The TCLUST Approach to Robust Cluster Analysis [*]

L.A. García-Escudero, A. Gordaliza,

C. Matrán and A. Mayo-Iscar

Departamento de Estadística e Investigación Operativa

Universidad de Valladolid. Valladolid, Spain[†]

Abstract

A new method for performing robust clustering is proposed. The method is designed with the aim of fitting clusters with different scatters and weights. A proportion α of contaminating data points is also allowed. Restrictions on the ratio between the maximum and the minimum eigenvalues of the groups scatter matrices are introduced. These restrictions make the problem to be well-defined guaranteeing the existence and the consistency of the sample estimators to the population parameters.

[*] Research partially supported by Ministerio de Ciencia y Tecnología and FEDER grant MTM2005-08519- C02-01 and by Consejería de Educación y Cultura de la Junta de Castilla y León grant PAPIJCL VA102A06.

[†] Departamento de Estadística e Investigación Operativa. Facultad de Ciencias. Universidad de Valladolid. 47002, Valladolid. Spain.

The method covers a wide range of clustering approaches, which arise depending on the strength of the chosen restrictions. Our proposal includes an algorithm for approximately solving the sample problem which takes advantage of the Dykstra's algorithm.

Key words: Robustness; Cluster Analysis; trimming; asymptotics; trimmed k-means; EM-algorithm; fast-MCD algorithm; Dykstra's algorithm.

Abbreviated title: Robust clustering based on trimming.

1 Introduction

Many statistical practitioners view the Cluster Analysis as a collection of mostly heuristic techniques for partitioning multivariate data. This view relies on the fact that most of the cluster techniques are not *explicitly* based on a probabilistic model. This could "...lead the naive investigator into believing that he or she did not make any assumption at all, and that the results therefore are 'objective'..." (Flury 1997, page 123). However, that objectiveness is far from the reality as long as most of the times the cluster's results are strongly affected by the chosen method and its performance is somehow very dependent on the assumed underlying probabilistic model for which the method is implicitly aimed to. For instance, when using k-means, we must keep in mind that this method is designed for clustering spherical groups of roughly equal sizes and, thus, the method is not reliable when the groups we are searching for depart strongly from this assumption. So in order to understand clustering methods and decide which method should be applied in a particular case it is interesting to determine appropriate models and develop methods specially tailored for these models.

That determination of appropriate models for clustering is even more important in the presence of noisy data or outliers. Without specifying a model, it is not clear what we understand by an observation following an "anomalous" behavior. For instance, it is not clear when a set of very scattered observations may be seen indeed as an extra proper group or merely as a background noise to be deleted (see Figure 1). Additionally, it is not obvious wether a small group of tightly joined outliers should be considered as a proper group instead of a contamination phenomenon. Finally, notice that the precise detection of the outliers is very important due to the serious troubles they introduce in standard clustering procedures (see, e.g., García-

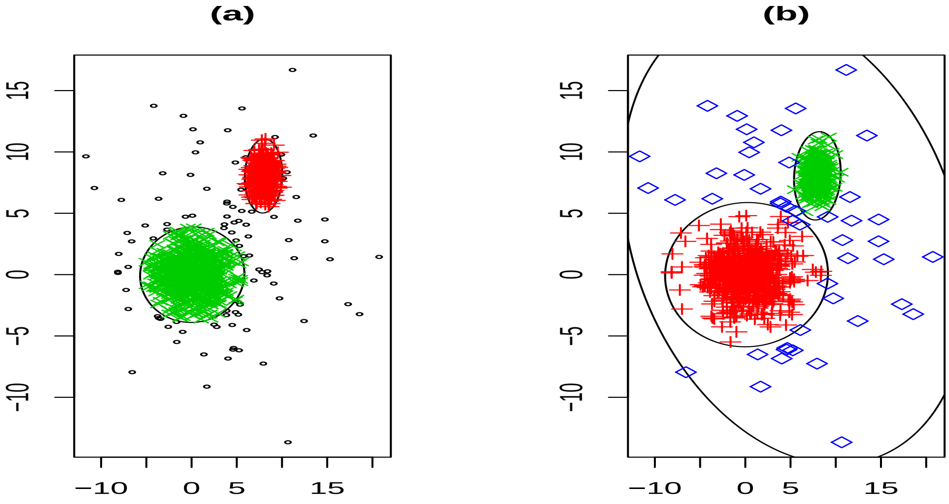

Figure 1: (a) Two groups with 10% of the observation discarded (trimmed points are the small circles). (b) Three groups partition with no observations discarded.

Escudero and Gordaliza 1999 and Hennig 2004) and also due to the appealing interest they could have by themselves after explaining why they depart from the general behavior.

Bock (2002) mentions in a recent Cluster Analysis review paper two model-based approaches which provide a "theoretical well-based clustering criterion" in presence of outliers:

(i) *Mixture modelling:* This first approach is based on considering mixture fittings. Some examples are the MCLUST and the EMMIX software. MCLUST (Fraley and Raftery 1998) allows the addition of a mixture component accounting for the "noise" (usually modelled by a uniform distribution in a convex set containing all the observations). McLachlan and Peel (2000)'s EMMIX tries to fit the noise resorting to mixtures of t distributions.

(ii) *Trimming approach:* The second approach assumes a known fraction α of outliers to be trimmed off and, later, the non-trimmed observations or "regular" data are split into k groups. Examples of this approach are the trimmed k-means (Cuesta-Albertos et al. 1997) and some recent proposals by Gallegos (2001, 2002) and Gallegos and Ritter (2005).

Notice that a "crisp" 0-1 approach is usually adopted in the second approach while the mixture approach generally returns some groups' ownership probabilities. However, the main difference between these two approaches intended for clustering in presence of outliers relies on the fact that *mixture modelling* approach tries to fit the outlying observations in the model while the outliers are completely discarded (being trimmed-off) in the *trimming* approach. The methodology presented in this paper is included within the second, "trimming", approach and all the comparisons will be made within this category of methods. The so-called "spurious-outlier" model commented below will serve as a common model for this approach.

The ability of trimming the least reliable observations has played historically a very important role as a way of providing robustness to many statistical procedures. Nevertheless, trimming in Cluster Analysis is not straightforward because no privileged directions there exist for searching outlying values and, most of the times, we even need to remove observations which fall between the groups ("bridge" data points). The first attempts of trimming in clustering appeared in Cuesta-Albertos et al. (1997). There, a modification of the k-means method (the most widely used non-hierarchical clustering method) is presented. The way to perform the trimming is called "impartial" as it is the sample itself which tells us the observations to be discarded. Moreover, García-Escudero and Gordaliza (1999) showed that the impartial trimming provides better results in terms of robustness than the consideration of different penalty functions in the k-means method (e.g., k-medoids).

The use of trimmed k-means involves a considerable drawback in how it implicitly assumes the same spherical covariance matrix for the groups (as classical k-means does). This justifies the recent extension for that method in Gallegos and Ritter (2005) through the trimmed determinant criterion. This approach retains the assumption on the equality of the covariance matrices, but it allows for a general expression of the common covariance matrix non necessarily spherical. There, it also appears a statistical clustering model with outliers called the *spurious-outlier model* extending the usual statistical clustering setup (Mardia et al. 1979) to consider the presence of a proportion α of noise. The likelihood function for the data set $\{x_1, ..., x_n\}$ is

$$\left[\prod_{j=1}^{k}\prod_{i\in R_j} f(x_i;\mu_j,\Sigma)\right]\left[\prod_{i\notin R} g_{\psi_i}(x_i)\right] \tag{1.1}$$

with $R = \cup_{j=1}^{k} R_j$ and $\#R = [n(1-\alpha)]$. The parameter k denotes the total number of groups, R_j contains the indexes of the "regular" observations assigned to group j and $f(\cdot;\mu,\Sigma)$ stands for the p.d.f. of the p-variate normal distribution with mean μ and covariance matrix Σ while g_{ψ_i}'s are some p.d.f.'s in $\mathbb{R}^p$. If $\Sigma = \sigma^2 \cdot I$ is chosen in (1.1) then we would be performing the trimmed k-means method. Gallegos and Ritter (2005) showed that the maximization of (1.1) reduces to the consideration of the regular part of the observations under just some reasonable assumptions for the g_{ψ_i}'s whenever the "non-regular" observations may be seen as merely "noise". In spite of that, the maximization of this classification likelihood is a computationally hard problem because of its combinatorial nature. Therefore, they also propose an algorithm in the spirit (both algorithms coincide when $k = 1$) of the fast-MCD in Rousseeuw and van Driessen (1999) for approximately maximizing (1.1).

The assumption for the equality of the groups' covariance matrices could be very restrictive in many contexts. Therefore, the next natural extension is to allow for different covariance matrices Σ_j's instead of a single Σ in expression (1.1). Unfortunately, this heterogeneous robust clustering problem is notably harder. It is easy to see the unboundedness of the proposed objective function, as each data point gives rise to a singularity on the edge of the parametric space. Moreover, a straight adaptation of the fast-MCD algorithm is not longer adequate. The inadequacy of a naïve adaptation follows from the disturbing presence of different groups' "scales" (we define the "scale" parameter for a covariance matrix Σ_j as $|\Sigma_j|^{1/p}$) which makes complex the global ordering of the observations around their closest centers through Mahalanobis distances (see, García-Escudero and Gordaliza 2006) when performing the so-called "concentration" steps. This is the reason why unrestricted algorithms frequently find clusters containing a few data points either very close together or almost lying in a lower dimensional space (Figure 2,a) and the application of some kind of restriction would allow us to obtain (perhaps) more interesting or informative partitions (Figure 2,b).

As a way of posing constraints to the heterogeneous robust clustering problem, Gallegos (2001, 2002) proposes normalizing the covariances to have unit determinant when computing the Mahalanobis distances in the "concentration" steps. This serves to avoid the harmful effect of the different scales and even so benefiting from the rationale behind the fast-MCD algorithm. Gallegos's procedure works nicely when the groups have similar scales, but it does not work so well when very different groups' scales are involved. Normalizing the covariances to have unit determinant can be very restrictive and, surely, such strong restrictions are not always needed. Moreover, it seems also adequate to incorporate the restrictions directly in the problem

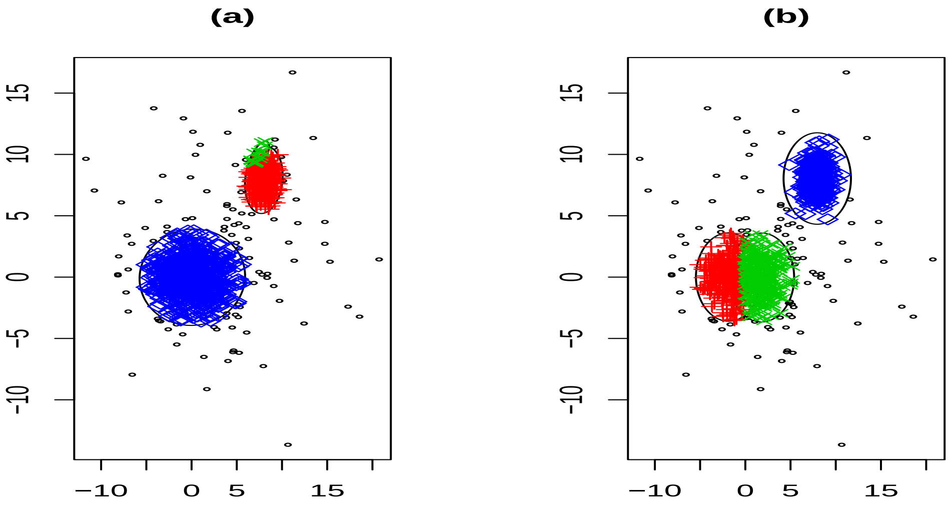

Figure 2: An unrestricted solution for the same data set in Figure 1 appears in (a) when $k = 3$ and $\alpha = .1$. Compare with a restricted partition, also when $k = 3$ and $\alpha = .1$, in (b). A description of the applied clustering methods will be given in Section 3.

statement instead of (artificially) appearing in the algorithm.

Having these things in mind, constraints for the heterogeneous robust clustering problem will be incorporated in this paper through an eigenvalues-ratio restriction. A constant c will control the strength of the posed restriction entailing a wide range of clustering problem depending on its value.

Finally, it can be shown that the heterogeneous robust clustering problem is even notably harder under the presence of different weights for the underlying groups. Hence, the introduction of some weight terms π_j's in (1.1) will also be considered for handling different groups' weights. In fact, it can be easily shown that do not include the weights in the objective function is equivalent to the constraint of equally weighted groups.

Since we assume that every point in the sample space $\mathbb{R}^p$ will be assigned to some group (or it is part of the sample space to be discarded), we will be able to define a population counterpart serving as a probabilistic benchmark. Existence results for both, the sample and the population problems, will be given in this Section 2. Moreover, consistency of the sample maximizers to the population ones under mild assumptions is proven. The proofs of these results make clearer the importance of the eigenvalues restrictions in order to guarantee these existence and consistency results.

In Section 3, we propose a feasible algorithm (TCLUST) for approximately solving the sample version of the problem. The algorithm may be seen as a Classification EM-algorithm (Celeux and Govaert 1992) where a kind of "concentration" step is also applied. The eigenvalues-ratio restrictions will be imposed by solving a constrained least squares problem. Dykstra (1983)'s algorithm may be applied for addressing that problem.

Finally, Section 4 shows a simulation study showing the gain provided by the proposed method with respect to other "trimming" proposals.

2 Robust Clustering with Eigenvalues-Ratio Restrictions

Suppose that $\{x_1, ..., x_n\}$ denotes the available data in some p-dimensional Euclidean space. Let $f(x; \mu, \Sigma)$ be the p.d.f. of a p-variate normal distribution $f(x; \mu, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp(-(x-\mu)'\Sigma^{-1}(x-\mu)/2)$ where $|\cdot|$ stands for the determinant. We denote a probability measure P acting over a function f by $Pf(\cdot) = \int f(x)dP(x)$.

2.1 Mathematical formulation

We start by modifying the "spurious-outlier" model considered in Gallegos and Ritter (2005). First, as mentioned before, we consider different scatter matrices Σ_i's as in Gallegos (2001, 2002). Moreover, we assume the presence of some underlying weights π_j's with $\sum_{j=1}^{k} \pi_j = 1$ associated to the distributions generating the set of "regular" observations. This leads us to the maximization of

$$\left[\prod_{j=1}^{k} \prod_{i \in R_j} \pi_j f(x_i; \mu_j, \Sigma_j) \right] \left[\prod_{i \notin R} g_{\psi_i}(x_i) \right], \tag{2.1}$$

with $R = \cup_{j=1}^{k} R_j$ and $\#R = n - [n\alpha]$. Additionally, the restrictions on the eigenvalues of the Σ_j's matrices will be later introduced in order to avoid singularities. As in Gallegos and Ritter (2005), if the g_ψ's satisfy the condition

$$\arg\max_{\mathcal{R}} \max_{\mu_j, \Sigma_j} \prod_{j=1}^{k} \prod_{i \in R_j} \pi_j f(x_i; \mu_j, \Sigma_j) \subseteq \arg\max_{\mathcal{R}} \prod_{i \notin \cup_{j=1}^{k} R_j} \max_{\psi_i} g_{\psi_i}(x_i)$$

where $\mathcal{R}$ stands for the set of all partitions of the indexes $\{1, ..., n\}$ onto k groups of regular observations, R, and a group containing the non-regular ones, with $\#R = n - [n\alpha]$, then we can avoid the non-regular contribution to the previous maximization problem. This condition easily holds under just some reasonable assumptions for the g_{ψ_i}'s whenever the "non-regular" observations are seen as merely "noise". For instance, the examples for g_{ψ_i}'s shown in the Gallegos and Gallegos and Ritter' papers can be trivially also considered here. We refer the interested reader to these papers for more details.

Additionally, for an easier statement of our problem, we will use some *assignment functions* z_j's telling us which class *every* point x in $\mathbb{R}^p$ is assigned to (not only the sample observations x_i's are classified). We follow a 0-1 "crisp" approach where x is

$$\text{assigned to the class } j \text{ if } z_j(x) = 1,$$

or

$$\text{it is being trimmed off if } z_0(x) = 1.$$

With these functions, assuming that the g_{ψ_i}'s may be omitted, we can raise again the problem in (2.1) to the maximization of

$$\prod_{i=1}^{n} \left[\prod_{j=1}^{k} \pi_j^{z_j(x_i)} f(x_i; \mu_j, \Sigma_j)^{z_j(x_i)} \right],$$

with z_j being 0-1 functions defined in the whole sample space verifying $\sum_{j=0}^{k} z_j(x_i) = 1$ and $\sum_{i=1}^{n} z_0(x_i) = [n\alpha]$. This statement of the problem, taking logarithms, leads us to the following general one:

> **Robust Clustering problem:** Given a probability measure P, we search for the maximization of:
>
> $$P \left[\sum_{j=1}^{k} z_j(\cdot) \big(\log \pi_j + \log f(\cdot; \mu_j, \Sigma_j) \big) \right], \qquad (2.2)$$
>
> made in terms of the assignment functions:
>
> $$z_j : \mathbb{R}^p \mapsto \{0, 1\}, \text{ such that } \sum_{j=0}^{k} z_j = 1 \text{ and } P z_0(\cdot) = \alpha,$$
>
> and the parameters $\theta = (\pi_1, ..., \pi_k, \mu_1, ..., \mu_k, \Sigma_1, ..., \Sigma_k)$ corresponding to weights $\pi_j \in [0, 1]$ with $\sum_{j=1}^{k} \pi_j = 1$, mean vectors

$\mu_j \in \mathbb{R}^p$ and symmetric positively definite $p \times p$-matrices Σ_j, $j = 1, ..., k$.

If P_n stands for the empirical measure, $P_n = 1/n \sum_{i=1}^n \delta_{x_i}$, we just need to replace P by P_n in the previous problem in order to recover the original sample problem (notice that, perhaps, $P_n z_0(\cdot) = \alpha$ can not be exactly achieved but this familiar fact will not be very important in our reasonings).

Let us finally introduce our restrictions on the eigenvalues of the covariance matrices. This type of restriction may be seen as an extension of those introduced by Hathaway (1985) for one-dimensional data. It allows to avoid the singularities introduced by the possibility of very different Σ_j's by controlling the ratio between the maximum and the minimum eigenvalue of these matrices:

> **(ER) Eigenvalues-Ratio restrictions:** We fix a constant $c \geq 1$ such that
>
> $$M_n/m_n \leq c$$
>
> for
>
> $$M_n = \max_{j=1,\ldots,k} \max_{l=1,\ldots,p} \lambda_l(\Sigma_j) \text{ and } m_n = \min_{j=1,\ldots,k} \min_{l=1,\ldots,p} \lambda_l(\Sigma_j)$$
>
> where $\lambda_l(\Sigma_j)$ are the eigenvalues of the matrices Σ_j, $j = 1, ..., k$ and $l = 1, ..., p$.

We denote by Θ_c the set constituted by the θ's which obey condition ER for a given c.

Notice that, the strongest possible restriction follows from setting $c = 1$. In this particular case, the proposed method may be viewed as a trimmed k-means method with weights. However, the main advantage of this approach relies on the fact that parameter c allows us to achieve certain (controlled) freedom in how we want to handle the different scatter of the groups. Figure 1 and 2 in Section 1 show the results of the application of the proposed methodology (by using the TCLUST algorithm described in Section 3) to a data set made up of three gaussian 2-dimensional clusters where the most scattered one accounts for 10% of the data. The result when $k = 2$, $\alpha = .1$ and $c = 5$ appears in Figure 1,(a). The result there is not very dependent on c as long as the two (main) groups are not too different in their eigenvalues once the most scattered group has been trimmed off. The values $k = 3$ and $\alpha = 0$ were considered in Figure 1,(b), with a large value for c ($c = 50$) which allows for the presence of the more scattered group. The values $k = 3$ and $\alpha = .1$ were applied in Figure 2. A rather large c (unrestricted) was chosen in Figure 2,(a) while a small $c = 1$ (restricted) was considered in Figure 2,(b).

Once our problem was stated, we must exclude in the subsequent analysis those probability distributions obviously unappropriate for this approach. This leads us to assume on the underlying distribution P the following mild condition which trivially holds if it is a continuous distribution or if it is the empirical measure P_n corresponding to a sample from an absolutely continuous distribution (for n large enough):

The distribution P is not concentrated on k points after removing a probability mass equal to α. (2.3)

To conclude this section, we will notably simplify our problem through an adequate reformulation. This will lead to express the assignment functions z_j's only in terms of θ. This new statement will be of capital importance for deriving later an algorithm to solve the sample counterpart of the problem. In order to state this result, some additional notation will be needed:

Given $\theta \in \Theta_c$, we consider some *discriminant functions* defined as

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

and

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

(notice that these functions appear when applying Bayes' rules in Discriminant Analysis). These functions will serve to determine which are the most "outlying" observations. For a fixed choice of θ, the smaller $D(x;\theta)$ for a given x is, the more outlying x will be supposed.

Using the previous definitions, for a given θ and a probability measure P, we define

$$G(\cdot;\theta,P) : u \in \mathbb{R} \mapsto P\Big[I_{[0,u]}(D(\cdot;\theta))\Big] \tag{2.4}$$

and

$$R(\theta,P) := G^{-1}(\alpha;\theta,P) = \inf_u\{G(u;\theta,P) \geq \alpha\}$$

(notice that if X is a random variable with distribution given by P then $R(\theta,P)$ is the α-quantile of the random variable $D(X;\theta)$).

With this notation, we have the following characterization for the z_j's functions:

Lemma 1 *For a probability measure P, using the discriminant functions $D_j(x;\theta)$, the Robust Clustering problem can be simplified to the maximization only on terms of θ of*

$$\theta \mapsto L(\theta, P) := P\left[\sum_{j=1}^{k} z_j(\cdot;\theta)\log D_j(\cdot,\theta)\right], \tag{2.5}$$

where the assignment functions are obtained from θ as

$$z_j(x;\theta) = I\{x : \{D(x;\theta) = D_j(x;\theta)\} \cap \{D_j(x;\theta) \geq R(\theta, P)\}\}$$

and

$$z_0(x;\theta) = 1 - \sum_{j=1}^{k} z_j(x;\theta).$$

In other words, we assign x to the class j with the largest discriminant function value $D_j(x;\theta)$ or x is trimmed off when all the $D_j(x;\theta)$'s (and consequently $D(x;\theta)$) are smaller than $R(\theta, P)$. To be more precise, a rule for breaking ties in the discriminant function values is also needed. For instance, the lexicographical ordering could be applied. The proof of Lemma 1 is straightforward and it will be omitted.

2.2 Existence

Our analysis begins by proving the existence of solutions for the proposed problem:

Consider a sequence $\{\theta_n\}_{n=1}^{\infty} = \{(\pi_1^n, ..., \pi_k^n, \mu_1^n, ..., \mu_k^n, \Sigma_1^n, ..., \Sigma_k^n)\}_{n=1}^{\infty}$ such that

$$\lim_{n\to\infty} L(\theta_n, P) = \sup_{\theta\in\Theta_c} L(\theta, P) = M > -\infty \tag{2.6}$$

(the boundedness from below for (2.6) can be easily obtained just considering $\pi_1 = 1$, $\mu_1 = 0$, $\Sigma_1 = I$, and setting the other weights as 0 with arbitrary choices of means and variances).

Since $[0,1]^k$ is a compact set, we can extract a subsequence from $\{\theta_n\}_n^\infty$ (that will be denoted like the original one) such that

$$\pi_j^n \to \pi_j \in [0,1] \text{ for } 1 \leq j \leq k, \tag{2.7}$$

and also satisfying for some $g \in \{0, 1, ..., k\}$ (a relabelling could be needed) that

$$\mu_j^n \to \mu_j \in \mathbb{R}^p \text{ for } 0 \leq j \leq g \text{ and } \min_{j>g} \|\mu_j^n\| \to \infty. \tag{2.8}$$

With respect to the scatter matrices, if the restriction ER is assumed, we can also consider a further subsequence verifying one (and only one) of these possibilities:

$$\Sigma_j^n \to \Sigma_j \text{ for } 1 \leq j \leq k, \tag{2.9}$$

$$M_n = \max_{j=1,\ldots,k} \max_{l=1,\ldots,p} \lambda_l(\Sigma_j) \to \infty, \tag{2.10}$$

or

$$m_n = \min_{j=1,\ldots,k} \min_{l=1,\ldots,p} \lambda_l(\Sigma_j) \to 0. \tag{2.11}$$

The next lemma will show the convergence of the covariance matrices by showing that only the convergence (2.9) is possible:

Lemma 2 *If ER holds and if P satisfies (2.3), then the convergences (2.10) or (2.11) for the covariance matrices are not possible. Therefore, we can find subsequences of Σ_j^n converging toward some matrices Σ_j, $j = 1, ..., k$.*

Proof: We will see that (2.10) or (2.11) would imply $\lim_{n\to\infty} L(\theta_n, P) = -\infty$.

Let $\lambda_{l,j}^n := \lambda_l(\Sigma_j^n)$ be the eigenvalues , $j = 1, ..., k$ and $l = 1, ..., p$, of the group covariance matrices and $v_{l,j}^n$ their associated eigenvectors with $\|v_{l,j}^n\| = 1$. We see that

$$\begin{aligned} L(\theta_n, P) &= P\Bigg[\sum_{j=1}^k z_j(\cdot;\theta_n)\big(\log \pi_j^n - \frac{p}{2}\log 2\pi - \frac{1}{2}\sum_{l=1}^p \log \lambda_{l,j}^n \\ &\qquad -\frac{1}{2}\sum_{l=1}^p (\lambda_{l,j}^n)^{-1}(\cdot - \mu_j^n)' v_{l,j}^n (v_{l,j}^n)'(\cdot - \mu_j^n)\big)\Bigg] \\ &\le P\Bigg[\sum_{j=1}^k z_j(\cdot;\theta_n)\big(\log \pi_j^n - \frac{p}{2}\log 2\pi - \frac{p}{2}\log m_n - \frac{1}{2}M_n^{-1}\|\cdot - \mu_j^n\|^2\big)\Bigg] \end{aligned} \tag{2.12}$$

If we assume that (2.10) holds, i.e. $M_n \to \infty$, then $m_n \to \infty$ by ER. Thus, we would have that $L(\theta_n, P) \to -\infty$ and this leads us to a contradiction with (2.6).

Now assume that (2.11) holds. We will need a technical result which guarantees that if P satisfies (2.3), then there exists a constant h such that

$$P\left[\sum_{j=1}^k z_j(\cdot;\theta_n)\|\cdot - \mu_j^n\|^2\right] \ge h > 0. \tag{2.13}$$

The proof of this result is based on an existence result for trimmed k-means (Cuesta-Albertos et al. 1997) and it is left to the Appendix.

Since $\log \pi_j^n \le 0$, the fact that $P[z_1(\cdot) + ... + z_k(\cdot)] = 1 - \alpha$ implies

$$L(\theta_n, P) \le (1-\alpha)\big(-\frac{p}{2}\log 2\pi - \frac{p}{2}\log m_n\big) - \frac{1}{2}M_n^{-1} P\left[\sum_{j=1}^k z_j(\cdot;\theta_n)\|\cdot - \mu_j^n\|^2\right].$$

Therefore, ER and (2.13) give

$$L(\theta_n, P) \le (1-\alpha)\big(-\frac{p}{2}\log 2\pi - \frac{p}{2}\log m_n\big) - \frac{1}{2}(cm_n)^{-1}h. \tag{2.14}$$

But this upper-bound in (2.14) tends to $-\infty$ as $m_n \to 0$. □

By applying the previous lemma, let us consider a subsequence verifying (2.7), (2.8) and (2.9). We will see that whenever the classes in the optimal partition have strictly positive probability masses we can guarantee the convergence of the centers μ_j^n. These result will be of key importance to understand the role played by the weights π_j's in this approach.

Lemma 3 *When ER and (2.3) hold, if every π_j in (2.7) verifies $\pi_j > 0$, $j = 1, ..., k$, then $g = k$ in (2.8) (i.e., the centers μ_j^n are not allowed to arbitrarily increase in norm).*

Proof: If $g = 0$, we can take a ball with center 0 and radius big enough $B(0, R)$ such that $P[B(0, R)] > \alpha$. We can thus easily see that

$$P\left[\sum_{j=1}^{k} z_j(\cdot;\theta_n)\|\cdot - \mu_j^n\|^2\right] \to \infty,$$

so that $L(\theta_n, P) \to -\infty$ from (2.12). Notice that condition ER has also been applied.

When $g > 0$, we prove first that

$$P\left[\sum_{j=g+1}^{k} z_j(\cdot;\theta_n)\right] \to 0. \tag{2.15}$$

This arises from the dominated convergence theorem taking into account that the sequence is obviously bounded by $1 - \alpha$, and the fact that

$$\{x : z_j(x;\theta_n) = 1\} \subseteq \{x : \max_{j=g+1,\ldots,k} D_j(x;\theta_n) \geq D_1(x;\theta_n)\} \tag{2.16}$$

for $j = g + 1, ..., k$, where the right-hand side converges toward the empty set, when n tends to ∞, due to (2.8) and (2.9).

We can now use (2.15) in order to get:

$$\begin{aligned}
&\lim_{n\to\infty} \sup L(\theta_n, P) \\
\leq\ & \lim_{n\to\infty} P\left[\sum_{j=1}^{g} z_j(\cdot;\theta_n)\big(\log \pi_j^n - \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_j^n| - \frac{1}{2}(\cdot - \mu_j^n)'(\Sigma_j^n)^{-1}(\cdot - \mu_j^n)\big)\right] \\
=\ & P\left[\sum_{j=1}^{g} z_j(\cdot;\tilde{\theta})\big(\log \pi_j - \frac{p}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(\cdot - \mu_j)'\Sigma_j^{-1}(\cdot - \mu_j)\big)\right],
\end{aligned}$$

where $x \mapsto z_j(x;\tilde{\theta})$ are the assignment functions which would be derived when working with g (instead of k) populations and $\tilde{\theta}$ being equal to a limit of the subsequence $\{\tilde{\theta}_n\}_{n=1}^{\infty} = \{(\pi_1^n, ..., \pi_g^n, \mu_1^n, ..., \mu_g^n, \Sigma_1^n, ..., \Sigma_g^n)\}_{n=1}^{\infty}$.

As $\sum_{j=1}^{g} \pi_j < 1$, the proof ends up by showing that we can change the weights $\pi_1, ..., \pi_k$ by

$$\pi_j^* = \frac{\pi_j}{\sum_{j=1}^{g} \pi_j} \text{ for } 1 \leq i \leq g \text{ and } \pi_{g+1}^* = ... = \pi_k^* = 0, \qquad (2.17)$$

(and properly modifying the assignment functions z_j's). This change produces a strict decrease in the objective function, leading to a contradiction with the optimality stated in (2.6). Thus, we conclude $g = k$. □

Proposition 1 (Existence) *If (2.3) holds for the probability measure P, then there exists some $\theta \in \Theta_c$ such that the maximum of (2.6) under the restriction ER is achieved.*

Proof: The existence of such as θ follows easily from Lemmas 2 and 3:

(i) If $\pi_j^n \to \pi_j > 0$ for $1 \leq j \leq k$, then the choice of θ is obvious.

(ii) Now assume that $\pi_j^n \to \pi_j > 0$ with $\pi_j > 0$ for $j \leq g$ and $\pi_j = 0$ for $g < j \leq k$. We define the weights π_j as

$$\pi_j = \lim_{n\to\infty} \pi_j^n \text{ for } j = 1, ..., g \text{ and } \pi_{g+1} = ... = \pi_k = 0.$$

> Analogously, take $\mu_j = \lim_{n\to\infty} \mu_j^n$ and $\Sigma_j = \lim_{n\to\infty} \Sigma_j^n$ for $j \le g$. The other μ_j's and Σ_j's may be arbitrarily chosen (but with the eigenvalues of the Σ_j's verifying the restriction possed by ER). □

Although we admit weights $\pi_j = 0$, this is not a drawback when taking $\log \pi_j$ because in this case $z_j(\cdot;\theta) \equiv 0$ and then set $\{x : z_j(x;\theta) = 1\}$ is empty. Notice that the presence of groups with zero weight does actually happen in practice. For instance, when $k = 2$, $c = 1$, $\alpha = 0$ and P is the $N(0,1)$ distribution in the real line, we can see that $\theta = (\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_1^2) = (1, 0, 0, \mu_2, 1, 1)$ is the optimal solution for every $\mu_2 \in \mathbb{R}$.

Recall that the objective function for the trimmed k-means method always improves when increasing k (see, Lemma 9 in the Appendix). The analysis of the optimal values for these objective functions leads to an useful criterion for deriving appropriate choices of the parameters k and α (García-Escudero et al. 2003). Here, the possible existence of groups with $\pi_j = 0$ would imply that the value of the objective function does not necessarily improve when increasing k. However, this property could be even more interesting in order to develop better techniques for choosing k and α.

2.3 Consistency

Given $\{x_n\}_{n=1}^\infty$ an i.i.d. random sample from an underlying (unknown) probability distribution P, let $\{\theta_n\}_{n=1}^\infty = \{(\pi_1^n, ..., \pi_k^n, \mu_1^n, ..., \mu_k^n, \Sigma_1^n, ..., \Sigma_k^n)\}_{n=1}^\infty \subset \Theta_c$ denote the sequence of sample estimators obtained by solving the problem (2.2) for the empirical measures $\{P_n\}_{n=1}^\infty$ with the eigenvalue restrictions possed by ER for a fixed constant $c \le 1$.

Section 2.2 shows that such sequence does always exist, for large enough n, whenever P is an absolutely continuous distribution verifying (2.3). Notice that although similar notation to that applied in previous section will be used, here the index n will indicate the dependence on a random sample of size n from P.

2.3.1 Boundedness of the sample estimators

We will see first that there exists a compact set $K \subset \Theta_c$ such that $\theta_n \in K$ for n sufficiently large with probability 1.

Lemma 4 *If P is an absolutely continuous distribution (thus verifying condition (2.3)) then the elements of the matrices Σ_j^n are uniformly bounded with probability 1.*

Proof: We see first that there exists a constant M' such that

$$L(\theta_n, P_n) \geq M' > -\infty \text{ } P\text{-a.e. for } n \geq n_0. \tag{2.18}$$

Due to the tightness of P_n, we can find a sequence of radii $\{r_n\}$ uniformly bounded by a constant r such that $P_n(B(0, r_n)) \geq 1 - \alpha$. Thus we trivially have

$$L(\theta_n, P_n) \geq \int_{B(0,r_n)} \Big(-\frac{p}{2} \log 2\pi - \frac{1}{2} \|x\|^2 \Big) dP_n(x)$$

with $\{I\{B(0,s)\}(\cdot)(-\frac{p}{2}\log 2\pi - \frac{1}{2}\| \cdot \|^2) : s \leq r\}$ being a Glivenko-Cantelli class and so (2.18) follows.

In order to see the boundedness of the elements of the covariance matrices, we need to prove that the maximum eigenvalue M_n does not tend to ∞ and the minimum eigenvalue m_n does not tend to 0. We start again from (2.12) with $P = P_n$, but the analogous of (2.13) follows from another trimmed k-means result shown in Appendix entailing the existence of a constant h' such that

$$P_n \left[\sum_{j=1}^{k} z_j(\cdot;\theta_n) \| \cdot - \mu_j^n \|^2 \right] \geq h' > 0 \tag{2.19}$$

whenever P satisfies (2.3). The proof follows similar steps to that of Lemma 2. If $M_n \to \infty$ or $m_n \to 0$ then we would have that $L(\theta_n, P_n) \to -\infty$, P-a.e., and (2.18) does not hold. □

Lemma 5 *If P is an absolutely continuous distribution, then we can choose empirical centers μ_j^n, $j = 1, ..., k$, such that their norms are uniformly bounded with probability 1.*

Proof: Assume on the contrary that for every sequence of optimal centers we can find a further subsequence (denoted as the original one) and satisfying (2.8) for some $g < k$. We will see that this is not possible by considering two possible alternatives:

(i) Let us assume, first, that there exist strictly positive limit points corresponding to the associated sequences of empirical weights $\{\pi_j^n\}_{n=1}^{\infty}$ for every $j = 1, ..., k$. I.e., there exist a subsequence such that

$$\pi_j^n \to \pi_j^0 > 0 \text{ for } j = 1, ..., k.$$

We can assume that $g \geq 1$ by a similar reasoning to that leading us to the obtention of the bound (2.18). Notice that the inclusion

(2.16), together with the P-a.e. boundedness of the sample covariance matrices, easily implies that

$$P_n \left[\sum_{j=g+1}^{k} z_j(\cdot;\theta_n) \right] \to 0.$$

As we did in the proof of Lemma 3, we can define a new sequence $\{\widetilde{\theta_n}\}$ obtained from $\{\theta_n\}$ such that the weights $\tilde{\pi}_j^n$ for $j = 1, ..., g$ are obtained after distributing the initial weights of the groups whose probability mass tend to 0 as in (2.17) and setting $\tilde{\pi}_j^n = 0$ for $j = g+1, ..., k$. Let us also consider $\tilde{\mu}_j^n = \mu_j^n$ for $j = 1, ..., g$ and some arbitrary but uniformly bounded centers $\tilde{\mu}_j^n$ for $j = g+1, ..., k$. We would see that

$$\lim_{n\to\infty} L(\widetilde{\theta_n}, P_n) < \lim_{n\to\infty} L(\theta_n, P_n),$$

eventually contradicting the optimal character of the $\{\theta_n\}$ sequence.

(ii) Let now assume that there exits a j such that the only limit points of the sequence $\{\pi_j^n\}_{n=1}^{\infty}$ are zeroes value (i.e., $\pi_j^n \to 0$ for $j = g+1, ..., k$ for $g < k$), then it is easy to see again, due to the P-a.e. boundedness of the covariance matrices, that we could replace the original centers for some uniformly bounded ones without decreasing $L(\cdot, P_n)$ for n large enough with probability 1. □

2.3.2 Consistency result

Let us start by stating the Glivenko-Cantelli property for two classes of functions:

Lemma 6 *Given a compact set K, the classes of functions:*

$$\mathcal{H}_1 := \big\{ I_{[u,\infty)}\big(D(\cdot;\theta)\big) : \theta \in K, u \geq 0 \big\} \tag{2.20}$$

and

$$\mathcal{H}_2 := \big\{ I_{[u,\infty)}\big(D(\cdot;\theta)\big) \sum_{j=1}^{k} z_j^*(\cdot;\theta) \log D_j(\cdot;\theta) : \theta \in K, u \geq 0 \big\} \tag{2.21}$$

are Glivenko-Cantelli classes, where $z_j^*(x;\theta) = I\{x : D(x;\theta) = D_j(x;\theta)\}$ (all the observations in $\mathbb{R}^p$ are assigned to some class without trimming by ussing the z_j^*'s).

Proof: The sets $\{x : D(x;\theta) \geq r\}$ are the union of k ellipsoids in $\mathbb{R}^p$. The functions $\log D_j(x;\theta) = \log \pi_j + \log f(x;\mu_j,\Sigma_j)$ are polynomials of degree 2 and the sets $\{x : z_j^*(x;\theta) = 1\}$ can be obtained through the intersection of subgraphs of polynomials of degree 2.

The Glivenko-Cantelli class properties are so easily obtained by using the methodology in section 2.6 of van der Vaar and Wellner (1996). $\square$

A technical result on $R(\theta;P) = G^{-1}(\alpha;\theta,P)$ with $G(\cdot;\theta,P)$ defined in (2.4) is also needed:

Lemma 7 *Let P be an absolutely continuous distribution with an strictly positive density function. Then, for every compact subset K, we have that*

$$\sup_{\theta\in K} |R(\theta;P_n) - R(\theta;P)| \to 0, \ \ P\text{-a.e. } . \tag{2.22}$$

Proof: For every $\delta > 0$, we can find an $\varepsilon > 0$ such that

$$G(R(\theta,P)+\delta,P,\theta) - G(R(\theta,P)-\delta,P,\theta) > \delta \text{ uniformly in } \theta \in K. \tag{2.23}$$

Otherwise, for a given δ, we could find a sequence $\{\theta_n\}_{n=1}^{\infty} \subset K$ satisfying

$$G(R(\theta_n, P) + \delta, P, \theta_n) - G(R(\theta_n, P) - \delta, P, \theta_n) < 1/n.$$

Thus, as the θ_n are included in the compact set K and $R(\theta_n, P) \leq M$, we could choose a further convergent subsequence (denoted as the original one) such that $\theta_n \to \theta_0 \in K$ and $R(\theta_n, P) \to r_0$ for a value $r_0 \geq 0$ (we apply that $R(\theta, P)$ are uniformly bounded for $\theta \in K$).

Notice that the conditions assumed on P imply that $u \mapsto G(u, P, \theta)$ is a continuous and strictly increasing function. Therefore, we would have that $G(r_0 + \delta, P, \theta_0) = G(r_0 - \delta, P, \theta_0)$, and this would contradict the strictly increasing character of $G(\cdot, P, \theta)$.

Now, if (2.22) did not hold, we could choose another sequence $\{\theta_n\}_{n=1}^{\infty}$ in K such that

$$|R(\theta_n, P_n) - R(\theta_n, P)| > \delta$$

for some $\delta > 0$. Therefore, by applying (2.23), we would see that $G(R(\theta_n, P_n), P, \theta_n)$ is smaller than $\alpha - \varepsilon/2$ or greater than $\alpha + \varepsilon/2$ infinitely often.

On the other hand, the Glivenko-Cantelli class property for $\mathcal{H}_1$ in (2.20) entails

$$\sup_{\theta \in K, u \geq 0} |G(u, P_n, \theta) - G(u, P, \theta)| \to 0, \ P\text{-a.e.},$$

and we obtain

$$|G(R(\theta_n, P_n), P, \theta_n) - G(R(\theta_n, P_n), P_n, \theta_n)| \to 0, \ P\text{-a.e.}.$$

But, $G(R(\theta_n, P_n), P_n, \theta_n) = \alpha$ thus it is not possible that $G(R(\theta_n, P_n), P, \theta_n)$ leaves the interval $[\alpha - \varepsilon/2, \alpha + \varepsilon/2]$ infinitely often. □

We can state now the main result in this section which it is the consistency result:

Proposition 2 (Consistency) *Assume that P has an strictly positive density function and that θ_0 is the unique maximum of (2.2) under the restriction ER. If $\theta_n \in \Theta_c$ denotes a sample version estimator based on the empirical measure P_n, then $\theta_n \to \theta_0$ almost surely.*

Proof: We have proved the existence of a compact set K such that $\theta_n \in K$ for $n \geq n_0$ with probability 1.

Notice that our objective function in the empirical case can be rewritten as:

$$L(\theta, P_n) = \int_{\{x:D(x,\theta)\geq R(\theta;P_n)\}} \left[\sum_{j=1}^{k} z_j^*(x;\theta)\log D_j(x;\theta)\right] dP_n(x).$$

Define new objective functions as

$$\widetilde{L}(\theta, P_n) = \int_{\{x:D(x,\theta)\geq R(\theta;P)\}} \left[\sum_{j=1}^{k} z_j^*(x;\theta)\log D_j(x;\theta)\right] dP_n(x),$$

(where $R(\theta, P)$ is fixed only depending on P and not depending on P_n).

We can see that

$$\sup_{\theta\in K} |L(\theta; P_n) - \widetilde{L}(\theta; P_n)| = o_P(1),$$

by using Lemma 7 and the fact that the integrand can be bounded from above and below from some constants uniformly for θ in the compact set K.

Finally, we can resort to the Glivenko-Cantelli property for the class of functions $\mathcal{H}_2$ in (2.21), and apply Theorem 3.2.3 in van der Vaar and Wellner (1996) to achieve the desired consistency. □

Remark 1 Notice that a uniqueness condition is needed in order to establish the consistency result. Unfortunately, this property does not always hold. For instance, think of a symmetric mixture P in the real line with two well-separated modes, a high trimming level and $k = 1$. The uniqueness property was already needed for establishing the same consistency result for the trimmed k-means and, even in this simpler case, the statement of general uniqueness results was difficult (see Remark 4.1 in García-Escudero et al. 1999). However, as in the trimmed k-means problem, we believe that it is quite rare to find a distribution where the uniqueness fails, when dealing with "reasonable" data for clustering and when parameters k and α have been properly chosen.

3 The TCLUST algorithm

The empirical problem presented in Section 2.1 has obviously a very high computational complexity. An exact algorithm seems to be not feasible even for moderate sample sizes. Thus the existence of an adequate algorithm for approximately solving the sample problem can be as important as the procedure itself. With this in mind, we propose the TCLUST algorithm, an EM-principle based algorithm, intended to search for approximate solutions. The EM algorithm is the usual method of obtaining a solution to the mixture likelihood problem (Dempster et al. 1977). Here, as we follow a "crisp" approach where each point is uniquely assigned to one cluster, a classification EM approach (Celeux and Govaert 1992) is preferable. Moreover, as trimmed observations are allowed, the rationale behind the fast-MCD (Rousseeuw and van Driessen 1999) and behind the trimmed k-means algorithm (García-Escudero et al 2003) will also underly. The restriction on the eigenvalues will be incorporated through Dykstra (1983)'s algorithm.

The TCLUST algorithm may be described as follows:

1. Randomly select starting values for the centers m_j^0's, the covariance matrices S_j^0's and the weights of the groups p_j^0's for $j = 1, ..., k$.

2. From the $\theta^l = (p_1^l, ..., p_k^l, m_1^l, ..., m_k^l, S_1^l, ..., S_k^l)$ returned by the previous iteration:

    2.1. Obtain $d_i = D(x_i, \theta^l)$ for the observations $\{x_1, ..., x_n\}$ and keep the set H having the $[n(1-\alpha)]$ observations with largest d_i's.

    2.2. Split H into $H = \{H_1, ..., H_k\}$ with $H_j = \{x_i \in H : D_j(x_i, \theta^l) = D(x_i, \theta^l)\}$.

    2.3. Obtain the number of data points n_j in H_j and their sample mean and sample covariance matrix, m_j and S_j, $j = 1, ..., k$.

    2.4. Consider the singular-value decomposition of $S_j = U_j' D_j U_j$ where U_j is an orthogonal matrix and $D_j = \text{diag}(\Lambda_j)$ is a diagonal matrix (with diagonal elements given by the vector Λ_j). If the full vector of eigenvalues $\Lambda = (\Lambda_1, ..., \Lambda_k)$ does not satisfy the eigenvalues-ratio restriction, obtain through Dykstra's algorithm a new vector $\tilde{\Lambda} = (\tilde{\Lambda}_1, ..., \tilde{\Lambda}_k)$ obeying the ER restriction and with $\|\tilde{\Lambda} - \Lambda^{-1}\|^2$ being as smallest as possible. Λ^{-1} denotes the vector made up by the inverse of the elements of the vector Λ. Notice that the ER restriction for Λ correspond exactly to the same ER restriction applied to Λ^{-1}.

2.5. Update θ^{l+1} using:

- $p_j^{l+1} \hookleftarrow n_j/[n(1-\alpha)]$
- $m_j^{l+1} \hookleftarrow m_j$
- $S_j^{l+1} \hookleftarrow U_j' \widetilde{D}_j U_j$ and $\widetilde{D}_j = \text{diag}(\tilde{\Lambda}_j)^{-1}$

3. Perform F iterations of the process described in step 2 (moderate values for F are usually enough) and compute the evaluation function $L(\theta^F; P_n)$.

4. Draw random starting values (i.e., start from step 1) several times, keep the solutions leading to minimal values of $L(\theta^F, P_n)$ and fully iterate them to choose the best one.

The computed (E-step) ‘a posteriori’ probabilities, $D_j(x_i, \theta^l) = p_j f(x_i; m_j, S_j)$, are converted to a discrete classification where we leave unassigned the proportion α of observations which are the hardest to classify. It is easy to see that this lead us to an optimal assignment.

We later obtain a new θ^{l+1} by maximizing (M-step) the conditional expectation once all untrimmed observation have been assigned to the groups. Proposition 3 guarantees that the presented algorithm can be applied for performing this maximization. Notice that the obtention of the optimal scatter matrices is decomposed into the search of the corresponding optimal eigenvalues and eigenvectors. For every choice of eigenvalues, the best eigenvectors choice simply follows from the unitary eigenvectors of the sample covariance matrix of the observations assigned to each group. This decomposition is somehow similar to that considered in Gallegos' proposal, where the "shapes" and the "scales" are separately handled.

If we see $D(x_i, \theta^l)$ as an inverse outlyingness measure for the observation x_i with respect to a choice of θ^l, then step 2 may be seen as some kind of "concentration" steps. García-Escudero and Gordaliza (2006) analyzes some other attempts for extending the "concentration" step principle to the heterogeneous robust clustering setup.

Recall that the random initialization scheme (step 1) and the final refinement (step 4) were very important in the fast-MCD algorithm. For initializing the procedure in the step 1, we have seen that simply randomly choosing k sample data points for the centers, k identity matrices for the covariances and the same weights for the groups (equal to $1/k$) provide reasonably starting values in most of the cases.

With respect to the eigenvalues-ratio restriction, we would need $\Lambda = (\Lambda_1, ..., \Lambda_k)$ with $\Lambda_j = (\lambda_{1,j}, ..., \lambda_{p,j})$ belonging to the cone $\mathcal{C}$, where

$$\mathcal{C} = \{(\Lambda_1, ..., \Lambda_k) \in \mathbb{R}^{p\times k} : \lambda_{u,v} - c \cdot \lambda_{r,s} \leq 0 \text{ for all } (u, v) \neq (r, s)\}. \quad (3.1)$$

If $\Lambda \notin \mathcal{C}$, we need to replace Λ^{-1} by $\tilde{\Lambda} \in \mathcal{C}$ with minimal $\|\tilde{\Lambda} - \Lambda^{-1}\|^2$. Dykstra's algorithm serves to approximately solve that constrained least squares problem when $\mathcal{C}$ is the intersection of the several closed convex cones by resorting to iterative projections onto the individual cones. Notice that $\mathcal{C}$ may be seen as the intersection of the cones

$$\mathcal{C}_h = \{(\Lambda_1, ..., \Lambda_k) \in \mathbb{R}^{p\times k} : \lambda_{u,v} - c \cdot \lambda_{r,s} \leq 0\}, h = (u, v, r, s),$$

and the projections onto the cones $\mathcal{C}_h$ are very fast to obtain. Thus a fixed number of individual projections may be done retaining the best attained solution after these iterations and satisfying the restrictions. Alternatively, quadratic programming based solutions (see, e.g., Goldfarb and Idnani 1983) to that constrained minimization may be explored.

Next result serves to formalize the appropriateness of the TCLUST algorithm:

Proposition 3 *If the sets $H_j = \{x_i : z_j(x_i) = 1\}$, $j = 1, ..., k$, are kept fixed, the maximum of (2.2) for $P = P_n$ can be obtained through the following steps:*

(i) Fixed μ_j and Σ_j, the best choice of π_j is $\pi_j = n_j/[n(1-\alpha)]$ where $n_j = \#H_j$.

(ii) Fixed Σ_j and the optimal values for π_j given in (i), the best choice for μ_j is the sample mean m_j of the observations in H_j.

(iii) Fixed the eigenvalues for the matrix Σ_j and the optimum values given in (i) and (ii) for π_j and μ_j, the best choice for the set of unitary eigenvectors are the unitary eigenvectors of the sample covariance matrix S_j of the observations in H_j.

(iv) With the optimal selections made in (i), (ii) and (iii), the best choice for the eigenvalues corresponds to the projection of the vector containing the inverse of the eigenvalues onto the cone $\mathcal{C}$ in (3.1).

Proof: Once the $z_j(x_i)$ for $i = 1, ..., n$ and $j = 0, ..., k$ are known values, the expression (2.2) can be written as

$$\sum_{j=1}^{k} \left[n_j \log \pi_j + \sum_{x_i \in H_j} \log f(x_i; \mu_j, \Sigma_j) \right], \tag{3.2}$$

and the assertions (i) and (ii) trivially hold.

Considering these optimal values for π_j and μ_j, together with the cyclic property of the trace, the maximization of (3.2) simplifies to the minimization of

$$\sum_{j=1}^{k} \left[\log |\Sigma_j| + \text{trace}(\Sigma_j^{-1} S_j) \right].$$

The matrices S_j and Σ_j can be decomposed into $S_j = U_j' D_j U_j$ and $\Sigma_j = V_j' E_j V_j$, where $D_j = \text{diag}(\Lambda_j)$ and $E_j = \text{diag}(\Xi_j)$ are diagonal matrices $\Lambda_j = (\lambda_{1,j}, ..., \lambda_{p,j})$ and $\Xi_j = (\xi_{1,j}, ..., \xi_{p,j})$, and U_j and V_j are orthogonal matrices. So, as $\log |\Sigma_j| = \log |E_j|$ and the eigenvalues E_j were fixed, the previous minimization problem can be further simplified to that of

$$\sum_{j=1}^{k} \text{trace}(\Sigma_j^{-1} S_j) = \sum_{j=1}^{k} \text{trace}(E_j^{-1} (U_j V_j')' D_j (U_j V_j')).$$

(the cyclic property of the trace is again applied). Denote $T_j = U_j V_j'$, and, rewrite

$$\text{trace}(E_j^{-1} T_j' D_j T_j) = \sum_u \sum_v \frac{\lambda_{u,j}}{\xi_{v,j}} \cdot t_{u\,v,j}^2 \ , \tag{3.3}$$

where $t_{u\,v,j}$ denotes the element (u, v) of the matrix T_j. As long as T_j is an orthogonal matrix, we have that $\sum_u t_{u\,v,j}^2 = 1$ and $\sum_v t_{u\,v,j}^2 = 1$. Therefore, the minimization of (3.3) may be seen as a linear programming problem like

$$\min \sum_{u,v} c_{u,v} \cdot x_{u,v} \text{ subject to } \sum_u x_{u,v} = 1, \sum_v x_{u,v} = 1 \text{ and } x_{u,v} \geq 1,$$

with known coefficients $c_{u,v}$ (notice that $\lambda_{u,j}/\xi_{u,j}$ are fixed coefficients because $\lambda_{u,j}$ depends on the data set at hand and the $\xi_{u,j}$ are supposed known values in (iii)). Although fractional solutions are possible, these solutions will never be basic feasible ones due to the particular statement of the linear programming problem (see, e.g., Papadimitriou and Steiglitz 1982, pg. 249). Consequently, the optimal solution corresponds to a "real matching" where the optimal $t_{u,v}^2$ are 0 or 1. Thus, T_j is a permutation matrix product of the orthogonal matrices U_j and V_j'. It is quite easy to see that the columns of the matrices U_j and V_j must provide the same set of unitary eigenvectors and, thus, the assertion (iii) is proven.

By applying (i),(ii) and (iii), we finally need to search for a vector $\Xi = (\Xi_1, ..., \Xi_k)$ minimizing

$$\sum_{j=1}^{k}\sum_{i=1}^{p}\left(\log \xi_{i,j} + \frac{\lambda_{i,j}}{\xi_{i,j}}\right) = \sum_{j=1}^{k}\sum_{i=1}^{p}\big(-\log \tilde{\lambda}_{i,j} + \lambda_{i,j}\cdot\tilde{\lambda}_{i,j}\big),\ \text{with } \tilde{\lambda}_{i,j} = 1/\xi_{i,j}. \tag{3.4}$$

As (3.4) is a convex function on the $\tilde{\lambda}_{i,j}$ and its unrestricted minimum is attained when $\tilde{\lambda}_{i,j} = \lambda_{i,j}^{-1}$, the minimization of (3.4) under the eigenvalues-ratio restriction possed by (3.1) leads us to the optimal choice of $\tilde{\Lambda}$ with minimal $\|\tilde{\Lambda} - \Lambda^{-1}\|^2$ and $\tilde{\Lambda} \in \mathcal{C}$. □

Remark 2 Alternatively, other methods can be defined by imposing restrictions on the ratio between the determinant of the groups' covariances instead of controlling the eigenvalues. Gallegos (2001, 2002)'s proposal scales the covariance matrices to have determinant ratio equal to 1 in the algorithm. Maronna and Jacovski (1974), in the untrimmed case $\alpha = 0$, consider that normalization as the only reliable "distance" for clustering multivariate data. The algorithm proposed here can be easily adapted for handling restrictions of this type. In this case, the cone would be

$$\mathcal{C}' = \{(\sigma_1, ..., \sigma_k) \in \mathbb{R}^k : \sigma_u - c \cdot \sigma_v \leq 0 \text{ for all } u \neq v\},$$

and the factorization in step 2.4 of the algorithm is $S_j = \sigma_j \cdot U_j$ with $|U_j| = 1$ and $\sigma_j = |S_j|^{1/p}$. If $c = 1$ in $\mathcal{C}'$, we would have an analogous of Gallegos's proposal with groups' weights.

Moreover, other procedures which have been used for avoiding pathological solutions in the heterogeneous robust clustering problem are based on adding different types of parameterizations for the covariance matrices (see, e.g., Scott and Symons 1971 or Banfield and Raftery 1993). Although that possibility has not been considered here, we believe that similar ideas (based on relaxing those parameterizations) could be interesting.

4 A simulation study

A simulation study has been carried out to compare the performance of the proposed robust clustering method with respect to other trimming approaches in the literature. Several data sets of size $n = 2000$ have been generated. Each data set consists of two simulated p-dimensional normally distributed clusters with centers $\mu_1 = (8, 0, 0, ..., 0)'$ and $\mu_2 = (0, 8, ..., 0)'$ and covariance matrices

$$\Sigma_1 = \text{diag}(1, a, 1, ..., 1) \text{ and } \Sigma_2 = \text{diag}(b, c, 1, ..., 1).$$

The constants a, b and c serve to control the true differences between the eigenvalues of the groups' covariance matrices leading us to the following cases:

(M1) $(a, b, c) = (1, 1, 1)$: *Spherical equally scattered groups.*

(M2) $(a, b, c) = (5, 1, 5)$: *Not spherical but the same covariance matrices for the groups.*

(M3) $(a, b, c) = (5, 5, 1)$: *Different covariance matrices but the same scale (equal determinant).*

(M4) $(a, b, c) = (1, 20, 5)$: *Groups with different scales.*

(M5) $(a, b, c) = (1, 45, 30)$: *Groups with different scales and a severe overlap.*

We consider 1800 "regular" data points and, in order to take into account the groups' weights, a proportion ρ of them are generated from the first normal distribution and a proportion $(1 - \rho)$ from the second. We also generate uniformly distributed data points in a parallelogram defined by the coordinatewise ranges of the regular data points. Using an acceptance-rejection algorithm, only points having squared Mahalanobis distances from μ_1 and μ_2 (using Σ_1 and Σ_2) greater than $\chi^2_{p,.975}$ are finally considered until reaching an amount of 200 clear outliers.

The following approaches searching for $k = 2$ groups and with a trimming proportion $\alpha = .1$ are tried:

(TkM) *Trimmed k-means (specially aimed to the case M1).*

(GR) *Gallegos and Ritter's method (specially aimed to the case M2).*

(G) *Gallegos's proposal (specially aimed to the case M3).*

(TCLUST) *The presented algorithm with an eigenvalues-ratio restriction $c = 50$.*

The same number of random initializations and "concentration" steps are considered for all methods.

Table 1 shows the average proportion of misclassified observations for $B = 1000$ independent random samples of size 2000 when $p = 2$ and 6 and the groups' weights are $\rho = 1/2$ and $1/3$.

Notice that all the methods work nicely under the underlying model they are specially aimed to. However, the proposed eigenvalues-ratio restriction method is the only method which is able to cope with the mixtures with very different scales (mixtures M4 and M5) and it seems to be less affected in the unequal groups' size case. Figure 3 shows the result of these four analyzed procedures applied to the same data set generated by the simulation scheme M5 when $p = 2$ and $\rho = 1/3$. The proposed method seems to be the only one able to distinguish between the least and the most scattered group even in this rather overlapped case.

APPENDIX

The following lemma was applied in the proofs of Lemmas 3 and 4. Its proof is based on some results in Cuesta-Albertos et al. (1997).

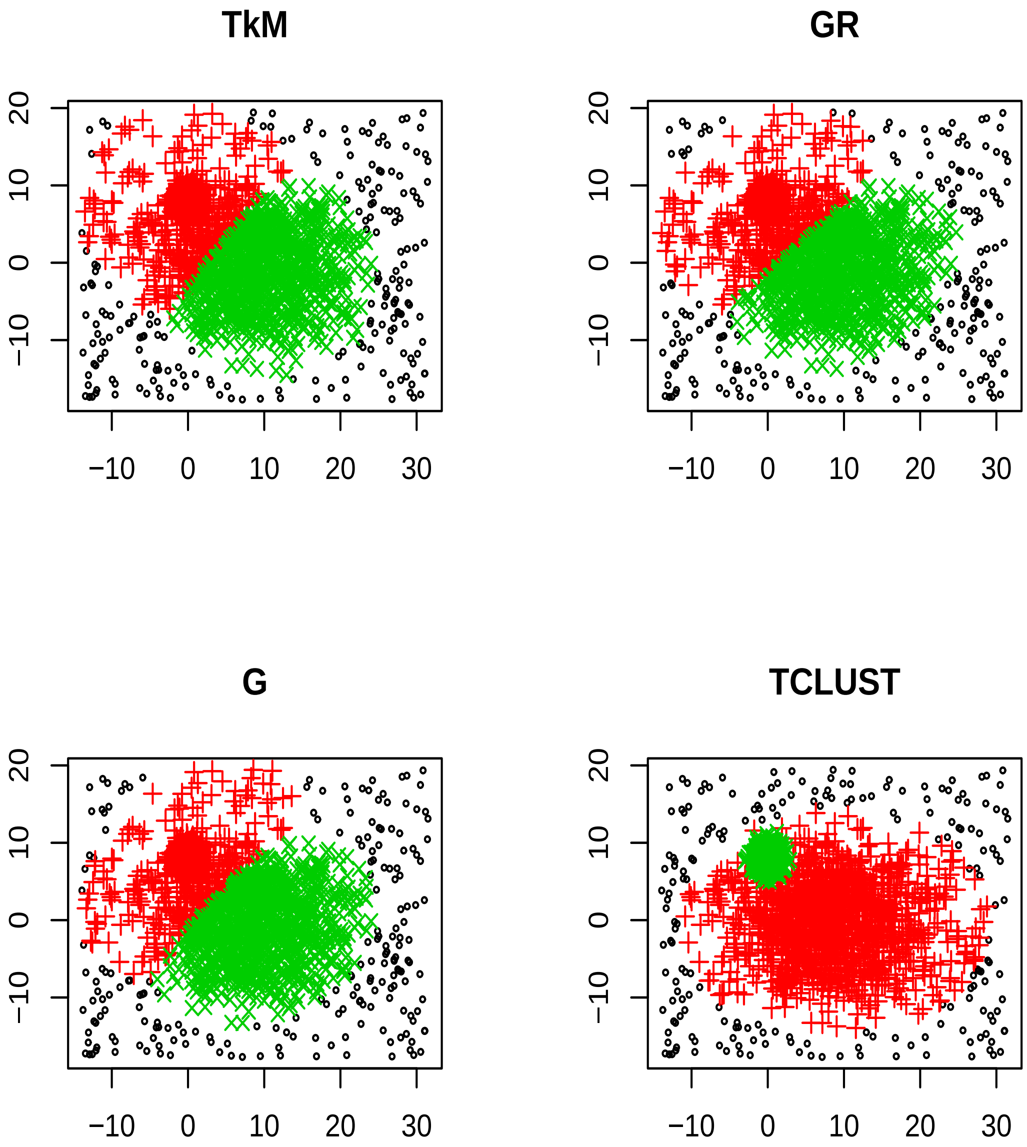

Figure 3: Clustering results when $k = 2$ and $\alpha = .1$ for a simulated data following the M5 scheme in the text with $p = 2$ and $\rho = 1/3$: Trimmed k-means (TkM); Gallegos and Ritter (GR); Gallegos (G) and the presented algorithm (TCLUST) with $c = 50$.

Lemma 8 *(i) If P satisfies condition (2.3) then there exists a constant $h > 0$ such that inequality (2.13) holds.*

(ii) If P is an absolutely continuous distribution satisfying condition (2.3), then there exists a constant $h' > 0$ such that (2.19) holds with probability 1 for n large enough.

Proof: (i) For a given probability measure P, the trimmed k-means problem is stated by the search of g points $\mu_1, ..., \mu_g$ in $\mathbb{R}^p$ and a Borel set B minimizing:

$$\min_{B:P(B)\geq 1-\alpha} \min_{\mu_1,\ldots,\mu_k} \frac{1}{P(B)} \int_B \inf_{1\leq j\leq k} \|x - \mu_j\| dP(x). \tag{4.1}$$

This paper contains an existence result which guarantees that problem (4.1) always has a solution, so it attains a minimum value that it can be denoted here by $V_{\alpha,k}$. Now, for every choice of θ, we can see that:

$$P\left[\sum_{j=1}^{k} z_j(\cdot;\theta)\|\cdot - \mu_j\|^2\right] \geq P\left[\sum_{j=1}^{k} z_j(\cdot;\theta) \inf_{1\leq j\leq k} \|\cdot - \mu_j\|^2\right] \geq (1-\alpha)V_{\alpha,k},$$

because $\cup_{j=1}^{k}\{x : z_j(x;\theta) = 1\}$ is a Borel set having probability greater or equal than $1 - \alpha$. Finally, we trivially see that $h := V_{\alpha,k} > 0$ whenever condition (2.3) holds for P.

(ii) The associated sample problem follows from (4.1) when considering $P = P_n$. Let us denote by $V_{\alpha,k}^n$ the minimum value attained by the objective function of this sample problem. Theorem 3.6 in Cuesta-Albertos et al. (1997) entails the consistency $V_{\alpha,k}^n \to V_{\alpha,k}$, P-a.e., and $V_{\alpha,k} > 0$ if P satisfies the condition (2.3). □

Lemma 9 *If $V_{\alpha,k}$ denotes the minimum value attained by (4.1) when k centers are allowed, then $V_{\alpha,k} \leq V_{\alpha,k+1}$.*

Proof: The proof of this result corresponds to Lemma 2.2. in Cuesta-Albertos et al. (1997). □

References

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

[2] Bock, H.-H. (2002), "Clustering methods: from classical models to new approaches," *Statistics in Transition*, **5**, 725-758.

[3] Celeux, G. and Govaert, A. (1992), "Classification EM algorithm for clustering and two stochastic versions",*Comput. Statit. Data Anal.,* **13**, 315-332

[4] Cuesta-Albertos, J.A., Gordaliza, A. and Matrán, C. (1997), "Trimmed k-means: An attempt to robustify quantizers," *Ann. Statist.,* **25**, 553-576.

[5] Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B*, **39**, 1-38.

[6] Dykstra, R.L. (1983), "An algorithm for restricted least squares regression," *J. Amer. Statist. Assoc.,* **78**, 837-842.

[7] Flury, B. (1997), *A first course in Multivariate Statistics,* Springer-Verlag New York.

[8] Fraley, C. and Raftery, A.E. (1998), "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The Computer J.,* **41**, 578-588.

[9] Gallegos, M.T. (2001), "Robust clustering under general normal assumptions", *preprint* avaliable at http://www.fmi.uni-passau.de/forschung/mip-berichte/MIP-0103.html.

[10] Gallegos, M.T. (2002), "Maximum likelihood clustering with outliers", in *Classification, Clustering and Data Analysis: Recent advances and applications*, K. Jajuga, A. Sokolowski, and H.H. Bock eds., 247-255, Springer-Verlag.

[11] Gallegos, M.T. and Ritter, G. (2005), "A robust method for cluster analysis," *Ann. Statist.*, **33**, 347-380.

[12] García-Escudero, L.A. and Gordaliza, A. (1999), "Robustness properties of k-means and trimmed k-means," *J. Amer. Statist. Assoc.,* **94**, 956-969.

[13] García-Escudero, L.A. and Gordaliza, A. (2007), "The importance of the scales in heterogeneous robust clustering," *Comput. Statit. Data Anal.,* **51**, 4403-4412.

[14] García-Escudero, L.A., Gordaliza, A. and Matrán, C. (1999), "A central limit theorem for multivariate generalized trimmed k-means," *Ann. Statist.,* **27**, 1061-1079.

[15] García-Escudero, L.A., Gordaliza, A. and Matrán, C. (2003), "Trimming tools in exploratory data analysis," *J. Comput. Graph. Statist.,* **12**, 434-449.

[16] Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming* **27**, 1-33.

[17] Hathaway, R.J. (1985), "A constrained formulation of maximum likelihood estimation for normal mixture distributions," *Ann. Statist*, **13**, 795-800.

[18] Hennig (2004), "Breakdown points for ML estimators of location-scale mixtures," *Ann. Statist.*, **32**, 1313-1340.

[19] Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London, New York, Toronto, Sydney, San Francisco.

[20] Maronna, R. and Jacovkis, P.M. (1974), "Multivariate clustering procedures with variable metrics," *Biometrics*, **30**, 499-505.

[21] McLachlan, G. and Peel, D. (2000), *Finite Mixture Models,* John Wiley Sons, Ltd., New York.

[22] Papadimitriou, C.H. and Steiglitz, K. (1982), *Combinatorial Optimization. Algorithms and Complexity*, Prentice-Hall, New Jersey.

[23] Rousseeuw, P.J. and Van Driessen, K. (1999), “A Fast Algorithm for the Minimum Covariance Determinant Estimator,” *Technometrics*, **41**, 212-223.

[24] Scott, A.J. and Symons, M.J. (1971), “Clustering based on likelihood ratio criteria,” *Biometrics,* **27**, 387-397.

[25] Van der Vaart, A.W. and Wellner, J.A. (1996), *Weak Convergence and Empirical Processes*, Wiley, New York.

Weights	Dimensions	Mixtures	TkM	GR	G	TCLUST
$\rho = 1/2$	$p = 2$	M1	.0150	.0146	.0150	.0152
		M2	.0450	.0198	.0203	.0200
		M3	.0430	.0429	.0200	.0205
		M4	.0879	.0679	.0645	.0199
		M5	.1484	.1461	.1466	.0346
	$p = 6$	M1	.0099	.0101	.0102	.0106
		M2	.0432	.0137	.0134	.0132
		M3	.0390	.0204	.0134	.0133
		M4	.1019	.0397	.0240	.0174
		M5	.1799	.1386	.0487	.0276
$\rho = 1/3$	$p = 2$	M1	.0137	.0137	.0138	.0135
		M2	.0480	.0184	.0186	.0183
		M3	.0425	.0437	.0200	.0202
		M4	.1054	.0664	.0694	.0212
		M5	.1993	.2002	.1957	.0400
	$p = 6$	M1	.0108	.0105	.0107	.0105
		M2	.0395	.0149	.0150	.0146
		M3	.0406	.0225	.0140	.0136
		M4	.1192	.0424	.0254	.0167
		M5	.2359	.1930	.1028	.0327

Table 1: Each entry represents the proportion of simulated observations that were misclassified by trimmed k-means (TkM), Gallegos-Ritter (GR), Gallegos (G), and the presented algorithm (TCLUST). Samples of size 2000 from five different 10%-contaminated mixtures of two p-variate normal distributions (M1, M2, M3, M4 and M5) were considered.