Supplementary Material to “A Reweighting Approach to Robust Clustering”

Francesco Dotto · Alessio Farcomeni*, ·
Luis Angel García-Escudero · Agustín Mayo-Iscar

Abstract In this supplementary material we report proofs for the theoretical results reported in the main paper and some additional simulation experiments.

1 Proofs

In this section some modifications to the main paper notation are introduced in order to better outline the proofs of the theoretical results stated in the main text.

Let $\theta = (\pi_1, \ldots, \pi_k, \pi_{k+1}, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k) \in \Theta$, where $\Theta$ is the considered parametric space. We define

$$D_{\theta}(x) = \min_{1 \leq j \leq k} D_{\theta}^j(x),$$

* To whom correspondence should be addressed.

Francesco Dotto
Dipartimento di Scienze Statistiche. Università di Roma “La Sapienza”.
Piazzale Aldo Moro, 5. 00185, Rome, Italy.
E-mail: francesco.dotto@uniroma1.it

Alessio Farcomeni
Dipartimento di Sanità Pubblica e Malattie Infettive. Università di Roma “La Sapienza”.
Piazzale Aldo Moro, 5. 00185 Rome, Italy.
E-mail: alessio.farcomeni@uniroma1.it

Luis Angel García-Escudero
Departamento de Estadística e Investigación Operativa Universidad de Valladolid.
Paseo de Belén, 7. 47011 Valladolid Spain
E-mail: lagarcia@eio.uva.es

Agustín Mayo-Iscar
Departamento de Estadística e Investigación Operativa
Universidad de Valladolid. Calle Ramón y Cajal, 7. 47005 Valladolid Spain
E-mail: agustin@med.uva.es
where $D_\theta(x) = d_\theta^2(x, \mu_j)$ is the Mahalanobis distance from the center $\mu_j$ and the scatter matrix $\Sigma_j$.

Given a fixed probability measure $P$, let us consider $G_\theta^P(u) = P[D_\theta(\cdot) \leq u]$ and its $\beta$ quantile $D_\theta^{P, \beta} = \inf_u \{G_\theta^P(u) \geq \beta\}$. If $\theta_{l-1}^P$ with

$$\theta_{l-1}^P = (\pi_{l-1}^1, \ldots, \pi_{l-1}^k, \pi_{l-1}^1, \ldots, \mu_{l-1}^k, \Sigma_{l-1}^1, \ldots, \Sigma_{l-1}^k),$$

are the values of the parameters at stage $l-1$, we consider the sets

$$A_{\theta_{l-1}^P}^{P,1-\alpha_1} = \{x | D_{\theta_{l-1}^P}(x) < D_{\theta_{l-1}^P}^{P,1-\alpha_1}\}$$

and

$$B_{\theta_{l-1}^P} = \{x | D_{\theta_{l-1}^P}(x) \leq \chi_{P,1-\alpha_1}^2\}$$

and

$$H_{\theta_{l-1}^P}^{P,1-\alpha_1} = \{x | D_{\theta_{l-1}^P}(x) = D_{\theta_{l-1}^P}^{P,1-\alpha_1}\} \cap A_{\theta_{l-1}^P}^{P,1-\alpha_1} \cap B_{\theta_{l-1}^P}.$$

The consistency factors for the scatter matrices are obtained as

$$\left(\left[\begin{array}{c} \eta_{P,1-\alpha_1} \\ \theta_{l-1}^P \\ \end{array}\right]\right)^{-1} = \frac{\eta_P(A_{\theta_{l-1}^P}^{P,1-\alpha_1} \cap B_{\theta_{l-1}^P})}{P(B_{\theta_{l-1}^P})}$$

if $P(A_{\theta_{l-1}^P}^{P,1-\alpha_1} \cap B_{\theta_{l-1}^P})/P(B_{\theta_{l-1}^P}) < 1$ and equal to $1$ otherwise. As done in the manuscript, $\eta_{\beta} = P(\chi_{P,2}^2 \leq \chi_{P,\beta}^2)/\beta$. Then, by using this notation and $I_A(\cdot)$ as the indicator function of set $A$, we have updated parameters:

$$\pi_{l}^j = P(H_{\theta_{l-1}^P}^{P,1-\alpha_1}),$$

$$\pi_{l+1}^j = 1 - P(B_{\theta_{l-1}^P}),$$

$$\mu_{l}^j = \int x I_{H_{\theta_{l-1}^P}^{P,1-\alpha_1}}(x) dP(x)$$

and

$$\Sigma_{l}^j = \left(\int x x' I_{H_{\theta_{l-1}^P}^{P,1-\alpha_1}}(x) dP(x) - \mu_{l}^j \mu_{l}^j'\right)_{\theta_{l-1}^P}^{P,1-\alpha_1}.$$
1.1 Proof of Theorem 1

The required bounds for the parameters have already been proved when \( l = 0 \) in García-Escudero et al. (2008). Notice that assuming an absolutely continuous distribution \( P \) automatically guarantees the PR condition in García-Escudero et al. (2008).

Let us also assume that the solution of that TCLUST population problem satisfies \( \pi^0_j > 0 \) for \( 1 \leq j \leq k \) (otherwise it is clear that \( k \) should have been decreased for clustering purposes). In order to apply an inductive reasoning, let us suppose that the parameters in \( \theta^{l-1}_P \) do satisfy the boundedness condition in the statement of Theorem 1. Given that \( P \) has a strictly positive density function, if \( \mu^{l-1}_j \neq \mu^{l-1}_{j2} \) for every \( j_1 \neq j_2 \), then it is trivial to prove that each \( H^{\alpha, \theta^{l-1}_P} \) contains a non empty open ball and consequently \( \pi^{j, \alpha, \theta^{l-1}_P} \) can be uniformly bounded from below by a strictly positive constant. The other bounds follow from the boundedness of the \( H^{\alpha, \theta^{l-1}_P} \) sets, which is a consequence of the bounds on the \( \theta^{l-1}_P \) parameters.

1.2 Proof of Theorem 2

As commented before, we are recovering the parameters in the iterative algorithm in the main text when the unknown probability measure \( P \) is replaced by the empirical measure \( P_n \). Therefore, we use the notation \( \theta^{l}_P \) for those parameters obtained from an i.i.d. random sample \( \{x_1, ..., x_n\} \) from \( P \).

Lemma A.4 and Lemma A.5 in García-Escudero et al. (2008) guarantee that there exists a compact set \( K \) satisfying \( \theta^{l}_P \in K \) for \( n > n_0 \) with probability 1. An inductive reasoning, similar to that applied in the proof of Theorem 1, would show that the \( H^{\alpha, \theta^{l}_P} \) are also uniformly bounded with probability 1. It may happen that one of these sets would have zero probability under \( P_n \). In that case, we just need to take \( \mu^{l}_j = \mu^{l}_j \) (recall that \( \mu^{l}_j \) was bounded because of the inductive reasoning applied) and take \( \Sigma^{l}_j \) equal to the zero matrix.

1.3 Proof of Theorem 3

In this proof, we will apply results from of Empirical Processes theory (see, e.g., van der Vaart and Wellner 1997) and the inductive reasoning again to prove the consistency of the sample parameters toward the population ones. Some technical lemmas are needed in this proof.

From García-Escudero et al. (2008) we know that the sample solution of the TCLUST method is consistent to its population counterpart. I.e., we have that

\[ \theta^{0}_{P_n} \rightarrow \theta^{0}_P, \; P\text{-almost surely.} \]
By assuming the consistency in the \((l - 1)\)-th iteration, i.e.

\[ \theta_{P_n}^{l-1} \rightarrow \theta_{P}^{l-1}, \quad P\text{-almost surely}, \quad (i) \]

we now have to prove the consistency for the \(l\)-th iteration.

If we use analogous notation as in the beginning of this supplementary document, we see that:

**Lemma 1** For a probability distribution \(Q\) in \(\mathbb{R}^p\) and \(\theta \in \Theta\), the sets \(A_{\theta}^Q\), \(B_{\theta}\) are contained in a Vapnik-Chervonenkis (VC) class of sets \(\Xi\), \(A_{\theta}^{Q,1-\alpha}\), \(B_{\theta}\) are contained in a VC class \(\Lambda\) and \(H_{\theta}^{Q,1-\alpha}\), \(1 \leq j \leq k\), are contained in a VC class of sets \(\Psi\). These classes are given by

\[
\Xi = \{U_{\theta_a}[\theta \in \Theta; a \in \mathbb{R}] \}
\]

\[
\Lambda = \{U_{\theta_a} \cap U_{\theta_b}[\theta \in \Theta; a, b \in \mathbb{R}] \}
\]

and

\[
\Psi = \{V_{\theta_j} \cap U_{\theta_a} \cap U_{\theta_b}[\theta \in \Theta; 1 \leq j \leq k; a, b \in \mathbb{R}] \},
\]

where

\[
U_{\theta_a} = \{x|D_{\theta}(x) \leq a\} \quad \text{and} \quad V_{\theta_j} = \{x|D_{\theta}(x) = D_{\theta}^j(x)\}.
\]

**Proof** Since \(D_{\theta}(x)\), for \(\theta \in \Theta\), is the minimum of \(k\) functions belonging to a finite dimensional subspace of functions, then \(\{D_{\theta}(x)[\theta \in \Theta]\) is a VC class by applying Lemmas 2.6.15 and 2.6.18 in van der Vaart and Wellner (1997). Analogously, \(\Xi, \Lambda\) and \(\Psi\) are VC classes of sets by application of Lemmas 2.6.15, 2.6.17 and 2.6.18 in the same reference. Sets \(A_{\theta}^{Q,1-\alpha}\) and \(B_{\theta}\) are contained in \(\Xi\) for \(\theta \in \Theta\). Their intersection \(A_{\theta}^{Q,1-\alpha} \cap B_{\theta}\) are contained in \(\Lambda\) and \(H_{\theta}^{Q,1-\alpha}\) are contained in \(\Psi\) for \(1 \leq j \leq k\).

**Lemma 2** Under the assumptions of Theorem 3 and assuming (i), we have

\[
D_{\theta_{P_n}^{l-1}}^{P,1-\alpha} \rightarrow D_{\theta_{P}^{l-1}}^{P,1-\alpha}, \quad P\text{-almost surely}.
\]

**Proof** In order to prove this lemma we need to show that:

\[
\sup_{\theta \in K}|D_{\theta_{P_n}^{l-1},1-\alpha} - D_{\theta_{P}^{l-1},1-\alpha}| \rightarrow 0, \quad P\text{-almost surely},
\]

in a compact set \(K \subseteq \Theta\). This follows exactly as in Lemma A.7 in García-Escudero et al. (2008), given the assumed convergence (i).

**Lemma 3** Under the assumptions of Theorem 3 and assuming (i), the following convergences hold

\[
\pi_{jP_n}^l \rightarrow \pi_{jP}^l, \quad \text{for} \quad j = 1, \ldots, k, k + 1, \quad \mu_{jP_n}^l \rightarrow \mu_{jP}^l \quad \text{and} \quad \Sigma_{jP_n}^l \rightarrow \Sigma_{jP}^l.
\]
Reweighting in robust clustering

Proof

Due to the Glivenko-Cantelli property of the classes $\Psi, \Lambda$ and $\Xi$ together with (i) and the consistency results for the quantiles in Lemma 2, we have

$$P_n \left( H_{\hat{\theta}_{Pn}}^{\frac{1-\alpha}{1}} \right) \to P \left( H_{\theta_{Pn}}^{\frac{1-\alpha}{1}} \right),$$

and, consequently $\pi^l_{jPn} \to \pi^l_{jP}$ for $1 \leq j \leq k$. Analogously, the consistency $\pi^l_{k+1Pn} \to \pi^l_{k+1P}$ follows from the convergence

$$P_n \left( B_{\theta_{Pn}}^{\frac{1-\alpha}{1}} \right) \to P \left( B_{\theta_{Pn}}^{\frac{1-\alpha}{1}} \right),$$

which is obtained in a similar fashion.

The Glivenko-Cantelli property for the class $t_{xI_{H}}h_p x_q|\ H_p \Psi$ and Lemma 2 entail $\mu^l_{jPn} \to \mu^l_{jP}, P$-almost surely.

Additionally, we have consistency for the correction factors:

$$c_{\theta_{Pn}}^{\frac{1-\alpha}{1}} \to c_{\theta_{Pn}}^{\frac{1-\alpha}{1}}, P$$

This last consistency is trivial given that $P_n \left( A_{\theta_{Pn}}^{\frac{1-\alpha}{1}} \cap B_{\theta_{Pn}}^{\frac{1-\alpha}{1}} \right) \to P \left( A_{\theta_{Pn}}^{\frac{1-\alpha}{1}} \cap B_{\theta_{Pn}}^{\frac{1-\alpha}{1}} \right),$

together with the convergence (ii) and the fact that $\eta^l_{\beta} = P \left( \chi^2_{p+2} \leq \chi^2_{p,\beta} \right) / \beta$

(see as a function on $\beta$) is a continuous function for $\beta \in (0, 1)$.

Therefore, given that the class $\{xH(x)|H \in \Psi\}$ is also a Glivenko-Cantelli class and the consistency of those $c_{\theta_{Pn}}^{\frac{1-\alpha}{1}}$ factors, we see that $\Sigma^l_{jPn} \to \Sigma^l_{jP}$

$P$-almost surely for $1 \leq j \leq k$. \(\square\)

The combination of all the above lemmas then allow us to argue in favor of consistency at each iteration $l = 0, \ldots, L$, by applying the inductive reasoning.

1.4 Proof of Theorem 4

Given a data sample $\{x_1, ..., x_n\}$, let us use the notation $L(\theta; \{x_1, ..., x_n\})$ for the target function (6) evaluated at $\theta = (\pi^l_1, ..., \pi^l_k, \mu^l_1, ..., \mu^l_k, \Sigma^l_1, ..., \Sigma^l_k) \in \Theta.$

We start with a technical lemma and where we are going to use the same notation as in Theorem 4:

**Lemma 4** Given $\theta^l_m = (\pi^l_1, ..., \pi^l_k, \mu^l_1, ..., \mu^l_k, \Sigma^l_1, ..., \Sigma^l_k)$ optimal parameters obtained when applying TCLUST with an eigenvalues ratio constraint $c \geq 1$ and an $\alpha_0$ trimming level to $X_m \cup Y_m$, let $R_m = \cup_{j=1}^k R^l_m$ be the set of non-trimmed observations and their optimal partition into $k$ clusters (which only depend on the sample $X_m \cup Y_m$ and on the optimal $\theta^l_m$ parameters). If there exists any $j$ such that

$$R^l_m \cap X_m \neq \emptyset \text{ and } R^l_m \cap Y_m \neq \emptyset,$$

(iii)
or if there exist some $j, j_1$ and $j_2$ with $j_1 \neq j_2$ such that

$$R^j_m \cap X^{j_1}_m \neq \emptyset \quad \text{and} \quad R^j_m \cap X^{j_2}_m \neq \emptyset,$$

(iv)

or if there exists any $j$ such that

$$\# \{R^j_m \cap Y_m \} \geq 2,$$

(v)

then

$$L(\theta_m; X_m \cup Y_m) \leq -h_1 \log(\lambda_m) - h_2 \min \left\{ b_m, c_m, d_m \right\},$$

where $h_1$ and $h_2$ are fixed (non-dependent on $m$) strictly positive constants,

$$\lambda_m = \min_{1 \leq j \leq k} \min_{1 \leq h \leq p} \lambda_h(\Sigma^j_m),$$

and $b_m, c_m$ and $d_m$ are those sequences defined in (7), (9) and (10) in the main manuscript.

Proof In this proof, we will use the notation

$$d(A, B) = \min_{x \in A} \min_{y \in B} \| x - y \|^2,$$

in such a way that inequalities (7) and (9) in the main paper can be rewritten as $d(X_m, Y_m) \geq b_m$ and as $d(X^{j_1}_m, X^{j_2}_m) \geq c_m$ for $j_1 \neq j_2$, respectively.

We have

$$L(\theta_m; X_m \cup Y_m) \leq \sum_{j=1}^k \# R^j_m \log(\pi^j_m) - [(n + r)(1 - \alpha_0)]^2 \log(2\pi)$$

$$- [(n + r)(1 - \alpha_0)]^2 \log \left( \min_{1 \leq j \leq k} \min_{1 \leq h \leq p} \lambda_h(\Sigma^j_m) \right)$$

$$- \frac{1}{2} \max_{1 \leq j \leq k} \max_{1 \leq h \leq p} \lambda_h(\Sigma^j_m),$$

where $T_m$ will be detailed later. The first two terms in the right-hand side are negative ones and they can be bounded from above by 0. The eigenvalues ratio constraint guarantees that

$$\max_{1 \leq j \leq k} \max_{1 \leq h \leq p} \lambda_h(\Sigma^j_m) \leq c \times \min_{1 \leq j \leq k} \min_{1 \leq h \leq p} \lambda_h(\Sigma^j_m).$$

Regarding $T_m$, if (iii) holds then $T_m$ can be chosen as

$$T_m = \max_{j=1, \ldots, k} \left\{ \max_{x \in R^j_m \cap X_m} \| x - \mu^j_m \|^2 + \max_{y \in R^j_m \cap Y_m} \| y - \mu^j_m \|^2 \right\}.$$

If (iv) holds then $T_m$ is chosen as

$$T_m = \max_{\{j, j_1, j_2\} \subseteq \{1, \ldots, k\} \text{ with } j_1 \neq j_2} \left\{ \max_{x \in R^{j_1}_m \cap X^{j_1}_m} \| x - \mu^{j_1}_m \|^2 + \max_{x' \in R^{j_2}_m \cap X^{j_2}_m} \| x' - \mu^{j_2}_m \|^2 \right\}.$$
Finally, if (v) holds then \( T_m \) is chosen as
\[
T_m = \max_{j=1,...,k} \left\{ \max_{y \neq y', y, y' \in Y_m} \{ ||y - \mu^m_j||^2 + ||y' - \mu^m_j||^2 \} \right\}.
\]

The triangular inequality (telling us that \( ||x - z||^2 + ||y - z||^2 \geq ||x - y||^2 \))
shows that
\[
T_m \geq \min_{x \in R_m \cap A_m} \| x - y \|^2 \geq d(X_m, Y_m) \geq b_m,
\]
in the first case, that
\[
T_m \geq \min_{x \in R_m \cap A_m} \| x - y' \|^2 \geq d(X_m', X_m) \geq c_m,
\]
for \( j_1 \neq j_2 \), in the second case, and that
\[
T_m \geq \min_{y \neq y', y, y' \in Y_m} \| y - y' \|^2 \geq d_m,
\]
in the third case. Therefore, in all these three cases, we have \( T_m \geq \min\{b_m, c_m, d_m\} \) and the final result is so obtained just by setting
\[
\lambda_m = \min_{1 \leq j \leq k} \min_{1 \leq h \leq p} \min \lambda_h \left( \Sigma^m_{h} \right),
\]
and by taking into account that both \( n_1 + \ldots + n_k = n \) and \( r \) remain fixed when moving \( m \). \( \square \)

Now, in order to proceed with the proof of Theorem 4, let us assume that
\( m_0 \in \mathbb{N} \), as that in the statement of this theorem, does not exist. In this case, a subsequence from \( \theta_n \), that will be denoted as the original one without lack of generality, can be extracted such that it satisfies either (iii) or (iv) or (v) in Lemma 4 for a sequence of indexes \( j = j(m) \), \( j_1 = j_1(m) \) and \( j_2 = j_2(m) \) with \( j_1 \neq j_2 \). The case \( \#(R^m_m \cap Y_m) = 1 \) is not a problem because the assumed upper bound on \( \alpha_0 \) in the statement of Theorem 4 would guarantee the existence of another \( j' \neq j \) such that either (iii) or (iv) or (v) holds when replacing \( R^m_m \) by this \( R^m_{j'} \). By applying Lemma 4, we would have for this subsequence that (iii) or (iv) or (v)
\[
L(\theta^*_m; X_m \cup Y_m) \leq -h_1 \log(\lambda_m) - h_2 \min\{b_m, c_m, d_m\} / c\lambda_m.
\]
If \( \lim_{m \to \infty} \lambda_m = \infty \), then \( \lim_{m \to \infty} L(\theta^*_m; X_m \cup Y_m) = -\infty \) trivially given that \( \min\{b_m, c_m, d_m\} / c\lambda_m \geq 0 \). We can easily see that condition (8) and the eigenvalues ratio constraint guarantee that \( \lim_{m \to \infty} \lambda_m \neq 0 \). Finally, if \( \lambda_m \) is uniformly bounded within \( [a_L, a_U] \), with \( 0 < a_L < a_U < \infty \), then \( \lim_{m \to \infty} L(\theta^*_m; X_m \cup Y_m) = -\infty \) too because \( \lim_{m \to \infty} \min\{b_m, c_m, d_m\} = \infty \).

On the other hand, we could have taken \( \hat{\theta}_m = (\hat{\pi}^m_1, \ldots, \hat{\pi}^m_k, \hat{\lambda}^m_1, \ldots, \hat{\lambda}^m_k, \hat{\mu}^m_1, \ldots, \hat{\mu}^m_k, \hat{\Sigma}^m_1, \ldots, \hat{\Sigma}^m_k) \) with \( \pi^m_j = 1/k, \mu^m_j \) being any observation in \( X_m' \) and \( \Sigma^m_j = I \) for \( j = 1,...,k \). This fact would contradict that \( \lim_{m \to \infty} L(\theta^*_m; X_m \cup Y_m) = -\infty \), which has already been proven to happen when that \( m_0 \) does not exist.
2 Additional Simulation Experiments

2.1 First Setting: \( k = 4 \) clusters

Within this additional setting we evaluated the performance of our reweighting method when there are \( k = 4 \) clusters. Similarly as in Section 4 of the main paper, the non-outlying part of the dataset comes from a mixture of four \( p \)-variate normal distributions \( \pi_1 N(\mu_1, \Sigma_1) + \pi_2 N(\mu_2, \Sigma_2) + \pi_3 N(\mu_3, \Sigma_3) + \pi_4 N(\mu_4, \Sigma_4) \) with centers \( \mu_1 = (0, 0, 0, \ldots, 0)' \), \( \mu_2 = (8, 0, 0, \ldots, 0)' \), \( \mu_3 = (16, 0, 0, \ldots, 0)' \) and \( \mu_4 = (24, 0, 0, \ldots, 0)' \) and covariance matrices:

\[
\Sigma_1 = I_p, \quad \Sigma_2 = \sqrt{\lambda} \begin{pmatrix}
1 & 1 & 1 & \ldots & 1 \\
1 & 2 & 2 & \ldots & 2 \\
1 & 2 & 3 & \ldots & 3 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 2 & 3 & \ldots & p
\end{pmatrix}
\]

and \( \Sigma_3 = \Sigma_4 = 1.44 \times \sqrt{\lambda} \begin{pmatrix}
1 & 1 & 1 & \ldots & 1 \\
1 & 1 & 2 & 1 \times 2 \times 2 \times \ldots \times 2 \\
1 & 1 & 2 & 1 \times 3 \times 3 \times \ldots \times 3 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1 & 2 & 1 \times 3 \times 4 \times \ldots \times 4 \times p
\end{pmatrix} \)

This means that \( |\Sigma_1| = 1 \) and \( |\Sigma_j| = 1.44^p \lambda \) for \( j = 2, \ldots, 4 \). Outlying observations are generated uniformly within hypercubes, but outliers with squared Mahalanobis distances from \( \mu_1, \mu_2, \mu_3 \) and \( \mu_4 \) using \( \Sigma_j \) for \( j = 1, 2, 3, 4 \) smaller than \( \chi^2_{p, \nu} \) are discarded. The operation is repeated until the desired proportion of \( \varepsilon \) outliers have been obtained. The parameter \( \nu \) 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 0.
- Two scales \( \lambda = 1 \) and 5
- Balanced clusters \( \pi_j = 0.25 \) for \( j = 1, 2, \ldots, 4 \) and unbalanced clusters \( \pi_1 = 0.2, \pi_2 = 0.2, \pi_3 = 0.25 \) and \( \pi_4 = 0.35 \).
- Two \( \nu \) values, \( \nu = 0.01 \) and \( \nu = 0.001 \)
- Two types of contamination: a symmetric one obtained sampling from a uniform distribution in the hypercube defined by the increased range of the non-contaminated part of the data and an asymmetric one obtained by sampling from a uniform distribution defined on \([-1, 20] \times [-7, -2] \times [-2, 2]^{p-2} \).

Figures 1, 2 and 3 report the results in case of \( \varepsilon = 0.10, 0.05 \) and 0, respectively.
Fig. 1 Simulations’ results. $k = 4$ groups and $\varepsilon = 0.10$ proportion of contaminated data.
Fig. 2 Simulations’ results. $k = 4$ groups and $\epsilon = 0.05$ proportion of contaminated data
Fig. 3  Simulations' results. $k = 4$ groups and no contamination.
2.2 Second setting: A direct comparison with the TCLUST methodology

This setting is aimed at comparing our reweighted methodology with the TCLUST method (García-Escudero et al. 2008) when this method is performed with the output values returned by the application of RTCLUST.

2.2.1 Fixed $\alpha$

As usual, we initialized our procedure with the TCLUST algorithm imposing two different initial trimming levels: $\alpha_0 = 0.33, 0.20$. Then we used the estimation of the proportion of outliers provided by the RTCLUST method to impose a trimming level for the TCLUST. As it can be seen in Figures 4, 5 and 6 the performance of these four procedures are pretty much the same in the estimation of the location parameters, while better results are achieved by the RTCLUST in the estimation of the scatter matrices. 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 $\mu_1 = (0, 0, 0, \ldots, 0)'$ and $\mu_2 = (8, 0, \ldots, 0)'$ and covariance matrices

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

This means that $|\Sigma_1| = 1$ and $|\Sigma_2| = \lambda$. As in previous simulation studies, outlying observations are generated uniformly within hypercubes, but outliers with squared Mahalanobis distances from $\mu_1$ and $\mu_2$ (using $\Sigma_1$ and $\Sigma_2$) smaller than $\chi^2_{p, \nu}$ are discarded. The operation is repeated until the desired proportion of $\varepsilon$ outliers have been obtained. The parameter $\nu$ 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 $\nu$ values, $\nu = 0.01$ and $\nu = 0.001$
- 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 $[-1, 20] \times [-7, -2] \times [-2, 2]^p$. 
Fig. 4 Simulations’ results. Results of the comparison with the TCLUST method with $\varepsilon = 0.10$ proportion of contaminated data. The trimming level imposed is the one estimated by the RTCLUST

2.2.2 Fixed $\alpha$ and $c$

Within this setting we provide an additional comparison with the TCLUST methodology when using the estimation of the proportion of outliers and the estimated eigenvalue ratio provided by the RTCLUST method to impose a trimming level and a bound for the eigenvalue ratio in TCLUST. Figures 4, 5 and 6 show that the performance of these four procedures are almost exactly the same, confirming that RTCLUST method is able to reach the same results as the TCLUST method and, moreover, does not require to fix these two tuning parameters which, especially in the applications can be cumbersome. The simulation scheme is exactly the same as previously described in Section 2.2.1.
Fig. 5 Simulations’ results. Results of the comparison with the TCLUST method with $\varepsilon = 0.10$ proportion of contaminated data. The trimming level imposed is the one estimated by RTCLUST.

References


Fig. 6 Simulations’ results. Results of the comparison with the texttttclust method with \( \varepsilon = 0.10 \) proportion of contaminated data. The trimming level imposed is the one estimated by RTCLUST.
Fig. 7 Simulations' results. Results of the comparison with the TCLUST method with $\varepsilon = 0.10$ proportion of contaminated data. The trimming level and the eigenvalue ratio imposed are the one estimated by the RTCLUST.
Fig. 8 Simulations’ results. Results of the comparison with the TCLUST method with \( \varepsilon = 0.10 \) proportion of contaminated data. The trimming level and the eigenvalue ratio imposed are the one estimated by the RTCLUST.
Fig. 9 Simulations’ results. Results of the comparison with the TCLUST method with $\varepsilon = 0.10$ proportion of contaminated data. The trimming level and the eigenvalue ratio imposed are the one estimated by the RTCLUST.