_1 and c_2 values, is shown in Figure 1,(b).

Different constrained clustering problems can be defined depending on the c_1 and c_2 values. For instance, we are ideally searching for spherical clusters when $c_2 = 1$. In fact, intermediate models between the EII (equal volume and spherical) and the VII (variable volume and spherical) Gaussian parsimonious parameterizations are handled with $1 < c_1 < \infty$ and $c_2 = 1$. Notice that the target function for the VII model ($c_1 = \infty$) is unbounded (defining a mathematically ill-posed problem again). On the other hand, the $c_1 = c_2 = 1$ case often yields a very constrained parametrization, which assumes spherical and equally scattered components.

Figure 2 shows the result of the proposed approach, in the classification likelihood case, with $k = 2$ and $c_2 = 1$ when applied to a data set made of two spherical components but with clearly different scatters. Figure 2,(a) uses $c_1 = 1$ while $c_1 = 10^{10}$ is used in (b). The dashed lines show the two circles including 95% of the probability mass and the solid lines represent the estimated circles through the model-based clustering approach. We are able to estimate more correctly the true clusters, so obtaining better cluster assignments with the more flexible model fitted in Figure 2,(b).

Figure 3 shows a summary of the type of cluster structures we are ideally searching for different combinations of the c_1 and c_2 values. The ellipses show equidensity contours of the fitted normal components.

2.2 Related approaches

The use of "ratio-type" constraints as in (3) and (4) goes back to the seminal paper of Hathaway (1985). In this direction, the ratio between the largest and the

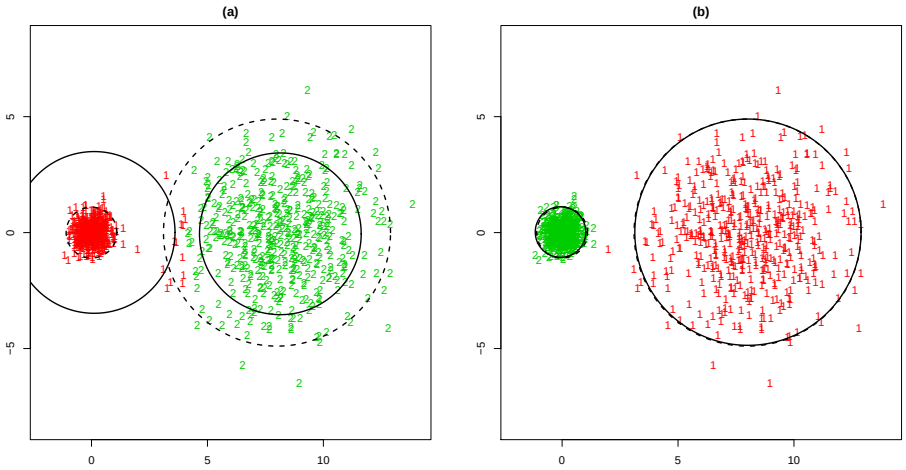

Fig. 2 A data set made of two circular components with different scatters (dashed lines show the contours including 95% of the probability mass). (a) Clustering results with $c_1 = c_2 = 1$ (b) Clustering results with $c_1 = 10^{10}$ and $c_2 = 1$. The solid lines represent the same estimated circles. Note that the solid lines in (b) completely overlap the dashed lines.

smallest of the $k \times p$ eigenvalues of the Σ_j matrices was forced to be smaller than a given fixed constant $c^* \geq 1$ (Ingrassia and Rocci 2007; García-Escudero et al. 2008, 2011, 2014b, 2015). This means that the maximization of (1) and (2) is done under the (more simple) constraint:

$$\max_{jl} \lambda_l(\Sigma_j) / \min_{jl} \lambda_l(\Sigma_j) \leq c^*, \tag{5}$$

where $\{\lambda_l(\Sigma_j)\}_{l=1}^p$ are the set of eigenvalues of the Σ_j matrix, $j = 1, ..., k$.

With this eigenvalue-ratio approach, we need a very high c^* value to be close to affine equivariance. Unfortunately, such a high c^* value does not always successfully prevent us from incurring into spurious solutions. Furthermore, the new determinant-and-shape constraints (based on $c_1 > 1$ and $c_2 = 1$) allow us to deal with spherical "heteroscedastic" cases, such as that in Figure 2 or in the top-right panel of Figure 3, whereas the eigenvalue ratio constraint with $c^* = 1$ can only handle the spherical "homoscedastic" case as in the top-left panel of Figure 3.

An interesting and closely related approach was given in Browne et al. (2013) where different constrained modifications of the Gaussian parsimonious models were proposed. These modifications include lower and upper bounds on the Σ_j matrices in such a way that all the $\lambda_l(\Sigma_j)$ eigenvalues, $j = 1, ..., k$ and $l = 1, ..., p$, are bounded to be within a fixed interval $[a, b]$. The eigenvalue truncation procedure in Ingrassia and Rocci (2007) is applied to impose that constraint in the associated CEM and EM algorithms. The constraints in Browne et al. (2013) certainly serve to prevent likelihood maximization degeneracies. However, the correct choice of the truncation interval $[a, b]$, because of the use of scatter matrices eigenvalues, strongly depends on the measurement scales of the input variables. Considering a separate treatment for the components "size" and "shape" parameters, together with working with "ratio-type" constraints, makes more flexible the specification

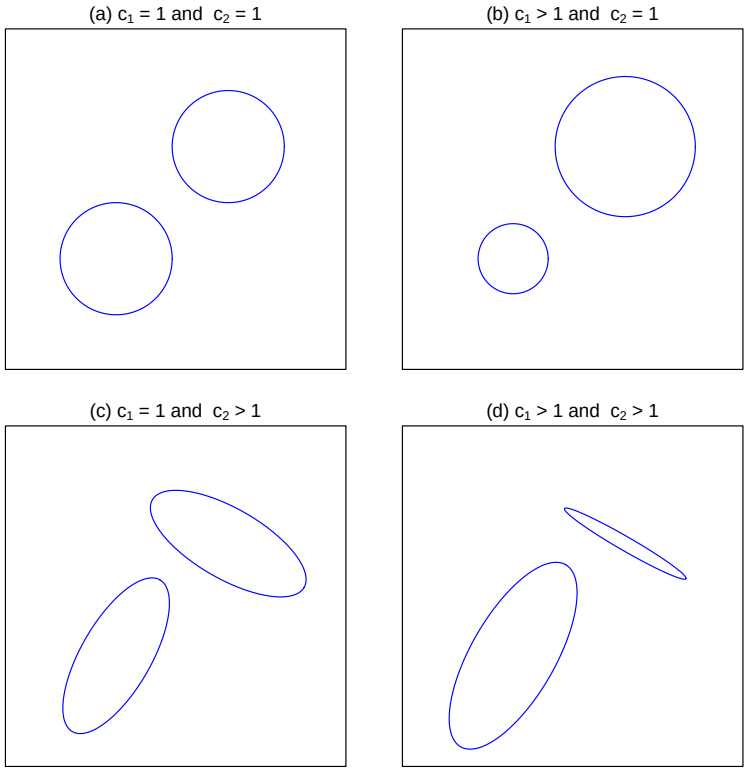

Fig. 3 Schematic plot showing the type of cluster components detected for different combinations of c_1 and c_2.

of the type of components that the user of the clustering procedure is particularly interested in.

An affine equivariant procedure for Gaussian mixture modeling has also been recently introduced in Rocci et al. (2018) by shrinking the Σ_j scatter matrices towards a pre-specified target matrix Ψ. Seo and Kim (2012) try to avoid singular and spurious solutions by taking out the k observations with the highest contributions to (2) through the k-deleted likelihood approach which also results in an affine equivariant procedure. These two procedures are clearly useful but they cannot provide the differentiated type of control on the "shapes" and "sizes" given by the proposed doubly-constrained methodology.

2.3 Theoretical results

Given an underlying theoretical probability model P, a population version of the doubly constrained likelihood maximizations can be defined. In this section, we present existence results for both, the theoretical and the sample, problems together with consistency results for the sample solutions towards the population values.

Given $\theta = (\pi_1, \cdots, \pi_k, \mu_1, \cdots, \mu_k, \Sigma_1, \cdots, \Sigma_k)$ (as defined in Section 1), let us introduce the functions

$$D_j(x;\theta) = \pi_j \varphi(x;\mu_j,\Sigma_j),$$

$$D(x;\theta) = \max\{D_1(x;\theta), ..., D_k(x;\theta)\},$$

and the sets

$$\Theta_{c^*} = \{\theta : \Sigma_1, ..., \Sigma_k \text{ satisfy constraint (5) for constant } c^*\}$$

and

$$\Theta_{c_1,c_2} = \{\theta : \Sigma_1, ..., \Sigma_k \text{ satisfy constraints (3) and (4) for constants } c_1 \text{ and } c_2\}.$$

The following Theorem 1 states the existence result under finite second order moment conditions.

Theorem 1 *If P is not concentrated at k points and $E_P\|\cdot\| < \infty$*

(a) then there exists some $\theta \in \Theta_{c_1,c_2}$ such that the maximum of

$$E_P\left[\log\left[\sum_{j=1}^{k} D_j(\cdot;\theta)\right]\right] \tag{6}$$

when θ is forced to be in Θ_{c_1,c_2} is achived.

(b) then there exists $\theta \in \Theta_{c_1,c_2}$ such that the maximum of

$$E_P\left[\sum_{j=1}^{k} z_j(\cdot;\theta)\log D_j(\cdot;\theta)\right], \tag{7}$$

with $z_j(x;\theta) = I\{x : D(x;\theta) = D_j(x;\theta)\}$, when θ is forced to be in Θ_{c_1,c_2} is achieved.

If we have $\{x_1, ..., x_n\}$ being the realization of a random sample of size n from distribution P, i.e. when replacing P by the empirical distribution $P_n = \sum_{i=1}^{n} \delta_{\{x_i\}}$, then we recover the sample maximizations (1) and (2) under constraints (3) and (4) exactly as proposed in Section 2. Therefore, Theorem 1 also guarantees the existence of the solution of these two empirical problems associated to the given sample $\{x_1, ..., x_n\}$.

We use the notation θ_0 for any constrained maximizer of the theoretical problem for the underlying distribution P and let $\theta_n = (\pi_1^n, \cdots, \pi_k^n, \mu_1^n, \cdots, \mu_k^n, \Sigma_1^n, \cdots, \Sigma_k^n)$ be the sequence of empirical solutions for the sequence of empirical sample distributions $\{P_n\}_{n=1}^{\infty}$ from P. The following result states consistency under similar assumptions as in Theorem 1 if the maximizer of the theoretical problem is assumed to be unique.

Theorem 2 *Let us assume that P is not concentrated at k points, that $E_P\|\cdot\| < \infty$ and that $\theta_0 \in \Theta_{c_1,c_2}$ is the unique constrained maximizer of (6), resp. (7), for P. If $\{\theta_n\}_{n=1}^{\infty}$ is a sequence of empirical maximizers of (1), resp. (2), when $\theta_n \in \Theta_{c_1,c_2}$ then $\theta_n \to \theta_0$ almost surely.*

The proofs of Theorem 1 and Theorem 2 follow the same lines as the proofs of Proposition 1 and Proposition 2 in García-Escudero et al. (2015) once we prove that, for every c_1 and c_2, there always exists c^* (depending on c_1 and c_2) such that $\Theta_{c_1,c_2} \subset \Theta_{c^*}$. The fact that $\Theta_{c_1,c_2} \subset \Theta_{c^*}$ allows us to control the scatter matrices in the proofs exactly as was done there.

Lemma 1 *For any pair of* $(c_1, c_2) \in [1, \infty) \times [1, \infty)$ *values, we have that* $\Theta_{c_1,c_2} \subset \Theta_{c^*}$ *when* $c^* = c_2^2 c_1^{1/p}$.

Proof: If $\theta \in \Theta_{c_1,c_2}$ then

$$\frac{\lambda_l(\Sigma_j)}{\lambda_{l'}(\Sigma_{j'})} \leq \frac{c_2|\Sigma_j|^{1/p}}{1/c_2|\Sigma_{j'}|^{1/p}} \leq c_2^2 c_1^{1/p},$$

for every $j, j' \in \{1, ..., k\}$ and $l, l' \in \{1, ..., p\}$.

To prove these inequalities, we take into account that

$$\frac{1}{c_2^p}|\Sigma_j| \leq (\lambda_l(\Sigma_j))^p \leq c_2^p|\Sigma_j|$$

if constraint (4) holds and that $|\Sigma_j|/|\Sigma_{j'}| \leq c_1$ if constraint (3) holds. □

Existence results in the population case can be proven only under the determinant constraints (3) if P is assumed absolutely continuous. However, the existence of the sample solution for a given sample $\{x_1, ..., x_n\}$ is not guaranteed, but it can be proven that sample solutions approximating the population one can be obtained when n tends to ∞.

3 An algorithm for determinants-and-shape constraints

3.1 "Optimal truncation"

In this subsection, we review the "optimal truncation" procedure introduced in Fritz et al. (2013) that was the keystone in the algorithms for implementing the eigenvalues ratio constraints in the robust model-based clustering approaches in García-Escudero et al. (2008) and García-Escudero et al. (2014a) and also for their (classical) non-robust counterparts in García-Escudero et al. (2011). This "optimal truncation" procedure is also going to play a very important role on how constraints (3) and (4) are imposed in the new doubly constrained algorithms.

Given a non-negative value d and a fixed constraining constant c, we consider its m-truncated value as

$$d^m = \begin{cases} d & \text{if } d \in [m, cm] \\ m & \text{if } d < m \\ cm & \text{if } d > cm \end{cases}.$$

Given $\{n_j\}_{j=1}^J \in \mathbb{N}^J$ and $\{d_{j1}, ..., d_{jL}\}_{j=1}^J \in [0, \infty)^{J \times L}$, let us define the "*optimal truncation*" operator

$$\mathsf{opt.trunc}_c\big(\{n_j\}_{j=1}^J; \{d_{j1}, ..., d_{jL}\}_{j=1}^J\big)$$

returning $\{d^*_{j1}, ..., d^*_{jL}\}_{j=1}^{J} \in [0,\infty)^{J\times L}$ with $d^*_{jl} = d_{jl}^{m_{\text{opt}}}$ for m_{opt} being the optimal threshold value obtained as

$$m_{\text{opt}} = \arg\min_{m} \sum_{j=1}^{J} n_j \sum_{l=1}^{L} \left(\log\left(d^m_{jl}\right) + \frac{d_{jl}}{d^m_{jl}} \right). \tag{8}$$

Note that if one of the input values is equal to 0 then the output is a sequence with $J \times L$ values all being equal to 0.

Obtaining the optimal threshold value only requires the maximization of a real valued function. Proposition 3.2 in Fritz et al. (2013) shows that m_{opt} can be efficiently obtained by doing only $2 \cdot J \cdot L + 1$ evaluations of the function in (8). The procedure can be fully vectorized, i.e., it can be obtained without loops.

3.2 Description of the algorithm

In this section, we modify the traditional CEM and EM algorithms to incorporate, in an optimal way, the determinant-and-shape constraints. The two algorithms follow similar steps and are presented in a unified fashion.

Let us denote the parameters at step s by $\theta^{(s)} = (\pi_1^{(s)}, \cdots, \pi_k^{(s)}, \mu_1^{(s)}, \cdots, \mu_k^{(s)}, \Sigma_1^{(s)}, \cdots, \Sigma_k^{(s)})$ and $D_j(x;\theta^{(s)}) = \pi_j^{(s)}\phi(x;\mu_j^{(s)},\Sigma_j^{(s)})$ for $j = 1, ..., k$.

1. *Initialization:* The procedure is initialized `nstart` times by randomly selecting different initial $\theta^0 = (\pi_1^0, \cdots, \pi_k^0, \mu_1^0, \cdots, \mu_k^0, \Sigma_1^0, \cdots, \Sigma_k^0)$ sets of parameters. A simple strategy for the initialization is to randomly select $k\times(p+1)$ observations and use them, after splitting them into k groups, to compute k initial μ_j^0 centers and k initial scatter matrices Σ_j^0. The procedure described in Step 2.2 has to be applied if the initial Σ_j^0 scatter matrices do not satisfy the required constraints.
2. *Iterative steps:* The following steps are executed until convergence (i.e., when $\|\text{vec}(\theta^{(s+1)}) - \text{vec}(\theta^{(s)})\| / \|\text{vec}(\theta^{(s)})\| < \varepsilon$ for a fixed ε) or until a fixed maximum number of iterations `iter.max` is reached.
   2.1. *Computing observation weights:* From $\theta^{(s)}$, observation weights $\tau_j(x_i;\theta^{(s)})$, $i = 1, ..., n$ and $j = 1, ..., k$, are computed as

$$\tau_j(x_i;\theta^{(s)}) = \begin{cases} 1 & \text{if } D_j(x_i;\theta^{(s)}) = \max\{D_1(x_i;\theta^{(s)}), ..., D_k(x_i;\theta^{(s)})\} \\ 0 & \text{if not} \end{cases},$$

   in the CEM algorithm. The H_j sets are defined by

$$H_j^{(s)} = \{i : \tau_j(x_i;\theta^{(s)}) = 1\}.$$

   On the other hand, for the EM algorithm, observation weights are computed as

$$\tau_j(x_i;\theta^{(s)}) = \frac{D_j(x_i;\theta^{(s)})}{\sum_{j=1}^{k} D_j(x_i;\theta^{(s)})}.$$

   2.2. *Updating parameters:* From these $\tau_j(x_i;\theta^{(s)})$ weights, we define

$$n_j^{(s+1)} = \sum_{i=1}^{n} \tau_j(x_i;\theta^{(s)}),$$

and the component weights are updated as

$$\pi_j^{(s+1)} = n_j^{(s+1)}/n.$$

Centres are updated as

$$\mu_j^{(s+1)} = \frac{1}{n_j^{(s+1)}} \sum_{i=1}^{n} \tau_j(x_i;\theta^{(s)}) x_i.$$

Updating the scatter estimates is not so easy.
We start from the weighted sample covariance matrices

$$S_j = \frac{1}{n_j^{(s+1)}} \sum_{i=1}^{n} \tau_j(x_i;\theta^{(s)})(x_i - \mu_j^{(s+1)})(x_i - \mu_j^{(s+1)})',$$

$j = 1, ..., k$, and their associated decompositions

$$S_j = d_j R_j D_j R_j', \tag{9}$$

where R_j is the orthogonal matrix of eigenvectors of S_j, $D_j = \text{diag}(d_{j1}, ..., d_{jp})$ with $|D_j| = 1$ and $d_j = |S_j|$.
We solve k "optimal truncation" problems with $J = 1$ and $L = p$

$$\mathsf{opt.trunc}_{c_2}\big(\{1\}; \{d_{j1}, ..., d_{jp}\}\big), \text{ for } j = 1, ..., k,$$

so obtaining k sets of values $\{d_{j1}^*, ..., d_{jp}^*\}_{j=1}^k$. We use them to get D_j^* matrices defined as $D_j^* = \text{diag}(d_{j1}^*, ..., d_{jp}^*)$.
We also compute

$$v_j = d_j \frac{\sum_{l=1}^{p} (d_{jl}^*)^{-1} d_{jl}}{p}$$

and obtain $\{d_j^*\}_{j=1,...,k}$ through the application of one additional optimal truncation problem with $J = k$ and $L = 1$ being the result of

$$\mathsf{opt.trunc}_{c_1^{1/p}}\big(\big\{n_j^{(s+1)}\big\}_{j=1}^k; \{v_1, ..., v_k\}\big).$$

We finally update $\Sigma_j^{(s+1)} = d_j^* R_j D_j^* R_j'$.
Note that the "optimal truncation" operator has been applied $k + 1$ times with different J and L values.

3. *Evaluate the target function:* After applying this iterative process, the likelihood target functions (1) or (2), depending on the adopted approach, are computed. The parameters yielding the highest value of this target function are returned as the algorithm's output.

Note that the proposed algorithm requires several random initializations. Due to the unboundedness of the target functions (1) and (2), considering "wisely-chosen" initializations has been proposed for CEM and EM algorithms in order to prevent them from being trapped into degenerate or spuriuous solutions. Considering several initializations, together with a careful monitoring of their evolutions in the iterative process, was proposed in Biernacki and Chretien (2003). Baudry and Celeux (2015) also analyzed the EM algorithm initialization problem and proposed useful initializing strategies. Instead, we have defined a mathematically

well-defined constrained maximization problem (see Section 2.3) and our plan is just to explore the constrained parametric space, as much as possible, so trying to truly maximize our target likelihood functions but under the required constraints. This exploration is done by considering several random initializations in the Step 1 of the proposed algorithm, and immediately applying the required constraints.

3.3 Justification of the algorithm

The proposed algorithm follows the main lines of classical CEM and EM algorithms. Given $\theta^{(s)}$ at stage s, optimal weights are obtained by using posterior probabilities in the EM version but converted into 0-1 in the CEM case. In the M-step, parameters $\theta^{(s+1)}$ are updated by performing a constrained maximization, on the π_j, μ_j and Σ_j parameters, of the "completed" (log-)likelihood

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_j(x_i;\theta^{(s)})\log\left(\pi_j\phi(x_i;\mu_j,\Sigma_j)\right). \tag{10}$$

The alternation of E- and M-steps monotonically increases the associated likelihood functions by ending in a local constrained maximum. Searching from several random starting parameters should serve to detect the global constrained maximum.

In the constrained maximization of (10), if $n_j = \sum_{i=1}^{n}\tau_j(x_i;\theta^{(s)})$ then it is easy to see that the best choice of π_j is $\pi_j^{(s+1)} = n_j/n$. It is also straightforward to see that the best choice for μ_j is given by

$$\mu_j^{(s+1)} = \frac{1}{n_j}\sum_{i=1}^{n}\tau_j(x_i;\theta^{(s)})x_i.$$

Standard arguments in Multivariate Analysis show that, after plugging those optimal $\pi_j^{(s+1)}$ and $\mu_j^{(s+1)}$ values into (10), the optimal unconstrained $\Sigma_j^{(s+1)}$ matrices are obtained through maximization over the Σ_j matrices of

$$-\sum_{j=1}^{k}\frac{n_j}{2}\left(\log|\Sigma_j| + \operatorname{trace}(\Sigma_j^{-1}S_j)\right), \tag{11}$$

where

$$S_j = \frac{1}{n_j}\sum_{i=1}^{n}\tau_j(x_i;\theta^{(s)})(x_i-\mu_j^{(s+1)})(x_i-\mu_j^{(s+1)})'. \tag{12}$$

With this idea in mind, we start from the decompositions of the S_j matrices as in (9) where R_j is again the orthogonal matrix of eigenvectors of S_j and D_j is a diagonal matrix with $|D_j| = 1$. This means that the diagonal elements of D_j, denoted as $\{d_{j1},...,d_{jp}\}$, have to satisfy $\prod_{l=1}^{p} d_{jl} = 1$ and that $|S_j| = d_j^p$. Analogously, let us consider similar decompositions of the Σ_j matrices as

$$\Sigma_j = \lambda_j\Omega_j\Gamma_j\Omega_j',$$

with exactly the same notation and properties as in Section 2.

As happens in the unconstrained case, we can easily see that the optimal Ω_j matrices are given by the R_j matrices. Note that rotation matrices do not affect the fulfillment of the required constraints. Consequently, the cyclic property of the "trace" operator allows us to reduce (11) to

$$-\sum_{j=1}^{k}\frac{n_j}{2}\left(\log\lambda_j^p+\lambda_j^{-1}d_j\sum_{l=1}^{p}\frac{d_{jl}}{\gamma_{jl}}\right). \tag{13}$$

The maximization of (13) over the λ_j values and the γ_{jl} values has to be done under the constraints:

$$\frac{\lambda_j^p}{\lambda_{j'}^p}\leq c_1,\frac{\gamma_{jl}}{\gamma_{jl'}}\leq c_2 \text{ and } \prod_{l=1}^{p}\gamma_{jl}=1 \text{ for every } j,j',l \text{ and } l'.$$

We are first going to see that the best γ_{jl} values can be optimally determined independently of the λ_j values. In order to see that, let us note that if $\{b_l^*\}_{l=1}^p$ minimizes

$$\sum_{l=1}^{p}\left(\log b_l+\frac{e_l}{b_l}\right)$$

on $\{b_l\}_{l=1}^p$ under the constraints $b_l/b_{l'}\leq c$ for $l\neq l'$, then the values

$$a_l^*=\frac{b_l^*}{\sqrt[p]{\prod_{l=1}^{p}b_l^*}},l=1,...,p,$$

serve to minimize

$$\sum_{l=1}^{p}\frac{e_l}{a_l},$$

on $\{a_l\}_{l=1}^p$ under the constraints $a_l/a_{l'}\leq c$, for $l\neq l'$, and the additional constraint $\prod_{l=1}^{p}a_l=1$.

Therefore, assuming that the λ_j values were fixed, we can trivially see that $\{\gamma_{jl}\}_{j=1,...,k}^{l=1,...,p}$ can be optimally determined as $\{d_{jl}^*\}_{j=1,...,k}^{l=1,...,p}$ where each set of $\{d_{jl}^*\}_{l=1,...,p}$ values is obtained by applying

$$\mathsf{opt.trunc}_{c_2}\big(\{1\};\{d_{j1},...,d_{jp}\}\big) \text{ for } j=1,...,k.$$

As this holds for any possible λ_j values, the optimal Γ_j matrix is then $R_j^*=\mathrm{diag}(d_{j1}^*,...,d_{jp}^*)$.

Once that the optimal Γ_j and Ω_j matrices are determined, we have to obtain the optimal λ_j values. If

$$v_j=d_j\frac{\sum_{l=1}^{p}(d_{jl}^*)^{-1}d_{jl}}{p},$$

then simple calculus shows that the function (13) can be rewritten as

$$p\times\left[-\sum_{j=1}^{k}\frac{n_j}{2}\left(\log\lambda_j+\frac{v_j}{\lambda_j}\right)\right].$$

The constraint in (3) imposes $d_j^p/d_{j'}^p \leq c_1$ or, analogously, $d_j/d_{j'} \leq c_1^{1/p}$. Therefore, we can apply again the "opt.trunc" transformation to obtain optimal $\{d_j^*\}_{j=1,...,p}$ values resulting from

$$\mathsf{opt.trunc}_{c_1^{1/p}}\big(\{n_j\}_{j=1}^k; \{v_1, ..., v_k\}\big).$$

After all these steps, the scatter matrices are optimally updated as

$$\Sigma_j^{(s+1)} = d_j^* R_j D_j^* R_j'.$$

Direct expressions for all the terms involved are ready available because an efficient method for carrying out the "opt.trunc" transformation is at hand (Fritz et al. 2013).

4 Extension to robust clustering

We propose extending this methodology by allowing for a fixed proportion α of trimmed observations. As in García-Escudero et al. (2008), we can search for optimal parameters and an optimal partition $\{H_0, H_1, ..., H_k\}$ of the indices $\{1, \cdots, n\}$ with $\#H_0 = \lceil n\alpha \rceil$ maximizing the *classification trimmed likelihood*

$$\sum_{j=1}^{k} \sum_{i \in H_j} \log\big(\pi_j \phi(x_i; \mu_j, \Sigma_j)\big). \tag{14}$$

Analogously, as proposed in Neykov et al. (2007) and García-Escudero et al. (2014a)), we can can also maximize the *trimmed mixture likelihood*

$$\sum_{i \in \{1,2,...,n\} \setminus H_0} \log\left(\sum_{j=1}^{k} \pi_j \phi(x_i; \mu_j, \Sigma_j)\right), \tag{15}$$

where $H_0 \subset \{1, \cdots, n\}$ with $\#H_0 = \lceil n\alpha \rceil$ again. A closely related approach is the use of the "improper" maximum likelihood (Coretto and Hennig 2016).

The same reasoning as in Section 2 shows that both target functions (14) and (15), without constraints, are unbounded. Therefore trimming alone is not able to entail robustness. Consequently, trimming has to be combined with appropriate scatter matrix constrains yielding a combined strategy for achieving robustness in model-based clustering.

Therefore, we introduce new robust model-based clustering approaches by maximizing the trimmed likelihood functions (14) and (15) but, again, under determinant-and-shape constraints given by (3) and (4).

Of course, the constrained maximization involved is a difficult problem but trimmed versions of the previous CEM and EM algorithms can be applied. The same arguments as in García-Escudero et al. (2008) and (Fritz et al. 2013) show that adding a "trimming step" to Step 2.1 in Section 3.2 is enough. That "trimming step" is often referred to as "concentration step" (Rousseeuw and Van Driessen 1999) in the high-breakdown point robustness literature.

To apply the "trimming step", let us compute

$$D(x_i; \theta^{(s)}) = \max\{D_1(x_i; \theta^{(s)}), ..., D_k(x_i; \theta^{(s)})\},$$

for the robustified CEM algorithm, and

$$D(x_i;\theta^{(s)}) = \sum_{j=1}^{k} D_j(x_i;\theta^{(s)})$$

for the robustified EM algorithm. We sort all these values as

$$D(x_{(1)};\theta^{(s)}) \leq ... \leq D(x_{(n)};\theta^{(s)}),$$

and set $\tau_j(x_i;\theta^{(s)}) = 0$, for every $j = 1, ..., k$, for all the i indexes such that

$$D(x_i;\theta^{(s)}) \leq D(x_{(\lceil n\alpha \rceil)};\theta^{(s)}).$$

It is straightforward to see that this is the optimal way of setting weights if a proportion $\lceil n\alpha \rceil$ of observations are allowed to be discarded (clearly those with the smallest contribution to the likelihood).

Figure 4 shows the result of applying the proposed trimming-based robustifed methodology, in the CEM case, with $k = 2$, $c_1 = 10^{10}$, $c_2 = 1$ and $\alpha = 0.1$. The $\alpha = 0.1$ trimming level allows us to discard a 10% proportion of uniformly distributed background noise that was added to a data set similar to that in Figure 1. Figure 4 shows that the clustering results are the same before and after applying the affine transformation.

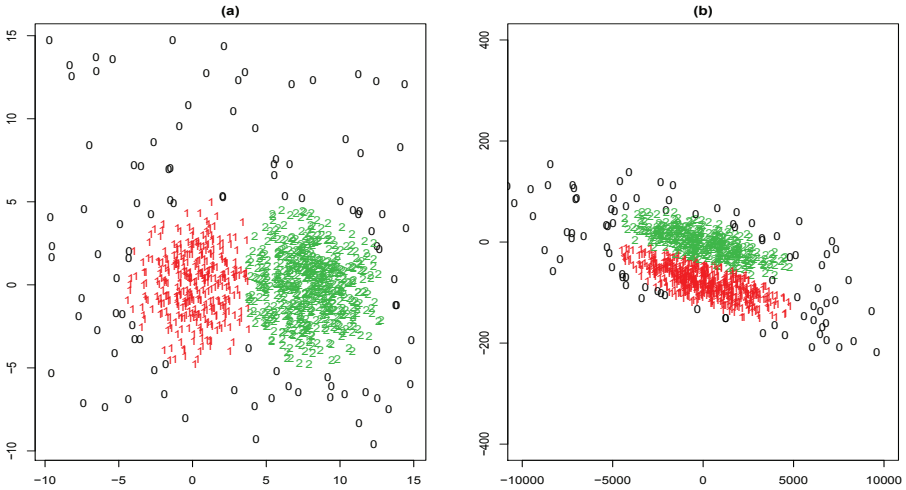

Fig. 4 Results of applying the proposed robustification with $\alpha = 0.1$, $c_1 = 10^{10}$ and $c_2 = 1$. (a) Original data set including 10% of background noise (b) The same data set after applying an affine transformation. Trimmed observations are represented by using "○" symbols in (a) and (b).

Existence and consistency results, similar to those in Theorem 1 and Theorem 2, can be proven for the trimmed setup just by replacing the second moment order condition $E_P\|\cdot\| < \infty$ by considering a strictly positive trimming level $\alpha > 0$. The consistency result also needs P to be absolutely continuous in the boundary of the

set including the nontrimmed mass of P. Only the proofs in García-Escudero et al. (2008) and García-Escudero et al. (2014b) have to be mimicked after taking into account Lemma 1. Avoiding moment conditions is interesting from the robustness point of view because, in this way, the existence and consistency results can be extended to the case of heavy-tailed underlying P distributions.

Moreover, the fact that $\Theta_{c_1,c_2} \subset \Theta_{c^*}$ for $c^* = c_2^2 c_1^{1/p}$ also allows direct extension to the Breakdown Point results in Section 3 of Dotto et al. (2018) to this new doubly constrained framework.

Theorem 3 *A breakdown point result similar to that in Theorem 4 in Dotto et al. (2018) also holds for determinant-and-shape constraints.*

The proof of this result easily follows from the same argument in the proof of Theorem 4 in Dotto et al. (2018).

An affine equivariant procedure in robust clustering was introduced in Gallegos and Ritter (2005) by assuming $\Sigma_1 = ... = \Sigma_k$ in the so-called "trimmed determinant criterion", robustifiying Friedman and Rubin (1967)'s procedure. Another attempt to achieve robustness in clustering throughout modified "concentration steps" was provided in Gallegos (2002). The components' sample covariance matrices S_j are transformed to have the same unitary determinant (i.e. replacing S_j by $S_j/|S_j|^{1/p}$) in the concentration step. This strategy is closely related to that of imposing constraint (3) with $c_1 = 1$. This type of common matrix standardization was previously suggested in Maronna and Jacovkis (1974) in the untrimmed case.

Another suggestion for robust affine equivariant clustering was introduced in Gallegos and Ritter (2009) (see also Ritter (2014)). The approach is simply based on running trimmed versions of the CEM and EM algorithms from many different random initializations without constraints. Afterward, all the many local minima detected are later inspected by promoting a trade-off between "good fit" (i.e., high values of (14) or (15)) and "scale balance". The "scale balance" is measured through the so-called HDBT ratio defined as the largest c such that $c\Sigma_j \preceq \Sigma_l$ for every $l \neq j$, where "$\preceq$" stands for the Löwner ordering for matrices. The HDBT ratio remains unchanged under affine transformation but a "constructive" algorithm, for a fixed c value, is not yet available. It is also interesting to note that, in this case, the $c = 1$ case does not reduce to the homoscedastic spherical case.

5 Simulation study

To illustrate the performance of the new type of constraints, we have conducted a simulation study. Simulated data sets are obtained starting from a "basic" two cluster structure. Data sets are made of n observations generated as a mixture of two p-variate normal components and with an average overlap of 0.01 and a maximum eigenvalue ratio for the scatter matrices of 1.1. This means that, initially, we are considering almost spherical clusters. The generation of these data sets is done through the `MixSim` method of Maitra and Melnykov (2010), as extended by Riani et al. (2015) and incorporated into the `FSDA` Matlab toolbox (Riani et al. 2012). The overlap is defined as a sum of pairwise misclassification probabilities. See more details in Riani et al. (2015).

We compare the performance of the proposed approach with respect to the more simple use of eigenvalues ratio constraints in (5). Figure 5 shows the Adjusted Rand Index (ARI) (Hubert and Arabie 1985) when applying "eigenvalues" and determinant-and-shape constraints and the classification likelihood approach in the same 100 randomly generated data sets from the basic (almost spherical) model with $n = 50$ and $p = 2$ denoted as $X_{50,2}$. The boxplots in the left panel of Figure 5 show the ARI values for different values of c^* when using eigenvalue ratio constraints. On the other hand, the panel on the right of this figure shows the same ARI boxplots but when applying the determinant-and-shape constraints when $c_1 = 1$ is kept fixed and c_2 varies. A large number of random intializations has been considered for each simulated data set. For a better comparison, the same number and exactly the same set of initializing values are considered for both types of constraints. To speed up the calculations the techniques of parallel processing have been adopted.

We observe very good performance, i.e. ARI values close to 1, when $c^* \simeq 1$ due to the added stability provided by the constraints. However, because of the small number of observations ($n = 50$), it also happens that sometimes the procedure ends up detecting almost degenerate "spurious" solutions for large c^* values. We can see that determinant-and-shape constraints uniformly provide more accurate cluster partitions regardless of the c_2 value when $c_1 = 1$.

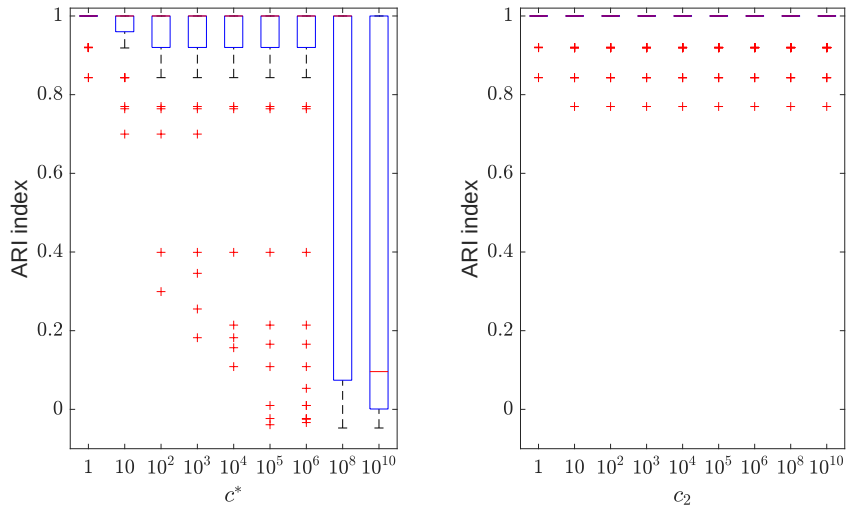

Fig. 5 $n = 50$, $p = 2$ *prior to the affine transformation:* The boxplots in the left panel shows ARI values when using eigenvalues ratio constraints for different values of the constraining constant c^*. The panel on the right shows the same ARI boxplots but when applying the determinant-and-shape constraints when $c_1 = 1$ and c_2 varies.

Starting from these basic data sets, we apply affine random transformation to them. To be more precise, all basic data matrices $X_{50,2}$ ($n = 50$ and $p = 2$) are

post-multiplied by matrices

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix},$$

where S is randomly generated from the uniform distribution $U(1, 101)$. This transformation serves to stretch the data set. Afterwards, we also consider a random rotation by post-multiplying $X_{50,2} \cdot A_1$ by the matrix

$$A_2 = \begin{pmatrix} \cos(\Theta) & -\sin(\Theta) \\ \sin(\Theta) & \cos(\Theta) \end{pmatrix},$$

where Θ is distributed as $U(0, 2\pi)$. The left panel of Figure 6 shows the results of applying eigenvalue ratio constraints for different c^* values for 100 different randomly generated $X_{50,2} \cdot A_1 \cdot A_2$ data sets. The plot on the right shows the results of the application of determinant-and-shape constraints when $c_1 = 1$ and c_2 varies. Again, the same random initializations are considered for both type of constraints.

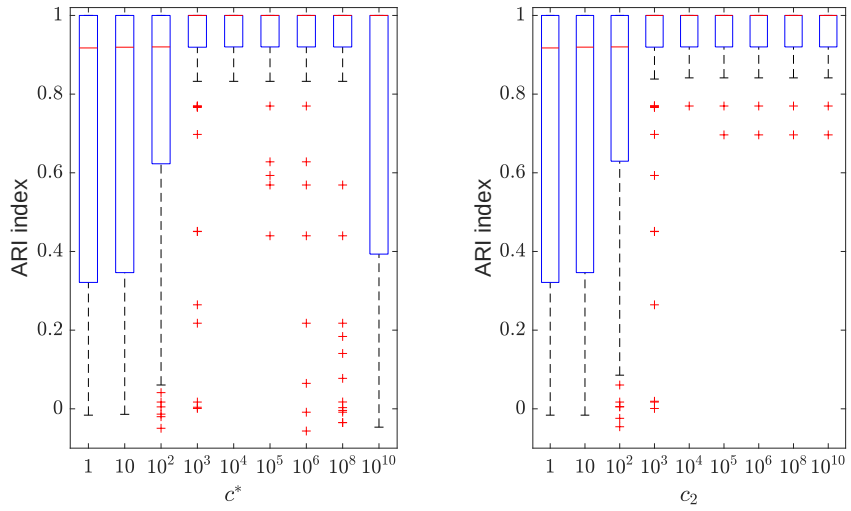

Fig. 6 $n = 50$, $p = 2$ *after affine transformation:* Left panel shows ARI boxplots when using eigenvalues ratio constraints for different values of the constraining constant c^*. Right panel shows the same ARI boxplots but when applying the determinant-and-shape constraints when $c_1 = 1$ and c_2 varies.

As expected, small c^* and c_2 values do not cope with the more elongated data structure and they are logically problematic with not very high ARI values. However, better ARI values are obtained in both cases when c^* and c_2 are greater than 100 (recall how the random A_1 matrices are obtained). However, it is important to note that, when applying eigenvalue ratio constraints, the large c^* values needed in order to deal with these more elongated structures also increase the chances that the algorithm gets trapped in spurious solutions. We can see several small ARI values in the left panel of Figure 6 when considering large c^* values for the

eigenvalue ratio constraint. These so small ARI values are not seen when considering the proposed determinant-and-shape constraints with $c_1 = 1$ and very large c_2 values.

To reinforce previous claims, the left panels of Figure 7 show some spurious components detected in the simulation study, "before and after" rotation, when using the eigenvalues-ratio constraint when $c^* = 10^6$. For the same data sets, the right panels of Figure 7 show how the detection of spurious components is avoided when we use the proposed methodology with $c_1 = 1$ and $c_2 = 10^6$.

It can be argued that increasing the sample size n reduces the chances of the EM algorithm becoming trapped in local spurious maxima. This is true, but the wrong detection of spurious solutions also arises, even with very high sample sizes, when the dimension p increases.

Figure 8 repeats the same study, but now with a higher sample size $n = 1000$ and a higher dimension $p = 10$ (almost spherical clusters and a controlled 0.01 average overlap). We can see that the ARI values are close to 1 for most of the 100 generated $X_{1000,10}$ data sets. However, a few spurious solutions can still be wrongly found with the eigenvalues constraints while fewer spurious solutions are found with the determinant-and-shape constraints.

We also consider the same affine transformation determined by matrices A_1 and A_2 on the first two dimensions of the basic data sets $X_{1000,10}$ while the remaining 8 coordinates are left unchanged. Figure 9 shows the results of the comparative study where no great differences can be noticed but still determinant-and-shape constraints seem to provide a slightly higher protection against spurious solutions.

Although not reported here, similar results have been obtained with other average overlaps in the two components.

6 Real data set example

The well-known "Swiss Bank Notes" data set in Flury and Riedwyl (1988) includes $p = 6$ measurements on 100 supposedly genuine and 100 counterfeit old Swiss 1000-franc bank notes. These measures quantify the size of the notes and the relative position of certain features within them. The data set is available, for instance, in the `mclust` and `tclust` contributed `R` packages.

We apply the eigenvalue and determinant-and-shape constraints with $k = 2$ trying to recover the two cluster structure (genuine and forged notes). However, the existence of 15 anomalous notes within the set of forged ones is also well-known (Flury and Riedwyl 1988) perhaps due to the presence of more than one forger. These are the notes numbered 111, 116, 138, 148, 160, 161, 162, 167, 168, 171, 180, 182, 187, 192 and 194 (in the `mclust` and `tclust` versions of this data set). Moreover, it is also known that note 70 is anomalous; it is assumed genuine but its measurements suggest a possible wrong assignment. Therefore $\alpha = 0.08$ can be considered as a sensible choice for the trimming level.

Figure 10,(a) shows the ARI results when applying eigenvalue and determinant-and-shape constraints for different values of the constraining parameters. The ARI values are obtained as if the set of these 16 notes were considered as a further different group (so having "genuine", "forged" and "anomalous" groups). The trimmed versions of the eigenvalue and determinant-and-shape constraints with $k = 2$ and

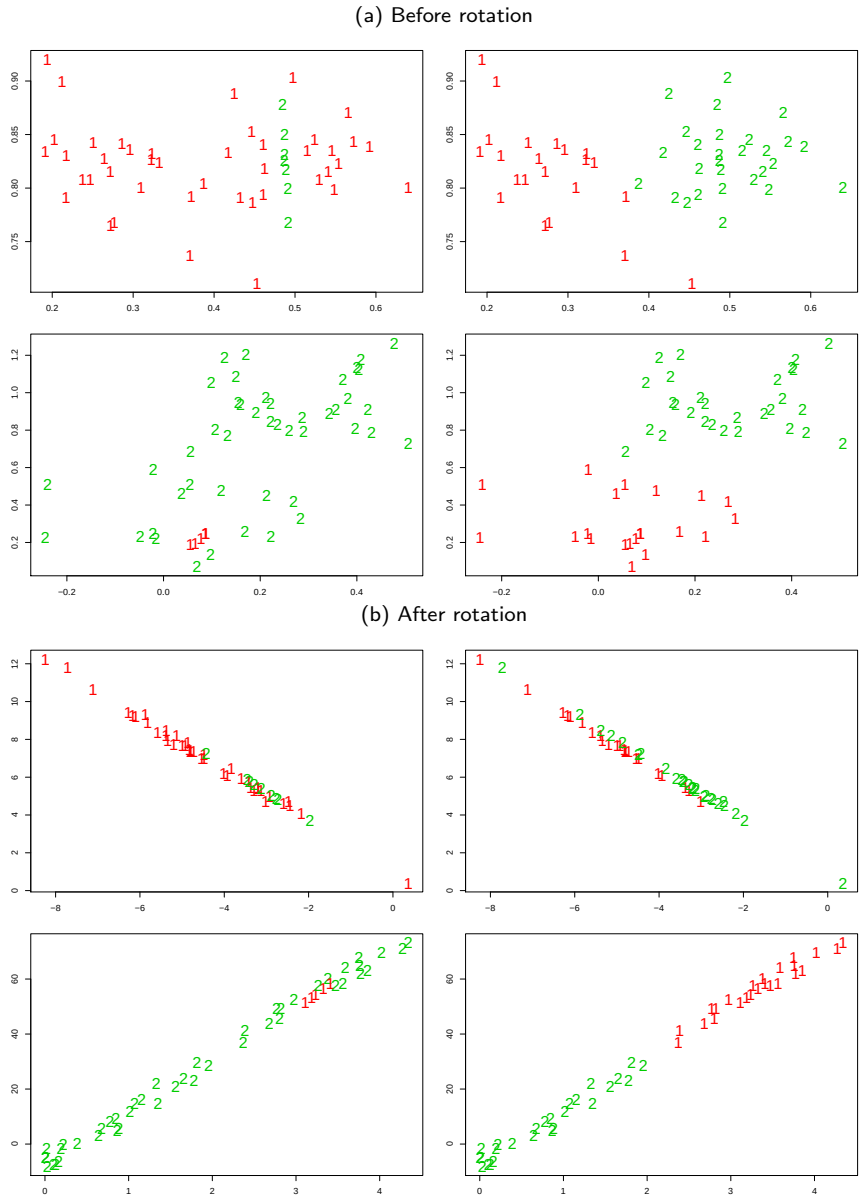

Fig. 7 Scatter plots for 4 different data sets included in the simulation study ("before and after" rotation) when $n = 50$ and $p = 2$. The results of applying the eigenvalue-ratio constraint with $c^* = 10^6$ are shown in the left panels and those resulting from the application of the determinant-and-shape constraint with $c_1 = 1$ and $c_2 = 10^6$ in the right panels.

$\alpha = 0.08$ are applied with a three "groups" partition (trimmed observations considered as a third group). The eigenvalue constraints are applied for different values of c^* and determinant-and-shape constraints with $c_1 = 1$ and different c_2 values.

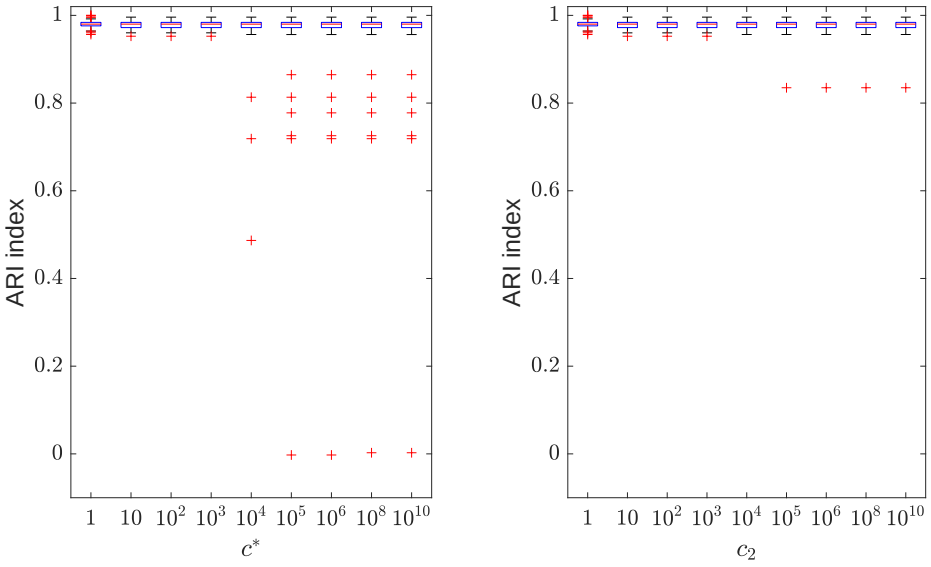

Fig. 8 $n = 1000$, $p = 10$ *prior to the affine transformation:* Left panel shows ARI boxplots when using eigenvalue ratio constraints for different values of the constraining constant c^*. Right panel shows the same ARI boxplots but when applying the determinant-and-shape constraints when $c_1 = 1$ and c_2 varies.

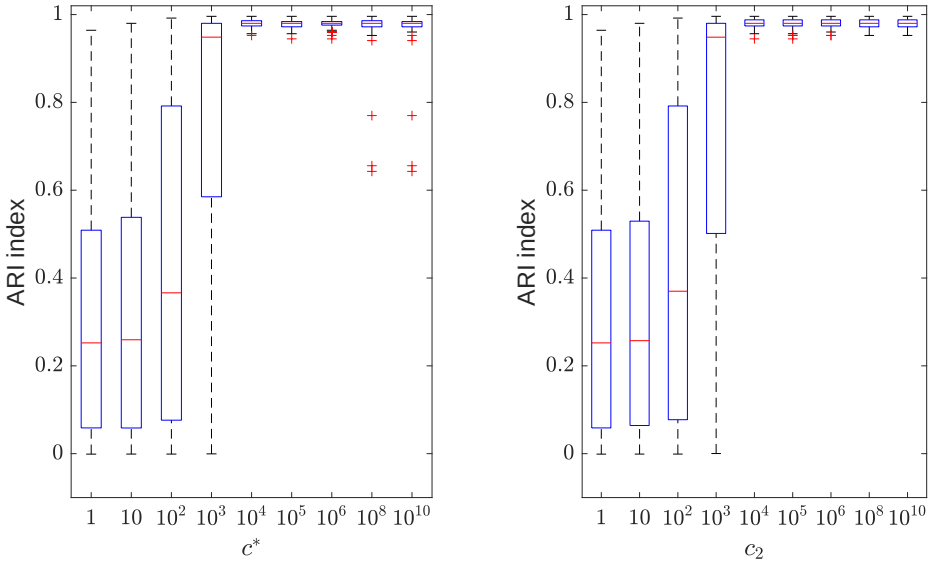

Fig. 9 $n = 1000$, $p = 10$ *after affine transformation:* The boxplots in the left panel show ARI values when using eigenvalue ratio constraints for different values of the constraining constant c^*. The panel on the right shows the same ARI boxplots but when applying the determinant-and-shape constraints when $c_1 = 1$ and c_2 varies.

We can see in Figure 10,(a) that the results are similar, with ARI values close to 1 in both cases, regardless of the type of constrains applied. Both approaches

recover the two main groups (genuine and forged bills) and detect the third set of notes.

In order to illustrate the possible troubles arising due to the lack of affine equivariance of the eigenvalues constraints, the same ARI monitoring procedure is applied to this data set but with the values of the fourth variable (distance of inner frame to the lower border) multiplied by a factor of 10000. Figure 10,(b) shows, how this change of scale in the fourth variable, is problematic for the eigenvalue constraints; high ARI values are found only when a very large c^* value is chosen. Note that the use of very large c^* values can also result in the detection of spurious clusters.

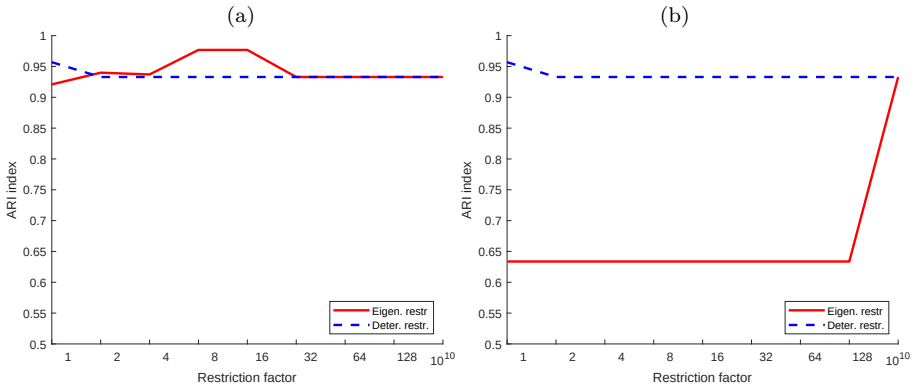

Fig. 10 *Swiss bank notes data set:* ARI values when varying the constraining constants for eigenvalue (solid line) and determinant-and-shape constraints (dashed line) with $k = 2$ and $\alpha = 0.08$. 16 outlying bills are considered as the "true" outliers. The original data set is considered in (a) and a modified data set is considered in (b) where the fourth variable is multiplied by 10000.

Figure 11 illustrates the results obtained by imposing the eigenvalue constraints in the modified data set when the fourth variable was again multiplied by 10000.

If we apply a robust procedure to the set of supposedly genuine notes, we find that notes 1 and 40 can be also declared as outliers. Figure 12 shows a pairs-plot of the "genuine" notes where notes numbered 1, 40 and 70 are denoted by symbols "A", "B" and "C", respectively. Similarly to "C", we can see that "A" and "B" may also be considered as outlying ones in some coordinates and, thus, the number of outlying observations could be increased to 18.

Therefore, we repeat the process of monitoring the ARI when moving the constraining constants for both types, the eigenvalue and determinant-and-shape, of constraints when $k = 2$ and $\alpha = 0.09$. The results are shown in Figure 13. We see that the results are very stable for the determinant-and-shape constraints regardless of the c_2 value and this stability remains even after transforming one of the coordinates in panel (b).

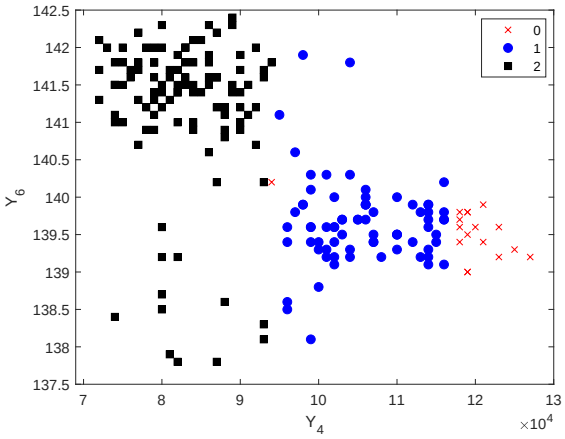

Fig. 11 Scatter plot of the fourth against the sixth variable and clustering results for eigenvalue constraints with $c^* = \{1, 2, ..., 128\}$, $k = 2$ and $\alpha = 0.08$ for the "modified" data set. Notice the measurement scale for the fourth variable.

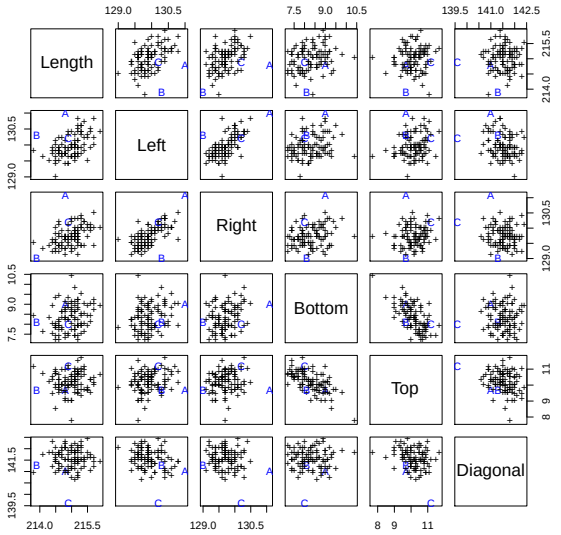

Fig. 12 Pairs plot for the "genuine" notes with notes numbered 1, 40 and 70 denoted by symbols "A", "B" and "C", respectively.

7 Conclusions and further directions

A doubly constrained approach has been presented for model-based clustering. Apart from the application of determinant constraints some further constraints on the shape matrix elements are also considered. Two constraining constants, c_1 and c_2, serve to control their strength. If the constant controlling the shape

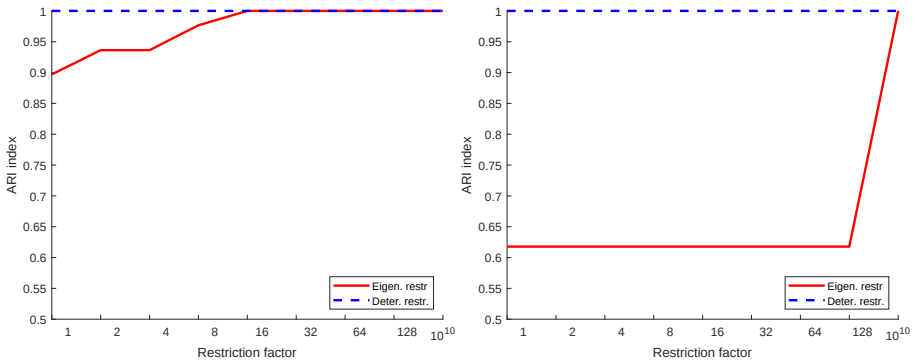

Fig. 13 *Swiss bank notes data set:* ARI values when varying the constraining constants for eigenvalues (solid line) and determinant-and-shape constraints (dashed line) with $k = 2$ and $\alpha = 0.09$. A set with 18 outlying notes is considered as "true" outliers. The original data set in (a) and the transformed set in (b).

matrix elements is chosen very large (for instance $c_2 = 10^{10}$) then the proposed methodology is almost affine equivariant, but certain protection against numerical issues is still provided with finite c_2 values. Existence and consistency results are presented for finite c_1 and c_2 values. We find almost spherical clusters, but with different scatters, when c_2 tends to 1. An extension to robust model-based clustering, adopting this double constraints, is introduced which incorporates a trimming step.

The lack of boundedness of the likelihood target function and the wrong detection of spurious local maxima are many times overridden by choosing appropriate initializing strategies for the CEM and EM algorithms (see, e.g., Biernacki and Lourme 2014, and the references therein). This idea provides good results in many cases. However, it is important to note that we are not maximizing any explicit target function anymore. Moreover, its practical performance becomes clearly dependent on the initialization procedure. For instance, we could face difficulties if the initializing procedure is not robust and the same could happen with the equivariance. On the other hand, the methodology presented in this work provides a "constructive" way of performing a mathematically well-defined constrained likelihood maximization with good robustness properties and a feasible and fast algorithm is available for its implementation. The routine which imposes the determinant- and shape constraints described in this paper has been implemented in function `restrdeter` of the FSDA toolbox (Riani et al. 2012) which is freely downloadable from the Mathworks file exchange platform.

The proposed methodology is very flexible because it can cope with very different types of data thanks to the flexibility that the tuning parameters entail. Although often the choices of the tuning parameters are motivated by the final clustering purposes, certain guidance about how to choose all of them is sometimes required. In that direction, for instance, the adaptation of modified BIC and ICL approaches to highlight sensible c_1 and c_2 values in the spirit of Cerioli et al. (2018) may be an interesting research line. In this way, some preliminary studies seem to show that sensible c_1 and c_2 values can be chosen as $\widehat{c}_1$ and $\widehat{c}_2$ resulting

from the minimization:

$$(\widehat{c_1}, \widehat{c_2}) = \arg\min_{c_1,c_2} \left\{-2CL_k^{c_1,c_2} + v_k^{c_1,c_2} \log n\right\}, \tag{16}$$

where $CL_k^{c_1,c_2}$ is the maximum value achieved in the constrained maximization of (1) for fixed c_1 and c_2 values, and where $v_k^{c_1,c_2}$ is a penalty term defined as:

$$v_k^{c_1,c_2} = kp + k - 1 + \underbrace{k\frac{p(p-1)}{2}}_{\text{rotation par.}} + \underbrace{(k-1)\left(1 - \frac{1}{c_1^{1/p}}\right) + 1}_{\text{determinant par.}} + \underbrace{k(p-1)\left(1 - \frac{1}{c_2}\right)}_{\text{shape par.}}.$$

Notice that larger values of c_1 or c_2 values yield more unrestricted Σ_j scatter matrices, such that more complex models are allowed to be fitted. The added penalty term serves to penalize a model complexity higher than needed. The previous proposal for $v_k^{c_1,c_2}$ takes into account the number of free parameters in the Ω_j matrices ("rotation pars."). By following an analogous philosophy as in Cerioli et al. (2018), the number of free parameters for the λ_j values ("determinant pars.") and for the elements of Γ_j ("shape pars.") are also recovered as limit cases corresponding to situations when c_1 or c_2 take values equal to 1 or go to ∞. A more deeply investigation of this proposal is an ongoing research work. If $L_k^{c_1,c_2}$ denotes the maximum value achieved in the constrained maximization of (2) then $L_k^{c_1,c_2}$ can be used instead of $CL_k^{c_1,c_2}$ in (16) when considering a mixture likelihood approach instead of the classification likelihood one.

It is also interesting to monitor the solutions obtained when moving c_1 and c_2 in a controlled manner. The solutions are often very stable to the different choices of c_1 and c_2 and not too many "essentially different solutions" are found once that pathological solutions associated to very large c_1 and c_2 values are excluded. In general, our proposal is just to consider a moderate c_1 value, and to set a large c_2 value if equivariance is a desired property or to set c_2 close to 1 if detecting almost spherical components is required. Examining the stability of the solutions, when moving c_1 and c_2, may be also useful to choose parameters as done in Riani et al. (2019). Additionally, graphical tools, as those in García-Escudero et al. (2011), can be also considered for determining k and α. All these proposals deserve further investigations.

Finally, the proposed methodology is not limited to normal mixtures, and other mixtures where a scatter matrix Σ is included in the definition of the k component-specific multivariate distributions, may benefit from its use. Some examples are mixtures of t distributions (Peel and McLachlan 2000; Andrews et al. 2018), contaminated normal distributions (Punzo and McNicholas 2016; Punzo et al. 2018), power exponential distributions (Zhang and Liang 2010; Dang et al. 2015), and leptokurtic-normal distributions (Bagnato et al. 2017).

References

Andrews J, Wickins J, Boers N, McNicholas P (2018) teigen: An R package for model-based clustering and classification via the multivariate t distribution. J Stat Softw 83:1–32

Bagnato L, Punzo A, Zoia MG (2017) The multivariate leptokurtic-normal distribution and its application in model-based clustering. Can J Stat 45:95–119

Banfield JD, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49:803–821
Baudry JP, Celeux G (2015) EM for mixtures - Initialization requires special care. Stat Comput 25:713–726
Biernacki C, Chretien S (2003) Degeneracy in the maximum likelihood estimation of univariate. Statistics & Probability Letters 61:373–382
Biernacki C, Lourme A (2014) Stable and visualizable Gaussian parsimonious clustering models. Stat Comput 24:953–969
Browne R, Subedi S, McNicholas P (2013) Constrained optimization for a subset of the Gaussian parsimonious clustering models, preprint available at https://arxiv.org/pdf/1306.5824.pdf
Celeux G, Govaert A (1992) A classification EM algorithm for clustering and two stochastic versions. Comput Stat Data Anal 14:315–332
Cerioli A, García-Escudero L, Mayo-Iscar A, Riani M (2018) Finding the number of normal groups in model-based clustering via constrained likelihoods. J Comput Graph Stat 27:404–416
Coretto P, Hennig C (2016) Robust improper maximum likelihood: Tuning, computation, and a comparison with other methods for robust Gaussian clustering. J Am Stat Assoc 111:1648–1659
Dang U, Browne R, McNicholas PD (2015) Mixtures of multivariate power exponential distributions. Biometrics 71:1081–1089
Day N (1969) Estimating the components of a mixture of normal distributions. Biometrika 56:463–474
Dotto F, Farcomeni A, García-Escudero L, Mayo-Iscar A (2018) A reweighting approach to robust clustering. Stat Comput 28:477–493
Flury B, Riedwyl H (1988) Multivariate Statistics, A Practical Approach. Cambridge University Press, Cambridge
Friedman H, Rubin J (1967) On some invariant criteria for grouping data. J Am Stat Assoc 63:1159–1178
Fritz H, García-Escudero L, Mayo-Iscar A (2013) A fast algorithm for robust constrained clustering. Comput Stat Data Anal 61:124–136
Gallegos M, Ritter G (2005) A robust method for cluster analysis. Ann Stat 33:347–380
Gallegos M, Ritter G (2009) Trimming algorithms for clustering contaminated grouped data and their robustness. Adv Data Anal Classif 10:135–167
Gallegos MT (2002) Maximum likelihood clustering with outliers. In: Jajuga K, Sokolowski A, Bock H (eds) Classification, Clustering and Data Analysis: Recent advances and applications, Springer-Verlag, pp 247–255
García-Escudero L, Gordaliza A, Matrán C, Mayo-Iscar A (2008) A general trimming approach to robust cluster analysis. Ann Stat 36:1324–1345
García-Escudero L, Gordaliza A, Matrán C, Mayo-Iscar A (2011) Exploring the number of groups in robust model-based clustering. Stat Comput 21:585–599
García-Escudero L, Gordaliza A, , Mayo-Iscar A (2014a) A review of robust clustering methods. Adv Data Anal Classif 8:27–43
García-Escudero L, Gordaliza A, Mayo-Iscar A (2014b) A constrained robust proposal for mixture modeling avoiding spurious solutions. Adv Data Anal Classif 8:27–43
García-Escudero L, Gordaliza A, Matrán C, Mayo-Iscar A (2015) Avoiding spurious local maximizers in mixture modeling. Stat Comput 25:619–633
García-Escudero L, Gordaliza A, Greselin F, Ingrassia S, Mayo-Iscar A (2018) Eigenvalues and constraints in mixture modeling: geometric and computational issues. Adv Data Anal Classif 12:203–233
Hathaway R (1985) A constrained formulation of maximum likelihood estimation for normal mixture distributions. Ann Stat 13:795–800
Hennig C, Liao TF (2013) How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. J R Stat Soc Ser C 62:309–369
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218
Ingrassia S, Rocci R (2007) Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. Comput Stat Data Anal 51:5339–5351
Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Annals of Mathematical Statistics 27:887–906

Maitra R, Melnykov V (2010) Simulating data to study performance of finite mixture modeling and clustering algorithms. J Comput Graph Stat 19:354– 376

Maronna R, Jacovkis P (1974) Multivariate clustering procedures with variable metrics. Biometrics 30:499–505

McLachlan G, Peel D (2000) Finite mixture models. Wiley Series in Probability and Statistics, New York

Neykov N, Filzmoser P, Dimova R, Neytchev P (2007) Robust fitting of mixtures using the trimmed likelihood estimator. Comput Stat Data Anal 52:299–308

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

Punzo A, McNicholas PD (2016) Parsimonious mixtures of multivariate contaminated normal distributions. Biometrical Journal 58:1506–1537

Punzo A, Mazza A, McNicholas PD (2018) Contaminatedmixt: An R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. J Stat Softw 85:1–25

Riani M, Perrotta D, Torti F (2012) FSDA: a Matlab toolbox for robust analysis and interactive data exploration. Chemometr Intell Lab Syst 116:17–32

Riani M, Cerioli A, Perrotta D, Torti F (2015) Simulating mixtures of multivariate data with fixed cluster overlap in FSDA library. Adv Data Anal Classif 9:461–481

Riani M, Atkinson A, Cerioli A, Corbellini A (2019) Efficient robust methods via monitoring for clustering and multivariate data analysis. Pattern Recognit 88:246–260

Ritter G (2014) Cluster analysis and variable selection. CRC Press, Boca Raton

Rocci R, Gattone S, Di Mari R (2018) A data driven equivariant approach to constrained Gaussian mixture modeling. Adv Data Anal Classif 12:235–260

Rousseeuw P, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223

Seo B, Kim D (2012) Root selection in normal mixture models. Comput Stat Data Analal 56:2454–2470

Zhang J, Liang F (2010) Robust clustering using exponential power mixtures. Biometrics 66:1078–1086