Chapter 1

Robustness and Outliers

L.A. García-Escudero, A. Gordaliza, C. Matrán, A. Mayo-Iscar and C. Hennig

Departamento de Estadística e Investigación Operativa.

Universidad de Valladolid and University College London.

Abstract

Unexpected deviations from assumed models as well as the presence of certain amounts of outlying data are common in most practical statistical applications. This fact could lead to undesirable solutions when applying non-robust statistical techniques. This is often the case in cluster analysis, too. The search for homogeneous groups with large heterogeneity between them can be spoiled due to the lack of robustness of standard clustering methods. For instance, the presence of (even few) outlying observations may result in heterogeneous clusters artificially joined together or in the detection of spurious clusters merely made up of outlying observations.

In this chapter we will analyze the effects of different kinds of outlying data in cluster analysis and explore several alternative methodologies designed to avoid or minimize their undesirable effects.

1.1 Introduction

Robustness in statistics refers to stable behavior of methodology under small changes of data or models. For example, a small percentage of outliers can have a large impact on many statistical techniques. Robustness is a desirable property for more or less general statistical methodology.

In cluster analysis, many methods, be they heuristic or model-based, may suffer from strong instability from various sources of outlying data, by which data is meant that does not belong to any clear cluster. Outlying data points can act as "connectors" or "bridge points" between different groups, or they can play a disaggregating role. They can exhibit some degree of internal grouping, leading to additional spurious clusters, but they can also be "radial" or "isolated".

For illustration, consider the artificial data set in Figure 1.1(a), made up of 3 main groups and a small fraction of "radial" outliers. If the well-known K-means method with $K = 3$ is applied, two intuitive main clusters are joined together and a spurious cluster is found, composed of outlying observations only. Moreover, only one single outlier placed in a remote position joins together the two main clusters in Figure 1.1(b) artificially when applying K-means with $K = 2$.

Similar problems arise when applying data mining techniques that use cluster analysis in "unsupervised learning" problems with large, complex and high dimensional data sets collected through automated processes, which often contain many outliers.

A common strategy is to view isolated outliers and small groups of outliers as "clusters on their own". This is quite logical since outlying observations are obviously separated from other existing data patterns. Thus, some statisticians simply recommend increasing the number of groups in order to detect and isolate possibly outlying data points. For instance, the case in Figure 1.1(b) could have been addressed by searching for $K = 3$ clusters instead $K = 2$. However, this strategy is not always the most sensible one. For instance, due to prior knowledge about the problem at hand, the cluster analysis user may know or fix in advance the number of clusters and may not be aware of the presence of outlying data. Furthermore,

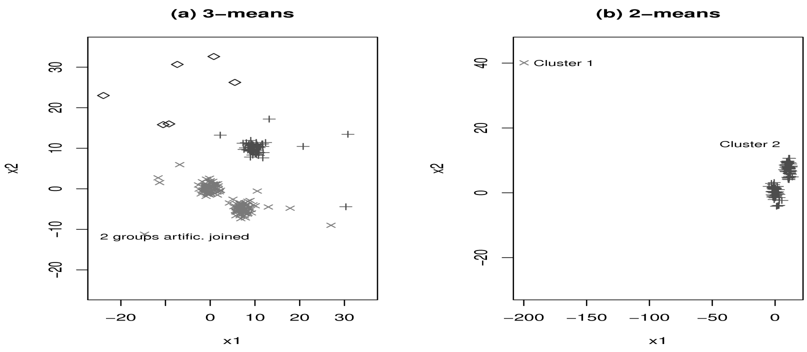

Figure 1.1: Effect of outliers in K-means cluster analysis: (a) Radial contamination (b) One single outlier.

in situations like the one shown in Figure 1.1(a), made up of a few main clusters and a certain amount of "radial" outliers or "background" noise, accommodating them with small clusters would require a very large number of clusters, which is tedious to interpret and may incur computational problems.

Note that the clustering problem does not have a general definition. The "correct" clustering of a data set depends on the application and the cluster concept of the user. In some applications it is fine to handle one or more small "clusters of outliers", in other applications only a small number of larger clusters is meaningful and outliers are better excluded from any cluster. It may also be appropriate to integrate them into the nearest clusters, but it is hardly desired to give them a strong impact on how the clearly well clustered non-outlying data points are partitioned. That there often is such an impact is the most serious robustness problem.

One type of outliers that can be very harmful in cluster analysis is "bridge points" located between the main clusters (see Figure 1.2). These bridge points are not outlying in any of

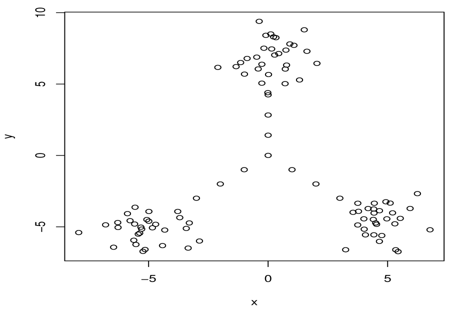

Figure 1.2: Outlying "bridge" points among 3 main clusters.

the coordinates and, thus, they cannot be detected with the use of the standard robust statistical tools existing in multivariate data analysis.

Another reason for the interest in robust cluster analysis techniques is the connection between cluster analysis and robust statistics, already pointed out in, e.g., Rocke and Woodruff (2012), Hennig (2002), García-Escudero et al. (2003), Hardin and Rocke (2004), Woodruff and Reiners (2004) and Schynsa et al. (2010). Firstly, as shown in previous examples, the need for robustness in cluster analysis is evident. Moreover, robust statistics in general can benefit from cluster analysis techniques, because often the most harmful outliers are those that appear clustered together (see Rocke and Woodruff (1996)), and cluster analysis is useful when handling groups of clustered outliers. Therefore, robust cluster analysis provides an appealing unifying framework for addressing both problems simultaneously.

Section 1.2 reviews some of the different robust cluster analysis approaches that have been proposed in the literature with some emphasis on trimming-based approaches. Subsection 1.2.6 summarizes different tools and approaches that have been employed to qualitatively and numerically measure robustness in cluster analysis. Section 1.3 discusses different software packages that can be applied in robust cluster analysis. Finally, in Section 1.4 some concluding remarks and other possible open research lines are presented.

Recent reviews of robust cluster analysis are García-Escudero et al. (2010) and Banerjee and Davé (2012). The former is more focused on a statistical approach whereas the latter is more focused on a "machine learning" approach.

1.2 Robustness in cluster analysis

1.2.1 $\mathcal{L}_1$ approaches

Given a sample $\mathcal{D} = \{x_1, ..., x_n\}$ in $\mathbb{R}^d$, the K-means method searches K point centers $\{\mathbf{m}_1, ..., \mathbf{m}_K\}$ in $\mathbb{R}^d$ and a partition $C_1, ..., C_K$ of $\mathcal{D}$ minimizing

$$\sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2. \tag{1.1}$$

Observations are then arranged into K clusters by assigning each observation to its closest center ($C_k = \{\mathbf{x}_i : \|\mathbf{x}_i - \mathbf{m}_k\| \leq \|\mathbf{x}_i - \mathbf{m}_l\|$ for every $l = 1, ..., K\}$).

As the sample mean (i.e., the 1-mean), the K-means method suffers from a serious lack of robustness. One single outlying data point placed arbitrarily far away from the others suffices to spoil K-means, see Figure 1.1(b). In fact, an extreme enough outlier will form a cluster on its own, so that, for fixed K, at least two original clusters (i.e., clusters found in the data set without outlier) will be merged.

Statistical methods based on least squares criteria are known to have poor robustness behavior. The lack of robustness of K-means could be explained by the least squares problem in (1.1). One could be tempted to try to robustify it by using the same arguments which led to the use of the sample median as a robust alternative to the sample mean in univariate location. Recall that the mean ($\mathcal{L}_2$ approach) minimizes over m the expression $\sum_{i=1}^{n}(x_i - m)^2$ whereas the median ($\mathcal{L}_1$ approach) minimizes the expression $\sum_{i=1}^{n} |x_i - m|$.

Starting from (1.1), this $\mathcal{L}_1$-approach leads to the K-medians and the Partitioning Around Medoids (PAM) methods (see Chapter **??** and references therein), essentially defined through minimization of

$$\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|$$

(the $\mathbf{m}_k$ may or may not be restricted to be members of the data set).

Unfortunately, the $\mathcal{L}_1$-approach only provides a rather modest robustification. The PAM method can resist the presence of an outlier such as in Figure 1.1(b) more often than K-means, but it breaks down if the number of outliers increases slightly or even if a single outlier is placed in a very remote position (see García-Escudero and Gordaliza (1999)). This also happens when considering other non-decreasing penalty functions $\Phi(\| \cdot - \mathbf{m}_k\|)$ in (1.1).

The unsuitability of the $\mathcal{L}_1$ attempt to derive robust clustering methods is analogous to what happens in the linear regression setup, where the robustness behavior of $\mathcal{L}_1$-regression is poor in presence of extreme leverage points (Rousseeuw and Leroy (1987)).

1.2.2 Approaches based on trimming

Among the first highly robust regression proposals were the Least Median of Squares (LMS) and the Least Trimmed Squares (LTS) method (Rousseeuw and Leroy (1987)). This motivates the adaptation of the trimming principle to cluster analysis. Trimming means that a certain "trimmed" fraction of the observations (the most outlying ones) are not clustered at all and not taken into account for clustering the remaining data. This is appropriate in many applications where clusters are interpreted as substantially meaningful if they are homogeneous and large enough, but where isolated points are not of interpretative use.

Trimming has a long history of providing robustness to statistical methods. For instance, the univariate trimmed mean removes a proportion $\alpha/2$ of the largest observations and a proportion $\alpha/2$ of the smallest ones before computing the sample mean. Extending this idea to cluster analysis is not straightforward. Most interesting cluster analysis applications are multivariate, and a natural geometrical order does not exist for choosing the most extreme observations. Moreover, the notion of "outlyingness" must be modified to include potential bridge points between clusters (recall Figure 1.2).

A sensible way to perform trimming is to let the data decide which observations should be trimmed in order to find an optimal clustering for the untrimmed ones. This is the idea behind the aforementioned LMS and LTS regression method as well as the MCD (Minimum Covariance Determinant) and MVE (Minimum Volume Ellipsoid) covariance matrix estimators, see Rousseeuw and Leroy (1987).

Let $\varphi(\cdot;\mu,\Sigma)$ be the probability density function of the d-variate normal distribution with mean μ and covariance matrix Σ. Robust cluster analysis methods with trimming can be introduced through maximization of the target function (weighted trimmed classification likelihood or trimmed penalized likelihood, see below)

$$\sum_{k=1}^{K}\sum_{\mathbf{x}_i\in C_k}\log\left(p_k\varphi\left(\mathbf{x}_i;\mathbf{m}_k,\mathbf{S}_k\right)\right), \tag{1.2}$$

where maximization is in terms of:

(i) K centers $\mathbf{m}_k$ in $\mathbb{R}^d$,

(ii) K symmetric positive definite $d\times d$ scatter matrices $\mathbf{S}_k$ (satisfying constraints to be detailed later),

(iii) K weights in $[0,1]$ satisfying $\sum_{k=1}^K p_k=1$, and,

(iv) a partition $C_0, C_1, \ldots, C_K$ of $\mathcal{D}$ with C_k containing the observations that are assigned to cluster k for $k=1,...,K$, and the set C_0 with cardinality $n_{C_0}=[n\alpha]$ to include the observations that are trimmed.

Note that summation goes from $k=1$ to K (not from $k=0$) and, thus, the observations in C_0 are not taken into account when evaluating the target function. This avoids the harmful influence of a fraction of at most α of outlying observations.

This approach is called "impartial trimming" in Gordaliza (1991) and Cuesta-Albertos et al. (1997). Gallegos (2002) and Gallegos and Ritter (2005) introduced an interesting probabilistic framework, called “spurious outliers model”, that justifies the consideration of this type of target function. This model states that the data points are independent, with a proportion of $1 - [n\alpha]$ points distributed according to one of K normal distributions defined by K different pairs of mean and covariance matrix, and the remaining $[n\alpha]$ (trimmed) points distributed according to $[n\alpha]$ different unspecified distributions $G_1, \dots, G_{[n\alpha]}$. In Gallegos and Ritter (2005) it is proved under some mild conditions on the set of admissible $G_1, \dots, G_{[n\alpha]}$ that maximizing (1.2) amounts to computing the maximum likelihood (ML) estimator for all the parameters of the K normal distributions and trimming the set of $[n\alpha]$ points assigned to but not depending on $G_1, \dots, G_{[n\alpha]}$. Such an ML estimator assumes that all p_k in (1.2) are the same (“classification likelihood”). It is known (e.g., Gallegos and Ritter (2005)) that this implicitly favors clusters of similar sizes, and allowing flexible p_k is a way to deal with this.

Using such theory, one can see that the trimming approach based on (1.2) uses the normal distribution as a “cluster prototype shape”, i.e., it is appropriate for clustering applications in which the clusters are expected or desired to have an approximately normal shape. This does not necessarily imply that the underlying “true” clusters have to be normal; however, non-normal clusters will be approximated by normal distributions and one may need more than K normal distributions to fit K non-normal “clusters”. It is possible to define similar approaches for other cluster prototype shapes, i.e., other families of distributions.

It is important to note that the maximization of (1.2) without any constraints on the scatter (covariance) matrices $\mathbf{S}_k$ is a mathematically ill-posed problem. To see this, just take $C_k = \{\mathbf{x}_i\}$ for any $\mathbf{x}_i \in \mathcal{D}$, $\mathbf{m}_k = \mathbf{x}_i$ and $\mathbf{S}_k$ with $\det(\mathbf{S}_k) \to 0$, which makes (1.2) unbounded. Different constraints on the scatter matrices have been considered in the literature. Different cluster analysis methods can be derived depending on whether equal weights ($p_1 = ... = p_K$) are assumed or not.

The first application of this impartial trimming approach was the trimmed K-means method in Cuesta-Albertos et al. (1997). This is the most constrained case, assuming equal weights $p_1 = ... = p_K$ and $\mathbf{S}_1 = ... = \mathbf{S}_K = s^2\mathbf{I}_d$, i.e., spherical clusters with equal within-cluster variation and (through $p_1 = ... = p_K$) a tendency to form clusters of similar sizes.

The maximization of (1.2) is simplified to the minimization of

$$\sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} \|\mathbf{x}_i - \mathbf{m}_k\|^2, \tag{1.3}$$

with $C_0, C_1, ..., C_K$ being a partition of $\mathcal{D}$ with $n_{C_0} = [n\alpha]$.

Figure 1.3 illustrates the ability of the trimmed K-means method to deal with different kinds of contamination. In both examples, parameter α is set to 0.1, that is, 10% of the data are trimmed (represented in this graph by circles). $K = 3$ main groups are detected in spite of the presence of radial contamination in Figure 1.3(a) and bridge points in Figure 1.3(b).

Problem (1.3) admits a population version. This is a a version of an estimator that is defined as a functional on the space of probability distributions instead of data sets in such a way that it generalizes the estimator, i.e., one obtains the estimator if the population version is evaluated at the empirical distribution of a data set. Well-behaved population versions of an estimator allow that the estimator is consistent for its own population version at a given distribution. In cluster analysis, population versions of methods such as K-means allow to partition the whole space of observations instead of a single data set. The population trimmed K-means can be viewed as a robustification of the "vector quantizers" used in signal processing, e.g., Gersho and Gray (1991). Cuesta-Albertos et al. (1997) established the existence without moment conditions of the sample and population trimmed K-means. Furthermore, the sample trimmed K-means centers are consistent estimators of the population ones. A central limit theorem is derived in García-Escudero et al. (1999). However, if the underlying distribution is a mixture of normal distributions (with or without additional outlier generating distributions), the population trimmed K-means centers are not identical to the means of the normal components for the same reasons as when using unrobustified K-means, see Section **??**.

An efficient algorithm for obtaining the trimmed K-means is available. The TRIMMED K-MEANS algorithm given by García-Escudero et al. (2003) is an extension of the classical Forgy (1965)'s K-MEANS algorithm, where some "concentration steps", similar to the ones used in the FAST-LTS and FAST-MCD algorithms (see Rousseeuw and Van Driessen (1999)), are applied.

```
Algorithm Trimmed K-Means
Input parameters $K$ and $\alpha$
  1. Draw $K$ random initial centers $\mathbf{m}_1^0, ..., \mathbf{m}_K^0$
  2. Concentration steps:
     (a) Keep the set $C$ of the $[n(1-\alpha)]$ observations closest to the centers $\mathbf{m}_1^l, ..., \mathbf{m}_K^l$.
     (b) Partition $C$ onto $K$ subsets $\{C_1, ..., C_K\}$, where $C_k$ contains the observations in $C$
         closer to the center $\mathbf{m}_k^l$ than to the other centers.
     (c) Update the centers $\mathbf{m}_1^{l+1}, ..., \mathbf{m}_K^{l+1}$ such that each center $\mathbf{m}_k^{l+1}$ is the sample mean
         of the observations in $C_k$. Go to (a) unless centers have not changed anymore.
  3. keep the optimal centers that minimize (1.3).
Output optimal centers and clustering.
```

Cuesta-Albertos et al. (1998) introduced a $\mathcal{L}_\infty$ version of trimmed K-means, called trimmed best K-nets. It is based on obtaining the K balls with minimal radii that cover a proportion $1-\alpha$ of the data. This approach may be seen as an extension of the LMS and MVE techniques to cluster analysis and it shares with them their slow rate of convergence (see Cuesta-Albertos et al. (2002)). Interpretation can be given through the excess-mass approach in Müller and Sawitzki (1991), i.e., a cluster concept according to which clusters are interpreted as connected sets of high density. Kumar and Orlin (2008) also consider an interesting extension of the MVE to cluster analysis. Jolion et al. (1991) use the MVE and a Kolmogorov-Smirnov test procedure to detect clusters in a data set sequentially.

Since K-means and trimmed K-means are based on the use of Euclidean distances, they aim at finding spherical groups of points with similar sizes. When researchers are interested in clusters that strongly depart from this assumption, these methods fail to identify them. For instance, an elongated group can be split into several clusters, or several elongated groups located close to each other can be joined together to constitute a single cluster. The search for clusters that have different sizes and different scatter structures leads to the "heterogeneous" cluster analysis problem, for which robustness issues should also be addressed.

With a heterogeneous robust cluster analysis perspective in mind, Gallegos and Ritter (2005) introduced the determinant criterion to robustify the proposal by Friedman and Rubin (1967). This implies that equal (but not necessarily spherical) scatter matrices $\mathbf{S}_1 = ... = \mathbf{S}_K = \mathbf{S}$ are assumed in (1.2). The criterion in Gallegos (2002) allows for different scatter matrices, but it implicitly assumes equal scales $\det(\mathbf{S}_1) = ... = \det(\mathbf{S}_K)$. Equal weights $p_1 = ... = p_K$ are assumed in both cases. The proposed algorithms (see details in Gallegos (2002) and Gallegos and Ritter (2005)) are similar to the Trimmed K-Means algorithm, but Mahalanobis distances are used when computing distances to point centers. Moreover, they reduce to the fast-MCD algorithm when $K = 1$. The algorithm in Gallegos (2002) constrains the $\mathbf{S}_k$ scatter matrices to have the same determinant in every concentration step. This idea was also considered in Maronna and Jacovkis (1974) as a sensible way to overcome the unboundedness of (1.2) in the untrimmed case. Considering the decomposition of $\mathbf{S}_k = s_k^2 \mathbf{U}_k$ with s_k^2 as an scale parameter and $\mathbf{U}_j$ with $\det(\mathbf{U}_k) = 1$ as the shape matrix,

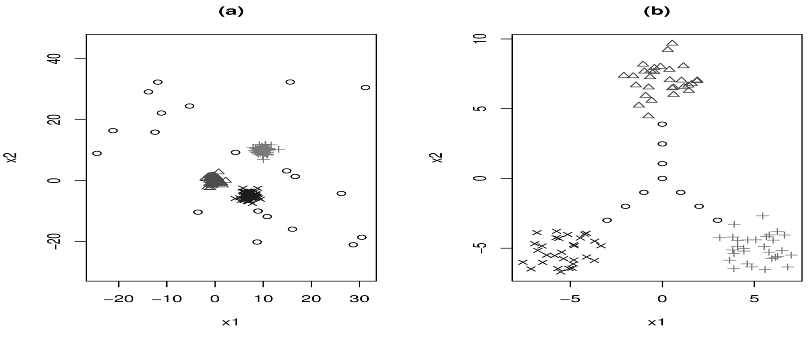

Figure 1.3: Trimmed K-means results with $K = 3$ and $\alpha = 0.1$ for two "contaminated" data sets (Trimmed points="○").

García-Escudero and Gordaliza (2007) stated that the proper estimation of the cluster scales s_k^2 is the most difficult task in (robust) cluster analysis (as long as α and K are fixed) and proposed an iterative procedure for estimating the scales based on a decreasing sequence of trimming levels.

In order to rely on more flexible constraints, we could try to extend Hathaway (1985)'s constraints to the heterogeneous robust cluster analysis framework. Consider the eigenvalues system of the cluster covariance matrices denoted by $\lambda_l(\mathbf{S}_k)$, for $k = 1, ..., K$ and $l = 1, ..., d$, and their maximum and minimum values

$$M_n = \max_{k=1,...,K} \max_{l=1,...,d} \lambda_l(\mathbf{S}_k) \text{ and } m_n = \min_{k=1,...,K} \min_{l=1,...,d} \lambda_l(\mathbf{S}_k).$$

We constrain the ratio M_n/m_n to be smaller than a fixed constant c. The use of this type of constraints together with group weights p_k's in (1.2) lead to the TCLUST approach in García-Escudero et al. (2008). The constant c controls the strength of the scatter constraint, yielding a weighted version of the trimmed K-means when $c = 1$, and an almost unrestricted procedure when using a large c value. The Tclust algorithm (García-Escudero et al. (2008) and Fritz et al. (2013a)) can be applied to solve this problem approximately. This is a Classification-EM type algorithm (see Celeux and Govaert (1992)) with concentration steps. The TCLUST approach admits a population version and a consistency result is available.

Gallegos and Ritter (2009b) proposed a different extension of Hathaway's constraints (Hathaway (1985)), called Hathaway-Dennis-Beale-Thompson (HDBT), to the heterogeneous robust cluster analysis setup, which defines their version of trimmed model-based clustering. The Löwner ordering on the space of symmetric matrices is considered to constrain the scatter matrices. This implies considering the following constraint:

$$\min_{1\leq l\leq d} \min_{1\leq h\neq j\leq K} \lambda_l(\mathbf{S}_h\mathbf{S}_j^{-1}) \geq \frac{1}{c} \text{ with } c \geq 1,$$

resulting in an affine equivariant cluster analysis procedure. Gallegos and Ritter (2010) deal with the unboundedness of the trimmed likelihood by controlling smaller cluster sizes.

Another application of trimming in heterogeneous robust cluster analysis is the MINO (minimum covariance determinant with outliers) approach by Rocke and Woodruff (2012) and Woodruff and Reiners (2004). Trimmed likelihoods like (1.3) are also considered in Markatou (2000) and Neykov et al. (2007). The main difference of these trimming approaches to TCLUST/HDBT is that TCLUST and HDBT place explicit constraints on the relations between cluster scatters whereas the other methods do not incorporate a mathematically consistent choice between multiple maximizers of the trimmed likelihood.

Careful monitoring of iterative trimming results is the basis of the "Forward Search" approach to robust cluster analysis (see Atkinson et al. (2006) and Atkinson and Riani (2007)). This starts from a large number of small initial potential cluster cores (trimming effectively the vast majority of points) and adds points successively. Monitoring the change of within cluster homogeneity over the point addition process can uncover the clustering structure and potential outliers.

Santos-Pereira and Pires (2002), Hardin and Rocke (2004) and Willems et al. (2009) pay special attention to how to define and analyze group Mahalanobis distances from a robust cluster analysis viewpoint. Fixed point clustering (Hennig (2002)) also applies Mahalanobis distances and concentration steps to discover clusters in a robust way.

1.2.3 Mixture approaches for robust clustering and modeling noise

Instead of discarding outlying observations, alternative approaches are based on fitting data not belonging to any cluster (so-called "noise", Fraley and Raftery (1998)) by mixture components. It is straightforward to obtain a "crisp" partition of $\mathcal{D}$ from a mixture fit, see Section **??**, including points classified as noise.

A popular approach is based on accommodating the noise a uniformly distributed (Poisson process) "noise component" as done, e.g., in Banfield and Raftery (1993), Dasgupta and Raftery (1998), Campbell et al. (1997), Campbell et al. (1999), Fraley and Raftery (1998) and Fraley and Raftery (2002). It are based on the maximization of log-likelihoods such as

$$\sum_{i=1}^{n} \log \left(p_0 \frac{1}{\mathrm{Vol}(V)} I_V(\mathbf{x}_i) + \sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i; \mathbf{m}_k, \mathbf{S}_k) \right), \tag{1.4}$$

where $p_k \in [0, 1]$ with $\sum_{k=0}^{K} p_k = 1$ and V is a set in $\mathbb{R}^d$ and $\mathrm{Vol}(V)$ denotes its volume.

Sometimes it is clear which set V should be considered (for instance, the area of a given image). Otherwise V could be considered as the convex hull of the data or the smallest rectangle whose sides are parallel to the axes or to the principal components, such that it contains all data points. For one-dimensional data all these amount to the range. All these are proposed in Fraley and Raftery (1998). The EM algorithm in Dempster et al. (1977) is used to estimate the unknown parameters in (1.3) once the set V is fixed. Another alternative approach is based on estimating V through ML estimation of a uniform distribution on V (Coretto and Hennig (2011)). All these methods provide some robustification but a single observation placed in a very remote position could still break down the clustering (Hennig (2004), Hennig (2008)).

Hennig (2004) also proposed the RIMLE (Robust Improper Maximum Likelihood Estimator), which is based on the maximization of an “improper” likelihood:

$$\sum_{i=1}^{n} \log \left(p_0 c + \sum_{k=1}^{K} p_k \varphi(\mathbf{x}_i; \mathbf{m}_k, \mathbf{S}_k) \right), \tag{1.5}$$

where c is a tuning parameter that ideally serves to classify as noise observations arising from areas where components account for density values smaller than c. There are clear connections of the RIMLE approach with the trimming approach discussed in Section 1.2.2. Coretto and Hennig (2010) show that this approach has good robustness properties through a simulation study (the cited results are for one-dimensional data, but can be generalized to multidimensional data). When maximizing (1.4) and (1.5), constraints on the scatter matrices $\mathbf{S}_j$ are again needed to avoid degeneration.

The main difference between these approaches and trimming is that the mixture-based techniques allow a smooth transition of the classification between outliers and clusters (and between different clusters) through the estimated posterior probabilities of the observations belonging to the clusters (normal mixture components) and the noise component. Although they are often used for obtaining a crisp clustering, smoothness still has a certain (usually small) impact on the parameter estimation of the normal components, which may change the resulting clustering. Choosing c in the RIMLE approach is analogous to choosing α for trimming.

A different way to tackle the problem of outlying data points uses heavy-tailed mixture components to accommodate them. Mixtures of t-distributions with or without estimation of the degrees of freedom have been considered in, e.g., McLachlan and Peel (2000) and Shoham (2002). This approach provides a valuable robustification, but Hennig (2004) showed again that a single extreme outlying value could cause breakdown. Greselin and Ingrassia (2010) consider appropriate scatter matrices constraints to avoid degeneration in such a setup.

Using mixtures of t-distributions for robustness implies a slightly different "clustering philosophy" than fitting normal mixtures. When normal mixtures are used for clustering, the normal distribution defines the "cluster prototype shape" and the idea behind adding a noise component is to capture points that cannot be assigned to any normal distribution sufficiently supported by the data. However, when mixtures of t-distributions are fitted for dealing with outliers, the idea is that only the core of a t-distribution is considered as "cluster", whereas points in the tails are interpreted as outliers, see McLachlan and Peel (2000), Sec. 2.5. This implies that outliers are still assigned to mixture components, even if they lie far away from all clusters.

Trimming principles were also considered to provide robustness when fitting mixtures. Such trimmed mixture likelihood methods are introduced in Neykov et al. (2007), Cuesta-Albertos et al. (2008) and Gallegos and Ritter (2009a).

Robustness in statistics is more general than dealing with outliers, although outliers often cause the worst robustness problems. Another interesting aspect in the mixture setup is what happens when normal mixtures are fitted to data where the clusters are actually non-normal. In some recent work (Tantrum et al. (2003),Hennig (2010),Baudry et al. (2010)) is was noted that when estimating the number of mixture components via the BIC (Sec.

??), normal mixtures can fit other distributions very well, but the estimated number of normal components will be higher than the number of interpretable clusters (depending on how exactly clusters are defined). The cited papers propose schemes to model clusters by sub-mixtures of normal distributions, i.e., the clusters are obtained by merging some of the normal components.

Finally, it is important not to confuse the use of the term robustness made in this chapter with that made in the Bayesian mixture modeling framework where it often means insensitivity to the choice of prior probability distributions.

1.2.4 Robust fuzzy cluster analysis

Whereas hard cluster analysis procedures are aimed at searching for sensible partitions of data into K disjoint clusters, fuzzy cluster analysis methods provide non-negative membership values of observations for clusters so that overlapping clusters are generated (see, e.g., Ruspini (1969), Dunn (1974), Bezdek (1981) and Hathaway and Bezdek (1993), Chapter **??**). Models based on mixture models as discussed in the previous section are not normally regarded as fuzzy clustering methods because the underlying mixture model assumes that every point was generated by a single mixture component, although they, too, enable "soft classification" through posterior probabilities of points to belong to clusters.

As before, it is widely recognized that robustness of fuzzy cluster analysis methods is important for many practical applications. The fuzzy cluster analysis community was perhaps the first one to take the robustness challenge in cluster analysis seriously. Intuitively, outliers tend to be approximately "equally remote" to all clusters and, thus, they could have similar (but not necessarily small) membership values for all clusters. For instance, membership values for outlying observations could be close to $1/K$ for all the clusters with K being the number of groups whenever membership values are assumed to sum up to 1. Moreover, the way that many fuzzy clustering procedures weight data points is related to the way incorporated by M-estimators in robust statistics. Davé and Krishnapuram (1997) gives a review on robust fuzzy cluster analysis.

One of the methods which is more widely considered in robust fuzzy cluster analysis is the fuzzy K-means method with "noise component" introduced in Davé (1991), which has a plethora of modifications. This method is based on a modification of the fuzzy K-means algorithm (Sec. **??**) that considers a fictitious $(K+1)$-th center at the same distance δ from every point in the data set. Observations that are associated to this fictitious center are denoted as belonging to the "noise cluster". This implies that points will have their highest membership degree for the noise cluster if they are further away than δ from any other cluster center. The choice of the parameter δ is not an easy task but some sensible heuristic rules have been proposed (see, e.g., Davé (1991), Davé and Sen (1997), Rehm et al. (2007)). The methods based on "noise distances" are also refereed as "noise clustering" in the literature. $\mathcal{L}_1$ approaches have been also considered in Krishnapuram et al. (1999) through the proposal of a fuzzy K-medoids algorithm.

A "least trimmed squares" fuzzy clustering method (closely related to the trimmed K-means in Section 1.2.2) was introduced in Kim et al. (1996). The proposed method is based on trimming a fixed fraction α of observations. The fixed trimming level controls the number of observations to be discarded in a different way than those based on defining a "noise distance" δ. Discarding a fixed fraction of data has also been considered in Klawonn (2004).

Possibilistic fuzzy clustering methods (Krishnapuram and Keller (1993)) address the problem of data with "noise"-observations in fuzzy clustering by relaxing the constraint that the membership values have to sum up to 1. Outlying observations could then receive arbitrarily low membership values for all clusters. Although possibilistic fuzzy clustering methods are more robust against outliers, they tend to produce coincident clusters (Barni et al. (1996)) and, therefore, it is better to see them as "mode-seeking" rather than "partitioning" procedures.

All these approaches inherit from fuzzy K-means their preference for spherical clusters, and so they are not well suited to detect clusters with very different shapes. Several procedures have been proposed to address heterogeneous fuzzy cluster analysis problems (see, e.g., Gustafson and Kessel (1979), Trauwaert et al. (1991) and Rousseeuw et al. (1996)). These methods can also be modified by including the possibility of trimming a fixed proportion α of the data. This would lead to a fuzzy version of the TCLUST procedure described in Section 1.2.2 as introduced in Fritz et al. (2013b). Robust fuzzy cluster analysis using t-distributions has been proposed in Chatzis and Varvarigou (2008). Łeski (2003) and Wu and Yang (2002) give further procedures for robust fuzzy clustering. More interesting references are in Banerjee and Davé (2012).

1.2.5 Robust non-model based cluster analysis

Some widely applied non-model based cluster analysis approaches suffer from serious robustness problems as well. For instance, anomalous observations often induce spurious "bumps" in non-parametric methods based on density estimation leading to the detection of artificial clusters; outlying data points can induce "chaining" effects on hierarchical cluster analysis methods, and so on.

Some attempts to improve the robustness of these non-model based cluster analysis methods can be found in the literature. For instance, "denoising" techniques have been proposed resorting to Voronoï methods in Allard and Fraley (1997), nearest-neighbor methods in Byers and Raftery (1998) and Minimum Spanning Trees in Jaing et al. (2001).

Some density-based cluster analysis methods like Cuevas et al. (2001) seem to have good behavior from a robustness point of view. This is due to the fact that points are considered "outliers" if they lie in regions with low density, so if (given a certain density estimator) points are only assigned to clusters in areas of high density, these can be expected to be unaffected by outliers. Such a technique will depend on the typical tuning for density estimators (e.g. a bandwith for kernels or a neighborhood size) and a density cutoff similar to the c required for the RIMLE in Section 1.2.3. Robustness can only be expected if these choices are not dominated by the outliers.

Cluster analysis methods based on statistical depths as in Ding et al. (2007) exist, too. Robust hierarchical cluster analysis proposals exist as well, such as in Lin and Chen (2005) and Balcan and Gupta (2010). Balcan and Gupta (2010) assume that there is a "good" clustering in the data, demanding that for all points in a subset of the data of size $(1-\nu)n$ all but αn (α very small) of their nearest neighbors belong to the same cluster, where ν and α are tuning constants. This means that all but $(1-\nu)n$ points are "strongly clustered" in the sense above, so that νn outliers are allowed. Their hierarchical algorithm starts by first partitioning the data set into so-called "blobs" of points, making sure that points that can be connected by having enough common neighbors are in the same blob. These blobs are then hierarchically clustered by agglomeration using a score function that considers clusters at the current agglomeration level as similar if the median distance between a point of one cluster and all points of the other one is small for at least half of the points of one of the clusters. The use of medians attempts to make this unaffected by potential outliers in blobs/clusters. Balcan and Gupta (2010) show that this produces a tree, a pruning of which is the assumed "good" clustering.

A number of other non-model based clustering methods, e.g., BIRCH, CURE, ROCK, "Chameleon" and DBSCAN, are also designed in order to provide protection against noise and outliers. Comments and useful references are given in Banerjee and Davé (2012).

1.2.6 Parameter-based measurement of robustness in cluster analysis

In cluster analysis, and generally in the statistical literature, many methods are advertised as "robust". Often this is motivated either by heuristic arguments why a method is supposedly unaffected by outliers, or by good results in simulation studies or example data sets with outliers. Although this is better than ignoring the robustness issue altogether, such arguments are unsatisfactory because it is unclear to what extent they generalize to situations other than those explicitly considered by the authors.

The measurement of the robustness of statistical procedures has been addressed through two main approaches: Huber's "minimax" approach and Hampel's infinitesimal approach (Huber (1981) and Hampel et al. (1986)). The first approach has only been applied in rather restricted setups, and up to our knowledge it has not been adapted cluster analysis. We will focus on the second one.

Hampel's approach rests on three fundamental pillars: qualitative robustness, the influence function, and the breakdown point. All these robustness measures were designed to be applied to statistical procedures viewed as functionals defined on a suitable space of probability measures. Qualitative robustness has to do with the continuity of the functional (population version of an estimator) at the assumed model, the influence function with its differentiability (therefore they both are "infinitesimal" concepts) and the breakdown point with the distance from the model to a singularity of the functional.

More precisely, qualitative robustness of an estimator T_n means that, for large enough n, if two distributions P and Q are close to each other (e.g. one being $(1-\epsilon)$ times the other plus a proportion of small ϵ of a distribution that generates extreme outliers) according to a suitable metric for probability measures, the distributions of T_n for i.i.d. samples from P and Q are close to each other as well.

Given an estimator functional (population version) T, its influence function at a distribution P is $\lim_{epsilon \to 0} T((1-\epsilon)P + \epsilon\delta_x) - T(P)\epsilon$ as a function of x. This formalizes the change of T under infinitesimal contamination by a one-point distribution at x. Infinitesimal robustness, requires this to be bounded.

The breakdown point complements the infinitesimal concepts by looking at to what minimum extent one has to contaminate P in order to drive a functional T as far as possible away from its value at P, i.e., to a singularity of the parameter space (often infinity, but it can be zero, e.g., for variances or mixture component proportions). The most popular version of the breakdown point looks at the infimum ϵ for distributions of the form $(1-\epsilon)P + \epsilon H$ so that breakdown happens, where H is chosen as destructive as possible (contamination model). This is for example zero for the arithmetic mean and $\frac{1}{2}$ for the median. Another version of the breakdown point is the "finite sample breakdown point" of an estimator (Donoho and Huber (1983)), formalizing what proportion of data points in a data set needs to be replaced (or added) in order to drive the estimator to a singularity. This is often easier to handle than the population-based version.

In cluster analysis, one of the advantages of model-based clustering methods is that clusters can be characterized by features that have a population counterpart. Thus, we can view clustering methods as statistical functionals of probability measures (the empirical and the population one) and then apply Hampel's classical approach to measure their robustness behavior.

The qualitative robustness notion is binary; it classifies the functionals as having this property or not. García-Escudero and Gordaliza (1999) proved that for distributions with uniquely defined trimmed K-means, the trimmed K-means are qualitatively robust whereas K-means and PAM/K-medoids are not. García-Escudero and Gordaliza (1999) also addressed the influence functions of K-means and trimmed K-means as estimates of the centers of the clusters. Influence functions were obtained for replacing the Euclidean norm in (1.3) by general penalty functions $\Phi(\| \cdot - \mathbf{m}_k \|)$, and it was shown that they are bounded when the derivative of this Φ-function is bounded. However, in cluster analysis (as in regression), it could happen that an estimator with bounded influence is nevertheless highly vulnerable to outliers. This is, for instance, the case of PAM as reflected through its breakdown point, which equals zero. Ruwet et al. (2012) showed that all statistical functionals involved in the TCLUST procedure, i.e. the centers, the scatter matrices and the weights, have bounded influence functions whenever they are uniquely defined for the underlying distribution P.

García-Escudero and Gordaliza (1999) pointed out that the finite sample breakdown point is very data dependent in the context of cluster analysis. For any clustering method incorporating location parameters, data sets can be designed in which the inherent structure is so unstable that a single point converging to infinity will eventually attract a cluster. Particularly if K clusters are fitted to a data set that can be fitted by $K-1$ clusters without much loss of quality, if one outlier is added one couldn't even say that a solution is clearly "wrong" or undesirable in which the Kth cluster is used to fit the outlier. Instability in cluster analysis may be caused by an unstable clustering structure of the data set as well as by choice of a non-robust method. Therefore one cannot hope that any reasonable clustering method has good breakdown behavior uniformly over all data sets. Non-robust methods such as K-means and PAM, however, have a zero breakdown point for all data sets.

García-Escudero and Gordaliza (1999) suggested dealing with "well clustered" data sets in order to measure and compare the breakdown behavior of cluster analysis methods. Gallegos and Ritter (2005), Gallegos and Ritter (2009b) and Gallegos and Ritter (2009a) formalized the notion of "well clustered" data sets through a cluster separation condition. Moreover, they introduced a restricted breakdown point notion on the class of data sets satisfying the separation property and applied this measure to the procedures introduced by them. Ruwet et al. (2013) studied the restricted breakdown values of the TCLUST procedure and compared them with other trimming based approaches. In all these papers, the restricted breakdown point, on the class of data sets satisfying the corresponding separation properties, is asymptotically equal to the trimming level α, whereas in the case of the untrimmed versions the restricted breakdown continues to be 0. Ruwet et al. (2013) shows that the separation criterion for the method in Gallegos and Ritter (2009b) is a bit more restrictive than the one for TCLUST as a price to be paid for the former's affine equivariance. In order to appreciate these results, note that one cannot hope for a better breakdown point in cluster analysis than the proportion of the smallest cluster (if breakdown is to be achieved by adding points) or even half that size (for a "replacement breakdown point"), because a sufficiently homogeneous set of outliers of this size will attract a fitted cluster in a stronger way than the original smallest cluster. Therefore, the trimming rate α should not be chosen larger than the smallest cluster of interest could be. Otherwise this cluster could be trimmed. This is different from classical situations such as regression or location estimation, where the best achievable breakdown point is often $\frac{1}{2}$.

Hennig (2004) shows that standard ML based on a normal mixture is not breakdown robust. ML under adding a uniform mixture component over the convex hull of the data or for t-mixtures have a very similar breakdown point behavior. Their (addition) breakdown point is $1/(n+1)$, which is the smallest possible value. Nevertheless, both the convex hull uniform noise method and ML for t-mixtures can be robust against one or more outliers of moderate size. They only break down under very extreme outliers. The RIMLE is proven to have a better (asymptotically nonzero) breakdown point under a condition making sure that K clusters can be fitted much better than $K-1$ components on the same data set, which takes the role of the above mentioned "good clustering" assumption.

All the results in the present section assume that the number of clusters K is fixed, which is required in order to make the parameter space and its boundaries well defined.

1.2.7 Assignment-based and other robustness measures

It is a matter of controversy to what extent parameter-based robustness measurement captures the robustness properties of cluster analysis methods appropriately. On one hand, robustness failure as flagged by these measures is certainly a serious problem. On the other hand, there are a number of robustness issues that do not show up as "breakdown" or "unbounded influence". Furthermore, these methods cannot be applied to clustering methods that are not based on estimation of statistical parameters, and they assume the number of clusters K to be fixed.

Hennig (2004) proposed an amendment of the (finite sample addition) breakdown point definition for mixture model parameters with K estimated, which declares the estimator as "broken down" for a data set with K clusters before addition of outliers, if the parameters of at most $K - 1$ components do not either reach a singularity or the components vanish. The idea is that loss of a component is treated as breakdown, whereas adding components is not a problem, particularly because it can be seen as appropriate that newly added points require new components to be fitted. The somewhat surprising but intuitively reasonable consequence of this definition is that if criteria such as the AIC or BIC (see Section **??**) are used to estimate K, all considered methods including ML for plain normal mixtures are perfectly robust against extreme outliers, and breakdown can only be achieved by closing gaps between clusters with added "bridge points". However, as already mentioned, increasing K with the number of outliers is not always the best possible solution because it could lead to a large value of K with radial contamination, which may result in computational trouble when fitting the mixture.

Back to fixed K, another concern regarding parameter-based robustness measurement is that not all serious robustness problems are caused by parameters diverging to the borders of the parameter space. One can imagine very different clusterings on the same data set implied by different parameters that are all finite and far away from the parameter space borders. According to the classical robustness concepts, this kind of problem should be captured by the influence function and qualitative robustness. However, these measurements may miss problems because of the discrete nature of cluster analysis.

Consider $K = 3$ (fixed) clusters fitted to a data set in which there are four clearly visible groups, about equispaced (and maybe a few outliers to be trimmed). Imagine that trimmed K-means (α chosen so that the outliers but nothing else are trimmed) merges two of the four clusters, because they may be a tiny little bit less separated from each other or a tiny little bit smaller than the others. Let us assume that this solution is unique, but the clustering of the data set is certainly ambiguous and a quite different partition, merging two other components, may achieve almost the same value of the objective function (1.3). According to García-Escudero and Gordaliza (1999), as explained above, the influence function is bounded, not indicating any robustness problem. However, the decision which clusters are merged may be switched, changing the solution completely, under adding a few points or even a single one in one or between two of the clusters, because this may just make the originally slightly worse partition slightly better with respect to (1.3). The problem is that in such situations what happens for infinitesimal contamination as formalized in the influence function is not a good approximation for what may happen under still quite small contamination, at least if the trimmed K-means objective function is close to non-uniqueness.

In order to address this kind of robustness problem, and also in order to enable robustness measurement for more general clustering methods, robustness measurement can be based on the assignment of points to the clusters, and on counting assignment changes between original and contaminated data.

Such approaches have been used for measuring clustering stability in the machine learning literature for some time, see von Luxburg (2010) for an overview. One key result is that the stability of K-means under random variations is strongly connected to the uniqueness of the solution to (1.1) in the data set, which can be related to the above discussion. Such results differ from the robustness approach in that worst case behavior is not taken into account.

Hennig (2008) defined general assignment-based robustness measurements for cluster analysis. The “dissolution point” is an adaptation of the breakdown point. Because in many data sets some clusters are more clear and more stable than others, a dissolution point is not defined for a whole data set but for a cluster (in principle one could define an overall dissolution point by minimizing over all clusters).

The concept is based on the Jaccard similarity for two data subsets C and D, $\gamma(C, D) = \frac{|C \cap D|}{|C \cup D|}$. The dissolution point, as the finite sample addition breakdown point, is based on looking at how many points have to be added to a data set in order to generate a worst case result, here to "dissolve" a cluster. A cluster C is called "dissolved" if in the clustering of the new data set after adding points there is no cluster D for which $\gamma(C, D^*) > \frac{1}{2}$, D^* being the intersection of D and the original data set. $\frac{1}{2}$ is not the worst possible value, but it is the maximum worst possible value over all clusters when comparing two partitions of the same data set, i.e., it is the smallest value which enables every cluster to dissolve.

Hennig (2008) shows dissolution robustness results for various methods, including a positive result for trimmed K-means, and equivalent results to Hennig (2004) for ML methods for mixtures, including a positive result for RIMLE. Results for K-means and PAM for fixed K are negative. There is also an analogous positive result to Hennig (2004) for estimating the number of mixture components by AIC or BIC. There are positive results for single and complete linkage hierarchical clustering for cutting the dendrogram at a fixed height (because this can isolate outliers safely), whereas cutting at a fixed number of clusters is not dissolution robust. All these results require "good clustering"-conditions as before.

Hennig (2007) uses the Jaccard similarity approach for empirical evaluation of robustness and stability of clusters in a given data set based on resampling and addition of artificial data, see also Section **??**.

Further, Hennig (2008) defines a method to be "isolation robust" if one cannot merge clusters bridging an arbitrarily large gap by adding a certain maximum number of points. It is argued that methods with fixed K can never be isolation robust, because if K is misspecified, instability of the partition can be achieved in data sets with arbitrarily large gaps between clusters. Mixture estimation with AIC and BIC and hierarchical clustering (single and complete linkage) cutting at fixed height are isolation robust. Estimating K is not isolation robust in general, though, because it is showed that the average silhouette width method to estimate K (see Chapter **??**) is not isolation robust.

Other assignment-based versions of robustness measurement exist. C. and Haesbroeck (2011) obtained influence functions for the classification error rates of K-means and PAM.

Results as those in Balcan and Gupta (2010) explained in Section 1.2.5 can also be interpreted as robustness results under specific "good clustering" conditions.

1.3 Software for robust cluster analysis

Most of the robust cluster analysis proposals cited in previous sections can be easily implemented with standard statistical software. We will focus on their readily available implementations in R (R Development Core Team (2010)) packages at the CRAN repository.

The $\mathcal{L}_1$ approach can be carried out with the `pam` function included in the `cluster` package. The `tclust` package implements the trimming approaches described in Section 1.2.2. For instance, the graphs shown in Figure 1.3 can be obtained through the use of function `tkmeans(X,k,alpha)`, where `X` is the data set, `k` is the number of clusters and `alpha` is the trimming proportion. The `tclust` function in this package allows to implement the heterogeneous robust cluster analysis procedures discussed in Section 1.2.2. Here, `restr` specifies the type of constraints applied to the scatter matrices and `restr.fact` is the value of the constant c. The parameter `equal.weights` specifies whether equal weights $p_1 = ... = p_k$ are assumed in (1.2) or not. The approach introduced in Gallegos (2002) is obtained by `restr="deter"` and `restr.fact=1`, and that introduced in García-Escudero and Gordaliza (2005) by `restr="sigma"`. The use of `restr = "eigen"` (constraints on the eigenvalues) serves to implement the TCLUST method in García-Escudero et al. (2008).

In order to illustrate the use of the `tclust` package, the well-known "Swiss bank notes" data set in Flury and Riedwyl (1988) will be used. This data set includes 6 measurements made on 100 genuine and 100 counterfeit old Swiss bank notes. The following code is used to obtain Figure 1.4:

```
R > data ("swissbank")
R > clus <- tclust (swissbank, k = 2, alpha = 0.08,
                        restr = "eigen", restr.fact = 15)
```

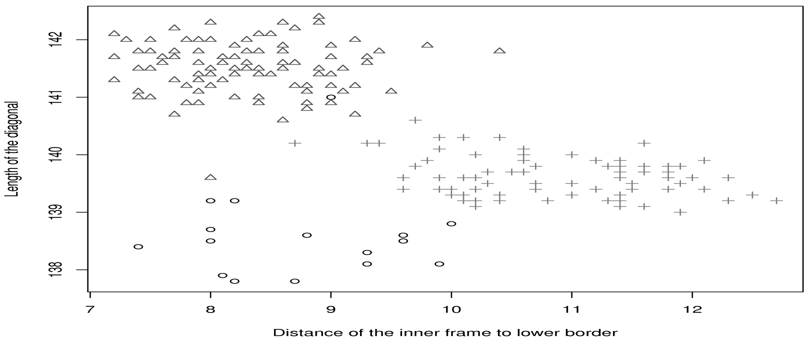

Figure 1.4: Clustering results with $k = 2$, $\alpha = 0.08$ and $c = 15$ for the "Swiss Bank notes" data set.

```
R > plot (swissbank[, 4], swissbank[, 6], pch = clus$cluster + 1)
```

`clus$cluster` returns a vector with the cluster assignments for all the observations in the data set, using the value "0" for the trimmed ones. One of the cluster essentially includes most of the "forged" bills whereas the other one includes most of the "genuine" ones. Within the 8% proportion of trimmed observations ($\alpha = 0.08$), we may identify a subset with 15 forged bills following a clearly different forgery pattern, which has been previously reported in the literature (see, e.g., Flury and Riedwyl (1988)). More details on the `tclust` package are found in Fritz et al. (2012).

The `tlemix` package Neytchev et al. (2012) implements the trimmed likelihood approach to robust cluster analysis, Neykov et al. (2007). The internal use of the `flexmix` package (Leisch (2004)) yields high flexibility that allows `tlemix` to be adapted to a large variety of statistical problems.

The mixture approach with noise component to robust cluster analysis can be carried out in R with the Fraley and Raftery (2012)'s `mclust` package. This package needs an initialization for the observations that are initially considered as noise through `initialization = list(noise=noiseInit)` with `noiseInit` being a `TRUE/FALSE` vector with the same length as the number of observations. The application of the EM algorithm often corrects an improper choice of this vector. One possibility to choose it is through the function `NNclean` in the `prabclus`-package (Hennig and Hausdorf (2012)), which implements the noise detection method proposed in Byers and Raftery (1998).

The fitting of mixtures of t-distributions can be done through the `EMMIX` package. Originally this was written in Fortran. An R version is available on Geoff McLachlan's web page McLachlan (2012).

1.4 Conclusion and open problems

The negative effect of outlying observations on many well-established clustering techniques is widely recognized. Therefore developing cluster analysis methods that are not severely affected by outliers or small deviations from the assumed models is of much interest. We have reviewed different robust cluster analysis approaches proposed in the literature.

There are still many open research problems. Some classical robustness measurements raise some issues when translated into the cluster analysis framework, and new meaningful and computable robustness measures are still of interest. The measures introduced in Hennig (2008) can be applied to far more methods than those featuring already in that paper. Specifically, more robustness analysis of methods involving the estimation of the number of clusters K, of nonparametric density-estimation based methods and of methods not based on probability models is required. More can be done as well regarding the systematic comparison of the various approaches by theory, as far as available, or simulation.

Another important problem that deserves careful attention and further research is the choice of the number of clusters K. This problem, which is generally notoriously difficult, is even harder in robust cluster analysis, because it is related to the task of proper estimation of the level of underlying contamination, required trimming α, respectively (Section 1.2.2). There is an inevitable element of user's choice involved, regarding the question from what size downward a homogeneous group of points should be treated as "grouped outliers" instead of a "cluster". A higher trimming level may "eat" smaller clusters, leading to a smaller number of clusters to be estimated. On the other hand, non-trimmed outliers might result in new clusters when choosing α too low. The user needs to decide whether and from what "borderline size" in the given application very small clusters are meaningful or to be ignored, be it because they are likely to be erroneous "contamination", or be it because small groups are not of interest.

Moreover, the simultaneous determination of the number of groups and the contamination level also depends on the type of clusters we are searching for and on the allowed differences between cluster's scatter matrices. For illustration, consider the data set in Figure 1.3(a). $K = 4$ may be chosen if we allow for very differently scattered clusters (the fourth cluster gathers all the observations far away from the three main clusters), but if clusters are required to be homogeneous, $K = 3$ plus a 10% contamination level makes more sense.

From our point of view, it is neither possible nor desirable to have unsupervised methodology that could simultaneously tell us the number of groups, contamination level and the maximum allowed difference between clusters for a given data set. These aspects interact, and need some amount of user tuning. It is more logical to think of a procedure that is able to return an appropriate cluster partition or a small list of tentative clustering partitions whenever at least one of these three ingredients (number of clusters, contamination level and allowed scatter differences) is available due to some specific knowledge about the problem at hand. The graphical tools described in García-Escudero et al. (2011) (see also Fritz et al. (2012)) have this aim. They rely on proper monitoring of the so-called "trimmed classification likelihoods" obtained from the maximum values achieved by (1.2) when varying K, α and constant c. The consideration of weights p_k in (1.2) is recommended for choosing the number of clusters as it was explained in Bryant (1991). García-Escudero et al. (2003) was a first attempt for this monitoring approach, considering trimmed K-means.

Ritter (2015) (who also provides a lot on asymptotic theory for methods based on trimming) suggests exploring many local optima of the objective function in connection with using various different covariance constraints. The so-called “Pareto solutions” are clusterings that appear optimal at least for a certain covariance constraint.

The monitoring of trimmed likelihoods with an over-fitting penalty term (“trimmed BIC”) was considered in Neykov et al. (2007). Baudry et al. (2010) monitor various numbers of clusters as well in their methodology to merge normal mixture components in order to fit mixtures with non-normal clusters by normal mixtures. Another monitoring technique is the “forward search” (see Section 1.2.2). All these authors agree that a dynamic sequence of images of the data (for instance, resulting from a sequence of different trimming levels) provides a more useful view of the data set than a single “static” view.

When adopting the mixture fitting approach to robust cluster analysis, the use of BIC-type criteria provides guidance for determining the number of clusters K and the contamination level p_0 in (1.4). The allowed differences in within-cluster scatter are controlled through the different parameterizations of the scatter matrices of the mixture components. However, it is important that the proper choice of the number of clusters in cluster analysis is not exactly equivalent to that of proper choice of the number of components in a mixture problem (see, e.g., Celeux and Soromenho (1996),Hennig (2010)). Hennig (2004) and Hennig and Coretto (2008) provide sensible ways to choose constant c for the RIMLE method. One approach is to find a c or contamination level that makes the non-noise part of the mixture as similar as possible to a normal mixture with the parameters estimated for the components in terms of the Kolmogorov- or another distance between distributions.

In the framework of mixture models, exploring the robustness of further types of mixtures of non-normal (e.g., skew) distributions (and robustifying them where necessary) is of current interest.

All the proposals discussed in this review concern the problem of clustering around centroids. There are interesting problems where observations are naturally clustered around linear and nonlinear manifolds, see also Section **??**. García-Escudero et al. (2009) show how the trimming principles presented in Section 1.2.2 can be adapted to robust clustering around linear subspaces through RLGA (robust linear grouping analysis). This method is available in the `lga` R-package (Harrington (2012)). Robust clusterwise regression techniques were introduced in Hennig (2002), Hennig (2003) and García-Escudero et al. (2010). Detecting specific shapes and objects in noisy images is a problem addressed by the pattern recognition community with procedures that often can be adapted to solve specific robust cluster analysis problems, e.g., Luan et al. (1998). Further robustification problems arise from clustering functional data (see García-Escudero and Gordaliza (2005) and Cuevas et al. (2007), Section **??**).

Acknowledgement: This research was partially supported by the Spanish Ministerio de Economía y Competitividad, grant MTM2014-56235-C2-1-P, and by Consejería de Educación de la Junta de Castilla y León, grant VA212U13.

References

Allard, D. and C. Fraley (1997). Nonparametric maximum likelihood estimation of features in spatial point processes using Voronoi tessellation. *Journal of the American Statistical Association 92*, 1485–1493.

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

Atkinson, A., M. Riani, and A. Cerioli (2006). Random start forward searches with envelopes for detecting clusters in multivariate data. In M. R. S. Zani, A. Cerioli and M. Vichi (Eds.), *Data Analysis, Classification and the Forward Search*, pp. 163–172.

Balcan, M. and P. Gupta (2010). Robust hierarchical clustering. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT), 2010*, pp. 282–294.

Banerjee, A. and R. N. Davé (2012). Robust clustering. *WIREs Data Mining and Knowledge Discovery 2*(1), 29–59.

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

Barni, M., V. Cappellini, and A. Mecocci (1996, aug.). Comments on "a possibilistic approach to clustering". *IEEE Transactions on Fuzzy Systems 4*(3), 393 –396.

Baudry, J.-P., A. Raftery, G. Celeux, K. Lo, and R. Gottardo (2010). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics 19*, 332–353.

Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, New York.

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

Byers, S. and A. Raftery (1998). Nearest neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association 93*, 577–584.

C., R. and G. Haesbroeck (2011). Impact of contamination on training and test error rates in statistical clustering analysis. *Communications on Statistics: Simulations and Computing 40*, 394–411.

Campbell, J., C. Fraley, F. Murtagh, and A. Raftery (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters 18*, 1539–1548.

Campbell, J., C. Fraley, D. Stanford, F. Murtagh, and A. Raftery (1999). Model-based methods for textile fault detection. *International Journal of Imaging Systems and Technology 10*, 339–346.

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

Celeux, G. and G. Soromenho (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification 13*, 195–212.

Chatzis, S. and T. Varvarigou (2008). Robust fuzzy clustering using mixtures of student's-t distributions. *Pattern Recognition Letters 29*, 1901–1905.

Coretto, P. and C. Hennig (2010). A simulation study to compare robust clustering methods based on mixtures. *Advances in Data Analysis and Classification 4*, 111–135.

Coretto, P. and C. Hennig (2011). Maximum likelihood estimation of heterogeneous mixtures of gaussian and uniform distributions. *Journal of Statistical Planning and Inference 141*, 462–473.

Cuesta-Albertos, J., L. García-Escudero, and A. Gordaliza (2002). On the asymptotics of trimmed best k-nets. *Journal of Multivariate Analysis 82*, 482–516.

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

Cuesta-Albertos, J., A. Gordaliza, and C. Matrán (1998). Trimmed best k-nets. a robustifyed version of a $\mathcal{L}_\infty$-based clustering method. *Statistics and Probability Letters 36*, 401–413.

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

Cuevas, A., M. Febrero, and R. Fraiman (2001). Cluster analysis: A further approach based on density estimation. *Computational Statistics and Data Analysis 36*, 441–459.

Cuevas, A., M. Febrero, and R. Fraiman (2007). Impartial trimmed k-means for functional data. *Computational Statistics and Data Analysis 51*, 4864–4877.

Dasgupta, A. and A. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association 93*, 294–302.

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

Davé, R. and R. Krishnapuram (1997). Robust clustering methods: A unified view. *IEEE Transactions on Fuzzy Systems 5*, 270–293.

Davé, R. and S. Sen (1997). Noise clustering algorithm revisited. In *Proceedings of the Biennial Workshop NAFIPS 1997*.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Ding, Y., X. Dang, H. Peng, and D. Wilkins (2007). Robust clustering in high dimensional data using statistical depths. *BMC Bioinformatics 8(Suppl 7):S8*.

Donoho, D. and P. Huber (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pp. 157–184. Wadsworth.

Dunn, J. (1974). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics 3*, 32–57.

Flury, B. and H. Riedwyl (1988). *Multivariate Statistics. A Practical Approach.* London: Chapman and Hall.

Forgy, E. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics 21*, 768–780.

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

Fraley, C. and A. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*, 611–631.

Fraley, C. and A. Raftery (2012). *mclust: Model-Based Clustering / Normal Mixture Modeling.* R package version 3.4.11.

Friedman, H. and J. Rubin (1967). On some invariant criterion for grouping data. *Journal of the American Statistical Association 63*, 1159–1178.

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

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

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

Gallegos, M. (2002). Maximum likelihood clustering with outliers. In K. Jajuga, A. Sokolowski, and H. Bock (Eds.), *Classification, Clustering and Data Analysis: Recent advances and applications*, pp. 247–255. Springer-Verlag.

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

Gallegos, M. and G. Ritter (2009a). Trimmed ML estimation of contaminated mixtures. *Sankhya (Series A) 71*, 164–220.

Gallegos, M. and G. Ritter (2009b). Trimming algorithms for clustering contaminated grouped data and their robustness. *Advances in Data Analysis and Classification 10*, 135–167.

Gallegos, M. and G. Ritter (2010). Using combinatorial optimization in model-based trimmed clustering with cardinality constraints. *Computational Statistics and Data Analysis 54*, 637–654.

García-Escudero, L., Gordaliza, A., and C. Matrán (1999). A central limit theorem for multivariate generalized trimmed k-means. *Annals of Statistics 27*, 1061–1079.

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

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

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

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

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

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

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

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

Gersho, A. and R. Gray (1991). *Vector Quantization and Signal Compression.* Springer, New York.

Gordaliza, A. (1991). Best approximations to random variables based on trimming procedures. *Journal of Approximation Theory 64*, 162–180.

Greselin, F. and S. Ingrassia (2010). Constrained monotone EM algorithms for mixtures of multivariate t distributions. *Statistics and Computing 20*, 9–22.

Gustafson, E. and W. Kessel (1979). Fuzzy clustering with a fuzzy covariance matrix. In *Proceedings of the IEEE lnternational Conference on Fuzzy Systems, San Diego, 1979*, pp. 761–766.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust statistics. The approach based on influence functions.* New York: Wiley series in probability and mathematical statistics: probability and mathematical statistics.

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

Harrington, J. (2012). *lga: Tools for linear grouping analysis.* R package.

Hathaway, R. (1985). A constrained formulation of maximum likelihood estimation for normal mixture distributions. *Annals of Statistics 13*, 795–800.

Hathaway, R. and J. Bezdek (1993). Switching regression models and fuzzy clustering. *IEEE Transactions on Fuzzy Systems 1*, 195–204.

Hennig, C. (2002). Fixed point clusters for linear regression: computation and comparison. *Journal of Classification 19*, 249–276.

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

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

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis 52*, 258–271.

Hennig, C. (2008). Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods. *Journal of Multivariate Analysis 99*, 1154–1176.

Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification 4*, 3–34.

Hennig, C. and P. Coretto (2008). The noise component in model-based cluster analysis. In L. S.-T. C. Preisach, H. Burkhardt and R. Decker (Eds.), *Data Analysis, Machine Learning and Applications, Studies in Classification, Data Analysis, and, Knowledge Organization,*, pp. 127–138. Springer, Berlin-Heidelberg.

Hennig, C. and B. Hausdorf (2012). *prabclus: Functions for clustering of presence-absence, abundance and multilocus genetic data.* R package version 2.2-4.

Huber, P. (1981). *Robust statistics.* New York: Wiley series in probability and mathematical statistics: probability and mathematical statistics.

Jaing, M., S. Tseng, and C. Su (2001). Two-phase clustering process for outlier detection. *Pattern Recognition Letters 22*, 691Â–700.

Jolion, J.-M., P. Meer, and S. Bataouche (1991, aug). Robust clustering with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*(8), 791 –802.

Kim, J., R. Krishnapurama, and R. Davé (1996). Application of the least trimmed squares technique to prototype-based clustering. *Pattern Recognition Letters 17*, 633–641.

Klawonn, F. (2004). Noise clustering with a fixed fraction of noise. In A. Lotfi and J. Garibaldi (Eds.), *Applications and Science in Soft Computing.* Springer, Berlin.

Krishnapuram, R., A. Joshi, and L. Yi (1999). A fuzzy relative of the *k*-medoids algorithm with application to web document and snippet clustering. In *Snippet Clustering, in Proc. IEEE Intl. Conf. Fuzzy Systems - FUZZIEEE99, Korea.*

Krishnapuram, R. and J. Keller (1993, may). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems 1*(2), 98 –110.

Kumar, M. and J. Orlin (2008). Scale-invariant clustering with minimum volume ellipsoids. *Computers and Operations Research 35*, 1017–1029.

Leisch, F. (2004). Flexmix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software 11*, 1–18.

Łeski, J. (2003). Towards a robust fuzzy clustering. *Fuzzy Sets and Systems 137*(2), 215 – 233.

Lin, C.-R. and M.-S. Chen (2005). Combining partitional and hierarchical algorithms for robust and efficient data clustering with cohesion self-merging. *IEEE Transactions on Knowledge and Data Engineering 17*, 145–149.

Luan, J., J. Stander, and D. Wright (1998). On shape detection in noisy images with particular reference to ultrasonography. *Statistics and Computing 8*, 377–389.

Markatou, M. (2000). Mixture models, robustness, and the weighted likelihood methodology. *Biometrics 356*, 483–486.

Maronna, R. and P. Jacovkis (1974). Multivariate clustering procedures with variable metrics. *Biometrics 30*, 499–505.

McLachlan, G. (2012). *R Version of EMMIX.* Brisbane, Australia.

McLachlan, G. and D. Peel (2000). *Finite mixture models.* New York: Wiley Series in Probability and Statistics.

Müller, D. and G. Sawitzki (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association 86*, 738–746.

Neykov, N., P. Filzmoser, R. Dimova, and P. Neytchev (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics and Data Analysis 52*, 299–308.

Neytchev, P., P. Filzmoser, R. Patnaik, A. Eisl, and R. Boubela (2012). *tlemix: Trimmed Maximum Likelihood Estimation.* R package version 0.1.

R Development Core Team (2010). *R: A Language and Environment for Statistical Com-*

puting. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rehm, F., F. Klawonn, and R. Kruse (2007). A novel approach to noise clustering for outlier detection. *Soft Computing 11*, 489–494.

Ritter, G. (2015). *Robust Cluster Analysis and Variable Selection.* Boca Raton, FL: CRC Press.

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

Rocke, D. and D. Woodruff (2012). Computational connections between robust multivariate analysis and clustering. In W. Härdle and B. Rönz (Eds.), *COMPSTAT 2002 Proceedings in Computational Statistics*, pp. 255–260. Heidelberg:Physica-Verlag.

Rousseeuw, P., L. Kaufman, and E. Trauwaert (1996). Fuzzy clustering using scatter matrices. *Computational Statistics and Data Analysis 23*, 135–151.

Rousseeuw, P. and A. Leroy (1987). *Robust Regression and Outlier Detection.* Wiley-Interscience, New York.

Rousseeuw, P. and K. Van Driessen (1999). A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics 41*, 212–223.

Ruspini, E. (1969). A new approach to clustering. *Information and Control 15*, 22–32.

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

Ruwet, C., L. García-Escudero, A. Gordaliza, and A. Mayo-Iscar (2013). On the breakdown behavior of robust constrained clustering procedures. *TEST 22*, 466–487.

Santos-Pereira, C. and A. Pires (2002). Detection of outliers in multivariate data: a method based on clustering and robust estimators. In W. Härdle and B. Rönz (Eds.), *Proceedings in Computational Statistics: 15th Symposium Held in Berlin, Germany.*, pp. 291–296. Physica-Verlag, Heidelberg.

Schynsa, M., G. Haesbroeck, and F. Critchley (2010). RelaxMCD: Smooth optimisation for the Minimum Covariance Determinant estimator. *Computational Statistics and Data Analysis 54*, 843–857.

Shoham, S. (2002). Robust clustering by deterministic agglomeration EM of mixtures of multivariate t distributions. *Pattern Recognition 55*, 1127–1142.

Tantrum, J., A. Murua, and W. Stuetzle (2003). Assessment and pruning of hierarchical model based clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 197–205. ACM, New York.

Trauwaert, E., L. Kaufman, and P. Rousseeuw (1991). Fuzzy clustering algorithms based on the maximum likelihood principle. *Fuzzy Sets and Systems 42*, 213–227.

von Luxburg, U. (2010). Clustering stability: An overview. *Foundations and Trends in Machine Learning 2*, 235–274.

Willems, G., H. Joe, and R. Zamar (2009). Diagnosing multivariate outliers detected by robust estimators. *Journal of Computational and Graphical Statistics 18*, 73–91.

Woodruff, D. and T. Reiners (2004). Experiments with, and on, algorithms for maximum likelihood clustering. *Computational Statistics and Data Analysis 47*, 237–253.

Wu, K.-L. and M.-S. Yang (2002). Alternative c-means clustering algorithms. *Pattern Recognition 35*(10), 2267 – 2278.