Received: 20 December 2023 | Revised: 10 April 2024 | Accepted: 24 May 2024

DOI: 10.1002/wics.1658

ADVANCED REVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

Robust clustering based on trimming

Luis A. García-Escudero | **Agustín Mayo-Iscar**

Department of Statistics and Operation Research and IMUVA, University of Valladolid, Valladolid, Spain

Correspondence
Luis A. García-Escudero, Department of Statistics and Operation Research and IMUVA, University of Valladolid, Valladolid, Spain.
Email: lagarcia@uva.es

Funding information
MCIN/AEI/10.13039/501100011033/FEDER, Grant/Award Number: PID2021-128314NB-I00

Edited by: Mia Hubert, Commissioning Editor and David Scott, Review Editor and Editor-in-Chief

Abstract

Clustering is one of the most widely used unsupervised learning techniques. However, it is well-known that outliers can have a significantly adverse impact on commonly applied clustering methods. On the other hand, clustered outliers can be particularly detrimental to (even robust) statistical procedures. Therefore, it makes sense to combine concepts from Robust Statistics and Cluster Analysis to deal with both clusters and outliers simultaneously through robust clustering approaches. Among the existing robust clustering techniques, we focus on those that rely on (impartial) trimming. Trimming offers the user an easy interpretation, as standard well-known clustering methods are applied after a fraction of the potentially most outlying observations is removed. This trimming approach, when combined with appropriate constraints on the clusters' dispersion parameters, has shown a good performance and can be implemented efficiently thorough available algorithms.

This article is categorized under:

Statistical Learning and Exploratory Methods of the Data Sciences > Clustering and Classification

Statistical and Graphical Methods of Data Analysis > Robust Methods

KEYWORDS

clustering, model-based clustering, robustness, trimming

1 | INTRODUCTION

Clustering involves the discovery of "crowds" of data points, while anomaly detection focuses on identifying data points located far from these crowds. Therefore, clustering and anomaly detection complement each other in some ways. Data anomalies, or outliers in other words, may negatively impact many well-known and widely applied clustering methods, while robust statistical procedures (see, e.g., Maronna et al., 2019) are aimed at resisting such anomalies, or more generally, resisting small deviations from the typically assumed models. On the other hand, clustered outliers are known to be especially harmful for (even robust) statistical procedures (Rocke & Woodruff, 1996), but they can usually be detected easily by clustering techniques. All the above claims serve to justify the interest of combining ideas from Robust Statistics and Cluster Analysis, as well as providing some insights into why "robust clustering" could serve as an appealing unifying framework for tackling these two problems simultaneously.

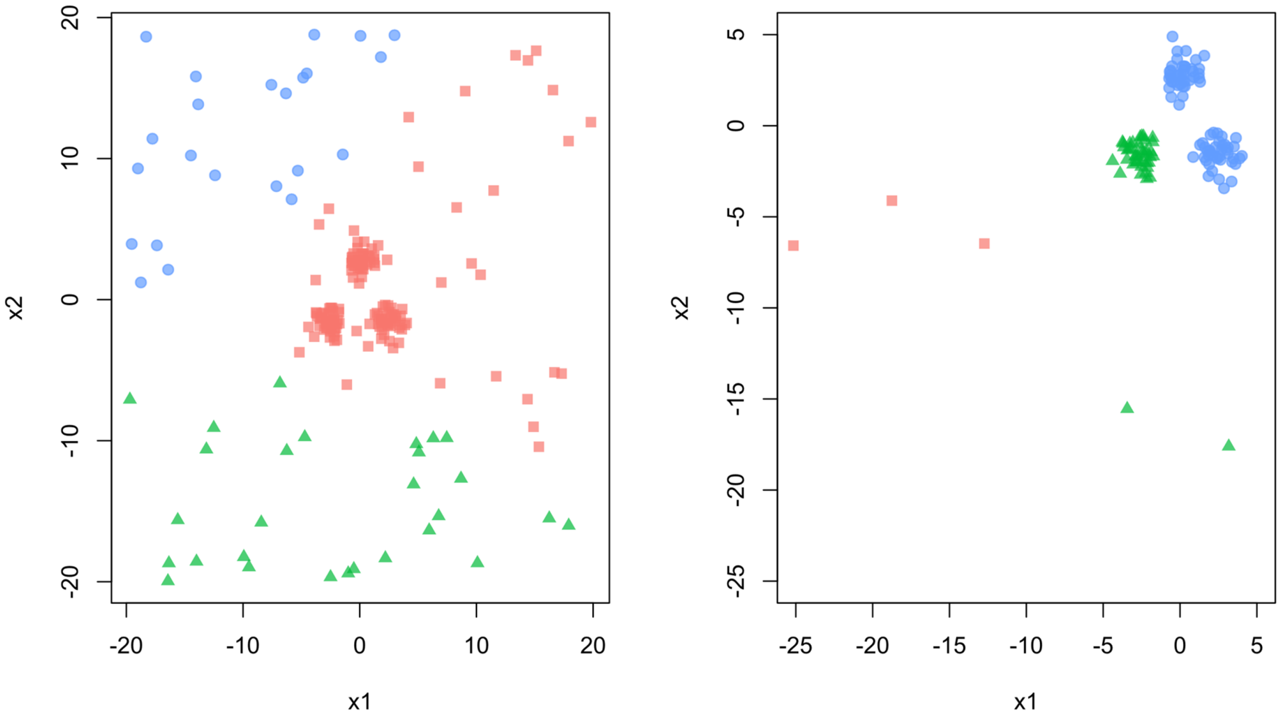

FIGURE 1 Results of applying 3-means in two contaminated data sets.

Figure 1 shows an example of the harmful effect of outliers when applying the classical 3-means methods in two simulated datasets that include $K = 3$ main clusters, whose centers are located in an equilateral triangle, with 40 observations each. 40% of background noise is added in Figure 1 (left) and 4% of remote outlying observations (right). We can see that main clusters are articially joined together by the 3-means method due to the effect of contamination.

It could be argued that increasing the number of clusters K could help to alleviate this lack of robustness. This idea makes sense, since outliers often tend to occur in (small) clusters of their own and because the mechanism that generates outlying observations could generate them consistently and sometimes grouped together. However, this idea is not always the optimal one, because the additional clusters found could be meaningless, with little interpretability; for instance, made up of a few isolated observations. Moreover, the user sometimes predefines the number of clusters required due to the final clustering purposes and is not aware of the presence of outliers which could negatively affect the preferred clustering technique.

Another possibility is to "model the noise" by, for instance, considering uniformly distributed components or components associated to heavily tailed distributions. This idea turns out to be extremely useful in many practical cases, but somehow implies making certain assumptions about the mechanism generating the contaminating data. This is problematic if actual noise strongly deviates from the assumed hypothesis and, in general, can be troublesome due to the unpredictible nature of contamination.

There are many other robust clustering proposals in the literature that can be successfully applied to cope with the presence of outliers (see, e.g., García-Escudero, Gordaliza, Matrán, & Mayo-Iscar, 2010; Banerjee & Davé, 2012; Ritter, 2015; Farcomeni & Greco, 2016 or García-Escudero, Gordaliza, Matrán, et al., 2016 and the references therein). However, we just focus on reviewing some of the existing approaches based on trimming.

Trimming is one of the oldest[1] methods to provide robustness to statistical techniques and relies on the idea of applying standard statistical tools once we have ensured that the observations most susceptible to being outliers have (hopefully) been discarded. Relying on well-known statistical methods, once trimmed observations are removed, makes it easier to understand and communicate what has been done to achieve robustness. A prototypical example of a trimming procedure is the univariate α-trimmed mean, which computes the mean after discarding the proportion $\alpha/2$ of the largest observations and the proportion $\alpha/2$ of the smallest ones. However, how to apply trimming in multivariate problems, and in clustering in particular, is not direct because, for instance, there is no natural order in multivariate data to declare the observations as "larger" or "smaller." Furthermore, the determination of the α fraction of observations most likely to be outliers should depend on the applied statistical technique. For instance, in the case of clustering, "bridge points" between clusters can be trimmed, even though they are not extreme observations in our sample.

For this trimming approach to clustering, we consider an "impartial" trimming. The term impartial means that it is the dataset itself that tells us which observations to trim, and not the choice of privileged or pre-specified trimming directions. This impartial trimming approach also underlies well-known and widely applied procedures for the so-called "high-breakdown point" (Rousseeuw & Leroy, 1987) robust proposals in regression, such as the LTS (Least Trimmed Squares) and the LMS (Least Median of Squares); as well as in multivariate location and scatter, such as the MCD (Minimum Covariance Determinant) and the MVE (Minimum Volume Ellipsoids).

Trimming techniques discard a proportion α of observations where it is supposedly more likely that any outlying observations can be found and they also highlight the fraction $1-\alpha$ of the most "relaible" observations. We are not claiming that all trimmed observations are outliers and, in fact, we can study their differences with respect to the non-trimmed ones and use that information to provide a final outlier determination.

2 | TRIMMED K-MEANS

A robust (impartial) trimming-based clustering procedure was given in the pioneering work by Cuesta-Albertos et al. (1997), where trimmed K-means were introduced as a trimmed extension of the classical K-means. Given a sample $x_1, x_2, \ldots, x_n$ in R^p, it was proposed to search for K location centers $m_1, m_2, \ldots, m_K$ in R^p (the trimmed K-means centers) and a partition $\{R_0, R_1, \ldots, R_K\}$ of the indices $\{1, 2, \ldots, n\}$ with R_0 including a proportion $[n\alpha]$ of indices and minimizing

$$\sum_{k=1}^{K} \sum_{i \in R_k} \|x_i - m_k\|^2.$$

Since the summation in the previous expression goes from 1 to K (not from 0 to K), the proportion α of observations in R_0 are then omitted when computing the previous target function.

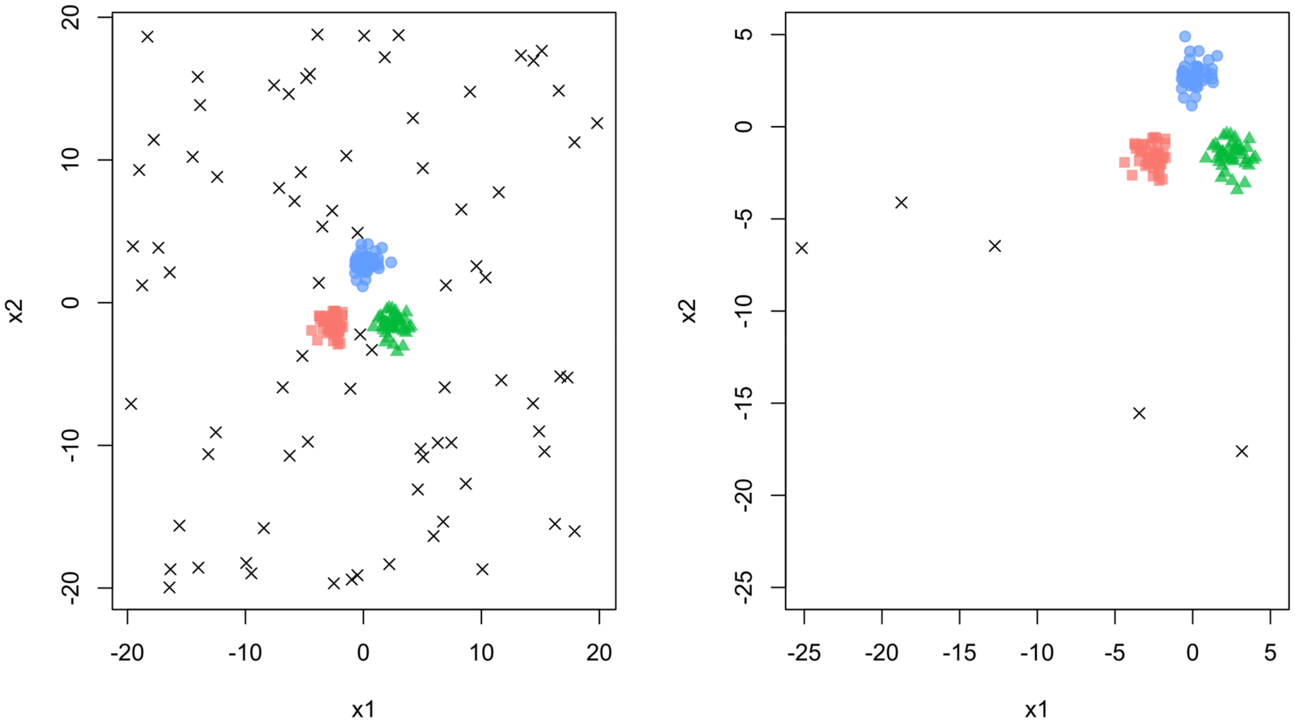

FIGURE 2 Results of applying trimmed 3-means for the data sets in Figure 1 when considering $\alpha = 0.40$ (left) and $\alpha = 0.04$ (right). Trimmed observations are denoted by crosses.

Figure 2 shows the results of applying trimmed 3-means for the same data sets as in Figure 1 when considering $\alpha = 0.40$ (left) and $\alpha = 0.04$ (right). The trimmed observations are shown by crosses. We can see how the three main clusters are correctly detected after trimming.

It is not difficult to prove the following two characterizing properties for the trimmed K-means: (i) the trimmed observations in the optimal R_0 are exactly the most distant fraction α of observations from their closest trimmed K-means centers. (ii) The optimal trimmed K-means centers are the standard K-means of the observations not in R_0. As a trivial consequence, the trimmed K-means reduces to the classical K-means if $\alpha = 0$.

In a similar fashion to how the median makes the sample mean more robust, one could think that removing the squared penalty in the K-means target function (the same one as for the trimmed K-means if $\alpha = 0$) would have served to achieve robustness throughout the so-called K-medians. However, such a robustification when $K > 1$ is known to be very timid and, in fact, just one single observation in a very remote position can take one of the K-median centers to be outside any bounded region in R^p (García-Escudero & Gordaliza, 1999). This lack of robustness of the K-medians is in some ways similar to what happens when removing the squared penalty in the ordinary least squares regression through the L_1-regression. Robust regression based on trimming, namely the LTS and LMS, were the first methods in regression that can resist a high positive number of adversarial contaminating observations. The trimmed K-means can be seen as a modification of the K-means in a similar fashion to how the LTS method modifies the ordinary least squares regression. The two previously commented properties, (i) and (ii), characterizing the trimmed K-means are, in fact, closely related to how LTS trims the proportion α of observations with the largest regression residuals, and how LTS applies standard least squares regression to the non-trimmed observations to produce the LTS regression fit. A fast-LTS algorithm based on so-called "concentration" steps was given in Rousseeuw and van Driessen (2000). The corresponding analogy of the concentration steps for trimmed K-means was introduced in García-Escudero et al. (2003). This algorithm is therefore a combination of the concentration step ideas behind the fast-LTS and the classical Lloyd's algorithm for K-means (Lloyd, 1982). We will discuss this in greater detail in Section 6. The extension of the LMS to cluster analysis is the trimmed best K-nets, introduced in Cuesta-Albertos et al. (1998).

García-Escudero and Gordaliza (1999) provides some insights into the robustness gain of trimmed K-means with respect to classical K-means. The formulation of breakdown-point notions for clustering, and consequently for robust clustering, is not straightforward, but some proposals for it are available (see a discussion of this in García-Escudero, Gordaliza, Matrán, et al., 2016).

Another extension of trimmed K-means is presented in Ibáñez et al. (2012), referred to as trimmed K-medoids. This extension involves impartial trimming but focuses on identifying K representative observations/objects in the sample, thus allowing the use of relational data where only measurements of dissimilarity between each pair of objects in our sample are available. Trimmed K-means, identified under the distinct name "K-means -," has also been considered in Chawla and Gionis (2013) along with an extension of the method to Bregman divergences.

3 | ELLIPTICALLY CONTOURED CLUSTERS

3.1 | Trimmed classification likelihoods: TCLUST

A clear drawback of the trimmed K-means, inherited from the classical K-means, is its preference for spherical and equally scattered clusters. In order to deal with more general types of clusters, and considering internal cluster variabilities and dependence structures, the TCLUST method in García-Escudero et al. (2008) can be applied. TCLUST considers a trimmed classification likelihood approach, assuming multivariate normally distributed components for the regular part of the data. Consequently, TCLUST searches for centers $m_1,\ldots,m_K$ in R^p, symmetric positive definite $p \times p$ matrices $S_1,\ldots,S_K$, weights $p_1,\ldots,p_K$ with $\sum_{k=1}^{K} p_k = 1$, a partition $\{R_0, R_1, \ldots, R_K\}$ of $\{1, 2, \ldots, n\}$ such that R_0 includes a proportion $[n\alpha]$ of indices, and maximizing

$$\sum_{k=1}^{K}\sum_{i \in R_k}\log(p_k\phi(x_i; m_k, S_k)), \qquad (1)$$

where $\phi(\cdot;\mu,\Sigma)$ is the probability density function of a p-variate normal with mean vector μ and covariance matrix Σ. A theoretical framework for the trimmed classification likelihoods is the "spurious outlier model" in Gallegos and Ritter (2005).

The original formulation in García-Escudero et al. (2008) included weights p_k, but these weights can be removed (or, analogously, set $p_k = 1/K$ for $k = 1,...,K$) if clusters with similar sizes are privileged. Another important ingredient in TCLUST is an "eigenvalues-ratio" constraint to control the relative size of the scatter matrices $S_1,...,S_K$. We comment on this type of restrictions in Section 3.3.

Figure 3 shows how TCLUST (right) is better suited to dealing with elliptical clusters, and different scatters, than trimmed K-means (left) for a simulated dataset when $K = 3$ and $\alpha = 0.1$.

It is not difficult to see that TCLUST with $K = 1$ (without the eigenvalues-ratio constraint) reduces to the MCD estimator. Hardin and Rocke (2004) considered a different extension of the MCD approach to clustering and Jobe and Pokojovy (2015) also employed MCD in a cluster-based outlier detection procedure.

An algorithm for applying TCLUST is available (Fritz et al., 2013a), by considering a modification of the classification EM algorithm (Celeux & Govaert, 1992), which incorporates trimming through concentration steps and imposing the eigenvalues-ratio constraint. Good robustness properties for TCLUST have been proven from the infinitesimal robustness point of view in Ruwet et al. (2012) and from the breakdown point of view in Ruwet et al. (2013).

3.2 | Trimmed mixture likelihoods

We can consider a trimmed mixture likelihood approach, where regular observations are assumed to be random realizations of a mixture of K normal distributions, which yields the maximization:

$$\sum_{i \notin R_0} \log\left(\sum_{k=1}^{K} p_k \phi(x_i; m_k, S_k)\right), \tag{2}$$

where R_0 again includes a proportion $[n\alpha]$ of observations. This trimmed mixture likelihood, combined with the eigenvalues-ratio constraint on the scatter matrices, was the proposal in García-Escudero et al. (2014). Trimmed mixture likelihoods were also considered in Neykov et al. (2007), Cuesta-Albertos et al. (2008) and Gallegos and Ritter (2009), as well as robust fixed point estimating equations, in a mixture framework, in Gonzalez et al. (2022).

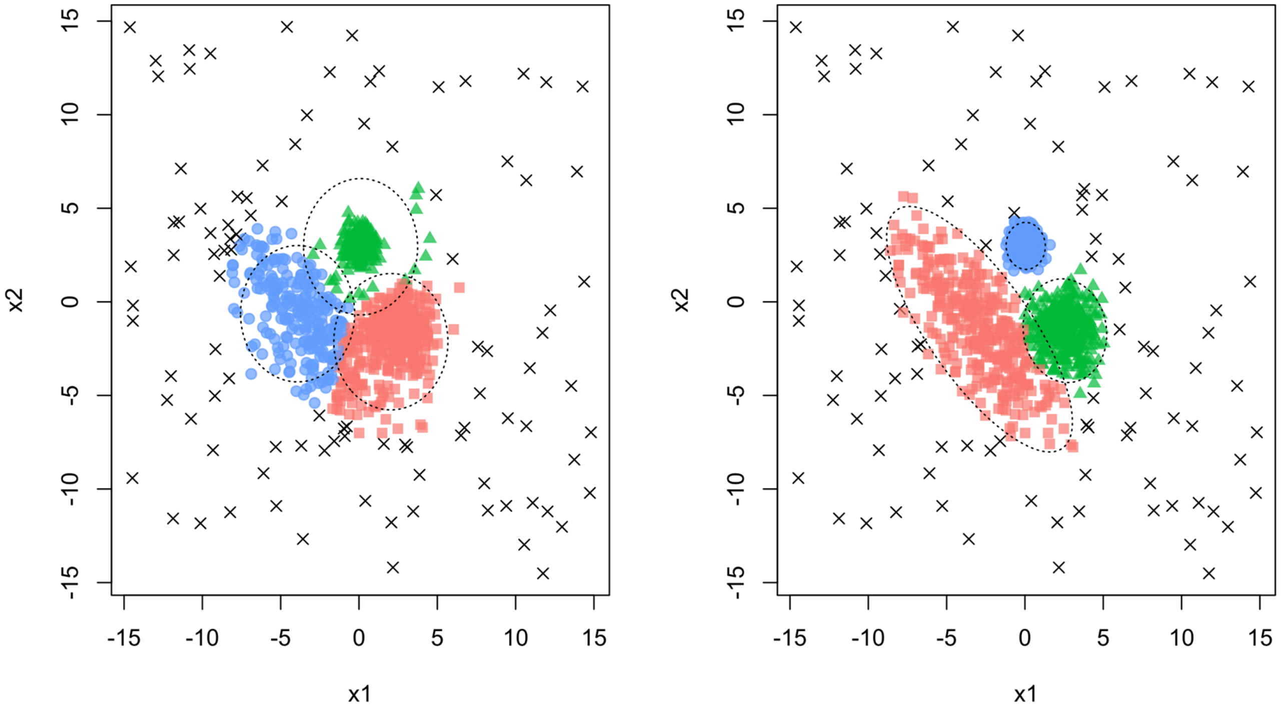

FIGURE 3 Trimmed 3-means (left) and TCLUST (right) for a dataset including nonspherical and nonequally scattered clusters.

The consideration of (trimmed) mixture likelihoods could be advantageous when data is generated from a probability mixture, and our objective is to estimate the corresponding mixture component parameters. This is due to the well-known fact that classification likelihoods produce biased estimators of these parameters (Bryant, 1991). Furthermore, posterior probabilities computed from these (robustly) estimated parameters enable us to define membership probabilities of observations to cluster components, thereby avoiding "crisp" 0–1 assignment decisions that may not be very useful when observations are not unequivocally assigned to any specific cluster.

Instead of using trimmed mixture likelihoods, which attempt to discard noisy observations from the estimating process, we can consider the addition of mixture components in order to "fit" or "accommodate" noisy data. With this viewpoint, using heavy-tailed components through mixtures of *t*-distributions (McLachlan & Peel, 2000) or mixtures of contaminated Gaussian distributions (Punzo & McNicholas, 2016) have been proposed. A uniformly distributed noise component (Banfield & Raftery, 1993), or even an improper constant density component (Coretto & Hennig, 2016; Hennig, 2004), are other possibilities to accommodate noisy observations. The well-known `mclust` package (Scrucca et al., 2023) incorporates the possibility of accommodating outliers within a uniform component. A novel procedure to initialize the noise component based on entropy has been introduced by Scrucca (2023). Another robustification, based on the generalization of Tyler's M-estimators to mixture modeling, can be found in Roizman et al. (2023).

3.3 | Restrictions

The target functions Equations (1) and (2) are unbounded without proper constraints on the scatter matrices $S_1,...,S_K$ when $K>1$. This can be easily seen by allowing one scatter matrix S_k to have a determinant made arbitrarily close to 0 and choosing its corresponding center m_k equal to one of the observations. Besides the theoretical weakness inherent in relying on mathematically ill-posed methods, from a more applied perspective, the typically used algorithms to maximize Equations (1) and (2) may become stuck in non-interesting local maxima without appropriate constraints. These non-interesting local maxima, with little statistical interest, are made of just a few almost colinear observations determining the fitted component and are known in the literature as "spurious solutions" (McLachlan & Peel, 2000). The undesired detection of spurious solutions can be seen as another lack of robustness issue in model-based clustering. In fact, very often, the effect of noise is precisely to favor the detection of spurious solutions (García-Escudero et al., 2014).

With this idea in mind, a main ingredient of TCLUST was to impose, for a fixed constant $c\geq 1$, the eigenvalues-ratio constraint:

$$\frac{\max_{k=1,...,K;j=1,...,p}\lambda_j(S_k)}{\min_{k=1,...,K;j=1,...,p}\lambda_j(S_k)}\leq c, \tag{3}$$

where $\{\lambda_j(\Sigma)\}_{j=1}^p$ denotes the set of p eigenvalues of a scatter matrix Σ. Constraint Equation (3) turns the constrained maximization of Equations (1) and (2) into mathematically well-defined problems for any finite c value and helps to avoid the detection of spurious solutions if very large values of c are excluded. The eigenvalue-ratio constraint is incorporated into the concentration steps in a very efficient manner through an eigenvalue truncation procedure (Fritz et al., 2013a).

Furthermore, note that Equation (3) controls the relative size of the scatter matrices $S_1,...,S_K$ and can therefore help to specify the type of cluster solutions that the user is particularly interested in. For instance, when choosing $c=1$, we will be constraining the scatter matrices to be $S_1=...=S_K=r\times \mathbf{I}_p$, where $\mathbf{I}_p$ denotes the $p\times p$ identity matrix and $r>0$ (thus assuming similar assumptions to trimmed K-means, unless the additional constraint $p_1=...=p_K$ in weights is not imposed). Figure 4 shows the results of applying TCLUST with $K=3$, $\alpha=0.1$ and $c=50$ (left) and with $K=6$, $\alpha=0.1$ and $c=2$ (right). The large $c=50$ value allows the detection of a more elongated cluster (with one notably larger eigenvalue). However, all eigenvalues must be very similar when $c=2$, which results in clusters close to sphericity and with similar scatters.

Other types of constraint have been considered in robust clustering based on trimming. For instance, imposing $S_1=...=S_K$ was proposed in Gallegos and Ritter (2005).

The eigenvalues-ratio constraints in Equation (3) lack affine equivariance, so it is recommended to scale the variables before their application. To approximate affine equivariance, while still being protected against spurious solutions, one can consider the "determinant-and-shape" constraints introduced in García-Escudero et al. (2020). Dotto and Farcomeni (2019) also explored a trimming approach to handle parsimonious parameterizations of the scatter matrices

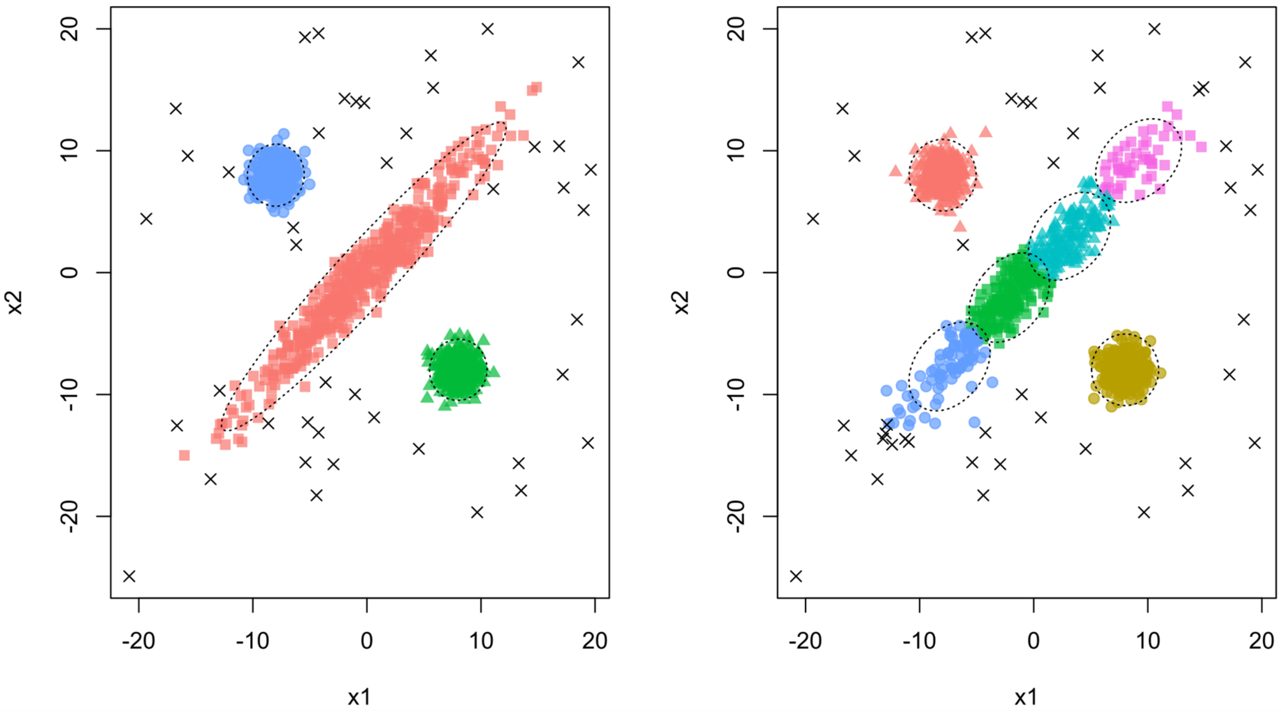

FIGURE 4 Results of applying TCLUST with $K=3$, $\alpha=0.10$ and $c=50$ (left) and $K=6$, $\alpha=0.10$ and $c=2$ (right).

(Celeux & Govaert, 1995) and identified scenarios where the eigenvalues-ratio constraints in Equation (3) could be omitted. More recently, García-Escudero et al. (2022) introduced a set of constraints that cover the well-known 14 parsimonious models in Celeux and Govaert (1995) as limit cases and which can be easily extended to robust clustering based on trimming.

An alternative application of trimming was proposed by Seo and Kim (2012) to prevent the detection of spurious solutions in the "k-deleted log likelihood" approach. This proposal suggests removing the k observations with the largest (log-)likelihood terms, rather than only trimming the observations with the smallest contributions to the likelihood.

More information about the interest of constraints and other possibilities to avoid spurious solutions in model-based clustering can be seen in García-Escudero et al. (2018).

4 | CLUSTERING AROUND AFFINE SUBSPACES

4.1 | Robust linear grouping

The methods seen so far could be described as generalizing the (trimmed) K-means method, in the sense that they are based on searching for clusters around K centroids. However, sometimes the presence of clusters is due to the existence of underlying relationships between the studied variables that, assuming linearity for them, would be associated with the problem of robust clustering around affine subspaces. This provides simultaneously robust cluster detection and dimensionality reduction, extending the type of dimensionality reduction that is done when applying the well-known Principal Components Analysis (PCA), but assuming the presence of K heterogenous subpopulations in our data.

The Robust Linear Grouping (RLG) algorithm in García-Escudero et al. (2009) is based on the search for K affine subspaces $h_1,\ldots,h_K$ with (intrinsic) dimensions $q_1,\ldots,q_K$ $(q_k<p)$ and a partition $\{R_0,R_1,\ldots,R_K\}$ of $\{1,\ldots,n\}$ such that R_0 includes a proportion $[n\alpha]$ of indices and minimizes

$$\sum_{k=1}^{K}\sum_{i\in R_k}\|x_i-\mathrm{Pr}_{h_k}(x_i)\|^2,$$

where $\Pr_h(x)$ denotes the orthogonal projection of x onto the affine subspace h. Although the original formulation in García-Escudero et al. (2009) assumed $q_1 = \ldots = q_K = p-1$, a more general statement is straightforward if we assume unequal intrinsic dimensions (not necessarily all equal to $p-1$). When $K=1$ and $\alpha=0$, RLG reduces to the classical PCA, and when $K=1$ and $\alpha>0$ to the LTS-PCA in Maronna (2005) (see also Croux et al., 2017). The RLG algorithm is a direct extension of the trimmed K-means algorithm, but replacing the distances to the centers by orthogonal distances to the affine subspaces, so K different PCA problems are solved in each step of updating affine subspaces.

4.2 | Robust clusterwise linear regression

In several problems, the existence of a privileged response variable that has to be explained by other explanatory ones is also common. Detecting K clusters around regression fits or clusterwise regression is a problem where lack of robustness is clearly expected. Recall that even ordinary least squares regression, which corresponds to the case $K=1$, is highly non-robust. Considering orthogonal errors, as in Section 4.1, does not seem the most suitable approach in clusterwise regression, and is more appropriate for considering the residuals typically applied in regression. This motivates the approach in García-Escudero, Gordaliza, Mayo-Iscar, and San Martín (2010) where, given values y_i for the response variable and the corresponding values for the p explanatory variables in $x_i = \left(x_{i1}, \ldots, x_{ip}\right)' \in R^p$, the proposal is to search for intercepts $\beta_0^1, \ldots, \beta_0^K$, slope vectors $b^1, \ldots, b^K$, weights $p_1, \ldots, p_K$ with $\sum_{k=1}^{K} p_k = 1$, variances of the error terms $\sigma_1^2, \ldots, \sigma_K^2$ and a partition $\{R_0, R_1, \ldots, R_K\}$ of $\{1, 2, \ldots, n\}$ with R_0 including a proportion $[n\alpha]$ of indices and minimizing

$$\sum_{k=1}^{K}\sum_{i\in R_k}\log\left(p_k\varphi\left(y_i;\beta_0^k+\left(b^k\right)'x_i,\sigma_k^2\right)\right), \tag{4}$$

where $\varphi(\cdot;\mu,\sigma^2)$ denotes the density of the univariate normal distribution with mean μ and variance σ^2. Constraints on the variances of the error terms can be added to avoid the detection of spurious clusters and ensure that the maximization problem is well-defined. The use of Equation (4), when removing constraints on the variances of the error terms σ_k^2, is a particular case of the general trimmed mixture likelihood approach in Neykov et al. (2007). Hennig (2003) also proposed another approach based on trimming by using an iterative fixed-point procedure.

It is well known that “bad leverage points” are a particular type of contamination that can be extremely harmful in regression. The trimming based on the maximization of Equation (4) does not necessarily provide protection against bad leverage points because they do not always result in large $y_i - \beta_0^k + \left(b^k\right)' x_i$ values. To address this issue, a “second trimming” step was proposed in García-Escudero, Gordaliza, Mayo-Iscar, and San Martín (2010). Torti et al. (2019) introduced a procedure for flexibly estimating that second level trimming, eliminating the requirement to fix it beforehand. Yao et al. (2014) also suggested fitting a mixture of regressions based on the t-distribution for the error terms, after adaptively trimming potential bad leverage points. However, a different solution to this problem was proposed in García-Escudero et al. (2017) arising from robustification through trimming and the constraints of the Cluster Weighted Model (CWM) introduced in Gershenfeld et al. (1999). In addition to taking advantage of the cases where the explanatory variables provide valuable additional information about clusters, robustification by trimming and constraints of the CWM avoids the (not very natural) second trimming to prevent the effect of bad leverage points, given that bad leverage points are atypical in the explanatory variables. This approach requires the specification of a single trimming level α to automatically decide the observations to be trimmed in a unified manner.

Figure 5 shows two simulated data examples where robust clusterwise linear regression can be applied with $K=2$, and where a 10% fraction of uniformly distributed background noise was added (first row), as well as a 10% fraction of pointwise contamination at the position $x=15$ and $y=10$ (second row). The direct results of fitting a $K=2$ mixture of regressions, by maximizing Equation (4) with $\alpha=0$ and no constraints, is shown in the first column, together with the results of maximizing Equation (4) with $\alpha=0.1$ and constraints in the second column and, finally, the results of the trimmed and constrained CWM in the third column. We can see that trimming and constraints are sufficient to handle the background noise case, apart from some misclassifications due to the “intersections” of the fitted regression lines. However, the problem is worst in the pointwise contamination case, since the observations corresponding to the pointwise contamination clearly act as bad leverage points even when applying trimming. On the other hand, the use of the trimmed CWM can deal perfectly with these two types of contamination and takes advantage of the clustering information from the x-values to avoid problems with the intersecting regression fits.

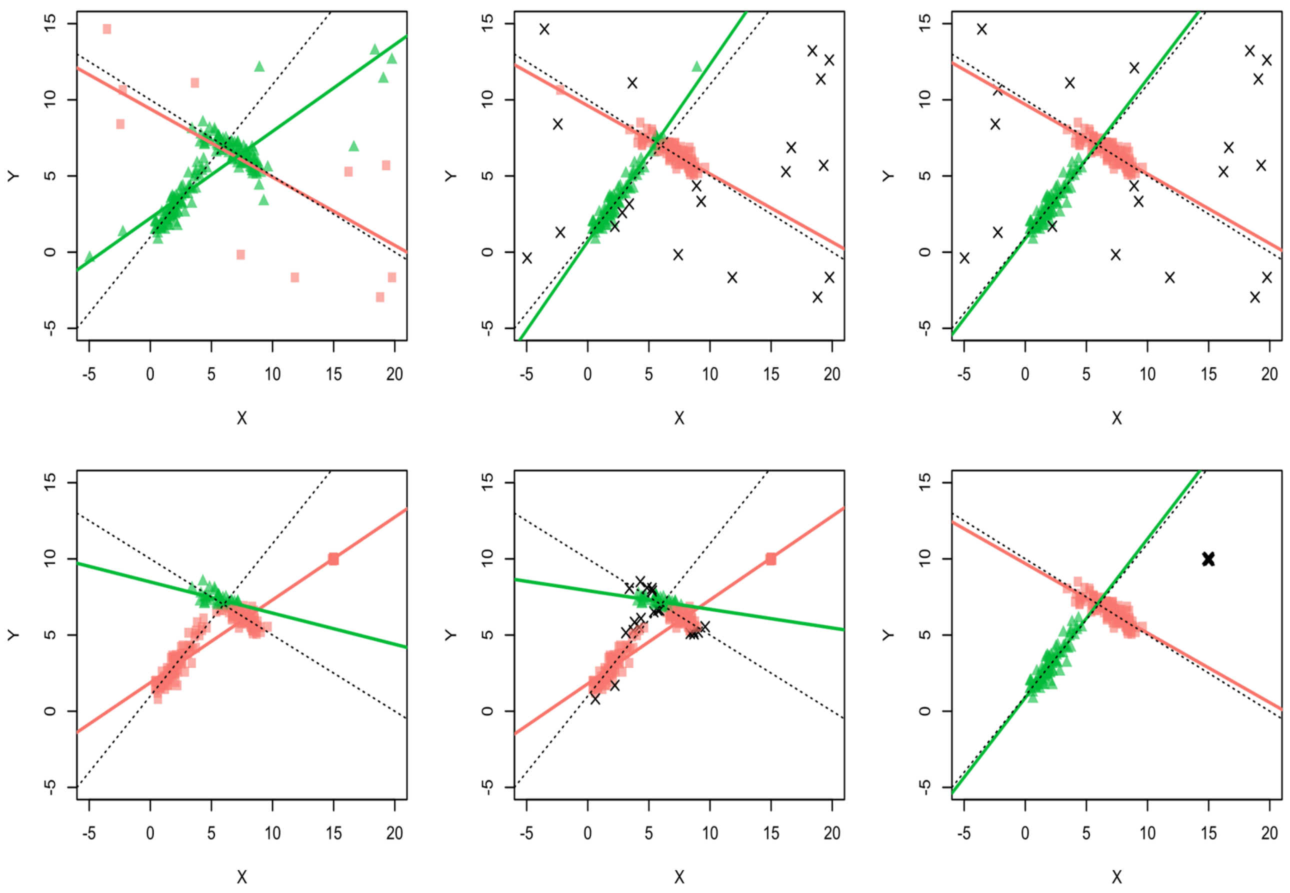

FIGURE 5 Two simulated datasets for clusterwise regression with uniformly distributed background noise (first row) and pointwise contamination at $x = 15$ and $y = 10$ (second row). The dotted lines are the real regression lines to be estimated. The clustering result and the fitted regression lines are shown with no trimming (first column), trimming and constraints (second column), and the trimmed and constrained CWM (third column).

5 | EXTENSIONS

Although a normality assumption has been consistently considered (implicitly or explicitly) in the previous sections, other distributional assumptions can be treated analogously. For instance, the trimmed likelihoods in Neykov et al. (2007) can be applied for mixtures of GLMs. Additionally, the use of trimming combined with assuming "contaminated normal" distributions for the components entails a distinction between mild and gross outliers, as can be seen in Farcomeni and Punzo (2020).

Fuzzy clustering (see, e.g., Ferraro & Giordani, 2020) is also an interesting clustering approach that can be made more robust through trimming. For instance, the "noise component" approach (Dave, 1991), which is a very popular method to deal with noisy data in fuzzy clustering, can be adapted to pre-specify the fixing of a fraction α of observations to be declared as noise. This can be seen as a fuzzy extension of trimmed K-means, while a more general fuzzy extension of the TCLUST method was given in Fritz et al. (2013b).

One of the biggest challenges that robust clustering must increasingly face is dealing with higher dimensional cases where, of course, outliers are also commonly encountered. Consequently, the necessity for dimension reduction techniques that also have a good robustness performance is clearly relevant. The trimmed version of the mixtures of Factor Analyzers in García-Escudero, Gordaliza, Greselin, et al. (2016) is an example of this. An extreme case of the higher dimensional problem is that of functional data. Garcia-Escudero and Gordaliza (2005) provides an impartial trimming procedure based on B-spline representations of the functions to be clustered, while Cuesta-Albertos and Fraiman (2007) applies impartial trimming directly by working on a functional Hillbert space. A model-based approach based on trimmed "pseudo" likelihoods and constraints was given in Rivera-García et al. (2019).

Another problem associated with higher dimensionality is cellwise contamination (Alqallaf et al., 2009). Trimming entire observations, denoted as $x_i = \left(x_{i1}, ..., x_{ip}\right)'$, or case-wise trimming, is too extreme when dimension p is large. Even a small proportion of outlying cells x_{ij} in our data set would result in the need to trim a substantial number of cases x_i whenever a large number of these observations include at least one outlying measurement in their p cells. A cellwise trimming approach was proposed in García-Escudero et al. (2021), which extends the methodology in Maronna and Yohai (2008) to $K > 1$ clusters. Farcomeni (2014) previously introduced the snipped K-means as a cellwise trimmed extension of the K-means.

Simultaneously finding clusters and selecting clustering variables by promoting sparsity is another way to deal with high-dimensionality. Kondo et al. (2016) proposes a robust and sparse K-means for this purpose and, more recently, Brodinová et al. (2019) proposed a methodology in that direction which does not require pre-specifying any trimming level in advance.

Trimming in co-clustering, that is, simultaneously clustering of rows and columns, has also been considered in Farcomeni (2009) and, more recently, in relation to the Latent Block Models, in Fibbi et al. (2023).

Finally, another application of trimming to robust clustering is presented in del Barrio et al. (2019), where trimmed K-barycenters in Wasserstein space are considered for the robust aggregation of clustering partitions. In this approach, the most anomalous clustering partitions are discarded when performing clustering aggregation.

6 | ALGORITHMS AND CODE

All the reviewed proposals for robust clustering based on trimming involve solving complex combinatorial problems, because all possible partitions of our sample $x_1, x_2, ..., x_n$ into $K+1$ subsets are considered, with one of these subsets including the fixed fraction α of observations to be trimmed. However, for this extremely complex computational problem, modifications of the classical EM and the classification EM algorithm, incorporating concentration steps, can often yield very good results. As already commented, this type of concentration steps was first introduced in Rousseeuw and van Driessen (1999) for the MCD estimator. The term "concentration" means that, in each step, the goal is to identify regions of the sample space where observations are more concentrated, making it less likely to find isolated outliers among them. It is worth noting that this concentration-step philosophy aligns well with the clustering goals because, in fact, the classical Lloyd's algorithm for K-means aims to move K-means centers to regions where observations are densely concentrated.

In each iteration of the proposed algorithms for trimming, a fixed fraction α of observations in the most "remote" regions (under the assumed model) is temporally discarded and they are not considered when updating the new parameters of the fitted components. This can be done in such a way that monotonic improvements in the associated target functions are guaranteed. As is common, for both robust and non-robust clustering methods, the associated target functions are far from being convex, and dealing with the existence of local optima can be a serious problem. To increase the chances of detecting the global optimum, it is common to use multiple random initializations of the parameters, which are further refined through concentration steps. The random initializations are typically obtained by using "elemental sets," which are the minimal number of observations needed to estimate the parameters for the K fitted components. When the dimension p increases, initializing these algorithms becomes a challenging task. For instance, selecting $K(p+1)$ random observations from the available data set is a common procedure for initializing parameters in TCLUST. It is important to note that $p+1$ observations in general position are required to initialize each of the K scatter matrices. Additionally, obtaining a suitable partition for these $K(p+1)$ random observations into K subsets is necessary to derive correct initializations of the components' location vectors and scatter matrices. Unfortunately, obtaining such a convenient subsample along with a correct partition becomes increasingly improbable as the dimension, p, increases. Therefore, there is still much work to be done to enhance these algorithms and increase the chances of finding the global optimum. For instance, optimized initialization strategies, as demonstrated in the $K=1$ case in Hubert et al. (2012), can be explored.

Algorithms following the concentration steps philosophy underlie the `tclust` package, available at the CRAN repository (Fritz et al., 2012). The `fsda` Matlab toolbox (Riani et al., 2012) can also implement several trimming approaches to clustering.

7 | CHOICE OF INPUT PARAMETERS

All the previously reviewed methodologies require the specification of different input parameters, among which the number of groups K and the trimming level α stand out.

The choice of the number of clusters K can be considered the quintessential problem in Cluster Analysis. This problem has seen and continues to see numerous proposed solutions, making it one of the most active research areas in this field. In robust clustering based on trimming, this problem is even more complex, given its dependence on the trimming level α. For instance, focusing the discussion on TCLUST, a high α can lead to trimming small clusters, thus requiring a smaller K. Moreover, the proper choice of K and α is also linked to the parameter c in Equation (3), which controls the differences allowed in the scatters, because a high value of c would allow the determination of a cluster including very scattered observations that might otherwise have been considered as background noise, thus requiring a higher α. In any case, TCLUST provides a flexible and unified mathematical framework to address the simultaneous choice of these three parameters K, α, and c, as well as analyzing their interrelationships.

It is important to note that the problem of choosing K in clustering is not a perfectly well-defined problem because the required number of clusters clearly depends on the practical application in which is intended to be used. To this end, let us go back to the dataset shown in Figure 4. A simple visual inspection of this dataset would surely suggest $K = 3$ distinct clusters with a small proportion of outlying observations, requiring a not too large α. Certainly, this would be a reasonable solution in many potential applications of clustering for this dataset and corresponds to the solution in Figure 4 (left) with $K = 3$, $\alpha = 0.03$, as well as a large value of c that allows the detection of the third most elongated cluster. However, let us assume, similarly to Hennig and Liao (2013), that we are dealing with a social stratification problem in which these observations represent families, and where the x_1-axis corresponds to the families' "income levels," and the x_2-axis reflects a measure of their "social status." Would it make sense for the most elongated cluster in Figure 4 (left) to include low-income and low-status social families clustered together with high-income and high-status social families? In that specific application, it would seem more reasonable to consider a value of c close to 1 (more spherical groups with more similar sizes), along with a higher number of groups K, as shown in Figure 4 (right). Consequently, even though we are dealing with a nominally "unsupervised" classification problem, it does not seem reasonable that the choice of the number of clusters K (and other parameters as the trimming level) could be a fully unsupervised procedure that does not require an even minimal user intervention to specify the type of clusters required. The use of constraints, such as those incorporated in TCLUST, could allow users to explicitly declare the type of clusters they are seeking.

Having fixed the parameter c, García-Escudero et al. (2011) proposed considering the so-called classification trimmed likelihood curves (ctlcurves) to assist the user in choosing K and α. The ctlcurves monitor the evolution of the maximum value of the objective function Equation (1) while varying K and α. The consideration of the weights p_k in Equation (1) is key because, in some cases, it is of interest to set some weights as $p_k = 0$ when the dataset does not suggest increasing K for particular c and α values (this possibility was already suggested in Bryant (1991) in the untrimmed case). Figure 6 shows the ctlcurves for the data set in Figure 4: Figure 6 (left) shows these curves when $c = 50$ and Figure 6 (right) when $c = 2$. We can see that it is clearly beneficial to increase K until $K = 4$ is reached, but not to increase K if $K > 4$ when $\alpha = 0$ and a large $c = 50$ are considered. This means that $\alpha = 0$ and $K = 4$ are sensible choices when $c = 50$, because one very scattered cluster gathers all the outlying observations, so that no trimming is needed. On the other hand, it is no longer convenient to increase K from $K = 3$ once an approximate 3% fraction of possible outliers has been trimmed. Therefore, $K = 4$ and $\alpha = 0$ or, alternatively, $K = 3$ and $\alpha = 0.03$ are both sensible parameter choices when $c = 50$ is pre-specified by the user. On the other hand, the very elongated cluster cannot be detected with a smaller value of $c = 2$ and, in that case, it is preferable that this more elongated elliptical component should be split into further clusters, thus requiring higher values of K, as can be seen in Figure 6 (right).

A precursor to the ctlcurves was the trimmed K-variogram in García-Escudero et al. (2003), which monitors changes in the trimmed K-means target function. This target function strictly decreases as K increases for any α, and the rates of decrease have to be analyzed.

As previously mentioned, another advantage of considering appropriate constraints on the scatter parameters is that well-defined optimization problems are considered. This is quite important when considering parameter selection criteria based on penalized likelihoods, such as BIC (Fraley & Raftery, 1998) or AIC (Biernacki et al., 2000). If the likelihood is unbounded (which happens for some of these methods without any constraint), the penalized likelihood is also unbounded, which, combined with the detection of spurious solutions, can be a very serious drawback for choosing parameters. The use of penalized criteria considering the strength of constraints on scatter parameters, as introduced in

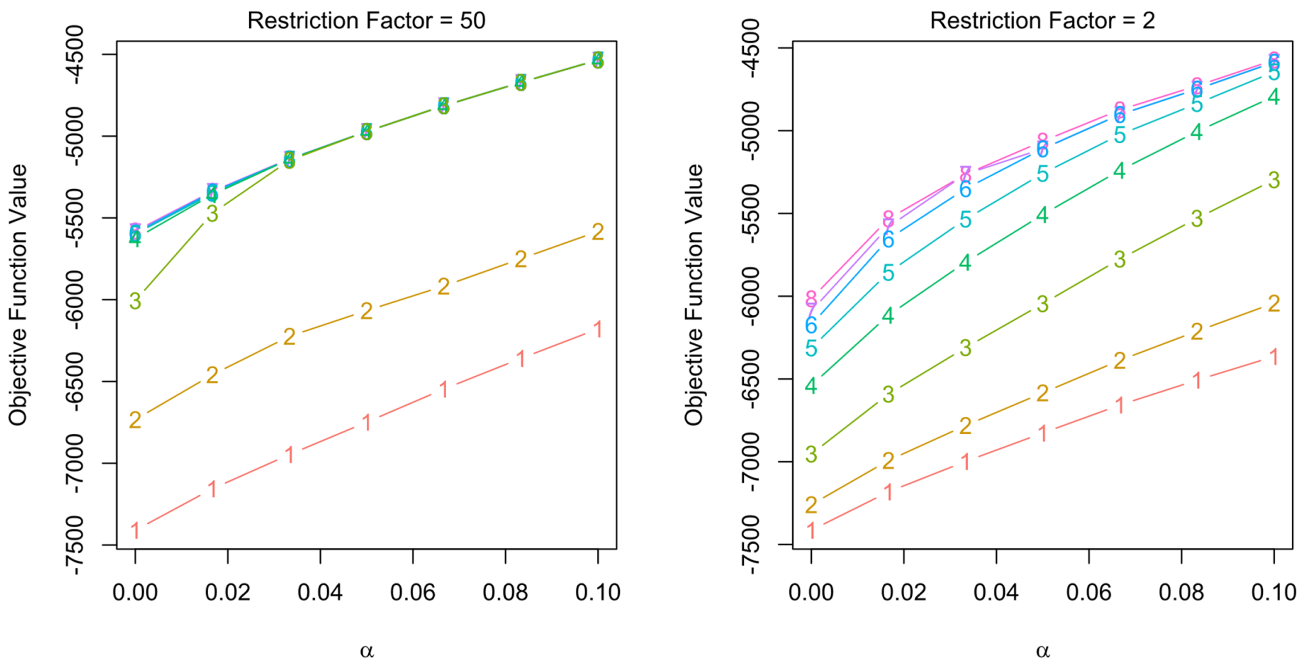

FIGURE 6 Application of ctlcurves to the dataset in Figure 4 when $c = 50$ (left) and $c = 2$ (right).

Cerioli et al. (2018), can be useful in overcoming this difficulty. There, so-called "carbike" plots were introduced to assist the user in choosing sensible c and K values. That proposal was formulated for the $\alpha = 0$ case, but a natural extension would be to incorporate the trimming level α in the definition of the penalized criterion. Some suggestions and proposals in this direction were made through the "trimmed BIC" in Neykov et al. (2007) and the "corrected BIC" in Gallegos and Ritter (2009). An alternative approach was suggested in Gallegos and Ritter (2018), wherein they proposed searching for sensible solutions related to Pareto points. These points act as a compromise between "scale balance" (interpreted as the difference between component scatters) and "model fit" (interpreted as the value attained in the defining likelihood target function) among all local maxima of the likelihood. Additionally, they proposed a procedure for their proper exploration.

Notice also that choosing α slightly higher than necessary is not necessarily a serious problem for methods based on trimming. The estimation of cluster parameters, especially those related to the location of the clusters, are not typically severely affected by considering a slightly higher than needed α. Even though some non-outlying observations are wrongly trimmed, concentration steps seek regions where observations are concentrated, which may still allow sensible estimated parameters to be obtained. Subsequently, we could attempt to recover incorrectly trimmed observations. For instance, a simple proposal to identify those wrongly trimmed observations when using TCLUST could be to resort to Mahalanobis distances, calculated robustly using parameters fitted with only untrimmed observations. This approach is aligned with that in Hardin and Rocke (2004). Building on that idea of starting with a high initial trimming level and sequentially recovering incorrectly wrongly trimmed observations, it is worth noting the proposals in Dotto et al. (2018). This approach involves a re-weighting scheme, where better estimators of the scatter matrix "scales" (a scatter matrix Σ is decomposed into $\Sigma = \sigma^2 U$ with $\sigma^2 = |\Sigma|^{1/p}$ as a scale parameter, and $U = \Sigma / |\Sigma|^{1/p}$ as a shape matrix) are progressively obtained through iterative steps, even though their initial values were not reliable due to a higher than needed trimming level. García-Escudero and Gordaliza (2007) shows the crucial role that the correct estimation of scale parameters plays in robust clustering. This re-weighting procedure can even correct wrong determinations of c or α. The use of re-weighting techniques is common in Robust Statistics (Lopuhaa & Rousseeuw, 1991). Another idea is to choose α so that the non-trimmed part of the data looks as close to a mixture of Gaussians as possible, as proposed in Coretto and Hennig (2016). Moreover, relying on a simple but highly robust procedure, such as trimmed K-means with a high α, can serve as a useful robust initialization for other types of clustering procedures (see, e.g., Mclachlan et al., 2006 or Cuesta-Albertos et al., 2008).

Many of the proposals for parameter selection are indeed based on monitoring by-products of the application of robust clustering when input parameters are systematically varied. This is in agreement with the view in Atkinson et al.

(2004) and Atkinson and Riani (2007) arguing that the "whole movie" provided by appropriate monitoring processes can be more informative than focusing on the "single frame" provided by one single application of any robust clustering technique. These last two references pertain to a "forward search" approach, which offers a form of adaptive trimming procedure that can be useful in robust clustering (see Atkinson et al., 2018, Riani et al., 2019 or Cerioli et al., 2019).

We would like to insist once more that clustering is not a well-defined problem, whose final solution should not depend at all on possibly different final clustering purposes. Therefore, our final goal could be just to produce a reduced list of sensible partitions, by selecting parameter combinations that may be of interest to the user and discarding parameter combinations that are likely not very interesting. An example of this approach would be, for instance, Torti et al. (2021) or Cappozzo et al. (2023).

Although we have focused on the problem of choosing the parameters K and α, some techniques require some other parameters to be fixed, such as the intrinsic dimensions of subspaces in $q_1,...,q_K$ in the case of the RLG. Therefore, procedures for selecting all of them are also required and it is important to note that all the involved parameters are likely to be interrelated.

8 | CONCLUSIONS

Different robust clustering techniques based on trimming have been summarized in this review. These techniques move from the trimmed K-means, as the simplest case, being an extension of the classical K-means, to more elaborate proposals in more complex scenarios. An impartial trimming approach has been followed, where the dataset itself and the clustering model assumptions automatically determine which observations to trim. The trimming approach can be combined with constraints on the clusters' scatter parameters to avoid detecting spurious clusters, which can be seen as another source of lack of robustness for clustering methods.

There are several open research problems, such as dealing with higher dimensionality. In higher dimensional problems, addressing the challenge of outlying measurements simultaneously with dimensionality reduction or the presence of noninformative variables, which add noise in the sense of not providing useful clustering information, is essential. Moreover, the possibility of cellwise trimming, which does not sacrifice too much information when only a few cells or individual measurements are atypical in one observation, also appears as a key research direction. Another problem that undoubtedly requires additional work is assisting the user in selecting the several parameters required for applying these methodologies. While we do not believe that fully automated parameter selection without any minimal user supervision can be provided, it does seem useful to narrow down this selection to a limited number of reasonable alternatives. All the techniques presented have a notably high computational complexity, and considering computational improvements that reduce computing times and enhance their effectiveness is also important, especially when dealing with high-dimensional problems or an increasing number of observations or clusters.

AUTHOR CONTRIBUTIONS

Luis A. García-Escudero: Conceptualization (equal); writing – original draft (equal). **Agustín Mayo-Iscar:** Conceptualization (equal); writing – original draft (equal).

ACKNOWLEDGMENTS

The authors want to thank our friends and colleagues Alfonso Gordaliza and Carlos Matrán for the continuous and fruitful collaboration. We also express our deep appreciation to other colleagues who have made substantial contributions to this trimming approach to robust clustering.

FUNDING INFORMATION

The authors acknowledge support by grant PID2021-128314NB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

RELATED WIREs ARTICLES

Model-based cluster analysis

Challenges in model-based clustering

Robust statistics: A selective overview and new directions

Cluster analysis: A modern statistical review

Anomaly detection by robust statistics

ORCID

Luis A. García-Escudero https://orcid.org/0000-0002-7617-3034
Agustín Mayo-Iscar https://orcid.org/0000-0003-0951-6508

ENDNOTE

[1] Thucydides in his "History of the Peloponnesian War (II 20, 3-4)" reported that the Plataeans, in 428 B.C., besieged by the Spartans, excluded extreme measurements when estimating the height of the walls that their enemies had built round their city. This reference to Thucydides was given in Cerioli et al. (2011), who thanked Spyros Arsenis and Domenico Perrotta for pointing it out.

REFERENCES

Alqallaf, F., van Aelst, S., Yohai, V. J., & Zamar, R. H. (2009). Propagation of outliers in multivariate data. *Annals of Statistics*, *37*(1), 311–331.

Atkinson, A. C., & Riani, M. (2007). Exploratory tools for clustering multivariate data. *Computational Statistics & Data Analysis*, *52*(1), 272–285.

Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. Springer.

Atkinson, A. C., Riani, M., & Cerioli, A. (2018). Cluster detection and clustering with random start forward searches. *Journal of Applied Statistics*, *45*(5), 777–798.

Banerjee, A., & Davé, R. N. (2012). Robust clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*, 29–59.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725.

Brodinová, Š., Filzmoser, P., Ortner, T., Breiteneder, C., & Rohm, M. (2019). Robust and sparse k-means clustering for high-dimensional data. *Advances in Data Analysis and Classification*, *13*(4), 905–932.

Bryant, P. G. (1991). Large-sample results for optimization-based clustering methods. *Journal of Classification*, *8*, 31–44.

Cappozzo, A., García-Escudero, L. A., Greselin, F., & Mayo-Iscar, A. (2023). Graphical and computational tools to guide parameter choice for the cluster weighted robust model. *Journal of Computational and Graphical Statistics*, *32*(3), 1195–1214.

Celeux, G., & Govaert, A. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, *14*(3), 315–332.

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*(5), 781–793.

Cerioli, A., Atkinson, A. C., & Riani, M. (2011). Some perspectives on multivariate outlier detection. In Ingrassia, S., Rocci, R., Vichi, M. (Eds.), *New perspectives in statistical modeling and data analysis: Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Berlin, Heidelberg.

Cerioli, A., Farcomeni, A., & Riani, M. (2019). Wild adaptive trimming for robust estimation and cluster analysis. *Scandinavian Journal of Statistics*, *46*(1), 235–256.

Cerioli, A., García-Escudero, L. A., Mayo-Iscar, A., & Riani, M. (2018). Finding the number of Normal groups in model-based clustering via constrained likelihoods. *Journal of Computational and Graphical Statistics*, *27*(2), 404–416.

Chawla, S., & Gionis, A. (2013). k-means -: A unified approach to clustering and outlier detection. In *Proceedings of the 2013 SIAM international conference on data mining* (pp. 189–197). Society for Industrial and Applied Mathematics.

Coretto, P., & Hennig, C. (2016). Robust improper maximum likelihood: Tuning, computation, and a comparison with other methods for robust Gaussian clustering. *Journal of the American Statistical Association*, *111*(516), 1648–1659.

Croux, C., Garcia-Escudero, L. A., Gordaliza, A., Ruwet, C., & San Martín, R. (2017). Robust principal component analysis based on trimming around affine subspaces. *Statistica Sinica*, *27*(3), 1437–1459.

Cuesta-Albertos, J. A., & Fraiman, R. (2007). Impartial trimmed *k*-means for functional data. *Computational Statistics & Data Analysis*, *51*(10), 4864–4877.

Cuesta-Albertos, J. A., Gordaliza, A., & Matrán, C. (1997). Trimmed *k*-means: An attempt to robustify quantizers. *Annals of Statistics*, *25*(2), 553–576.

Cuesta-Albertos, J. A., Gordaliza, A., & Matrán, C. (1998). Trimmed best *k*-nets: A robustified version of an L_∞-based clustering method. *Statistics & Probability Letters*, *36*(4), 401–413.

Cuesta-Albertos, J. A., Matrán, C., & Mayo-Iscar, A. (2008). Robust estimation in the normal mixture model based on robust clustering. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(4), 779–802.

Dave, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, *12*(11), 657–664.

del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., & Mayo-Íscar, A. (2019). Robust clustering tools based on optimal transportation. *Statistics and Computing*, *29*(1), 139–160.

Dotto, F., & Farcomeni, A. (2019). Robust inference for parsimonious model-based clustering. *Journal of Statistical Computation and Simulation*, *89*(3), 414–442.

Dotto, F., Farcomeni, A., García-Escudero, L. A., & Mayo-Iscar, A. (2018). A reweighting approach to robust clustering. *Statistics and Computing*, *28*(2), 477–493.

Farcomeni, A. (2009). Robust double clustering: A method based on alternating concentration steps. *Journal of Classification*, *26*, 77–101.

Farcomeni, A. (2014). Snipping for robust *k*-means clustering under component-wise contamination. *Statistics and Computing*, *24*(6), 907–919.

Farcomeni, A., & Greco, L. (2016). *Robust methods for data reduction*. CRC Press.

Farcomeni, A., & Punzo, A. (2020). Robust model-based clustering with mild and gross outliers. *Test*, *29*(4), 989–1007.

Ferraro, M. B., & Giordani, P. (2020). Soft clustering. *Wiley Interdisciplinary Reviews: Computational Statistics*, *12*(1), e1480.

Fibbi, E., Perrotta, D., Torti, F., van Aelst, S., & Verdonck, T. (2023). Co-clustering contaminated data: A robust model-based approach. *Advances in Data Analysis and Classification*, *18*, 121–161. https://doi.org/10.1007/s11634-023-00549-3

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, *41*(8), 578–588.

Fritz, H., García-Escudero, L. A., & Mayo-Iscar, A. (2012). tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, *47*(12), 1–26.

Fritz, H., García-Escudero, L. A., & Mayo-Iscar, A. (2013a). A fast algorithm for robust constrained clustering. *Computational Statistics and Data Analysis*, *61*, 124–136.

Fritz, H., García-Escudero, L. A., & Mayo-Iscar, A. (2013b). Robust constrained fuzzy clustering. *Information Sciences*, *245*, 38–52.

Gallegos, M. T., & Ritter, G. (2005). A robust method for cluster analysis. *The Annals of Statistics*, *33*(1), 347–380.

Gallegos, M. T., & Ritter, G. (2009). Trimming algorithms for clustering contaminated grouped data and their robustness. *Advances in Data Analysis and Classification*, *3*, 135–167.

Gallegos, M. T., & Ritter, G. (2018). Probabilistic clustering via Pareto solutions and significance tests. *Advances in Data Analysis and Classification*, *12*, 179–202.

García-Escudero, L. Á., & Gordaliza, A. (1999). Robustness properties of *k*-means and trimmed *k*-means. *Journal of the American Statistical Association*, *94*(447), 956–969.

Garcia-Escudero, L. A., & Gordaliza, A. (2005). A proposal for robust curve clustering. *Journal of Classification*, *22*(2), 185–201.

García-Escudero, L. A., & Gordaliza, A. (2007). The importance of the scales in heterogeneous robust clustering. *Computational Statistics and Data Analysis*, *51*(9), 4403–4412.

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

García-Escudero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S., & Mayo-Iscar, A. (2017). Robust estimation of mixtures of regressions with random covariates, via trimming and constraints. *Statistics and Computing*, *27*(2), 131–147.

García-Escudero, L. A., Gordaliza, A., Greselin, F., Ingrassia, S., & Mayo-Iscar, A. (2018). Eigenvalues and constraints in mixture modeling: Geometric and computational issues. *Advances in Data Analysis and Classification*, *12*, 203–233.

García-Escudero, L. A., Gordaliza, A., & Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, *12*(2), 434–449.

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

García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, *4*(2), 89–109.

García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2011). Exploring the number of groups in robust model-based clustering. *Statistics and Computing*, *21*(4), 585–599.

García-Escudero, L. A., Gordaliza, A., Matrán, C., Mayo-Iscar, A., & Hennig, C. M. (2016). Robustness and outliers. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 653–678). Serie Chapman & Hall/CRC Handbooks of Modern Statistical Methods.

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

García-Escudero, L. A., Gordaliza, A., Mayo-Iscar, A., & San Martín, R. (2010). Robust clusterwise linear regression through trimming. *Computational Statistics and Data Analysis*, *54*(12), 3057–3069.

García-Escudero, L. A., Gordaliza, A., San Martín, R., van Aelst, S., & Zamar, R. (2009). Robust linear clustering. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *71*(1), 301–318.

García-Escudero, L. A., Mayo-Iscar, A., & Riani, M. (2020). Model-based clustering with determinant-and-shape constraint. *Statistics and Computing*, *30*, 1363–1380.

García-Escudero, L. A., Mayo-Iscar, A., & Riani, M. (2022). Constrained parsimonious model-based clustering. *Statistics and Computing*, *32*(2), 1–15.

García-Escudero, L. A., Rivera-García, D., Mayo-Iscar, A., & Ortega, J. (2021). Cluster analysis with cellwise trimming and applications for the robust clustering of curves. *Information Sciences*, *573*, 100–124.

Gershenfeld, N., Schoner, B., & Metois, E. (1999). Cluster-weighted modelling for time-series analysis. *Nature*, *397*(6717), 329–332.

Gonzalez, J. D., Maronna, R., Yohai, V., & Zamar, R. (2022). Robust model-based clustering. *Journal of Data Science, Statistics, and Visualisation*, *2*(6), 1-29.

Hardin, J., & Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, *44*(4), 625–638.

Hennig, C. (2003). Clusters, outliers, and regression: Fixed point clusters. *Journal of Multivariate Analysis*, *86*(1), 183–212.

Hennig, C. (2004). Breakdown points for maximum likelihood-estimators of location-scale mixtures. *Annals of Statistics*, *32*(4), 1313–1340.

Hennig, C., & Liao, T. F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *62*(3), 309–369.

Hubert, M., Rousseeuw, P. J., & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, *21*(3), 618–637.

Ibáñez, M. V., Vinué, G., Alemany, S., Simó, A., Epifanio, I., Domingo, J., & Ayala, G. (2012). Apparel sizing using trimmed PAM and OWA operators. *Expert Systems with Applications*, *39*(12), 10512–10520.

Jobe, J. M., & Pokojovy, M. (2015). A cluster-based outlier detection scheme for multivariate data. *Journal of the American Statistical Association*, *110*(512), 1543–1551.

Kondo, Y., Salibian-Barrera, M., & Zamar, R. (2016). RSKC: An R package for a robust and sparse k-means clustering algorithm. *Journal of Statistical Software*, *72*(5), 1–26.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*, 129–137.

Lopuhaa, H. P., & Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, *19*(1), 229–248.

Maronna, R. (2005). Principal components and orthogonal regression based on robust scales. *Biometrics*, *47*(3), 264–273.

Maronna, R., & Yohai, V. (2008). Robust low-rank approximation of data matrices with elementwise contamination. *Technometrics*, *50*(3), 295–304.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with R)*. John Wiley & Sons.

McLachlan, G., & Peel, D. A. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics.

Mclachlan, G. J., Ng, S.-K., & Bean, R. (2006). Robust cluster analysis via mixture models. *Austrian Journal of Statistics*, *35*(2&3), 157–174.

Neykov, N., Filzmoser, P., Dimova, R., & Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, *4*(1), 299–308.

Punzo, A., & McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, *58*(6), 1506–1537.

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

Riani, M., Perrotta, D., & Torti, F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, *116*, 17–32.

Ritter, G. (2015). *Cluster analysis and variable selection*. CRC Press.

Rivera-García, D., García-Escudero, L. A., Mayo-Iscar, A., & Ortega, J. (2019). Robust clustering for functional data based on trimming and constraints. *Advances in Data Analysis and Classification*, *13*(1), 201–225.

Rocke, D. M., & Woodruff, D. L. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, *91*(435), 1047–1061.

Roizman, V., Jonckheere, M., & Pascal, F. (2023). A flexible EM-like clustering algorithm for noisy data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(5), 2709–2721.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley Series in Probability and Statistics.

Rousseeuw, P. J., & van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*(3), 212–223.

Rousseeuw, P. J., & van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. In W. Gaul, O. Opitz, & M. Schader (Eds.), *Data analysis. Studies in classification, data analysis, and knowledge organization*. Springer.

Ruwet, C., García-Escudero, L. A., Gordaliza, A., & Mayo-Iscar, A. (2012). The influence function of the TCLUST robust clustering procedure. *Advances in Data Analysis and Classification*, *6*(2), 107–130.

Ruwet, C., García-Escudero, L. A., Gordaliza, A., & Mayo-Iscar, A. (2013). On the breakdown behavior of the TCLUST clustering procedure. *Test*, *22*(3), 466–487.

Scrucca, L. (2023). Entropy-based anomaly detection for Gaussian mixture modeling. *Algorithms*, *16*(4), 195.

Scrucca, L., Fraley, C., Murphy, T. B., & Raftery, A. E. (2023). *Model-based clustering, classification, and density estimation using mclust in R*. Chapman and Hall/CRC.

Seo, B., & Kim, D. (2012). Root selection in normal mixture models. *Computational Statistics & Data Analysis*, *56*(8), 2454–2470.

Torti, F., Perrotta, D., Riani, M., & Cerioli, A. (2019). Assessing trimming methodologies for clustering linear regression. *Advances in Data Analysis and Classification*, *13*(1), 227–257.

Torti, F., Riani, M., & Morelli, G. (2021). Semiautomatic robust regression clustering of international trade data. *Statistical Methods & Applications*, *30*, 863–894.

Yao, W., Wei, Y., & Yu, C. (2014). Robust mixture regression using the *t*-distribution. *Computational Statistics & Data Analysis*, *71*, 116–127.

How to cite this article: García-Escudero, L. A., & Mayo-Iscar, A. (2024). Robust clustering based on trimming. *WIREs Computational Statistics*, *16*(4), e1658. https://doi.org/10.1002/wics.1658