A Reweighting Approach to Robust Clustering

Francesco Dotto, Alessio Farcomeni*,
Luis Angel García-Escudero and Agustín Mayo-Iscar
Sapienza - University of Rome and University of Valladolid

Abstract

An iteratively reweighted approach for robust clustering is presented in this work. The method is initialized with a very robust clustering partition based on an high trimming level. The initial partition is then refined to reduce the number of wrongly discarded observations and substantially increase efficiency. Simulation studies and real data examples indicate that the final clustering solution is both robust and efficient, and naturally adapts to the true underlying contamination level.

Key Words: MCD; trimming; robustness.

1 Introduction

Trimming approaches in statistics provide robustness by considering outlier-free subsamples extracted from the data. Observations outside these subsamples are discarded. Examples include the Minimum Volume Ellipsoid, the Minimum Covariance Determinant, the Forward Search. See Rousseeuw (1985); Rousseeuw and van Driessen (1999); Butler *et al.* (1993); Riani *et al.* (2009); Cerioli *et al.* (2014).

The loss in fixing a trimming level α is not symmetric: if it is too low, outliers can completely spoil the solution. If it is too high, a loss of efficiency (which is usually less problematic than the first scenario) is incurred. For this reason, a preventive (higher than needed) trimming level is often considered. This could

*To whom correspondence should be addressed.

result in a high number of non-outlying observations which are wrongly trimmed, and loss of efficiency in subsequent statistical analyses. Carefully tuning the trimming level may be cumbersome in several applications, and the final results may be dependent on a subjective choice of this tuning parameter. A popular solution in robust statistics is to resort to reweighting methodologies.

To fix ideas, we start reviewing an example of the use of this approach in the simpler framework of multivariate robust location and scatter matrix estimation. Given a sample $\{x_1, ..., x_n\} \subset \mathbb{R}^p$ and T and S being any robust location and scatter estimators for this sample, robust Mahalanobis distances are defined as

$$d_i = d_S(x_i, T) = \sqrt{(x_i - T)' S^{-1} (x_i - T)}$$

for $i = 1, ..., n$. For instance, Rousseeuw and Leroy (1987) proposed considering T as the center of the ellipsoid with the smallest volume (MVE) that contains a fraction $1 - \alpha_0$ of the observations (a high trimming level $\alpha_0 \simeq 0.5$ was indeed proposed) and S as the scatter matrix determined by the same ellipsoid and multiplied by a correction factor to be consistent at multivariate Gaussian distributions. Alternatively, estimators T and S based on MCD estimation can be also used (defined from the fraction $1-\alpha_0$ of observations whose sample covariance matrix has the smallest possible determinant). Reweighting of each observation x_i is usually based on the Mahalanobis distance through $w_i = v(d_i)$, with $v(\cdot)$ being a non-increasing function. The weights w_i allow us to compute (one-step) reweighed location and scatter estimators which have good robustness performance and better efficiency behavior just by considering weighted sample means and weighted sample covariances. See Lopuhaa (1999) for a detailed discussion on the properties of reweighted estimators. The approach could be then iterated (e.g., Cerioli (2010)).

A very simple and widely applied approach is to use binary weights. Given initial T and S (robust) location and scatter matrices estimators and their associated Mahalanobis distances $d_i = d_S(x_i, T)$, we can simply use

$$w_i = 1 \text{ if } d_i \leq \sqrt{\chi^2_{p,\alpha_L}} \text{ and } w_i = 0 \text{ otherwise.} \tag{1.1}$$

We use the notation $\chi^2_{p,\beta}$ for a $1 - \beta$ quantile of the χ^2_p distribution and α_L is taken as a positive value close to 0. This allows to recover some of the wrongly trimmed observations, which could have not been taken into account when computing T and S, by assuming a normal distribution for the non-outlying part of data.

In this work we are focused on robust clustering. There are several approaches to robust clustering that are based on trimming (see, e.g., Cuesta-Albertos *et al.*

(1997), Hennig (2003), Gallegos and Ritter (2005), Neykov *et al.* (2007), García-Escudero *et al.* (2008) and other references included in García-Escudero *et al.* (2010)). For a detailed review, see Farcomeni and Greco (2015) and Ritter (2014). Robust clustering methods based on trimming return a fraction $1 - \alpha_0$ of outlier-free observations which are assigned to the different clusters. A high number of wrongly trimmed observations (due to the consideration of high initial preventive α_0 trimming levels) could be a major problem as researchers usually would like to assign as many observations as possible to a cluster. Failure to assign a clean observation to a cluster might be associated with practical consequences. For instance in marketing research not assigning a potential buyer to a his/her appropriate cluster is associated to loss of the revenue associated with the future transaction. Our proposal is to use reweighting ideas to reduce as much as possible, in a data driven fashion, the trimming proportion in robust clustering applications.

The proposed methodology is initialized with a large trimming level α_0 which -hopefully- guarantees the detection of a proportion $1 - \alpha_0$ of outlier-free observations in the most central regions of each cluster. These observations can be seen somehow as the *cores* of the clusters. Starting from the cores, the initial (high) trimming level α_0 is repeatedly decreased by including wrongly trimmed observations which are close to these cores, and updating estimates. In this iterative process better estimates of the cluster scatter matrices, cluster proportions, and the contamination level are consecutively obtained. Providing efficient estimates of these parameters is helpful to detect the outliers and, consequently, avoid their insertion in the final set of the clustered data eventually stopped prior to reaching the small trimming levels that would include outliers in estimation sets. Our proposal, to be better detailed below, can be seen as an extension of the procedure presented in García-Escudero and Gordaliza (2007) where the final trimming level had to be determined manually.

Figure 1 shows the result of applying the proposed methodology to two simulated datasets. The first one shown in panel (a.1) is the result of simulating a mixture of two normal components with no contamination. In panel (b.1) 10% of the observations are replaced by outlying data points. A more detailed description of the simulation scheme will be given in Section 3. Panels (a.2) and (b.2) show the results of TCLUST (García-Escudero *et al.*, 2008) with $\alpha_0 = 0.33$ trimming. Several wrongly trimmed observations can be seen, but also that the TCLUST procedure successfully identifies cluster cores. Finally, panels (a.3) and (b.3) show the results of the proposed methodology, which we name RTCLUST, which in both cases adapts well to the true underlying contamination.

In the previous example, we applied TCLUST (García-Escudero *et al.*, 2008)

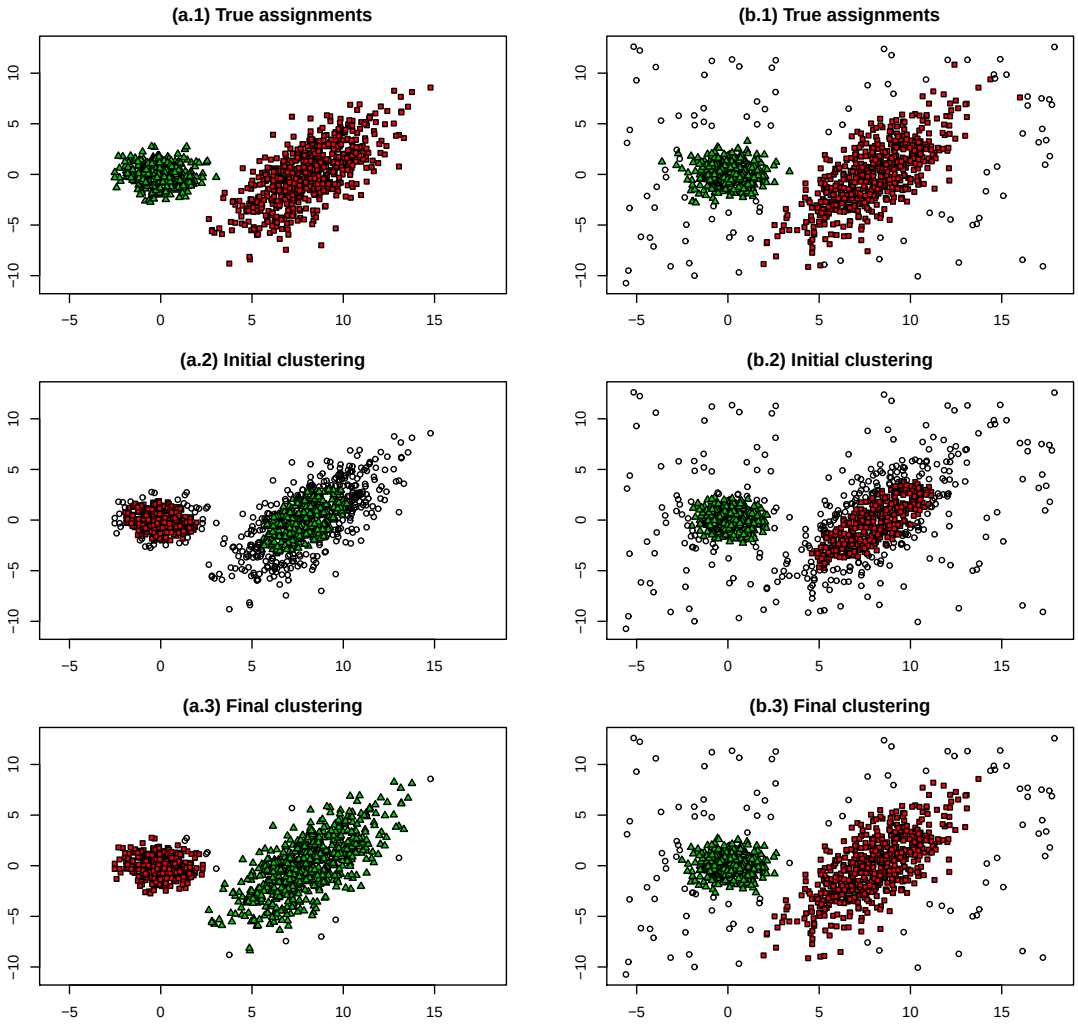

Figure 1: Two simulated data sets with their true assignments in (a.1) and (b.1). The result of TCLUST with $\alpha_0 = 0.33$ in (a.2) and (b.2). The final assignments obtained after applying the proposed methodology are given in (a.3) and (b.3). Noisy data and trimmed are denoted by $\circ$ in all graphs throughout the manuscript.

as the initial robust clustering technique to initialize the proposed methodology. TCLUST is a robust clustering method whose performance depends on three input parameters: the number of clusters k, the trimming proportion α, and a constraint on the maximal ratio of eigenvalues of scatter matrices c. The latter will be discussed in more detail below, and is used to guard against the occurrence of

spurious solutions (e.g., clusters with zero or infinite variance in some direction).It shall be underlined that the proposed reweighting methodology can be initialized from any other robust clustering method.

The underlying idea is that using an initial very robust estimator would make the procedure be resistant to a very high proportion of outliers (i.e., have a breakdown point of α_0). On the other hand, iteratively decreasing the trimming level would make the procedure almost as efficient as the non-robust counterparts. A similar idea but with a different rationale was proposed in Hardin and Rocke (2004), where an initial solution is improved based on a scaled F approximation to the distribution of Mahalanobis distances (see also Hardin and Rocke (2005)). We will compare in simulations below.

It is important to stress that while we will estimate the contamination level, and evaluate masking and swamping, what we are proposing is *not* a method to simultaneously perform robust clustering and outlier detection. We aim at obtaining robust and efficient estimates of partitions and model parameters. Outlier detection should then be based on robust estimators, but should be performed separately based on formal rules (see e.g. Cerioli and Farcomeni (2011) for a general discussion on this point).

The outline of the paper is as follows. The proposed methodology is presented in Section 2 together with some illustrative examples and guidelines about its practical use. A simulation study is given in Section 3. Examples on a benchmark data set and on an original study on exploring the status of food security in the world are given in Section 4. Finally, Section 5 gives concluding remarks.

2 Methodology

2.1 Proposed algorithm

Let us assume that the number of clusters k is known in advance but the proportion of observations π_j in each cluster is unknown and the true contamination level π_0 is also unknown. Non-outlying observations come from a mixture of k normally distributed components, and contamination might be present in our data. We also loosely make the assumption that the components are not too much overlapping.

We will consider a sequence of decreasing trimming levels $\alpha_0 > \alpha_1 > ... > \alpha_L$ with α_0 being an initial preventive (i.e., surely higher than needed) trimming level and α_L is a value close to 0 that can be interpreted in a similar fashion as parameter α_L in (1.1). Let us denote $\pi_1^l, ..., \pi_k^l$ the estimates of the cluster

proportions and π_{k+1}^l the proportion of contamination, for each trimming level α_l, with

$$\sum_{j=1}^{k+1} \pi_j^l = 1.$$

The center and scatter matrices estimates for the normally distributed components, in the iteration l, are denoted by $\mu_1^l, ..., \mu_k^l$ and $\Sigma_1^l, ..., \Sigma_k^l$.

By using this notation, the proposed methodology is described as follows:

1. *Initialization:* A very robust clustering is used to initialize, obtaining initial $\pi_1^0, ..., \pi_k^0, \pi_{k+1}^0, \mu_1^0, ..., \mu_k^0$ and $\Sigma_1^0, ..., \Sigma_k^0$. We propose considering TCLUST with a high trimming level α_0 as initializing method. Let $f(\cdot; \mu, \Sigma)$ denote the p.d.f. of the p-variate normal distribution. TCLUST is based on the double maximization of

$$\sum_{j=1}^{k} \sum_{i \in H_j} \log(\pi_j f(x_i; \mu_j, \Sigma_j)) \tag{2.1}$$

with respect to parameters ($\mu_j \in \mathbb{R}^p$, Σ_j p.s.d. matrices and $\sum_{j=1}^k \pi_j = 1$) and possible partition $H_0 \cup H_1 \cup ... \cup H_k$ of $\{x_1, ..., x_n\}$, with $\#H_1 + ... + \#H_k = [n(1 - \alpha_0)]$. A proportion α_0 of data is discarded in (2.1). Maximization of (2.1) is not a well-defined mathematical problem unless spurious solutions are avoided through constraints. We use a constraint on the estimated scatter matrices is required, that is,

$$\frac{\max_{j=1,...,k} \max_{h=1,...,p} \lambda_h(\Sigma_j)}{\min_{j=1,...,k} \min_{h=1,...,p} \lambda_h(\Sigma_j)} \leq c \tag{2.2}$$

where $\{\lambda_j(\Sigma)\}_{j=1}^p$ is the set of eigenvalues of matrix Σ and c is a fixed positive constant such that $c \geq 1$. We propose using a small c value to prevent us from detecting "spurious" clusters in this initializing step. For any α_0 and c, the optimal parameters solving the constrained maximization are considered as the initial $\pi_1^0, ..., \pi_k^0$, $\mu_1^0, ..., \mu_k^0$ and $\Sigma_1^0, ..., \Sigma_k^0$ parameters. Given that we are only considering the core of the clusters, it is not possible to obtain a reliable estimation of the contamination level and, thus, we prefer just initializing $\pi_{k+1}^0 = 0$.

Once again, we point out that other possible robust clustering methods can be applied for this initializing step. For instance, methods derived from the

maximization of (2.1) with different constraints on the Σ_j matrices and/or removing the π_j weights can be used. See Cuesta-Albertos *et al.* (1997), Hennig (2003), Gallegos and Ritter (2005) or Neykov *et al.* (2007) among others. If the π_j weights are removed then we may consider $\pi_1^0 = ... = \pi_k^0 = 1/k$ to initialize the procedure.

2. *Reweighting process:* Consider $\alpha_l = \alpha_0 - l \cdot \varepsilon$ with $\varepsilon = (\alpha_L - \alpha_0)/L$ for $l = 1, ..., L$

   2.1 *Update proportions:* Given $\pi_1^{l-1}, ..., \pi_k^{l-1}, \pi_{k+1}^{l-1}, \mu_1^{l-1}, ..., \mu_k^{l-1}$ and $\Sigma_1^{l-1}, ..., \Sigma_k^{l-1}$ from the previous step, let us consider

   $$D_i = \min_{1 \le j \le k} d^2_{\Sigma_j^{l-1}}(x_i, \mu_j^{l-1}) \tag{2.3}$$

   and sort these values as $D_{(1)} \le ... \le D_{(n)}$. Take the sets

   $$A = \{x_i : D_i \le D_{([n(1-\alpha_l)])}\} \text{ and } B = \{x_i : D_i \le \chi^2_{p,\alpha_L}\}$$

   (note that α_L is used to define the set B). Now, use the distances in (2.3) to obtain a partition $A \cap B = \{H_1, ..., H_k\}$ with

   $$H_j = \left\{ x_i \in A \cap B \ : \ d_{\Sigma_j^{l-1}}(x_i, \mu_j^{l-1}) = \min_{q=1,...,k} d_{\Sigma_q^{l-1}}(x_i, \mu_q^{l-1}) \right\}.$$

   We estimate, at this stage, the contamination level as

   $$\pi_{k+1}^l = 1 - \frac{\#B}{n}.$$

   If $n_j = \#H_j$ and $n_0 = n_1 + ... + n_k$ (notice that n_0 is not necessarily equal to $[n(1-\alpha_l)]$) then the proposed estimations at this stage of the cluster weights are

   $$\pi_j^l = \frac{n_j}{n_0}\big(1 - \pi_{k+1}^l\big). \tag{2.4}$$

   2.2 *Update locations and scatters:* We update the cluster centers by taking μ_j^l equal the sample mean of the observations in H_j. To update the scatter matrices estimates, we start from S_j^l equal to sample covariance matrix of the observations in H_j.

Additionally, covariance estimates need to be corrected by considering correcting factor defined as

$$c_j = \left(\eta_{\frac{n_0}{n(1-\pi^l_{k+1})}}\right)^{-1} \text{ if } \frac{n_0}{n(1-\pi^l_{k+1})} < 1$$

and

$$c_j = 1 \quad \text{if } \frac{n_0}{n(1-\pi^l_{k+1})} \geq 1$$

where $\eta_\beta = P\big(\chi^2_{p+2} \leq \chi^2_{p,\beta}\big)/\beta$ and $\beta = \#H_j/n\pi_j$.
We finally update the scatter matrices as

$$\Sigma^l_j = S^l_j \cdot c_j.$$

3. *Output of the algorithm:* $\mu^L_1, ..., \mu^L_k$ and $\Sigma^L_1, ..., \Sigma^L_k$ are the final parameters estimates for the normal components. From them, final assignments are done by computing

$$D_i = \min_{1\leq j\leq k} d^2_{\Sigma^L_j}(x_i, \mu^L_j),$$

for $i = 1, ..., n$. Observations assigned to cluster j are those in H_j with

$$H_j = \left\{ x_i : d_{\Sigma^L_j}(x_i, \mu^L_j) = \min_{q=1,...,k} d_{\Sigma^L_q}(x_i, \mu^L_q) \text{ and } D_i \leq \chi^2_{p,\alpha_L} \right\}$$

and the trimmed observations are observations not assigned to any of these H_j sets (i.e., those observations with $D_i > \chi^2_{p,\alpha_L}$).

Step 2.1 is targeted at keeping outliers outside $A \cap B$, while increasing the trimming size in a controlled fashion. Alongside, better parameter estimates are obtained by increasing the active sample size. In the step 2.2 we use well-known correction factors (see, e.g. Liu *et al.*, 1999) to inflate the covariance matrix estimates based on trimmed data. These guarantee consistency at the normal model. components. At each stage the fraction of observations in the central region of group j is $n_j/n\pi^l_j = n_0/(n(1-\pi^l_{k+1}))$, where $n\pi^l_j$ is an estimate of the total number of observations in group j.

The proposed RTCLUST procedure is not computationally intensive at all. The most time-consuming part is computation of the initial estimates, which is done only once.

Remark 1 *More sophisticated rules for discarding outliers, for instance, based on using the Beta distribution or multiple testing corrections could have been tried (Cerioli, 2010; Cerioli and Farcomeni, 2011). However, for sake of clarity of presentation, we have preferred the simpler use of a rule just based on* χ^2_{p,α_L}. *There is still room for improvement regarding better detection of outlying observations.*

Remark 2 *Sometimes, we could be interested in forcing some "a priori" constraints like those in (2.2) to the final estimated clusters scatter matrices. In this case, constraints can be forced by truncating the scatter matrices eigenvalues in the updating step 2.2 as done in Fritz* et al. *(2013).*

2.2 Illustrative examples

The two component normal mixture shown in panels (b.1) of Figure 1 account for 36% and 54% of the observations, respectively, while a 10% of not "very overlapped" contamination is added. The scatter matrix for the first component is Σ_1 equal to the identity matrix and Σ_2 is a scatter matrix with $|\Sigma_2| = 20$ and eigenvalues equal to 11.708 and 1.708. This means that the "true" eigenvalue ratio for these two scatter matrices is equal to 11.708. A more detailed description of the process generating this data set will be given in Section 3. We will use this data set in order to illustrate the lack of dependence of the final solution on the initializing trimming level α_0 and on the initial value of the restriction factor c when TCLUST is used as initializing procedure. Figure 2 shows the evolution of the determinants of the scatter matrices, i.e. $\{|\Sigma_j^l|\}_{l=0}^L$ for $j = 1, 2$ in panel (a), and the evolution of the estimated contamination level and estimated cluster sizes, i.e. $\{\pi_j^l\}_{l=0}^L$ for $j = 0, 1, 2$ in (b). These evolutions are studied for different values of $\alpha_0 = 0.3$, 0.25, 0.2 and 0.15 and it is always considered the same (wrong) eigenvalue ratio constraint value $c = 5$ for the TCLUST method as initializing procedure. We can see that the final output is not very dependent on the initializing trimming level and that the output estimated parameters are very close to the true ones we want to estimate (i.e., $|\Sigma_1| = 1$ and $|\Sigma_2| = 20$ for the cluster scatter matrices determinants and $\pi_0 = 0.1$, $\pi_1 = 0.36$ and $\pi_2 = 0.54$ for the contamination level and cluster sizes).

Analogously, the same type of study was made to analyze the possible dependence on the initializing choice of c. The results are shown in Figure 3 where c values equal to 1, 10 and 20 were chosen. Recall that the true eigenvalue ratio for the considered scatter matrices was exactly equal to 11.708 (which is not equal to any of the c initializing values tried). We can see again that the obtained results

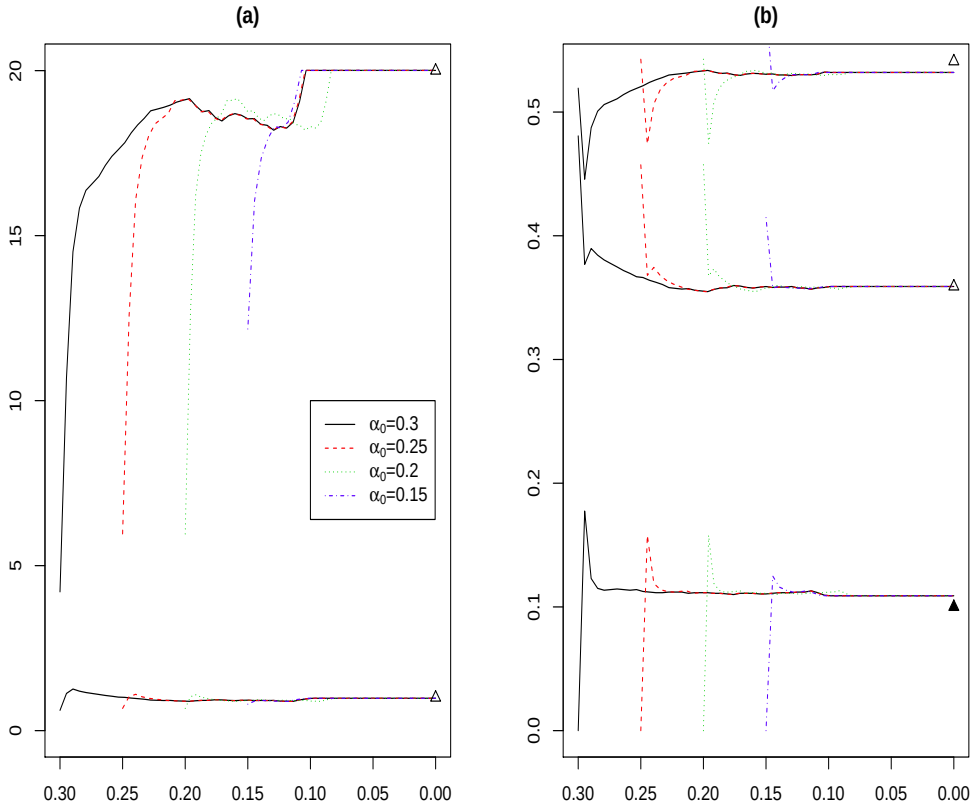

Figure 2: Evolution of $|\Sigma_j^l|$ in (a) and of π_j^l in (b) for different initial α_0 values ($\alpha_0 = .3, .25, .2$ and $.15$) for the data set shown in Figure 1 (b.1). The up-triangle symbols are the true parameters to be estimated.

are accurate and that they are not very dependent on the initial c value.

It is also important to note, in Figure 2 and Figure 3, that no great changes are noticeable in the estimated parameters when the procedure approximately reaches the true contamination level. This is because, we count on quite accurate estimators of the parameters of the normal distributions components throughout μ_j^l and Σ_j^l when $\alpha_l \approx 0.1$. Due to their effect the set $A \cap B$ defined in Step 2.1 remains essentially the same and equal to the set having all the regular (non-noisy) observations already included. On the other hand, one-step procedures only take into account the information from truncated sub-samples corresponding to central regions in the normal components. From this central regions, it is not so easy to have very accurate parameters estimations for the normal components parameters.

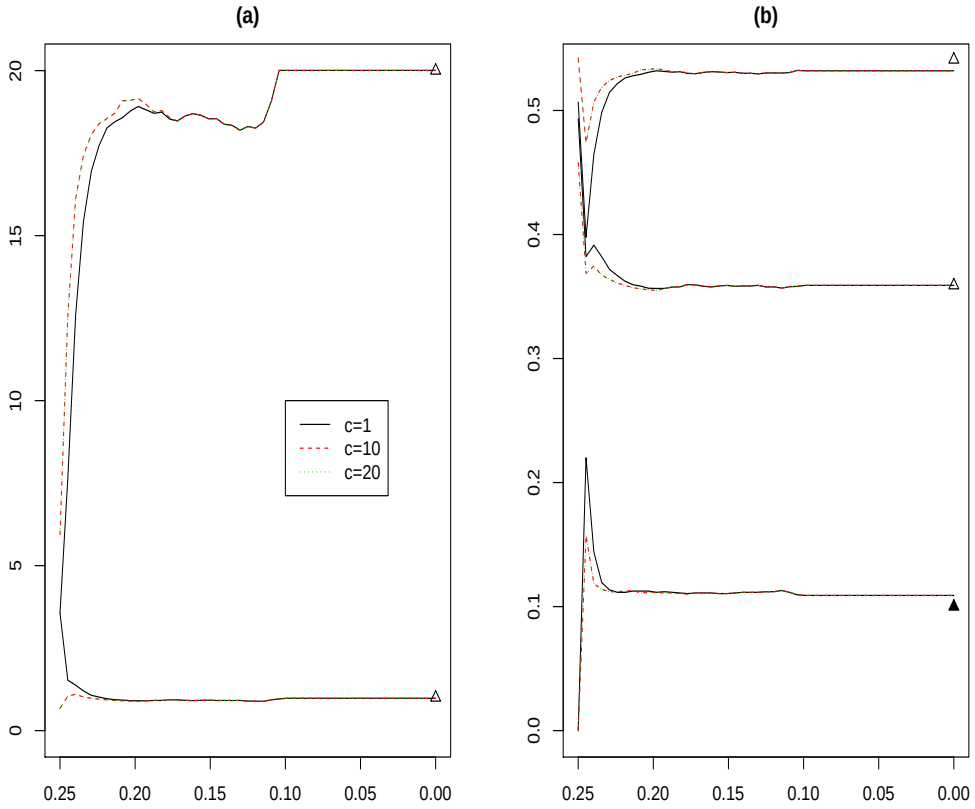

Figure 3: Evolution of $|\Sigma_j^l|$ in (a) and of π_j^l in (b) for different initial c values ($c = 1$, 10 and 20 while the true c needed was 11.71) for the data set shown in Figure 1 (b.1). The up-triangle symbols are the true parameters to be estimated.

To reinforce our previous claims, we will illustrate the advantages of the proposed iterative trimming procedure with respect to one-step reweighting approaches even in the $k = 1$ case. When $k = 1$, the reweighted MCD is clearly one of the most popular robust location and scatter estimator. After considering an initial large trimming level α_0, reweighting is done to increase efficiency as described in Section 1.

Figure 4 is based in a simulated data set of size $n = 1000$ generated from a bivariate normal distribution accounting for 73% of the data (the bulk of data), a 24% amount of pointwise contamination placed at $(4, 8)$ (labeled with an "arrow" symbol) and 3% of background contamination. Figure 4,(b) shows the result of applying the reweighted MCD approach in Section 1 by using the "robustbase"

package in R available in the CRAN repository with the default initial trimming level $\alpha_0 \simeq 0.5$ and $\alpha_L = 0.01$ and the function "tolEllipsePlot" (from "robustbase") to plot the 0.99 tolerance ellipses (the classical and the MCD-based robust ones). Despite there exists a "good" initial sub-population including more than half of the observations, the final estimation is very distorted by the added pointwise contamination as can be seen in 4,(b). On the other hand, Figure 4,(a) shows how the proposed iterative trimming resists very well this pointwise contamination.

Finally, an additional important parameter for the proposed methodology is α_L. In all the shown illustrative examples, the same $\alpha_L = 0.01$ has been take. The α_L parameter has to do with the quantile in the χ^2_p distribution and it plays the same role as in all analogous reweighting methods. For instance, $\alpha_L = 0.01$ means that around 1% of the observations are wrongly discarded when we have normal components without contamination. The smaller the α_L the lesser is the number of proportion of wrongly trimmed observations but higher is the risk of incorporating near outlying observations.

3 Simulation study

We now study the performance of the previously described procedure when applied to several (contaminated) mixtures of Gaussian distributions. Additionally, we detail how the data sets used in previous sections have been simulated in the illustrative examples.

The non-outlying part of the dataset comes from a mixture of two p-variate normal distributions $\pi_1 N(\mu_1, \Sigma_1) + \pi_2 N(\mu_2, \Sigma_2)$ with centers $\mu_1 = (0, 0, 0, ..., 0)'$ and $\mu_2 = (8, 0, ..., 0)'$ and covariance matrices

$$\Sigma_1 = I_p \text{ and } \Sigma_2 = \sqrt[p]{\lambda} \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & 2 & \cdots & 2 \\ 1 & 2 & 3 & 3 & \cdots & 3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & 4 & \cdots & p \end{pmatrix}.$$

This means that $|\Sigma_1| = 1$ and $|\Sigma_2| = \lambda$.

To generate outliers we fix an hypercube where each dimension includes the range of the non-contaminated data. Outlying observations are generated uniformly within this hypercube, but outliers with squared Mahalanobis distances

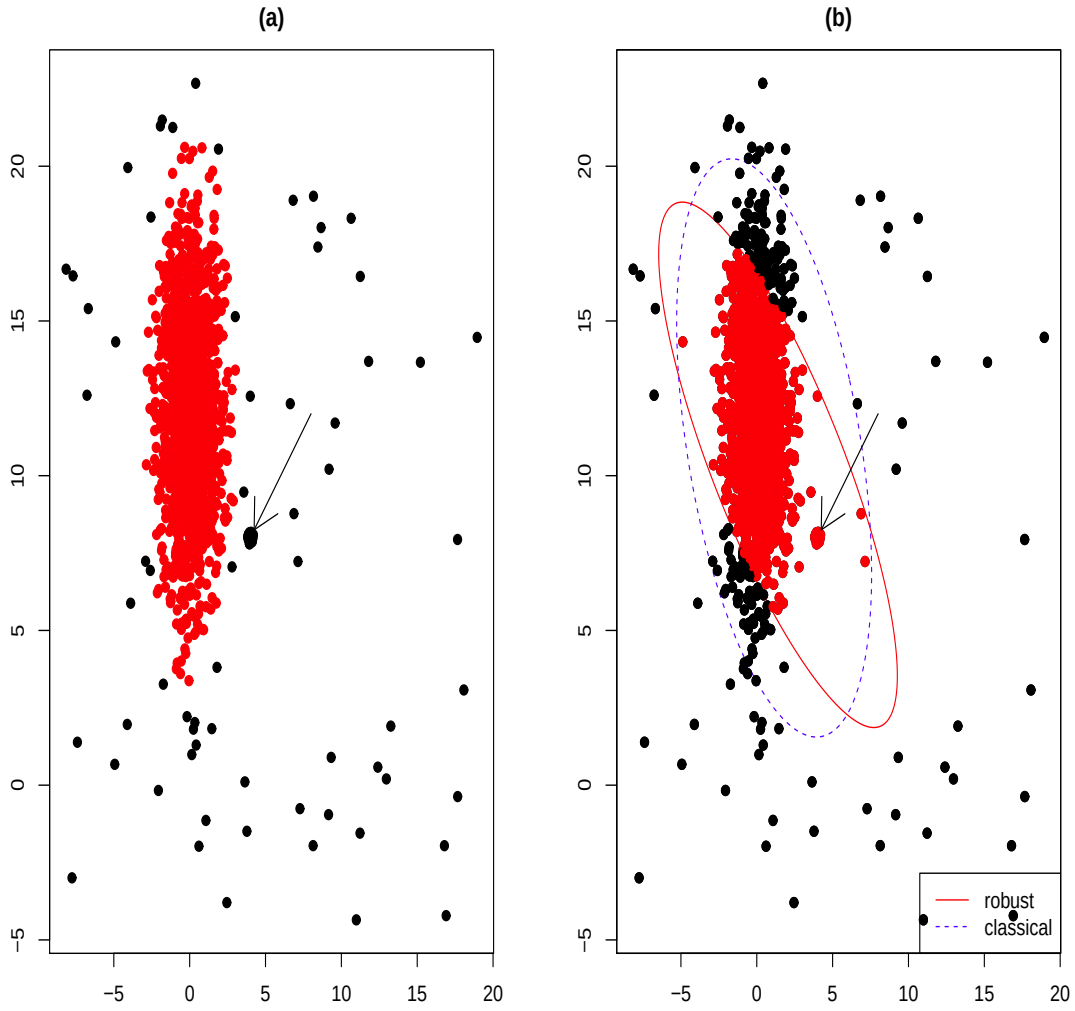

Figure 4: (a) The proposed iterative reweighting procedure when $k = 1$ started from $\alpha_0 = 0$ and $\alpha_L = 0.01$ (b) The (traditional) reweighted MCD started from $\alpha_0 = 0$ and $\alpha_L = 0.01$. Trimmed points are the black points.

from μ_1 and μ_2 (using Σ_1 and Σ_2) smaller than $\chi^2_{p,\nu}$ are discarded. The operation is repeated until the desired proportion of ε outliers have been obtained. The parameter ν controls how far away contaminated data points are.

We generate data sets of size $n = 1000$ under all possible combinations of the following scenarios:

- Three data dimensions: $p = 2, 4$ and 6
- Three contamination levels $\varepsilon = 0.10$, 0.05, and zero.
- Two scales $\lambda = 1$ and 5
- Balanced clusters $\pi_j = 0.5$ for $j = 1, 2$ and unbalanced clusters $\pi_1 = 0.4$ and $\pi_2 = 0.6$
- Two ν values, $\nu = 0.01$ and $\nu = 0.005$
- Two types of contamination: a symmetric one obtained sampling from a uniform distribution in the hypercube defined by the range of the non-contaminated part of the data and an asymmetric one obtained by sampling from a uniform distribution defined on $[-3, 0]\times[-7, -2]\times[-2, 2]^{p-2}$, which is closer to the second cluster than to the first.

The case $\varepsilon = 0$ is used to evaluate efficiency of the proposed methodology when applied to clean data.

Regarding the illustrative examples in Figure 1 we generated two datasets once from a bivariate normal distribution, fixing $\lambda = 20$, $\pi_1 = 0.4$ $\pi_2 = 0.6$, with symmetric contamination and $\nu = 0.01$. A contamination level $\varepsilon = 0$ was used in (a.1) and $\varepsilon = 0.10$ in (b.1).

We compare the performance of the following robust clustering proposals:

- `rtclust33` and `rtclust20`: The proposed iterative reweighting approach started from TCLUST with initial trimming levels $\alpha_0 = 0.33$ and $\alpha_0 = 0.2$
- `HR33` and `HR20`: a one-step version of the procedure by Hardin and Rocke (2004) started from TCLUST with initial trimming levels $\alpha_0 = 0.33$ and $\alpha_0 = 0.2$
- `HR-it33` and `HR-it20`: the iterated and adapted version of Hardin and Rocke (2004) started from TCLUST with initial trimming levels $\alpha_0 = 0.33$ and $\alpha_0 = 0.2$
- `tclust33`, `tclust20`, `tclust10` and `tclust05`: TCLUST with fixed trimming levels $\alpha_0 = 0.33, 0.2, 0.1$ and 0.05

The same value $\alpha_L = 0.01$ was used for RTCLUST and Hardin and Rocke's methods. For iterative procedures we fixed $L = 20$. The TCLUST procedure was included with with trimming levels which could be higher or the correct one. The same eigenvalue restriction factor $c = 12$ is always applied when using TCLUST (in the initialization of RTCLUST and in the direct application of TCLUST). Note that $c = 12$ could be smaller or larger than the true eigenvalue ratio, depending on p and λ.

The Hardin and Rocke's methods are clustering algorithms based on the MCD philosophy. These methods are going to be initialized in this simulation study with exactly the same TCLUST robust clustering initial solution used for RTCLUST. Indeed Hardin and Rocke (2004) commented in their work that "any" robust clustering solution can be used and we have seen that TCLUST always provides quite sensible initial solutions for all the considered data sets in the simulation study. In fact, TCLUST always removes all noisy observations (together with others wrongly trimmed ones) with these high trimming levels ($\alpha_0 = 0.33$ and 0.2). Let $\mu_1^0, ..., \mu_k^0$, $\Sigma_1^0, ..., \Sigma_k^0$ and $H_0^0, H_1^0, ..., H_k^0$ being the solution obtained by applying the TCLUST method. The Hardin and Rocke's approach proposes cut-off values to declare outliers based on the approximation

$$\frac{k_j(m_j - p + 1)}{pm_j} d^2_{\Sigma_j^0}(x_i, \mu_j^0) \sim F_{p,m_j-p+1}, \tag{3.1}$$

where $k_j = \eta_{\beta_j}$ is a correction factor (as that used in Section 2.1) with

$$\beta_j = \tilde{h}_j/n_j \text{ for } \tilde{h}_j = \#H_j^0$$

and

$$n_j = \#\big\{x_i : d_{\Sigma_j^l}(x_i, m_j^l) = \min_{q=1,...,k} d_{\Sigma_q^l}(x_i, m_q^l)\big\}$$

and m_j is the approximated degrees of freedom for the associated Wishart distribution (see details in Hardin and Rocke (2004) and Hardin and Rocke (2005)). "HR33" and "HR20" apply directly the cut-off values in (3.1) to the observations in the H_j^0 sets while "HR-it33" and "HR-it20" refine these H_j^0 sets until stabilization by applying the iterative steps described in Section 3.3 of Hardin and Rocke (2004).

For all the 96 different data scenarios, we generated the data 500 times and evaluated the performance of the methods in terms of:

- Mean Square Error for estimation of the mean vectors μ_1 and μ_2, indicated in the plot with MSE_μ.

- Mean Square Errors associated to the logarithm of the eigenvalue ratio, indicated in the plots with MSE_Σ. We decided to report the error associated to this quantity since this ratio is forced in the initialization step to be smaller than a fixed constant $c = 12$ to avoid spurious maximizers. Nevertheless, as already commented, this is not necessarily the true eigenvalue ratio and we want to see how far the final estimated ratio is with respect the true one given that the proper estimation of the cluster scales play a key role in the detection of outliers.

- The estimated contamination level $\hat{\varepsilon}$.

- *Swamping:* the proportion of non outlying observations that are trimmed

- *Masking:* the proportion of outliers that are not trimmed

Figures 5 and 6 summarize the simulation results obtained when $\varepsilon = 0.05$ and 0.1, respectively. Figures are separated in five row panels, one for each performance measure, and three column panels, one for each data dimensionality p. Given that there are several settings, in order to summarize the results in a concise way we do not distinguish among them further and just report the average performance measures all together. Note that some values exceed the scale of the plots, as identified by the upward triangle symbols.

The iterative reweighing procedure efficiently estimates the mean vector and the covariance matrix in every data scenario. In all cases we see small MSE values, and not much variability, meaning that results do not depend on the simulation setting considered. The MSE values are smaller than those obtained when applying TCLUST with large trimming values as 0.20 and 0.33. Moreover, MSE is even slightly better than what obtained with an oracle TCLUST whose trimming level is exactly equal to the true contamination level ε. This happens for two reasons. The first is that reweighting can adapt well to the positioning of the outliers, therefore flexibly trimming more or less as needed within each replicate. The second is that TCLUST is based on a sometimes wrong eigenvalue ratio constraint value $c = 12$. RTCLUST does not have further constraints and therefore can exceed this value when needed.

As far as estimation of the contamination level $\hat{\varepsilon}$ is concerned, RTCLUST provides very stable results in all simulation scenarios, with a systematic slight overestimation of ε. On the other hand, the procedures based on Hardin and Rocke approach may underestimate contamination levels in a remarkable way. The swamping proportion is small for all reweighting approaches but masking

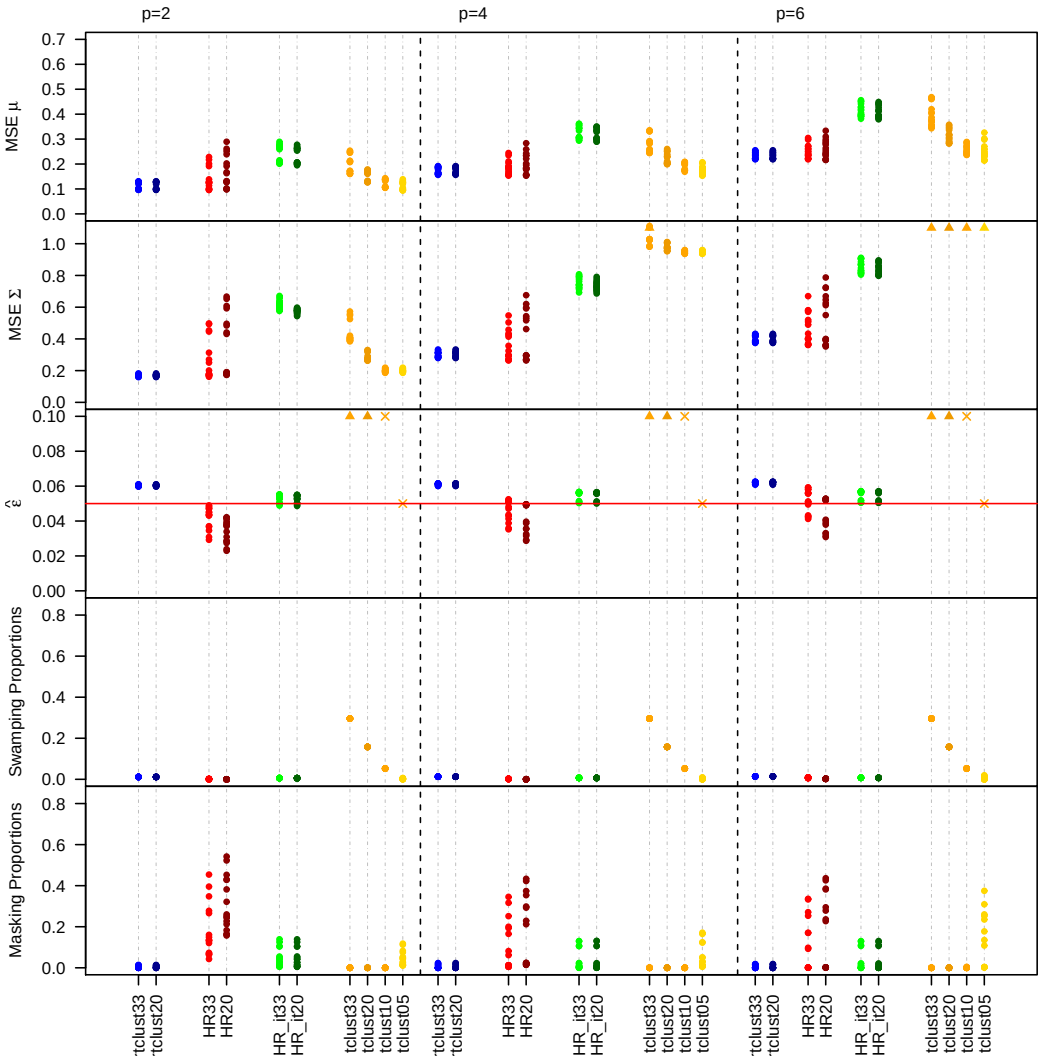

Figure 5: Results when $\varepsilon = 0.05$. Every procedure is labeled as explained in the text. Values appearing in the Figure that are fixed in advance (e.g the trimming level for the `tclust` method) are plotted with the symbol "$\times$" while when the considered value exceeds the scale of the plot we used a "▲"

proportions can be very high in some scenarios with Hardin and Rocke's proposals. Underestimation of the contamination level is clearly more harmful than overestimation, as outliers included in the estimation set might break down the estimates. We believe that the problem with the Hardin and Rocke's approach is

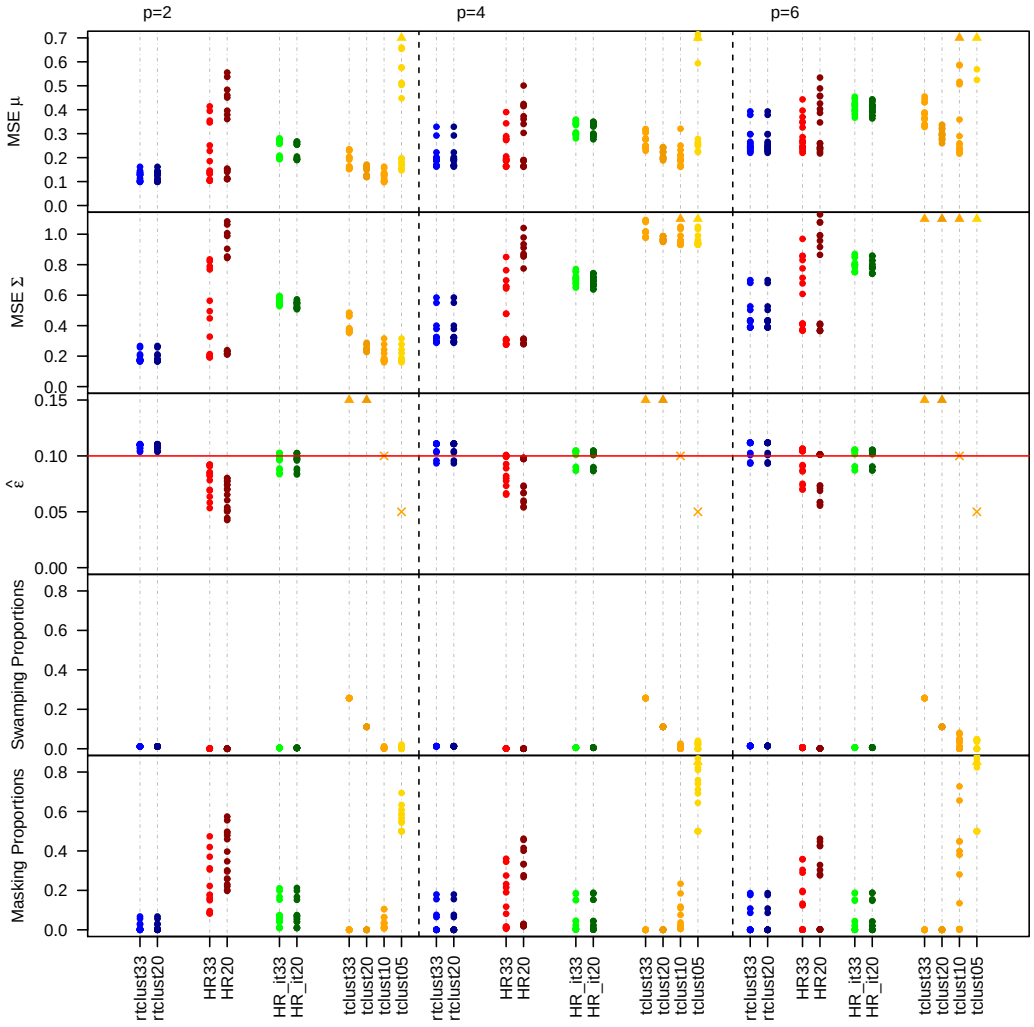

Figure 6: Results when $\varepsilon = 0.10$. Every procedure is labeled as explained in the text.

within the correction factor, which exploits an estimator of the fraction of observations in each cluster. The latter might not be resistant to outliers in our experience.

We end this section by comparing the performance of these methods in the non contaminated $\varepsilon = 0$ case. This is reported in Figure 7.

We can see that the iteratively reweighting approach exhibits a very good performance in terms of providing small MSE values. We can also see that the (non-iterated) Hardin and Rocke's approaches are very competitive in this non-

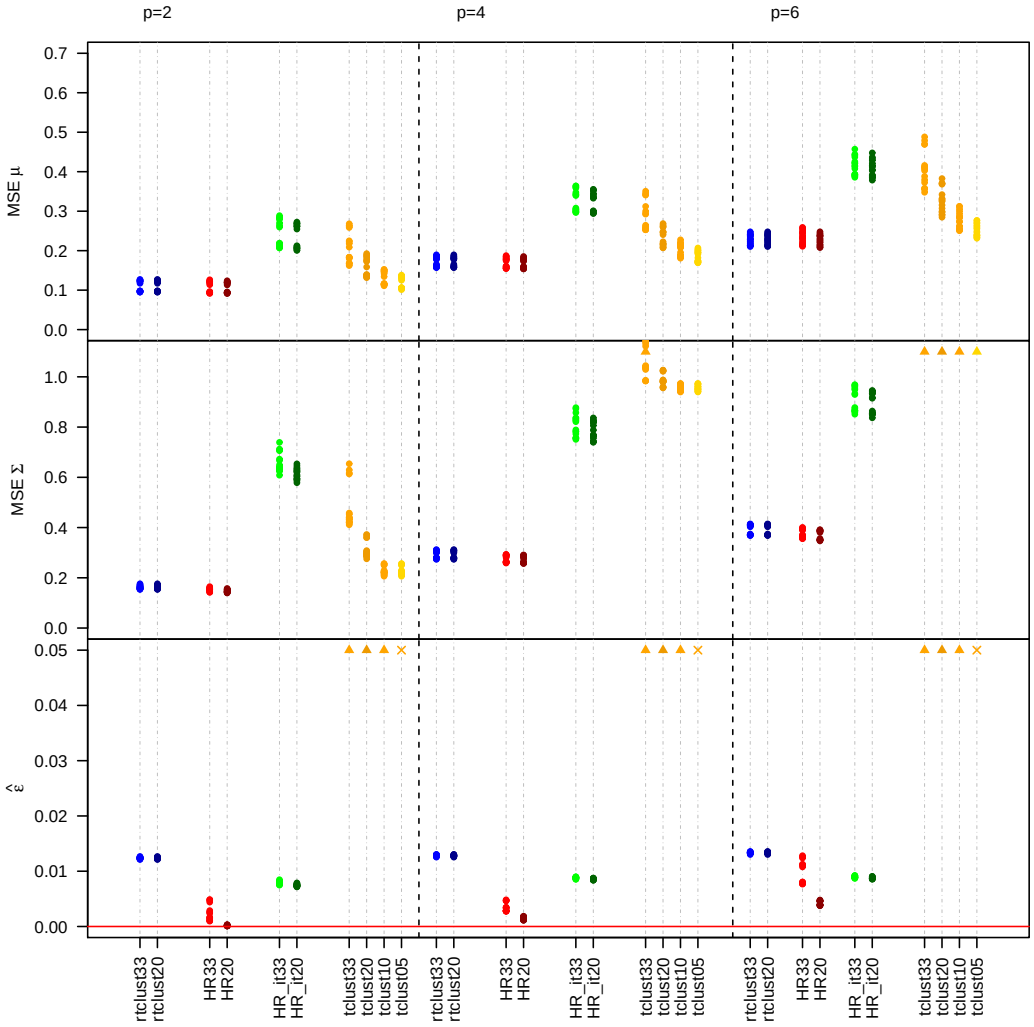

Figure 7: Simulation results study under no contamination ($\varepsilon = 0$).

contaminated $\varepsilon = 0$ case. RTCLUST wrongly discards a limited proportion of observations, about 1%. This is not so surprising as $\alpha_L = 0.01$ in this section.

4 Real data examples

4.1 Swiss Bank Notes

In this section we apply the proposed iterative reweighting approach to the 6-dimensional "Swiss Bank Notes" data set presented in Flury and Riedwyl (1988) which describes certain features in 200 printed Swiss 1000-franc bank notes divided in two groups: 100 genuine and 100 counterfeit notes. This is a well known benchmark data set. In Flury and Riedwyl (1988), it is pointed out that the group of forged bills is not homogeneous since 15 observations arise from a different pattern and are, for that reason, outliers. Figure 8,(a) shows a scatterplot of the fourth ("Distance of the inner frame to lower border") against the sixth variable ("Length of the diagonal") with the classification of bills given in Flury and Riedwyl (1988) by using symbols "G" for the genuine bills and "F" for the forged ones. The previously commented 15 "anomalous" forged bills are surrounded by circles in this graph. Figure 8,(b) shows the results of applying TCLUST with a high trimming level $\alpha_0 = 0.33$ and $c = 12$. We can see that all these 15 outlying points are successfully discarded and observations in the "cores" of the genuine and forged bills are correctly found. However, due to the use of this high trimming level, many observations are also discarded apart from the 15 clear outliers. We have surrounded these "probably wrongly" trimmed observations by square symbols. Finally, Figure 8,(c) shows the results of applying the proposed iterative trimming approach starting from the TCLUST's solution in (b) with $\alpha_L = 0.001$. We can see that the proportion of "probably wrongly" trimmed observations reduces to 4 (also surrounded by square symbols). One of these 4 observations is a genuine bill which clearly exhibits certain anomalous behavior in these two plotted variables and we could also see that the other 3 (wrongly) trimmed observations analogously seems to exhibit slight deviations in some of the (non-plotted) variables.

We have used a smaller $\alpha_L = 0.001$ value in this real data example. If $\alpha_L = 0.01$ then 7 wrongly trimmed observations (instead of 4) are obtained. As stated in the introduction, RTCLUST is not an outlier detection method. Estimates of the clusters location and scatter matrices do not change notably with the choice of α_L, which makes RTCLUST a good choice for robust clustering and parameters estimation for this data set. Formal rules for outlier detection could be then based on RTCLUST robustly estimated parameters.

We conclude with an analysis based on $k = 1$. As half of the bank notes are genuine ones, one could think that setting $k = 1$ and trimming 50% of the

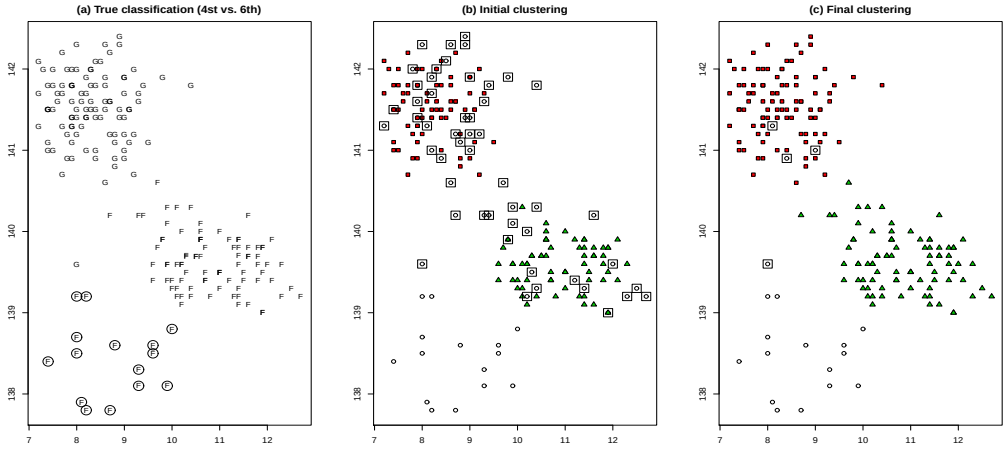

Figure 8: Fourth against the sixth variable of the Swiss Bank Notes data set. (a) `G` stands for genuine bills, `F` for forged ones and 15 bills listed in Flury and Riedwyl (1988) as anomalous ones are surrounded by $\circ$ symbols. (b) The initial TCLUST solution with $\alpha_0 = 0.33$ (c) Final solution when applying the proposed iterative approach. Trimmed observations not coinciding with those in Flury and Riedwyl's list are surrounded by $\square$ symbols.

observations would identify them. Use of TCLUST with $k = 1$, $?_0 = 0.5$ and $c = 12$ (which is the default value of c fixed in the `tclust` package in Fritz *et al.* (2012)) successfully identifies 96 genuine bills (out of the 100 non-trimmed observations). The standard application of RTCLUST, started from this TCLUST solution with $\alpha_0 = .5$ and $\alpha_L = 0.001$, returns a final set with 102 notes which includes 98 genuine bills. Therefore, RTCLUST is well-suited to discover, in an automatized way, the genuine observations. On the other hand, use of MCD through the well-known `robustbase` package with $\alpha = 0.5$ returns 103 bills (i.e., the largest integer less than or equal to $(n + p + 1)/2$ as the "best" subset found and used for computing the raw estimates. Surprisingly, only 42 out of these 103 observations are genuine ones. Additionally, things become even worse when applying the default consistency correction factor for the covariance matrix estimation and the use of (1.1) with $\alpha_L = 0.025$, as this finally leads to 176 notes used for robust estimation.

4.2 Food Security Data

In this section we apply the proposed procedure to an original and very recent data set on an investigation of the status of food insecurity in the world in 2014. Food security is defined by the Committee on World Food Security of United Nations as when people

> *at all times, have physical, social and economic access to sufficient safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life.*

For reviews see Godfray *et al.* (2010) and Jones *et al.* (2013).

In 2014, the Gallup Organization conducted a World Poll based on a questionnaire given to a representative sample of about 1000 adults from each of several areas in the world. Areas mostly correspond to countries, while in some cases countries have been split in different areas (e.g., Congo has been split in two, Brazzaville and Kinshasa areas). The Gallup World Poll (GWP) answers are then routinely summarized by Gallup into thematic indeces, which are evaluated for each polled subject and could be used to make comparisons across countries. A detailed description of the GWP can be found at
`http://www.gallupworldpoll.com/content/24046/About.aspx`.
In 2014 the usual GWP questionnaire has been augmented with eight questions, in partnership with the Voices of the Hungry (VoH) project of the Food and Agriculture Organization (FAO) of the United Nations. These questions were aimed at evaluating specifically a new index, the Food Insecurity Experience Scale (FIES). A very challenging issue that has been tackled by the VoH team is the standardization of the FIES score over different cultures and languages. Details on how this was performed are given in Cafiero *et al.* (2016). A more general discussion is provided in Ballard *et al.* (2013); Cafiero *et al.* (2014).

We have obtained the individual standardized FIES scores, in addition to the rest of GWP data for 2014. Data have been aggregated at country level, taking sampling weights into account. Our aim is to cluster and identify outlying countries, and secondly to evaluate the discriminating power of FIES after taking into account information collected by the other indices. Our final data set, aggregated over subjects, is therefore made of $n = 127$ countries and $p = 6$ indeces. These are Food Insecurity Experience Scale, Civic Engagement Index, Struggling Index, Food Security Index, Corruption Index, Youth Development Index. The aim of each index is rather self-explanatory from its name. Details can be found in Gallup (2015) and on the GWP website.

In order to explore the number of groups we use the `ctlcurves` of García-Escudero *et al.* (2011), which for different values of k show the log-likelihood at convergence of TCLUST, as a function of α and k. They can be used to determine both the number of groups and the optimal trimming level. The `ctlcurve` for the FIES data is reported in Figure 9.

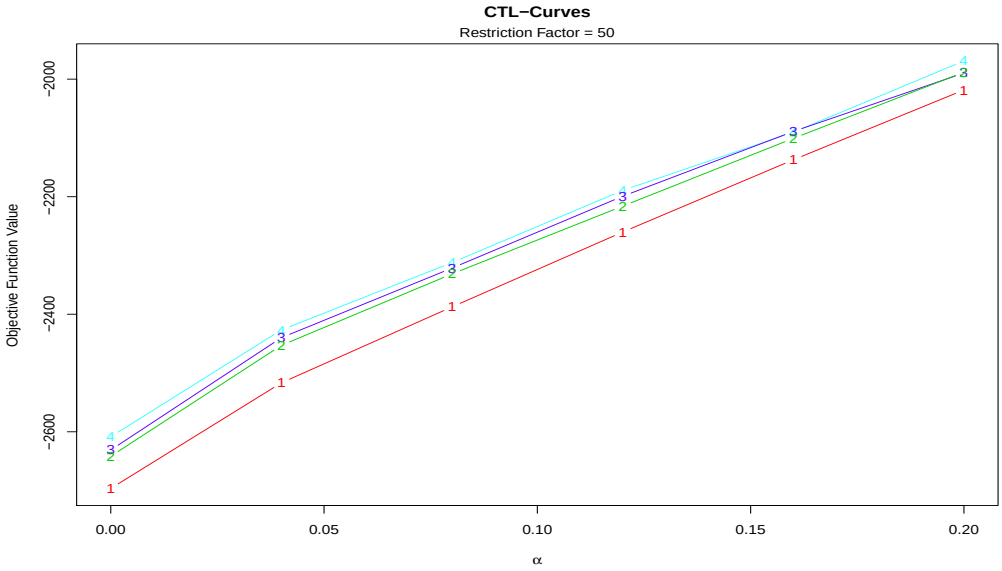

Figure 9: `ctlcurve` plot for the FIES data.

As sometimes happens, Figure 9 clearly indicates that there should be $k > 2$ groups, but it is unclear as with respect to the choice between $k = 3$ and $k = 4$. Additionally, it is definitely not conclusive with respect to the optimal trimming level α, which here is a parameter of interest as it is connected with the number (and identity) of outlying nations. The final estimates depend on the choice of α. In this example, RTCLUST can be seen as an automatic way of choosing the optimal trimming level, as the one balancing between robustness and efficiency. For the proposed methodology we do not need to specify α. We have applied our method both based on $k = 3$ and $k = 4$. As with $k = 4$ two groups are not very separated, we prefer $k = 3$ and report only those results for reasons of space. We run `rtclust` with $k = 3$, initial trimming level $\alpha_0 = 0.2$, $\alpha_L = 0.001$.The results are remarkably stable with respect to the tuning parameters. Nine countries (7.1%) are flagged as outlying, 13 are classified in group 1, 95 in group 2, and 10 in group 3. The cluster profiles (cluster means) and raw measurements for the

outlying countries are reported in Table 1. It shall be noted that groups 1 and 3 are of similar size as the group of outliers. Countries in groups 1 and 3 are very similar though and close to the reported profiles, while outliers are provably scattered, or have extremal values at least in one of the dimensions considered.

Table 1: Cluster profiles and measurements for the outlying countries. FIES: Food Insecurity Experience Scale. CE: Civic Engagement. St: Struggling. FS: Food Security. Co: Corruption index. YD: Youth Development. C-j: j-th cluster profile.

	FIES	CE	St	FS	Co	YD
C-1	-0.34	44.64	55.19	69.66	45.53	75.79
C-2	0.13	31.42	63.24	53.99	74.33	58.61
C-3	0.41	22.55	63.94	52.51	67.97	44.90
Myanmar	-0.95	66.84	85.80	13.21	53.33	85.48
Sweden	-0.64	43.22	48.08	76.68	37.39	59.67
Georgia	-0.43	21.24	60.49	41.31	30.85	67.56
New Zealand	-0.12	57.98	55.52	67.02	40.50	66.94
Paraguay	0.06	17.43	81.05	88.26	66.31	40.64
Rwanda	0.27	13.12	69.74	61.54	9.29	84.48
Cambodia	0.90	26.93	62.30	20.61	73.53	86.70
South Sudan	3.81	35.17	51.53	35.22	58.29	49.97
Haiti	5.04	35.33	51.47	43.66	57.24	32.07

It can be seen that the three clusters are well separated in terms of all of the items considered. The first cluster is characterized by the lowest food insecurity (and largest food security), corruption and struggling, and by the largest civic engagement and youth development. Sadly, only a minority of countries are assigned to cluster 1. The third cluster is characterized by largest food insecurity, lowest civic engagement and youth development. No differences are seen in terms of struggling and FS index between clusters 2 and 3. Finally, not surprisingly the corruption index is higher in the slightly more developed countries belonging to cluster 2 than in those in cluster 3. The outliers are easily explained, as for instance Haiti and South Sudan have an extremely high FIES. Sweden might belong to cluster 1, but its corruption is so low and its food security (however measured) is so large that it is outlying. All other outliers have at least one measurement in complete disagreement with the three clusters. A special note regards Myanmar, where there might have been problems with the questionnaire and with the sampling, and whose measurements therefore might not be completely reliable.

It shall be noted that the new FIES score is able to separate very well the three clusters, while Gallup's FS score only discriminates between the first and the other two. Other evidence in favor of the added value of FIES is that if we remove it and repeat the analysis the average silhouette width decreases by about 4%.

5 Conclusions and further directions

We have presented an iteratively reweighed approach that can recover wrongly trimmed observations when applying robust clustering procedures based on a high (preventive) trimming levels. This approach also makes easier the use of the TCLUST robust clustering method by diminishing its influence on the initial trimming level and on the chosen value for the eigenvalue ratio constraint. RTCLUST has two advantages over TCLUST: first, a sometimes not easily chosen tuning parameter, the trimming level, does not need to be perfectly specified in advance and the same happens for the eigenvalue ratio constraint value c. Secondly, it conjugates high robustness (as it can resist to an α_0 proportion of outliers) with high efficiency (as under no or little contamination the proportion of discarded observations will be much lower than α_0). The simulation study and the real data example also show how this methodology could be useful in practical applications. There is still room for further work. Formal theoretical properties could be explored. As commented in Remark 1, the outlier labeling process at each iteration could also be refined. We have applied very simple thresholds based on the χ^2 approximation for the Mahalanobis distances. More accurate procedures could be obtained, for instance, by considering small sample approximations or correcting for the multiple testing when labeling outliers (see, e.g., Cerioli (2010); Cerioli and Farcomeni (2011)). The multiple testing approach to reweighting might be tweaked to yield a simultaneous robust estimation and outlier detection method. The proposed methodology assumes that the number of groups k is known in advance. Estimating a correct k value is an important, but difficult too, problem. In fact, this is an ill-posed problem because the total number of groups depends on the type of clusters we are searching for or on what we understand by noise. For instance, a set made up with several disperse observations can be seen as a proper group with a large scatter or it can also be seen as background noise. Therefore, searching for the proper number of groups k would require making some subjective choices specifying all these aspects somehow. Another interesting open research line has to do with the extension of this iteratively reweighing approach for mixture modeling. This could be useful in order to address severe overlaps

among groups.

Acknowledgments

Research partially supported by the Spanish Ministerio de Economía y Competitividad, grant MTM2014-56235-C2-1-P, and by Consejería de Educación de la Junta de Castilla y León, grant VA212U13. We are grateful to Gallup, Inc. and the Voices of the Hungry project, FAO, for access to the GWP/FIES data.

References

T.J. BALLARD, A.W. KEPPLE, AND C. CAFIERO (2013). The food insecurity experience scale: developing a global standard for monitoring hunger worldwide. *Tech. rep.*, Food and Agriculture Organization of the United Nations, Rome.

R.W. BUTLER, P.L. DAVIES, AND M. JHUN (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, **21**, 1385–1400.

C. CAFIERO, H. R. MELGAR-QUINONEZ, T. J. BALLARD, AND A. W. KEPPLE (2014). Validity and reliability of food security measures. *Annals of the New York Academy of Sciences*, **1331**, 230–248.

C. CAFIERO, M. NORD, S. VIVIANI, M. E. DEL GROSSI, T. J. BALLARD, A. W. KEPPLE, M. MILLER, AND C. NWOSU (2016). Methods for estimating comparable rates of food insecurity experienced by adults throughout the world. *Tech. rep.*, Food and Agriculture Organization of the United Nations, Rome.

A. CERIOLI (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, **105**, 147–156.

A. CERIOLI AND A. FARCOMENI (2011). Error rates for multivariate outlier detection. *Computational Statistics and Data Analysis*, **55**, 544–553.

A. CERIOLI, A. FARCOMENI, AND M. RIANI (2014). Strong consistency and robustness of the forward search estimator of multivariate location and scatter. *Journal of Multivariate Analysis*, **126**, 167–183.

J.A. CUESTA-ALBERTOS, A. GORDALIZA, AND C. MATRÁN (1997). Trimmed k-means: an attempt to robustify quantizers. *Annals of Statistics*, **25**, 553–576.

A. FARCOMENI AND L. GRECO (2015). *Robust Methods for Data Reduction*. CRC Press.

B. FLURY AND H. RIEDWYL (1988). *Multivariate Statistics. A Practical Approach*. Chapman and Hall, London.

H. FRITZ, L.A. GARCÍA-ESCUDERO, AND A. MAYO-ISCAR (2012). tclust: An R package for a trimming approach to cluster analysis. *J Stat Softw*, **47**.

H. FRITZ, L.A. GARCÍA-ESCUDERO, AND A. MAYO-ISCAR (2013). A fast algorithm for robust constrained clustering. *Computational Statistics and Data Analysis*, **61**, 124–136.

M.T. GALLEGOS AND G. RITTER (2005). A robust method for cluster analysis. *Annals of Statistics*, **33**, 347–380.

GALLUP (2015). *Worldwide Research Methodology and Codebook*. Gallup, Inc., Washington, D.C.

L.A. GARCÍA-ESCUDERO AND A. GORDALIZA (2007). The importance of the scales in heterogeneous robust clustering. *Computational Statistics and Data Analysis*, **51**, 4403–4412.

L.A GARCÍA-ESCUDERO, A. GORDALIZA, C. MATRÁN, AND A. MAYO-ISCAR (2008). A general trimming approach to robust cluster analysis. *Annals of Statistics*, **36**, 1324–1345.

L.A GARCÍA-ESCUDERO, A. GORDALIZA, C. MATRÁN, AND A. MAYO-ISCAR (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, **4**, 89–109.

L.A GARCÍA-ESCUDERO, A. GORDALIZA, C. MATRÁN, AND A. MAYO-ISCAR (2011). Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, **21**, 585–599.

H. C. J. GODFRAY, J. R. BEDDINGTON, I. R. CRUTE, K. HADDAD, D. LAWRENCE, J. F. MUIR, J. PRETTY, S. ROBINSON, S. M. THOMAS, AND

C. TOULMIN (2010). Food security: the challenge of feeding 9 billion people. *Science*, **327**, 812–818.

J. HARDIN AND D.M. ROCKE (2004). Outlier detection in the multiple cluster setting using the Minimum Covariance Determinant estimator. *Computational Statistics and Data Analysis*, **44**, 625–638.

J. HARDIN AND D.M. ROCKE (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, **14**.

C. HENNIG (2003). Clusters, outliers and regression: fixed point clusters. *Journal of Multivariate Analysis*, **83**, 183–212.

A.D. JONES, F.M. NGURE, G. PELTO, AND S.L. YOUNG (2013). What are we assessing when we measure food security? A compendium and review of current metrics. *Advances in Nutrition*, **4**, 481–505.

R. Y. LIU, J. M. PARELIUS, AND K. SINGH (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference. *The Annals of Statistics*, **27**, 783–858.

H. P. LOPUHAA (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *The Annals of Statistics*, **27**, 1638–1665.

N. NEYKOV, P. FILZMOSER, R. DIMOVA, AND P. NEYTCHEV (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics and Data Analysis*, **52**, 299–308.

M. RIANI, A. ATKINSON, AND A. CERIOLI (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society (Series B)*, **71**, 447–466.

G. RITTER (2014). *Robust cluster analysis and variable selection*. CRC Press.

P. J ROUSSEEUW (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, **8**, 283–297.

P. J. ROUSSEEUW AND K. VAN DRIESSEN (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

P.J. ROUSSEEUW AND A.M. LEROY (1987). *Robust Regression and Outlier Detection*. Wiley-Interscience, New York.