Wild adaptive trimming for robust estimation and cluster analysis

Wild adaptive trimming for robust estimation and cluster analysis

ANDREA CERIOLI[†],
ALESSIO FARCOMENI[‡],
MARCO RIANI[†]
[†]Department of Economics and Management, University of Parma
[‡]Department of Public Health, Sapienza - University of Rome Italy

Abstract

Trimming principles play an important role in robust statistics. However, their use for clustering typically requires some preliminary information about the contamination rate and the number of groups. We suggest a fresh approach to trimming that does not rely on this knowledge and that proves to be particularly suited for solving problems in robust cluster analysis. Our approach replaces the original K-population (robust) estimation problem with K distinct one-population steps, which take advantage of the good breakdown properties of trimmed estimators when the trimming level exceeds the usual bound of 0.5. In this setting we prove that exact affine equivariance is lost on one hand, but on the other hand an arbitrarily high breakdown point can be achieved by "anchoring" the robust estimator. We also support the use of adaptive trimming schemes, in order to infer the contamination rate from the data. A further bonus of our methodology is its ability to provide a reliable choice of the usually unknown number of groups.

Key words: Breakdown Point; Forward Search; MCD; Outliers; Robust Clustering

1 Introduction

Trimming principles play an important role in robust statistics and allow to solve complex problems in the analysis of contaminated multivariate data (see, e.g., Clarke and Schubert, 2006; Cuesta-Albertos et al., 2008; García-Escudero et al., 2008; Ritter, 2014; Farcomeni and Greco, 2015). Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a random sample of v-variate observations from a population with distribution function $G(x)$. We assume that $G(x)$ is an unknown element within a family $\mathfrak{G}$ of distribution functions such that

$$\mathfrak{G} = \{G(x) : G(x) = (1-\epsilon)G_0(x) + \epsilon G_1(x); \;\; x \in \mathbb{R}^v; \;\; \epsilon \in [0, 1)\}, \qquad (1)$$

where $G_0(x)$ is the distribution function of the "good" part of the data, i.e. $G_0(x)$ represents the postulated null model, $G_1(x)$ is the contaminant distribution belonging

to a class $\mathfrak{C}$ of distributions and ϵ is the contamination rate. Although it is not necessary to define $\mathfrak{C}$ as a specific parametric family, some regularity conditions on it are often assumed (see, e.g., Cuesta-Albertos et al., 2008; Cerioli et al., 2013).

In the one-population version of model (1) it is common to take

$$G_0(x) = \Phi_{\mu,\Sigma}(x), \tag{2}$$

where $\Phi_{\mu,\Sigma}(x)$ is the distribution function of a v-variate normal random variable with mean μ and dispersion Σ. The trimmed estimators of μ and Σ take the form

$$\tilde{\mu}_\alpha = \frac{1}{W_\alpha} \sum_{i=1}^{n} w_{i,\alpha} x_i, \tag{3}$$

$$\tilde{\Sigma}_\alpha = \frac{\zeta_\alpha}{W_\alpha} \sum_{i=1}^{n} w_{i,\alpha} (x_i - \tilde{\mu}_\alpha)(x_i - \tilde{\mu}_\alpha)', \tag{4}$$

where $\alpha \in (0, 0.5)$, either $w_{i,\alpha} = 0$ or $w_{i,\alpha} = 1$, $W_\alpha = \sum_{i=1}^{n} w_{i,\alpha}$, and

$$\zeta_\alpha = \frac{1-\alpha}{G_{\chi^2_{v+2}}(\chi^2_{v,1-\alpha})} \tag{5}$$

is a scaling factor ensuring consistency of $\tilde{\Sigma}_\alpha$ when $\epsilon = 0$. In (5), we define $G_{\chi^2_v}(\cdot)$ to be the distribution function of a χ^2_v random variable, while

$$\chi^2_{v,1-\alpha} = G^{-1}_{\chi^2_v}(1-\alpha) \tag{6}$$

is its $(1-\alpha)$-th quantile. Computation of the binary weights $w_{i,\alpha}$, $i = 1, \ldots, n$, is sketched in §2.1 and §2.2 for two alternative trimmed estimators. In all cases, the weights $w_{i,\alpha}$ are defined in such a way that $W_\alpha = \lfloor (1-\alpha)n \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. The number α thus gives the trimming level, i.e. the proportion of observations discarded by the robust procedure. The squared distances

$$d^2_{i,\alpha} = (x_i - \tilde{\mu}_\alpha)' \tilde{\Sigma}^{-1}_\alpha (x_i - \tilde{\mu}_\alpha) \qquad i = 1, \ldots, n \tag{7}$$

are used for outlier identification and, more generally, for robustly ordering multivariate data (Atkinson et al., 2004; Hubert et al., 2008; Riani et al., 2009; Cerioli, 2010).

In the standard approach to trimming α must be fixed in advance, thus requiring some a priori information on the degree of contamination in model (1). Otherwise, the usual suggestion is to choose $\alpha = \epsilon^* - 1/n$, where

$$\epsilon^* = \lfloor (n - v + 1)/2 \rfloor / n \approx 0.5 \tag{8}$$

is the maximal value of the (replacement) breakdown point of $\tilde{\mu}_\alpha$ and $\tilde{\Sigma}_\alpha$ (Davies, 1987; Lopuhaä and Rousseeuw, 1991). Under model (1), this choice corresponds to the assumption that at least 50% plus another $\lceil (v-1)/2 \rceil$ observations, where $\lceil \cdot \rceil$ is the ceiling function, come from $G_0(x)$, i.e. that

$$\epsilon < \frac{1}{2}. \tag{9}$$

Condition (9) is very natural in one-population models for $G_0(x)$, like (2) (see, e.g., Rousseeuw and Leroy, 1987, p. 14). However, choosing $\alpha = \epsilon^* - 1/n$ makes computation of the trimmed estimators $\tilde{\mu}_\alpha$ and $\tilde{\Sigma}_\alpha$ virtually useless when there is more than one "good" population and the goal is to robustly cluster the observations in $\mathcal{X}$ according to these populations. In a multi-population structure for $G_0(x)$, we typically assume that the "good" data come from

$$G_0(x) = \sum_{k=1}^{K} \pi_k \Phi_{\mu_k, \Sigma_k}(x), \qquad \pi_k > 0, \qquad \sum_{k=1}^{K} \pi_k = 1, \tag{10}$$

where K is the unknown number of populations, μ_k and Σ_k are population-specific parameters, and $\pi_1, \ldots, \pi_K$ are the unknown mixing proportions. It is then possible to identify the largest population in $G_0(x)$ through estimates (3) and (4), and the associated robust distances (7), only in the unlikely situation where $\max_k \pi_k \gg 0.5$. Indeed, a simple example where trimming methods fail to detect a multi-population structure even if $\max_k \pi_k \approx 0.6$ is provided by Atkinson et al. (2004, pp. 372–373). A related qualitative comment is made by Huber and Ronchetti (2009, p. 21).

The goal of this work is to suggest a fresh approach to trimming that allows computation of the trimmed estimators $\tilde{\mu}_\alpha$ and $\tilde{\Sigma}_\alpha$ also when $G_0(x)$ follows (10) with $\max_k \pi_k \leq 0.5$. Our methodology is particularly suited for solving problems in robust cluster analysis, whose aim is to identify individual membership to the populations that originate mixture (10). This is achieved by replacing a K-population (robust) estimation step, at the heart of the available model-based clustering algorithms, with K distinct one-population steps, which take advantage of (3) and (4). Specifically, our proposal consists in computing the trimmed estimators $\tilde{\mu}_\alpha$ and $\tilde{\Sigma}_\alpha$ with $\alpha \geq 0.5$, and possibly much larger than the usual bound $(\epsilon^* - 1/n)$. We name our method *wild trimming*, since we suggest to trim much more than it is customary following (8). Indeed, our trimming level can as large as the highest value of α for which $\tilde{\Sigma}_\alpha$ is positive definite. We also strongly support the use of adaptive procedures (in the sense described by Huber and Ronchetti, 2009, p. 8), where the trimming level is not fixed in advance, but different values of α are used and the best one is selected from the data, thus yielding a closer agreement between α in (3) and (4), and ϵ in (1). Therefore, the output of our methodology is a robust procedure where not all units in $\mathcal{X}$ are classified into groups, since we discard the observations that are believed to come from the contaminant distribution $G_1(x)$, and where neither the number of groups K nor the trimming level α are specified a priori.

In spite of its simplicity, we believe that the idea behind wild trimming has not gained the popularity that it deserves. One motivation lies in the often implicit assumption that the "good" population should correspond to the majority of data. Instead, our point of view is different and we define outlyingness of a multivariate observation with respect to a specific point, say x_0, ideally sampled from $G_0(x)$. In our framework also the bulk of the data can become anomalous if sufficiently far from x_0. Wild trimming thus provides a very natural approach, since condition (9) is not required. Although the robustness properties of the estimators obtained with $\alpha \geq 0.5$ turn out to be far from trivial, they are intuitively appealing and obviously do not contradict the well known findings for the case $\alpha < 0.5$, such as the fundamental bound (8). One excep-

tion towards the application of wild trimming ideas is provided by the Forward Search, which has shown good potential for performing robust cluster analysis (Atkinson et al., 2004, 2017), but mainly in an exploratory context. Our work thus puts robust clustering through the Forward Search within a statistically principled framework where the robustness properties of the algorithm are made explicit. However, our goal is more ambitious and through our methodology we aim at performing robust estimation and cluster analysis under the same umbrella, with data-driven selection of both K and α. This task is clearly not possible in the standard approach to trimming, where $\alpha < 0.5$, and is also problematic through the available robust clustering techniques, which require some a priori information about the features of models (1) and (10).

The paper is structured as follows. In §2 we introduce the two multivariate trimmed estimators that we use in our work. In §3 we obtain the robustness properties of these estimators under wild trimming and we define a new class of estimators having the required breakdown properties. The suggested robust divisive clustering method is detailed in §4. We show the practical advantages of our proposal in §5, where we also provide comparisons. Some concluding remarks are given in §6. Further examples are provided in the Supplementary Material.

2 Multivariate (adaptive) trimming

2.1 The Minimum Covariance Determinant

For trimming level α, the Minimum Covariance Determinant (MCD) subset of $\mathcal{X}$ is defined as the subsample of $h_\alpha = \lfloor (1-\alpha)n \rfloor$ observations whose covariance matrix has the smallest determinant. Let $\iota_\alpha^\flat = \{i_1, \ldots, i_{h_\alpha}\}$ denote the set of the indices of the observations belonging to this subset. The MCD estimators of μ and Σ (Rousseeuw and Leroy, 1987, p. 262–265) are then defined by (3) and (4), with weights

$$\begin{aligned} w_{i,\alpha} &= 1 \qquad \text{if} \qquad i \in \iota_\alpha^\flat \\ &= 0 \qquad \text{otherwise,} \end{aligned} \tag{11}$$

and $W_\alpha = h_\alpha$. The MCD estimators are consistent under very general conditions on $G(x)$ (Butler et al., 1993; Cator and Lopuhaä, 2012). They also attain the breakdown bound (8) when $\alpha = \epsilon^* - 1/n$. To increase efficiency, while keeping a high breakdown point, a one-step reweighting scheme is often used. Reweighted estimators are computed by giving weight 0 to observations for which the squared robust distance (7) exceeds a threshold value, defined in terms of a new trimming level $\alpha^* \in (0, \alpha)$ and such that $\alpha^* \ll \alpha$. The reweighted MCD (RMCD) estimates are then obtained through (3) and (4), but now with weights

$$\begin{aligned} w_{i,\alpha^*} &= 1 \qquad \text{if} \qquad d^2_{i,\alpha} \leq d^2_{\alpha^*} \\ &= 0 \qquad \text{otherwise,} \end{aligned} \tag{12}$$

and scaling factor $\zeta_{\alpha^*} = (1-\alpha^*)/\{G_{\chi^2_{v+2}}(\chi^2_{v,1-\alpha^*})\}$. A popular choice in (12) is $\alpha^* = 0.025$, so that $d^2_{\alpha^*}$ is the $(1-\alpha^*) = 0.975$-th quantile of the distribution of the squared robust distances (7). When the asymptotic distribution of such distances is

considered $d^2_{\alpha^*} = \chi^2_{v,0.975}$, but more accurate approximations exist (Hardin and Rocke, 2005; Cerioli, 2010). The RMCD estimates can thus be seen as the result of a two-step adaptive trimming procedure, computed using two different subsets of $\mathcal{X}$. Each of these subsets is defined by a specific trimming level: α in the initial step and α^* in the reweighting stage. An advantage is that the number of units declared to be outliers by distances (7), and thus discarded in the second subset, can provide information on the contamination rate ϵ in model (1). A fully adaptive procedure based on the MCD should extend the reweighting scheme to a decreasing sequence of trimming levels, starting from α, and monitor the resulting changes in the parameter estimates (see Riani et al., 2014; Cerioli et al., 2017). A similar approach is exploited in the Forward Search.

2.2 The Forward Search

The Forward Search (FS) is a flexible general method for detecting anomalies in structured data (Atkinson et al., 2004). Given a sample of n observations and a generating model for them, the method starts from a subset of cardinality $m_0 \ll n$, which is robustly chosen to contain observations coming from the postulated model. This subset is used for fitting the model and suitable deviance measures are computed. The subsequent fitting subset is then obtained by taking the m_0+1 observations with the smallest deviance measures. The algorithm iterates this fitting and updating scheme until all the observations are used in the fitting subset, thus yielding the classical statistical summary of the data. Therefore, the FS applies a decreasing sequence of trimming levels $\alpha_0 > \alpha_1 > \ldots > \alpha_L > 0$, with

$$\alpha_0 = 1 - \left\lfloor \frac{m_0}{n} \right\rfloor, \tag{13}$$

$$\alpha_l = \alpha_{l-1} - \frac{1}{n}, \qquad l = 1, \ldots, L, \tag{14}$$

and $L = n - m_0$. Clearly, $\alpha_L = 1/n$, while in the last step of the FS we have $\alpha_{L+1} = \alpha_L - 1/n = 0$ and no trimming is performed. The typical initialization in a multivariate framework is with $m_0 = v+1$ observations in (13), so that $L = n-v-1$. A slightly larger value of m_0 is sometimes selected to improve numerical stability of the initial estimates.

At step $l = 1, \ldots, L$ of the FS, the trimmed estimators (3) and (4) are computed with weights

$$\begin{aligned} w_{i,\alpha_l} &= 1 \qquad \text{if} \qquad i \in \iota^\dagger_{\alpha_l} \\ &= 0 \qquad \text{otherwise,} \end{aligned} \tag{15}$$

where $\iota^\dagger_{\alpha_l} = \{i_1, \ldots, i_{m_l}\}$ is the set of the indices of the $m_l = \lfloor (1-\alpha_l)n \rfloor$ observations that form the l-th fitting subset. Specifically, $\iota^\dagger_{\alpha_l}$ is obtained by taking the units with the m_l smallest squared distances

$$d^2_{i,\alpha_{l-1}} = (x_i - \tilde{\mu}_{\alpha_{l-1}})' \tilde{\Sigma}^{-1}_{\alpha_{l-1}} (x_i - \tilde{\mu}_{\alpha_{l-1}}), \tag{16}$$

computed from the estimates with trimming level α_{l-1}. The initial set $\iota^{\dagger}_{\alpha_0}$, of cardinality $m_0 = \lfloor (1-\alpha_0)n \rfloor$, is instead defined through an exogenous criterion, such as the intersection of robust bivariate projections, or the optimization of a robust objective function on subsets of m_0 observations. In §4 we adopt a random sampling strategy that proves to be suitable for clustering purposes.

The presence of observations deviating from the null model can be displayed through pictures that monitor relevant quantities along the search, such as the squared robust distances (16) and their order statistics. For instance, if only $n_0 < n$ units actually belong to the postulated population, we typically observe a peak in the monitoring plot of the minimum (squared) distance outside the fitting subset, when this subset only contains the n_0 "good" observations and the first outlier is about to enter. A formal procedure for precise identification of the contaminated observations is developed by Riani et al. (2009) when (2) is the null model. Under the same assumption, Cerioli et al. (2014) show that the FS estimators are both consistent and robust, while Johansen and Nielsen (2016a,b) provide a general asymptotic theory for the FS in regression.

3 Robustness properties under wild trimming and a new class of estimators

Unless otherwise stated, in what follows we make use of the replacement version of the breakdown point (BP) of an estimator, which is defined as the smallest fraction of outliers that can take the estimate over all bounds. A well known result of Lopuhaä and Rousseeuw (1991, Th. 2.1) is that the maximal BP of any translation equivariant location estimator cannot exceed $\lfloor (n+1)/2 \rfloor / n$. An estimator of location $t(\cdot)$ is translation equivariant if $t(x+c) = t(x) + c$. Instead, we say that an estimator is *quasi*-translation-equivariant if it is translation equivariant within a subspace (see Proposition 2 below).

The result of Lopuhaä and Rousseeuw (1991) corresponds to the intuitive statement that we should trim at most a portion $\alpha = \lfloor (n-1)/2 \rfloor / n$ of the observations, in order to get rid of the possible contaminants and to base our estimate on a subset of at least $\lfloor n/2 \rfloor + 1$ "good" data points. However, in this paper we use trimming levels much larger than 50% and it is natural to wonder what are the breakdown properties of our location estimators. There are only two possibilities: either our estimators are translation equivariant and their BP is $\lfloor (n+1)/2 \rfloor / n$ even if the trimming level is much larger than $\lfloor (n-1)/2 \rfloor / n$, or they can achieve a BP much larger than 50% but they are not translation equivariant. The latter claim holds. Formally, we define a class of trimmed *quasi*-translation-equivariant location estimators whose breakdown point can be arbitrarily higher than 50%, depending on the chosen level of trimming.

To do so, we fix a point $x_0 \in \mathbb{R}^v$. We define an *anchored* class of estimators, say $t_{x_0}(\cdot)$, which correspond to the original MCD (or FS, or other trimmed) estimator. Given a sample of v-variate observations $\mathcal{X} = \{x_1, \ldots, x_n\}$, the anchored estimator

of the location parameter μ for trimming level α is then

$$t_{x_0}(\mathcal{X}) = \frac{1}{\lfloor n(1-\alpha) \rfloor} \sum_{i=1}^{\lfloor n(1-\alpha) \rfloor} x_i^{\ddagger}, \tag{17}$$

where $x^{\ddagger}_{\lfloor n(1-\alpha) \rfloor} = \{x_1^{\ddagger}, \ldots, x^{\ddagger}_{\lfloor n(1-\alpha) \rfloor}\}$ is the minimizer of the objective function among all the subsets of $\lfloor n(1-\alpha) \rfloor$ points of $\mathcal{X}$ whose convex hull contains x_0. We remark that in this paper we indicate with the term "objective function" the objective function of any estimator of interest, be it MCD, FS or any another high-breakdown multivariate estimator; and with the term "solution" to the objective function the estimator itself. The anchored estimator arises as a solution to an anchored objective function, that is, a constrained objective function that we now formally define. Specifically, we look for solutions of the objective function subject to

$$\exists\, (\lambda_1, \ldots, \lambda_{\lfloor n(1-\alpha) \rfloor}) \;:\; \lambda_i \geq 0 \text{ and } \sum_{i=1}^{\lfloor n(1-\alpha) \rfloor} \lambda_i = 1 \text{ and } ||x_0 - \sum_{i=1}^{\lfloor n(1-\alpha) \rfloor} \lambda_i x_i^{\ddagger}|| = 0, \tag{18}$$

for a norm $||\cdot||$.

Some results for anchored trimmed estimators follow. A key issue in their derivation is that any point within the support of $G_0(x)$ also belongs to the convex hull of the points in the support. Additionally, convex hulls are non-decreasing: for every two sets A and B, where $A \subseteq B$, the convex hull of A is a subset of the convex hull of B.

Proposition 1 *If x_0 belongs to the convex hull of the points in $\mathcal{X}$, the anchored trimmed estimator exists. If x_0 belongs to the interior of the support of $G_0(x)$, the probability that the anchored trimmed estimator of location exists converges to the unity with the sample size.*

Proof. By the non-decreasing property of convex hulls, the convex hull of any subset of size $\lfloor n(1-\alpha) \rfloor$ is a subset of the convex hull of $\mathcal{X}$. Consider the set of all subsets of size $\lfloor n(1-\alpha) \rfloor$ and call h_j the convex hull of the j-th subset, and h the convex hull of $\mathcal{X}$. It is straightforward to check that $\cup h_j \subseteq h$. Therefore, if x_0 belongs to the convex hull of $\mathcal{X}$, there must exist at least one subset of size $\lfloor n(1-\alpha) \rfloor$ whose convex hull contains x_0. To see the second part, suppose that $\mathcal{X}_{\lfloor n(1-\alpha) \rfloor} = \{x_1, \ldots, x_{\lfloor n(1-\alpha) \rfloor}\}$ is sampled from $G_0(x)$. The probability that any x_0 in the interior of the support belongs to the convex hull of $\mathcal{X}$ obviously converges to the unity. □

For any x_0 in the interior of $G_0(x)$, existence is often very likely even for small n. As few as two or three well placed points are enough regardless of v. It can also be argued that the solution is unique if x_0 is in the interior of a unimodal and elliptically contoured G_0, in view of Proposition 1 and the uniqueness results in Davies (1987) and Butler et al. (1993). Additionally, x_0 does not need to be chosen in advance, as in the case of the FS (see §4, where we adopt a sampling strategy for a data-driven choice of x_0), or it can be easily tuned if no solution is found for an initial choice. While computation of a convex hull is rather computationally expensive (being in general $O(n^{v/2}+1)$), checking whether any x_0 belongs to the convex hull of a set of n points

has quadratic complexity. Hence, if x_0 is fixed, one could verify in advance whether the anchoring point belongs to the convex hull of $\mathcal{X}$ before computing robust estimates.

The following result assumes that the anchoring point is fixed.

Proposition 2 *The anchored trimmed estimator of location is quasi-translation-equivariant. Formally, for any collection of points $\mathcal{X} = \{x_1, \ldots, x_n\}$ and a fixed $x_0 \in \mathbb{R}^v$ within their convex hull there exists a set $A(x_0, \mathcal{X}) \subseteq \mathbb{R}^v$ such that if $t(x) + c \in A(x_0, \mathcal{X})$, then $t(x + c) = t(x) + c$.*

Proof. If x_0 belongs to the convex hull of $\mathcal{X}$, there exists at least one anchored estimator by Proposition 1. Let $x^{\ddagger}_{\lfloor n(1-\alpha) \rfloor}$ denote this solution. Fix any vector of constants, say $b \in \mathbb{R}^v$. By the properties of the objective function, if x_0 belongs to the convex hull of $x^{\ddagger}_{\lfloor n(1-\alpha) \rfloor} + b$, then

$$t_{x_0}(\mathcal{X} + b) = t_{x_0}(\mathcal{X}) + b. \tag{19}$$

If x_0 does not belong to the convex hull of $x^{\ddagger}_{\lfloor n(1-\alpha) \rfloor} + b$, then $x^{\ddagger}_{\lfloor n(1-\alpha) \rfloor} + b$ is not an admissible solution. Consequently, it will either happen that another subset of $\mathcal{X} + b$ is the solution to the anchored estimator problem, or that no solution exists. In both cases, the property of translation equivariance is lost. □

Denote with $C_{\mathcal{X}}$ the convex hull of $\mathcal{X}$. From the proof of the previous proposition it can be seen that

$$A(x_0, \mathcal{X}) = \left\{ b \ : x_0 \in C_{x^{\ddagger}_{\lfloor n(1-\alpha) \rfloor} + b} \right\}.$$

We now discuss a restricted version of the BP, suitable for anchored estimators.

Proposition 3 *For a fixed α, consider the substitution of $(\lfloor n\alpha \rfloor)$ points of $\mathcal{X}$ such that the chosen anchoring point x_0 belongs to the convex hull of the remaining $\lceil n(1-\alpha) \rceil$. Then, the anchored estimator of location* (17) *cannot break down. Consequently, the BP (restricted to substitution of certain subsets of $\mathcal{X}$) is equal to $(\lfloor n\alpha \rfloor + 1)/n$, where $\alpha \in (0, 1)$ is the trimming level.*

Proof. It is immediate to see that $(\lfloor n\alpha \rfloor + 1)/n$ is an upper bound. Suppose that we replaced $\lfloor n\alpha \rfloor + 1$ observations with arbitrary values. By definition, at least one of these values would not be discarded, hence leading to breakdown of the anchored estimator. To see that $(\lfloor n\alpha \rfloor + 1)/n$ is also a lower bound, fix a collection of (non-contaminated) points $\mathcal{X} = \{x_1, \ldots, x_n\}$, with $x_i \in \mathbb{R}^v$. Let $T = \sup_{x \in \mathcal{X}} ||t_{x_0}(x)||$. Replace at most $\lfloor n\alpha \rfloor$ points of $\mathcal{X}$ with arbitrary points to obtain $\mathcal{Y}$. Since there are $\lfloor n(1-\alpha) \rfloor$ points of the original sample $\mathcal{X}$, and by assumption these form a convex hull containing x_0, there exists at least one solution for $t(\mathcal{Y})$, where $t(\cdot)$ is the unconstrained version of $t_{x_0}(\cdot)$. This solution is such that $||t(\mathcal{Y})|| < T$ by the properties of $t_{x_0}(\cdot)$. Consequently, after anchoring there are one or more possible solutions to the anchored objective function, at least one of which is bounded. The bound depends only on the original sample points. It only remains to show that the anchored estimator $t_{x_0}(\mathcal{X})$ chooses one of the bounded solutions, which follows since the unbounded solutions either correspond to unbounded objective functions, or any $x_0 \in \mathbb{R}^v$ does not belong to their convex hull. □

An important remark about anchored estimators is that they are not translation equivariant. A sort of hard shrinkage towards x_0 happens: as soon as a translation moves $\mathcal{X}$ far enough from x_0, the translation equivariant solution is mapped in a bounded set close to x_0. This is illustrated in Figure 1, where we set $\alpha = 80\%$. In the top left panel, the MCD and anchored MCD coincide (blue dot). In the top right panel, data are translated by a small amount and still the anchor x_0 (red cross) is within the convex hull of the optimal MCD subset. If we translate by a larger amount, the anchor is now poorly chosen and the MCD (green X) does not coincide with the anchored MCD anymore (bottom left panel). A data-dependent choice of x_0, such as the one adopted in §4, will help to prevent this situation. The fact that anchoring is useful is illustrated in the bottom right panel, where a cluster of outliers is added. The MCD is now within this cluster, as its points have a very small scatter, while the anchored MCD is not affected.

Another important remark concerns the seemingly restrictive assumption that substitution of points in $\mathcal{X}$ is restricted to a subset which guarantees existence of a convex hull containing x_0. It shall be noted though that the substituted points are (as usual) replaced by arbitrary points. Additionally, as said before, x_0 need not be chosen in advance and can hence be tuned to guarantee existence. Once the anchored estimators exist, they cannot break down. For this reason, it can be readily shown that the *addition* breakdown point – which is based on adding contaminating points to the data set, instead of replacing them: see Hennig (2004) – is not restricted.

Theorem 1 *Fix $g > 0$ as the number of points to be trimmed. Suppose that the chosen anchoring point x_0 belongs to the convex hull of $\mathcal{X}$. Then, the anchored estimator of location* (17) *has an addition breakdown point of $g/(n+g)$.*

Proof. Suppose that $g+1$ contaminated points are added. Since only g can be trimmed, at least one contaminated point will be included in the estimators and possibly lead to break down. Therefore an upper bound for the addition breakdown point is $g/(g+n)$. Note now that, since the n points originally included in the data $\mathcal{X}$ are not modified or removed, by assumption on x_0 the anchored estimator exists with at least one solution such that

$$|t_{x_0}(\mathcal{X})| \leq T$$

for a certain finite value T which depends only on the original data. It only remains to show that the anchored estimator $t_{x_0}(\mathcal{X})$ chooses one of the bounded solutions, which is straightforward as the unbounded solutions either have unbounded objective functions, or any $x_0 \in \mathbb{R}^v$ does not belong to their convex hull. □

A fairly similar discussion can be put forward for affine equivariant estimators. Recall for instance that an estimator of location $t(\cdot)$ is affine equivariant if $t(Ax+c) = At(x) + c$. Davies (1987) shows that the BP of any affine equivariant covariance estimator is at most given by (8), which might be much smaller than 50%. The trimming level obviously has an upper bound of $\lfloor (n-v-1) \rfloor / n$ to guarantee that the estimated covariance matrix is positive definite. If we use a trimming level $\alpha \in (0, 1-(v+1)/n)$, a reasoning along the lines of the proof of Proposition 3 can be used to show that the BP of the anchored estimator of scatter is equal to $(\lfloor n\alpha \rfloor + 1)/n$.

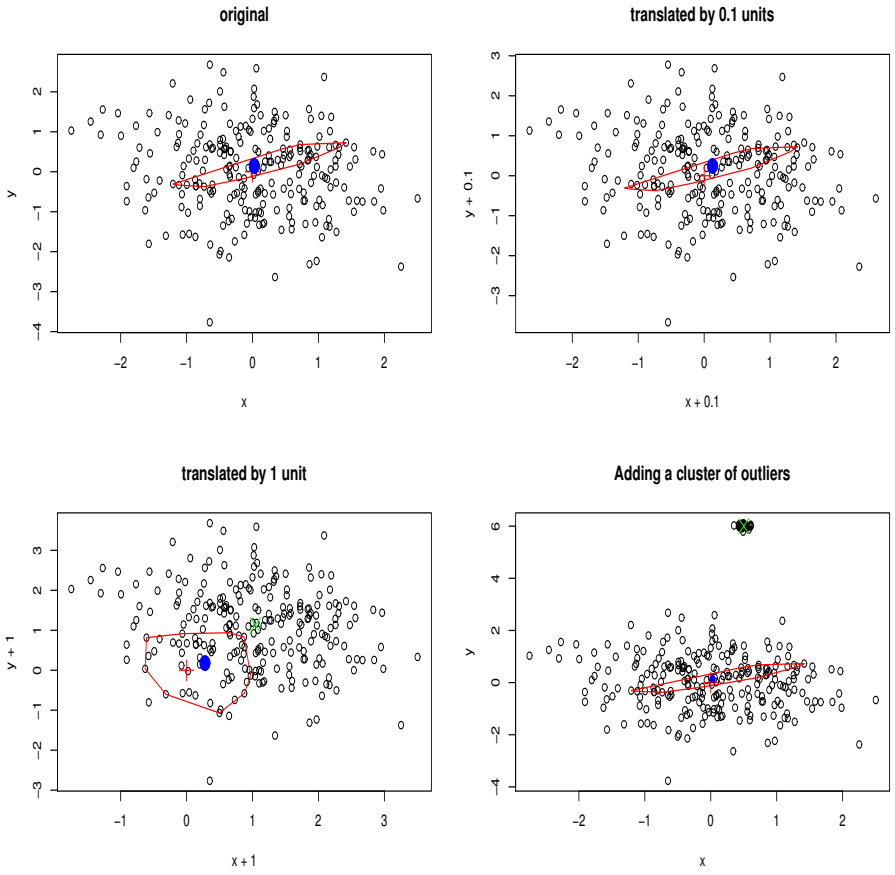

Figure 1: Anchoring estimators in action. The red cross indicates the anchoring point. The blue dot is the anchored estimator, surrounded by the convex hull (containing the anchoring point x_0). When this differs from the MCD, the MCD is shown as a green X.

The anchored estimator of scatter is found along the same lines as the anchored estimator of location: given a sample of v-variate observations $\mathcal{X} = \{x_1, \ldots, x_n\}$ and trimming level α, the anchored estimators of the location parameter μ and scatter Σ are

$$t_{x_0}(\mathcal{X}) = \frac{1}{\lfloor n(1-\alpha)\rfloor} \sum_{i=1}^{\lfloor n(1-\alpha)\rfloor} x_i^{\ddagger}, \;\; T_{x_0}(\mathcal{X}) = \frac{\zeta_\alpha}{\lfloor n(1-\alpha)\rfloor} \sum_{i=1}^{\lfloor n(1-\alpha)\rfloor} (x_i^{\ddagger} - t_{x_0}(\mathcal{X}))(x_i^{\ddagger} - t_{x_0}(\mathcal{X}))', \tag{20}$$

where $x^{\ddagger}_{\lfloor n(1-\alpha)\rfloor} = \{x_1^{\ddagger}, \ldots, x^{\ddagger}_{\lfloor n(1-\alpha)\rfloor}\}$ is the subset of $\lfloor n(1-\alpha)\rfloor$ points of $\mathcal{X}$ minimizing the objective function among all subsets whose convex hull contains x_0.

Theorem 2 *Assume contaminating points are in general position, that is, they do not lie in a lower dimensional space. For a fixed α, consider the substitution of $(\lfloor n\alpha\rfloor)$ points of $\mathcal{X}$ such that the chosen anchoring point x_0 belongs to the convex hull of the remaining $\lceil n(1-\alpha)\rceil$. For $\alpha \in (0, 1-(v+1)/n)$, the anchored estimator of scatter has (restricted) BP $(\lfloor n\alpha\rfloor + 1)/n$.*

Proof. The fact that $(\lfloor n\alpha\rfloor + 1)/n$ is an upper bound is shown as in the proof of Proposition 3. In fact, if $\alpha = 1-(v+1)/n$, no positive definite solution exists. For smaller values of α, if we replace $(\lfloor n\alpha\rfloor + 1)$ observations with arbitrary values, at least one of these values would not be discarded. Consequently, all solutions would become unbounded. To see that $(\lfloor n\alpha\rfloor + 1)/n$ is also a lower bound, we need to show two things. First, we need to show that the minimum covariance determinant cannot be arbitrarily increased. To see this, replace at most $\lfloor n\alpha\rfloor$ points of $\mathcal{X}$ with arbitrary points to obtain $\mathcal{Y}$. Since there are $\lfloor n(1-\alpha)\rfloor$ points of the original sample $\mathcal{X}$, and by assumption there exists at least one anchored solution $\hat{\Sigma}$, there exist at least one solution with bounded determinant. The anchored estimator chooses one of the bounded solutions as the unbounded solutions either have unbounded objective functions, or any $x_0 \in \mathbb{R}^v$ does not belong to their convex hull. Secondly, we need to show that the minimum covariance determinant cannot be arbitrarily decreased. This is straightforward from the assumption that $\mathcal{X}$ and contaminating points are in general position (see detailed reasoning on this point in Davies (1987); Lopuhaä and Rousseeuw (1991)). □

As in Theorem 1, it is possible to show that the addition BP for $\hat{\Sigma}$ is unrestricted.

4 Wild adaptive trimming for robust cluster analysis

Our main strategy for performing robust cluster analysis through wild (adaptive) trimming follows the general principle that several analyses from more than one starting point are necessary to reveal the clustering structure. When the data come from model (10), the trimmed estimators computed with $\alpha < 0.5$ typically use observations from several clusters and may thus fail to identify the different populations. On the other hand, starting with a subset of observations belonging to the same population would lead to high values of the robust distances (7) for the observations from other clusters, which are then detected as outliers. Empirical evidence of this behaviour has been shown in an exploratory context (see, e.g., Atkinson et al., 2004, 2017), while the robustness properties of §3 provide a more general theoretical justification.

With a slight abuse of notation, in what follows we let $G_0(x)$ denote the distribution function of one of the normal mixture components in (10), to be taken as the target population with parameters μ and Σ, instead of the distribution of the whole mixture. That is, we now let

$$G_0(x) = \Phi_{\mu,\Sigma}(x)$$

as in (2), with

$$\mu = \mu_{k^*}, \qquad \Sigma = \Sigma_{k^*}$$

for a given $k^* \in \{1, \ldots, K\}$. Correspondingly, we take $G_1(x)$ to be the (unspecified) distribution function mixturing all the other components and the contaminated observations. It is then crucial to use trimming levels in the computation of (3) and (4) that could lead to estimation of μ and Σ, and thus to identification of $G_0(x)$, under model (1). This task is clearly made possible by the adoption of a wild trimming approach, and by the fact that, as noted, $G_0(x)$ now identifies a single population.

We suggest a divisive clustering approach that splits the K-population estimation problem defined by (10) into K one-population steps, which take advantage of the good breakdown properties of wild trimming estimators and do not force all units to be classified. In an adaptive framework, we start the trimming procedure from values of α much larger than 0.5, typically including only $v + 1$ observations or a slightly larger number, in the first estimation step. The details of our robust divisive clustering procedure are described below. For concreteness we refer to wild adaptive trimming through the FS, but any other procedure that shares the same properties (e.g., based on the MCD) could be potentially used.

1. **Computation of minimum Mahalanobis distances from random starts.** We perform R forward searches starting from R random subsets of size $m_0 = v+1$. For all the searches and each $m_l = \lfloor (1 - \alpha_l)n \rfloor$, $l = 0, \ldots, L$, we control the ratio between the maximum and the minimum eigenvalue of the estimated covariance matrix. We impose that this ratio is smaller than a certain threshold, say c, in order to avoid the detection of spurious groups due to the presence of almost collinear points. We discard the searches (whose number, say, is r) for which this condition is not fulfilled. For each of the $R-r$ remaining searches and each subset size m_l, we store the value of the minimum Mahalanobis distance of the units not belonging to the fitting subset:

$$d_{min}(m_l) = \min \sqrt{d^2_{i,\alpha_l}}, \qquad i \notin \iota^{\dagger}_{\alpha_l}, \tag{21}$$

where, for $l = 0, \ldots, L$, d^2_{i,α_l} is defined as in (16). In all the examples that follow we set $R = 500$, although a larger number of random searches should be performed in the case of big data sets. Indeed, it usually suffices that the majority of points of the starting subset belong to $G_0(x)$ in order to obtain high values of the robust distances (7), and thus of (21), for the observations not belonging to this population. The probability of randomly selecting $\eta_0 = \lfloor m_0/2 \rfloor + 1$ observations from a sample of N units drawn from $G_0(x)$ is $\binom{N}{\eta_0}/\binom{n}{\eta_0}$, which is often not exceedingly small when $m_0 = v + 1$ and n is of the order of just a few hundreds. Furthermore, in the FS we observe that the method is often able

to recover from a bad starting point, by replacing contaminated observations in the fitting subset with others coming from $G_0(x)$ (see, e.g., Atkinson et al., 2004). This phenomenon, which is called "interchange", clearly enhances the diagnostic power of (21) and increments the probability of detecting $G_0(x)$ with a fixed number of random starts, as our empirical examples show in §5.

2. **Envelope calibration.** If all the units in $\iota^{\dagger}_{\alpha_l}$ come from $G_0(x)$ and the units not in $\iota^{\dagger}_{\alpha_l}$ come from different populations (separated from $G_0(x)$), the observation giving rise to $d_{min}(m_l)$ will be an outlier and its distance will be large if compared to the distances of the units in $\iota^{\dagger}_{\alpha_l}$. We thus need to compare the value of $d_{min}(m_l)$ with the quantiles of its distribution for each step $l = 0, \ldots, L$. The simulation results given in Atkinson et al. (2006) show that the envelopes of $d_{min}(m_l)$ starting from random subsamples are much larger than those based on robust initialization when $m_l < n/3$, especially in the case of extreme quantiles. However, this difference decreases as l increases and becomes negligible for subsets of size $m_l > n/2$. A reliable approximation to envelopes based on the theory of order statistics is given in Riani et al. (2009), but only for the case of robust initialization.

   Since we are interested in values of m_l which may be much smaller than $n/2$, we have simulated the 1%, 50%, 75%, 90%, 95%, 99%, 99.9%, 99.99%, 99.999% and 99.9999% percentiles of $d_{min}(m_l)$, for $l = 0, \ldots, L$, under the normal model $G_0(x) = \Phi_{\mu,\Sigma}(x)$ and using 1,000,000 random initializations. In what follows we call these envelopes the *null envelopes*, since they correspond to the situation where only one population exists, i.e. (10) holds with $K = 1$. They have been obtained for each value of $v \leq 10$ and a grid of values of n from 50 to 2000. If n is not in the grid we use linear interpolation between the two adjacent values.

3. **Convex hull constraint and pruning.** The observations entering in the first steps of the FS are those more likely to come from $G_0(x)$. It is thus natural to adopt a data-dependent anchoring procedure and to fix the anchor x_0 as a point which lies inside the convex hull of the observations belonging to the fitting subset $\iota^{\dagger}_{\alpha_l}$. In this way the anchor will not be too far from the true mode of $G_0(x)$, at least when $G_0(x)$ belongs to the elliptical family, until $\iota^{\dagger}_{\alpha_l}$ remains free of contaminated observations. Therefore, we find α_l as the trimming level prior to the first exceedance of the null envelope of $d_{min}(m_l)$ for a certain probability level, say ς. As we see in §5, in the first steps just after the random start the values of $d_{min}(m_l)$ may be above the 99% threshold due to random fluctuations and the number of different trajectories diminishes considerably as m_l increases. In fact, all random starts of size $m_0 = v+1$ initialized with more than $\lfloor v/2 \rfloor + 1$ observations from $G_0(x)$, as soon as the subset size grows, tend to remove the observations from the other groups and include other observations from $G_0(x)$. Therefore, after a few steps, the $R - r$ searches naturally anchor themselves to a few centroids and a few covariance matrices. In what follows we call $m^{\dagger}$ the subset size for which we start to impose the "anchoring". More precisely,

$$m^{\dagger} = \max\{(\min m_l : d_{min}(m_l) > d_{\varsigma}(m_l)), \texttt{mingrsize} + 1\} - 1,$$

where $d_\varsigma(m_l)$ is the ς-quantile of $d_{min}(m_l)$ and `mingrsize` is the minimum group size that we are willing to tolerate. For all $R-r$ trajectories we monitor whether the units which are progressively included for $m_l > m^\dagger$ satisfy the convex hull constraint (18). Therefore in practice there is no need to choose a particular anchoring point, such as the centroid or the median of the units belonging to subset at step $m^\dagger$, because the satisfaction of the convex hull constraint ensures we anchor so to obtain an estimator which cannot break down (see Proposition 3). As a result of this procedure we prune some of the trajectories of $d_{min}(m_l)$, i.e. their monitoring ends much earlier than at the final step $l = L$ because the convex hull constraint fails to be satisfied. Let $m^\dagger + a_t$ be the final subset size for trajectory t $(t = 1, \ldots, R-r)$. In the examples which follow, in order to obtain a satisfactory degree of pruning, we take $\varsigma = 0.9999$. It shall be noted that identical results can be obtained if we increase the confidence band to 99.9999%, or decrease it to 99%. In the latter case, however, the number of random starts which must be used has to be considerably increased. Otherwise, we may lose trajectories which completely end up into one group but which, due to random fluctuations, are pruned before the peak due to loss of anchorage because the convex hull is too "small".

4. **Extreme exceedance of null envelopes.** The envelopes of $d_{min}(m_l)$ are pointwise because their confidence level is referred to a fixed subset size m_l. In the adaptive trimming framework of the FS we potentially make many comparisons, one for each value of m_l. Our detection rule thus needs to allow for simultaneity. In order to avoid random exceedances, we select all the pruned $R-r$ trajectories of $d_{min}(m_l)$ for which there is an exceedance of a very extreme threshold of the null envelope, say $d_{\varsigma^*}(m_l)$. Let $d^{(t)}_{min}(m_l)$ be the trajectory for the t-th pruned search. Then, m^*_t is the first subset size for which

$$d^{(t)}_{min}(m_l) > d_{\varsigma^*}(m_l) \qquad m_l = m^\dagger, m^\dagger+1, ..., m^\dagger+a_t, \qquad t = 1, 2, \ldots, N-r. \tag{22}$$

   In what follows we take $\varsigma^* = 0.999999$.

5. **Divisive split.** Given that our purpose is to identify and remove first the group which is most remote from the others, among the searches for which condition (22) is verified we take the one (say the t^*) for which

$$rs_{j^*} = \arg\max_t rs_t, \tag{23}$$

   where

$$rs_t = \max_{m_l} \frac{d^{(t)}_{min}(m_l) - d_{0.5}(m_l)}{d_{0.99}(m_l) - d_{0.5}(m_l)}. \tag{24}$$

   We then assign the $m_l - 1$ observations which form $\iota^\dagger_{\alpha_{l-1}}$ to a tentative group.

6. **Iteration of previous steps.** The previous steps (1-5) are iterated until with the units which are left out we end up with one of the two following cases: A) their number is smaller than `mingrsize`. B) we do not observe any exceedance of the extreme null envelope.

7. **Robust tree.** At the end of the procedure we display the binary splits and the resulting clusters by a tree-like structure. In the vertical axis of the tree we show the distance level rs_{j^*} (see Equation 23) in which the various groups are formed (it is also possible to use a rescaled version of rs_{j^*} in case one wants to standardize the results over different datasets). One additional bonus of the suggested procedure is that it enables us to immediately appreciate the degree of separation (overlapping) of each group with the remaining part of the sample. Clearly the higher is the value of internal cohesion of a group with respect to the rest, the greater is the value of rs_{j^*}. The outliers and the other units not assigned to any of the tentative groups are left aside, thus making the tree robust.

We note that, by considering a crude rule like (22), we are likely not to consider all units belonging to a particular group. Therefore, a successive reweighting step (which in the spirit of the FS can be performed adaptively; see also Dotto et al. (2017)) is necessary for refining the tentative groups which have been found. A preliminary proposal, rooted in an exploratory framework, is described in (Atkinson et al., 2004, p. 369). However, in our divisive procedure, the fact of leaving out the units which are at the boundary of a particular group helps to detect the remaining groups in the successive steps of the procedure. The null envelopes obtained after removal of the first cluster, being based on a number of observations larger than the number of units belonging to the remaining populations, will be flatter and tighter, thus increasing the probability of exceedance of the extreme envelope in the central part of the algorithm.

5 Robust divisive clustering in action

5.1 Geyser data

In order to illustrate the performance of our divisive clustering procedure, we start by considering the "geyser data set", a well known application in robust clustering (see, e.g., García-Escudero and Gordaliza, 1999). This bivariate data set, obtained from the Old Faithful Geyser, contains the eruption length and the length of the previous eruption for 271 eruptions of this geyser. Both variables are measured in minutes. The data show the presence of three main groups. Close to the origin there is also a small group of "short followed by short" eruptions, which are not very common (6 observations, i.e. 2.2% of the sample size).

The left panel of Figure 2 shows the forward plot of the trajectories of minimum Mahalanobis distances computed from 500 random starts, together with 1%, 50%, 99% and 99.9999% simulation envelopes from model (10) with $K = 1$. To provide a comparison with previous applications of the FS, in this plot we do not anchor the estimator and all trajectories are monitored up to $m_l = n$. In presence of a homogenous population of size n, all the R trajectories of $d_{min}(m_l)$ rapidly converge to a single trajectory which tends to remain inside the envelopes up to $m_l = n$. In this situation the shape of the envelopes of $d_{min}(m_l)$ looks like the prows of viking longships, i.e. they are virtually horizontal in the centre of the plot and rapidly increase as m_l tends to n due to the inclusion in the final steps of the observations coming from the tails of the distribution. On the contrary, if model (10) holds with $K > 1$ and if there is a group with

size (say) $n \times 0.2$, for all the searches which start in this group we observe a rapid increase in the trajectory of $d_{min}(m_l)$ as m_l tends to $n \times 0.2$. The same also typically happens for the searches such that the $m_0 = v + 1$ initial observations contain at least $\lfloor v/2 \rfloor + 1$ units from the group, due to the interchange of units from other groups as the subset size grows. After this subset size, as observations from other groups join $\iota^{\dagger}_{\alpha_l}$, we are likely to observe a sudden decrease in the trajectory of $d_{min}(m_l)$ and its values will be even below the lower quantiles of the null envelopes for a single homogenous population of size n. It is an evidence of the extreme effect that contamination by different populations can produce on parameter estimates. The same effect will lead to a "loss of anchorage" when we impose the convex hull constraint (18). The plot in the left panel of Figure 2 reveals some trajectories which go persistently above the extreme 99.9999% threshold around subset size $m_l = 90$. Similarly, around $m_l = 170$ we can observe two trajectories which go outside the extreme envelope and, to a minor extent, another trajectory which spends some time above this threshold just before $m_l = 200$. All trajectories converge into one from $m_l = 230$ onwards. It is interesting to notice that before all the trajectories converge into one there is a persistent exceedance of the lower threshold. The same phenomenon is also visible around $m_l = 100$, shortly after the first peak. In this example, given that the first exceedance of the extreme envelope is when $m^{\dagger} = 76$, for each of $R - r$ trajectories we impose the convex hull constraint from this step and find $m^{\dagger} + a_t$ ($t = 1, 2, \ldots, R - r$). The right panel of Figure 2 shows the pruned trajectories. For example, for the 3 trajectories exceeding the extreme envelope the constraint is fulfilled up to $m^{\dagger} + a_1 = 107$, $m^{\dagger} + a_2 = 107$ and $m^{\dagger} + a_3 = 97$. In order to understand which is the group which is most remote from the others we compute index rs_t given in equation (24), select the trajectory associated with the maximum value and select the step where first exceedance of extreme envelope takes place. In this case this leads to the identification of a first tentative group made up of $m^* = 83$ units with a corresponding value of $rs_{j^*} = 9.71$

It is interesting to notice that the number of searches which end up with the three trajectories is 152, 144 and 177, respectively. Therefore, starting from 500 random initializations in almost 95% of the times we have reached trajectories which collapse into just one group. This also implies that in this example it was not necessary to consider as many as 500 random starts to elucidate the existence of three distinct groups.

The random start FS procedure is repeated using the remaining 188 units, after removal of tentative Group 1. The left panel of Figure 3 shows the new pruned trajectories of $d_{min}(m_l)$ with the corresponding null envelopes based on $n = 188$. The same argument described above leads to the identification of a second tentative group, again of size 83 with a corresponding value of $rs_{j^*} = 9.23$. The results of the third iteration of our procedure are displayed in the right panel of Figure 3, obtained after removal also of tentative Group 2. The pruned trajectories of $d_{min}(m_l)$, with null envelopes based on $n = 105$ units, now lead to the identification of a third tentative group of size 84 with $rs_{j^*} = 4.56$. These three steps leave us with 21 unassigned units without apparent structure. Therefore, the procedure terminates and we set $K = 3$.

The binary splits and the resulting clusters are displayed in a tree-like structure in Figure 4, while our robust tentative clustering of these data is given in Figure 5. The latter plot, given that it has on the vertical axis the value of rs_{j^*}, not only shows the order in which the groups are found but also reflects their degree of compactness.

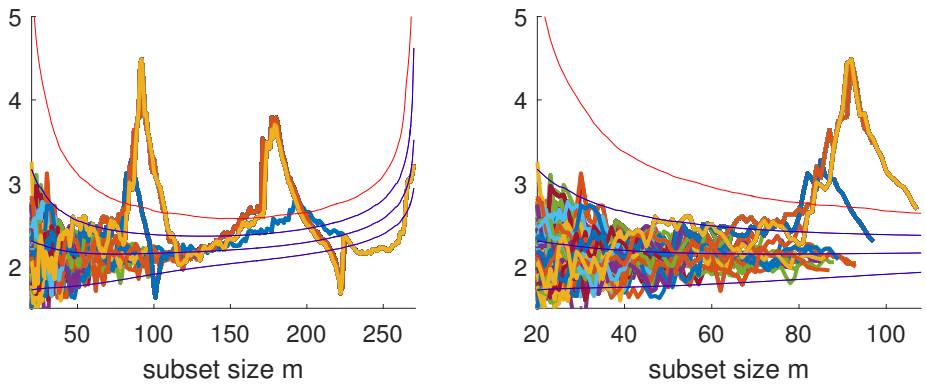

Figure 2: Geyser data. Left panel: forward plot trajectories of minimum Mahalanobis distances from 500 random starts monitored up to $m_l = n$ with 1%, 50%, 99% and extreme 99.9999% null envelopes. Right panel: trajectories of minimum Mahalanobis distances pruned after anchoring the estimators and imposing the convex hull constraint. In the right panel the horizontal axis goes up to $110 \approx \max(m^\dagger + a_t) = 107$).

Indeed, the group on the top left of Figure 5 (which is found in the first divisive step) appears to be the most compact one, while the cluster on the top right (which is found in the third divisive step) is the most dispersed one. The units not assigned to any of the three main clusters correspond to some borderline observations and to the peculiar set of "short followed by short" eruptions. We do not insist in labelling these eruptions as outliers, or as representatives of a fourth, more uncommon, population. We believe that the final interpretation will strongly depend on subject-matter knowledge and on the purposes of the study. The important statistical finding from our robust cluster analysis is that they do not belong to the three main populations that our method identifies.

5.2 M5 data

The geyser data set originates from well separated populations, with the possible addition of markedly different outliers. Another application with similar features (to Swiss banknotes) is described in the Supplementary Material. We now show the performance of our divisive procedure in a case with highly overlapping populations. The M5 data were introduced by García-Escudero et al. (2008) for assessing some trimming-based robust clustering methods. The data (shown in Figure 6) are obtained from three normal bivariate distributions with fixed centers but different scales and proportions. One of the components strongly overlaps with another one. To these data a uniform noise contamination is sometimes added. However, in this application we concentrate on the "uncontaminated version" of the data set with $n = 1800$, since our main interest is not on the effect of widespread noise.

The top left panel of Figure 7 shows the monitoring of the trajectories of minimum Mahalanobis distances from 500 random starts without pruning. This plot dis-

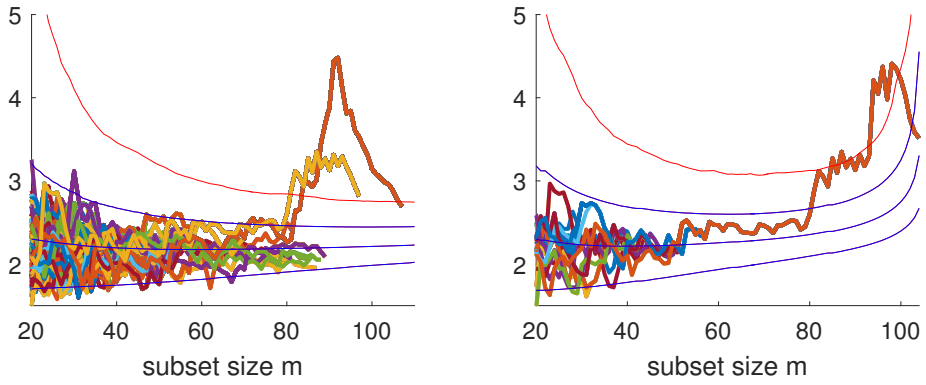

Figure 3: Geyser data. Left panel: forward plot of pruned trajectories of minimum Mahalanobis distances after removing the units from the first tentative group. Right panel: forward plot of pruned trajectories of minimum Mahalanobis distances after removing the units from tentative Groups 1 and 2.

plays three distinct trajectories around $m_l = 400$. It is interesting to see that around $m_l = 500$ there is a dip below the lower envelope for one trajectory before being absorbed by another one which stands above the upper envelope. This data set has also been analyzed by Atkinson et al. (2015), who manually choose a particular step in the random start procedure and compare the units inside the fitting subset for the various trajectories. The top right panel of Figure 7, which shows the pruned trajectories of the same minimum Mahalanobis distances, avoids this manual choice because it enables us to appreciate that there is only one trajectory which exceeds the extreme threshold and which terminates when $m_l = 427$. All the other trajectories are pruned much earlier and are completely inside the null envelopes. This is an indication that the remaining trajectories refer to searches which include observations from other groups which are outside the convex hull of the reference group and therefore we lose the good propoerties of the anchored estimator. Considering the first exceedance of the extreme envelope leads to the identification of a first tentative group of 308 observations. The two bottom panels show the pruned trajectories of minimum Mahalanobis distances, again from 500 random initializations, in the two successive steps of the divisive procedure. Also in this case pruning enables to identify just one trajectory outside the extreme envelope and leads in a natural way to the identification of the underlying group. The divisive procedure is detailed in Figure 8, while the left panel of Figure 9 shows the scatter plot of the original data with the $K = 3$ tentative groups which have been found and the unclassified units. For clarity of interpretation, in Figure 9 the 99% confidence ellipses have been added using the centroids and the covariance matrices of the estimated groups. The overall performance of our method is very good, with only 3.6% of the assigned observations clustered in the wrong group.

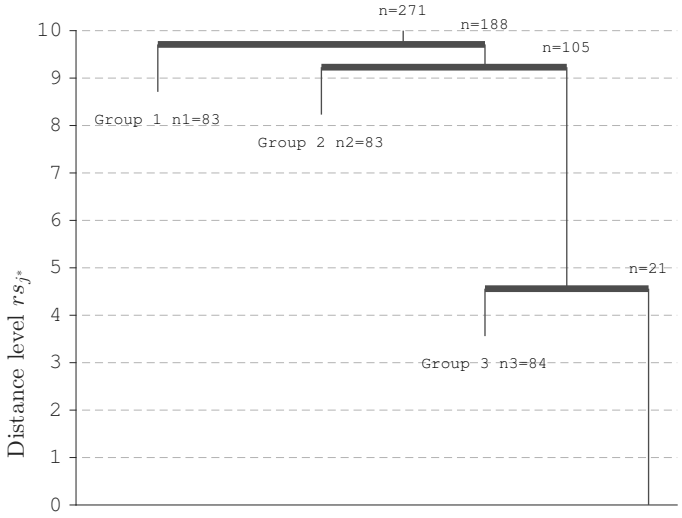

Figure 4: Geyser data. Robust tree showing the details of the divisive procedure. The 21 unassigned units clearly stand apart in the last split.

5.3 Comparison with TCLUST

We conclude our empirical analysis by comparing the results of our divisive procedure with those obtained through the TCLUST methodology of García-Escudero et al. (2008), Fritz et al. (2012) and Ruwet et al. (2013). TCLUST is a robust model-based clustering technique that relies on constrained maximization of a trimmed version of a generalized classification likelihood function. The constraint is that

$$\frac{\max_{k=1}^{K} \tilde{\lambda}_{1,k}}{\min_{k=1}^{K} \tilde{\lambda}_{v,k}} \leq c, \tag{25}$$

where $\tilde{\lambda}_{1,k}$ and $\tilde{\lambda}_{v,k}$ denote the largest and the smallest eigenvalue of $\tilde{\Sigma}_k$, respectively, and $c \geq 1$ is a fixed constant specified by the user. Values of c close to 1 favour solutions with spherical clusters, while high values of c tend to produce one (or more) large and dispersed cluster possibly overlapping with the other groups.

It is important to emphasize that, unlike our method, TCLUST requires to specify in advance a number of important parameters: the number of groups K, the level of trimming α to be used in the generalized classification likelihood function and the eigenvalue ratio restriction (25). Our adaptive-trimming divisive procedure finds instead suitable values of K and α from the data, while (25) is not needed because we just fit one population at a time. Therefore, we can see this comparison as a worst-case scenario for our technique, since we are not taking advantage of the prior information which is used to initialize TCLUST.

In order to apply TCLUST to the data sets considered in this paper, we use the number of clusters which we have found in an automatic way and we set α equal to

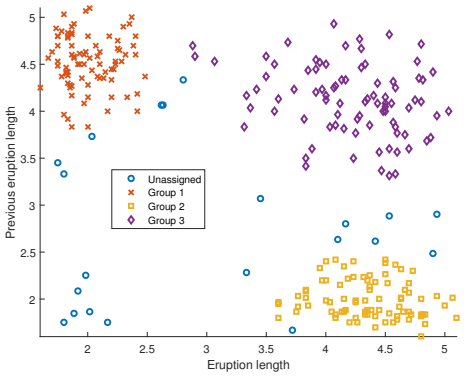

Figure 5: Geyser data. Robust tentative clustering into $K = 3$ main groups after the divisive procedure.

the proportion of units which are left unclassified by our procedure. We also fix the restriction factor $c = 100$ in order to be able to find elongated clusters, such as the one present in the M5 data set. For the geyser data and the Swiss banknotes (analyzed in the Supplementary Material), the results from TCLUST are similar to those obtained with our method. Indeed, the adjusted Rand index between the two clusterings is 0.95 for the geyser data and 0.88 for the Swiss banknote data. This means that the performance of our divisive procedure is virtually equivalent to that of TCLUST, provided that the latter is properly tuned through appropriate a priori information. The outcome is somewhat different in the case of the M5 data set, for which the right panel of Figure 9 shows the three-group robust clustering obtained by TCLUST. Although the overall performance of this clustering is very good (the misclassification rate is 2.4%), it is clear that TCLUST underestimates the size of Population 3. The key issue for explaining such a poorer performance in the identification of a dispersed group is the rigid use of the trimming level α made by TCLUST. Since in this example there are no outliers, fixing $\alpha \gg 0$ leads TCLUST to trim all the observations that lie at the border of the more dispersed population, which is not instead the case for our procedure. Through our adaptive trimming approach, we are able to detect the population borders in a flexible way and without penalizing uncommon structures too much. On the contrary, TCLUST treats the farthest observations from Population 3, and in particular those lying above Population 1, as uniform noise to be trimmed. We may thus expect an improvement in the performance of TCLUST if the method could be embedded in an adaptive trimming framework similar to that considered in our work.

6 Concluding remarks

This work is motivated by the requirement of robust and efficient procedures for clustering multivariate data generated by mixture model (10) with additional contamination. Our approach replaces the original K-population (robust) estimation problem with K distinct (robust) one-population steps, which take advantage of the good break-

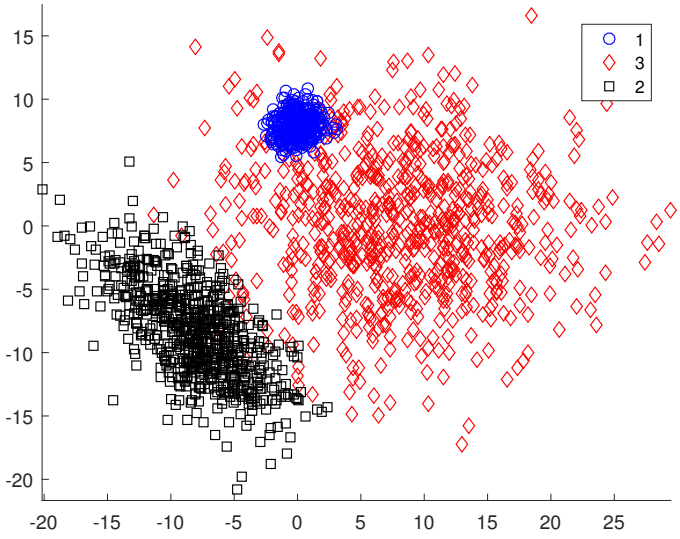

Figure 6: M5 data. Scatter plot of the two variables. Cluster 1 is very tight and lies completely inside cluster 2. Clusters 2 and 3 also overlap.

down properties of trimmed estimators. For this purpose, we have studied the theoretical behaviour of trimmed estimators when the trimming level exceeds the usual bound of 0.5, thus relaxing the familiar condition that at least half of the data should correspond to the main population. We have shown that exact affine equivariance must be lost, but it is a reasonable price to be paid in order to achieve arbitrarily high breakdown for the resulting trimmed estimators. This conclusion parallels similar findings in other situations where contamination produces only a minority of "good" observations, as in the case of cell-wise contamination (see, e.g., Farcomeni, 2014a,b; Agostinelli et al., 2015; Rousseeuw and Van den Bossche, 2017). We also support the use of adaptive trimming schemes, in order to explore the effect of different levels of trimming and to find a sensible trade-off between robustness and efficiency. A further bonus of our methodology is its ability to provide a reliable choice of the usually unknown number of groups that correspond to genuine populations in model (10).

We have provided empirical evidence that our technique can perform well even when there is considerable overlap among the groups. The price that we pay for separating the groups when they partially overlap is to trim a bit more than necessary in steps (d) and (e) of our divisive procedure. Even if our trimming approach is adaptive and provides a good trade-off between robustness and efficiency, there is the need of additional theoretical work in order to find a stopping rule that guarantees the required simultaneous test size when testing for exceedances of null envelopes. Similarly, precise estimation of the contamination rate in (1) and of the mixing proportions in (10) are still open issues. A refined estimate of the population sizes, as well as a refined identification of group membership, could be obtained by adding a confirmatory step to the tentative clustering that we obtain by our divisive procedure. The confirmatory

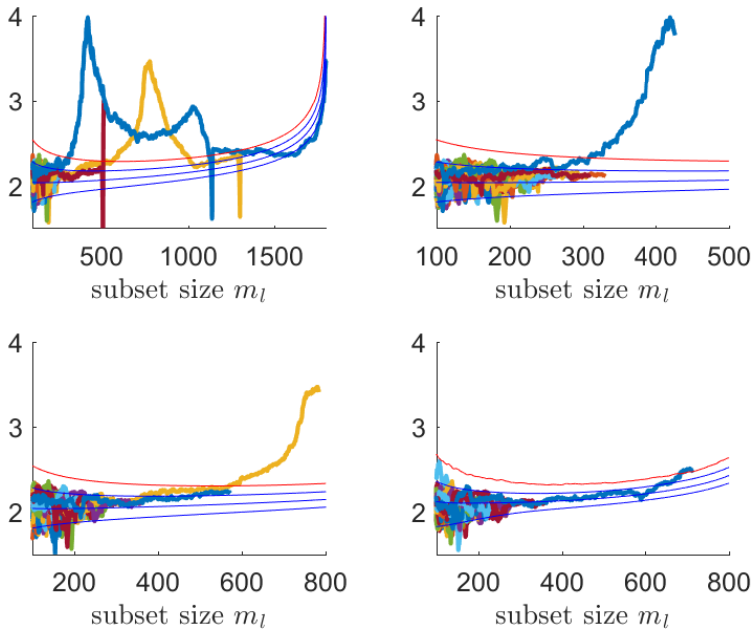

Figure 7: M5 data. Top left panel: trajectories of minimum Mahalanobis distances from 500 random starts without pruning, together with 1%, 50%, 99% and 99.9999% envelopes. Top right panel: forward plot of the same pruned trajectories. Bottom left panel: forward plot of pruned trajectories of minimum Mahalanobis distances after removing the units from the first tentative group. Bottom right panel: forward plot of pruned trajectories of minimum Mahalanobis distances after removing the units from the first two tentative groups.

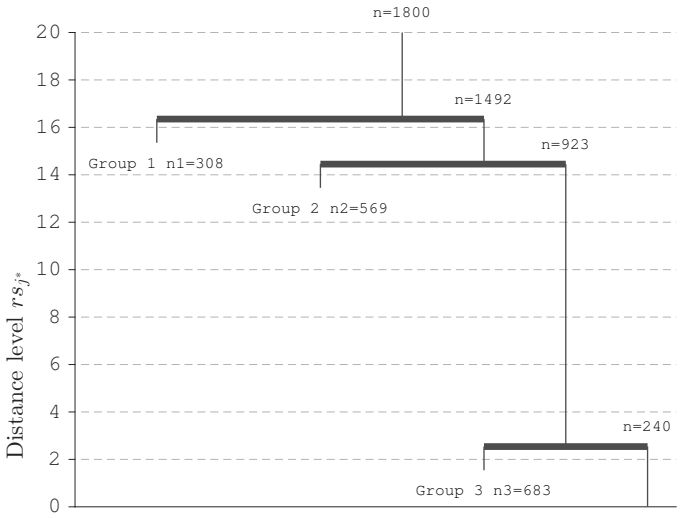

Figure 8: M5 data. Robust tree showing the details of the divisive procedure.

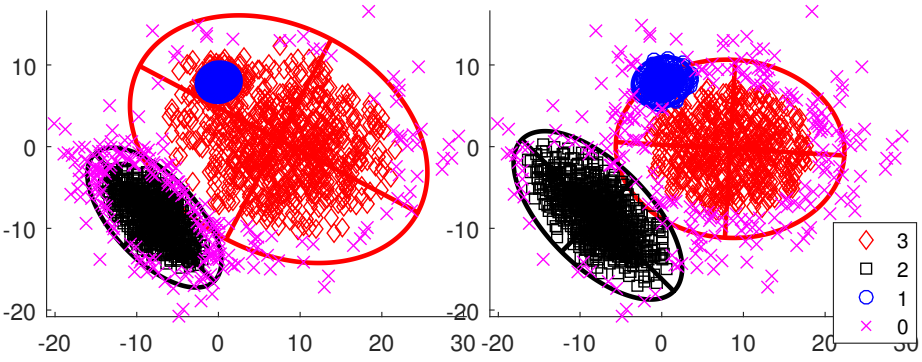

Figure 9: M5 data. Left panel: tentative clustering from the divisive procedure. Right panel: tentative clustering from TCLUST when K is set equal to 3 and the same trimming level is used as in the divisive procedure. In both panels 99% confidence ellipses have been added using the centroids and the covariance matrices of the estimated groups.

step could also help to separate small and concentrated groups of contaminated observations from background noise (Hennig and Liao, 2013; Coretto and Hennig, 2016), as well as to highlight the relationship between our procedure and robust fitting of mixture models. Both these topics are the subject of ongoing research.

Acknowledgements

The authors are grateful to the Editor, an Associate Editor and three anonymous referees for very detailed and thoughtful suggestions. They also thank Gunter Ritter for discussion on a previous draft.

References

Agostinelli, C., A. Leung, V. J. Yohay, and R. H. Zamar (2015). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test 24*, 441–461.

Atkinson, A. C., A. Cerioli, G. Morelli, and M. Riani (2015). Finding the number of disparate clusters with background contamination. In B. Lausen, S. Krolak-Schwerdt, and M. Böhmer (Eds.), *Data Science, Learning by Latent Structures, and Knowledge Discovery*, pp. 29–42. Heidelberg: Springer-Verlag.

Atkinson, A. C., M. Riani, and A. Cerioli (2004). *Exploring Multivariate Data with the Forward Search*. New York: Springer–Verlag.

Atkinson, A. C., M. Riani, and A. Cerioli (2006). Random start forward searches with envelopes for detecting clusters in multivariate data. In S. Zani, A. Cerioli, M. Riani, and M. Vichi (Eds.), *Data Analysis, Classification and the Forward Search*, pp. 163–171. Berlin: Springer-Verlag.

Atkinson, A. C., M. Riani, and A. Cerioli (2017). Cluster detection and clustering with random start forward searches. *Journal of Applied Statistics*, DOI:10.1080/02664763.2017.1310806.

Butler, R. W., P. L. Davies, and M. Jhun (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics 21*, 1385–1400.

Cator, E. A. and H. P. Lopuhaä (2012). Central limit theorem and influence function for the MCD estimators at general multivariate distributions. *Bernoulli 18*, 520–551.

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

Cerioli, A., A. Farcomeni, and M. Riani (2013). Robust distances for outlier-free goodness-of-fit testing. *Computational Statistics and Data Analysis 65*, 29–45.

Cerioli, A., 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.

Cerioli, A., M. Riani, A. C. Atkinson, and A. Corbellini (2017). The power of monitoring: how to make the most of a contaminated multivariate sample. *Statistical Methods and Applications*, DOI: 10.1007/s10260–017–0409–8.

Clarke, B. and D. Schubert (2006). An adaptive trimmed likelihood algorithm for identification of multivariate outliers. *Australian and New Zealand Journal of Statistics 48*, 353–371.

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

Cuesta-Albertos, J. A., C. Matrán, and A. Mayo-Iscar (2008). Trimming and likelihood: Robust location and dispersion estimation in the elliptical model. *The Annals of Statistics 36*, 2284–2318.

Davies, P. L. (1987). Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices ellipsoid estimator. *The Annals of Statistics 15*, 1269–1292.

Dotto, F., A. Farcomeni, L. A. García-Escudero, and A. Mayo-Iscar (2017). A reweighting approach to robust clustering. *Statistics and Computing in press*, https://doi.org/10.1007/s11222–017–9742–x.

Farcomeni, A. (2014a). Robust constrained clustering in presence of entry-wise outliers. *Technometrics 56*, 102–111.

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

Farcomeni, A. and L. Greco (2015). *Robust Methods for Data Reduction*. Boca Raton: Chapman and Hall/CRC.

Fritz, H., L. A. García-Escudero, and A. Mayo-Iscar (2012). tclust: An R package for a trimming approach to Cluster Analysis. *Journal of Statistical Software 47*, https://www.jstatsoft.org/v47/i12.

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

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

Hardin, J. and D. M. Rocke (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics 14*, 910–927.

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

Hennig, C. and T. Liao (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Applied Statistics 62*, 309–369.

Huber, P. J. and E. M. Ronchetti (2009). *Robust Statistics. Second Edition*. Hoboken: Wiley.

Hubert, M., P. J. Rousseeuw, and S. Van Aelst (2008). High-breakdown robust multivariate methods. *Statistical Science 23*, 92–119.

Johansen, S. and B. Nielsen (2016a). Analysis of the Forward Search using some new results for martingales and empirical processes. *Bernoulli 22*, 1131–1183.

Johansen, S. and B. Nielsen (2016b). Asymptotic theory of outlier detection algorithms for linear time series regression models (with discussion). *Scandinavian Journal of Statistics 43*, 321–348.

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

Riani, M., A. C. Atkinson, and A. Cerioli (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B 71*, 447–466.

Riani, M., A. Cerioli, A. C. Atkinson, and D. Perrotta (2014). Monitoring robust regression. *Electronic Journal of Statistics 8*, 646–677.

Ritter, G. (2014). *Robust Cluster Analysis and Variable Selection*. Boca Raton: Chapman and Hall/CRC.

Rousseeuw, P. J. and A. M. Leroy (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Rousseeuw, P. J. and W. Van den Bossche (2017). Detecting deviating data cells. *Technometrics*, DOI: 10.1080/00401706.2017.1340909.

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