Dissolution point and isolation robustness: robustness criteria for general cluster analysis methods

Christian Hennig

Department of Statistical Science, UCL,
Gower St.,
London, WC1e 6BT,
United Kingdom,
chrish@stats.ucl.ac.uk

September 22, 2005

Abstract

Two robustness criteria are presented that are applicable to general clustering methods. Robustness and stability in cluster analysis are not only data dependent, but even cluster dependent. Robustness is in the present paper defined as a property of not only the clustering method, but also of every individual cluster in a data set. The main principles are: (a) dissimilarity measurement of an original cluster with the most similar cluster in the induced clustering, (b) the dissolution point, which is an adaptation of the breakdown point concept to single clusters, (c) isolation robustness: given a clustering method, is it possible to join, by addition of g points, arbitrarily well separated clusters?

Results are derived about k-means, k-medoids (k estimated by average silhouette width), trimmed k-means, mixture models (with and without noise component, with and without estimation of the number of clusters by BIC), single and complete linkage.
AMS 2000 subject classification: Primary 62F35; secondary 62H30.
Keywords: breakdown point, model-based cluster analysis, mixture model, trimmed k-means, average silhouette width

1 INTRODUCTION

Stability and robustness are important issues in cluster analysis. As a motivation, Figure 1 shows the 7-means clustering of a 4-dimensional data set of 80 images that

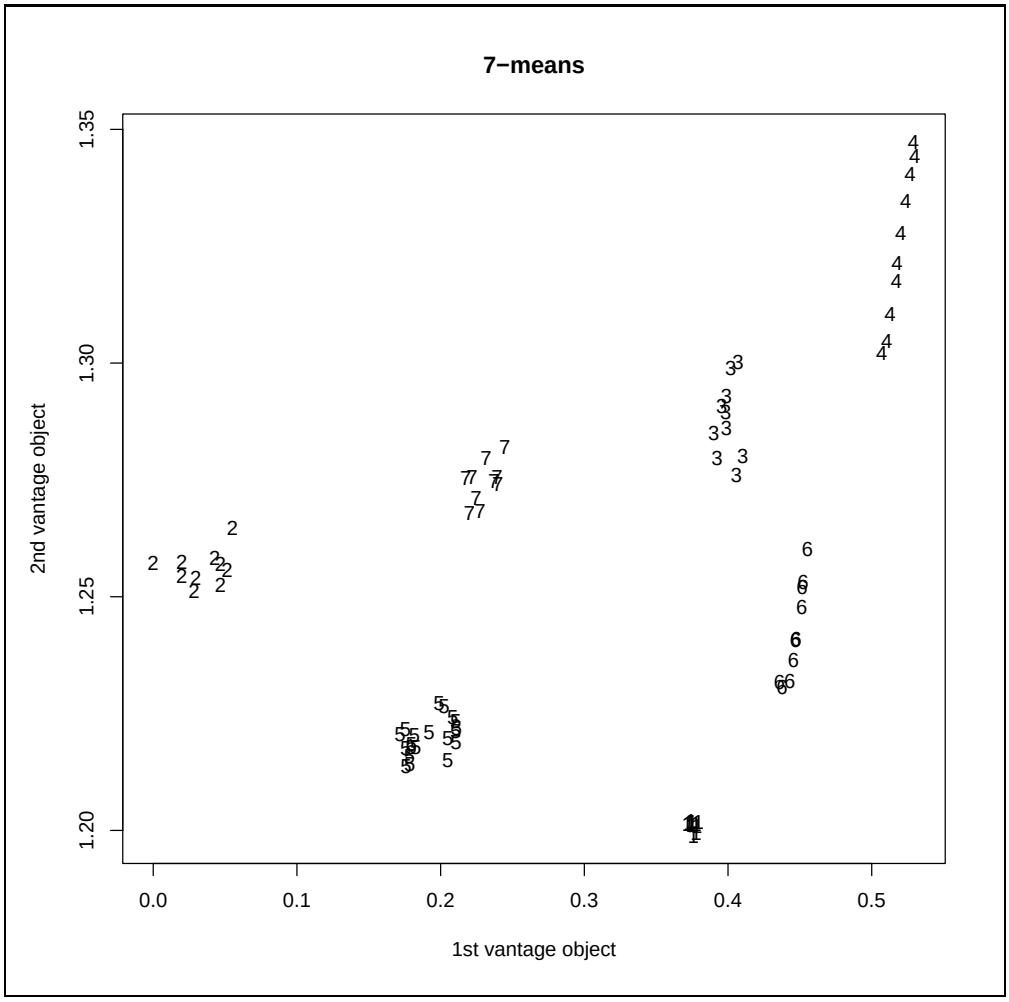

Figure 1: First two variables of 80-images data set with 7-means clustering.

are screen captures of movies (only the first two variables are shown in all figures). The data set has been obtained by first defining a visual distance measure between the images. Then the data have been embedded in the 4-dimensional space by choosing four so-called "vantage objects" and taking their distances to all objects as variables (see Hennig and Latecki 2003 for the full procedure). The images are from 8 different scenes, and therefore there is a "true clustering" (cluster 5 in Figure 1 consists of the images of two scenes). Originally, the data set consisted of 100 images from 10 scenes. The images from the two omitted scenes are very different from the other images. Four of them are included in Figure 2, and they are used to illustrate the effect of "realistic" outliers on a clustering. The clustering on the right side is further discussed in Section 3.4.

If only one of the outliers shown as "cluster 2" in Figure 2 is added to the data set on the left side, the 7-means solution reserves one cluster for the outlier and merges

the well separated clusters 5 and 7 on the left side. This could be interpreted as a kind of breakdown or "dissolution" of at least cluster 7 (cluster 5 consists of 20 points and still has the majority in the merged cluster).

The 8-means solution on the 80-images data splits cluster 4 into two parts instead of separating the two scenes underlying cluster 5. With additional outlier, 8-means generates the clustering of Figure 1 plus one cluster for the outlier, which seems to be an adequate clustering. Here, the splitting of cluster 4 into two halves is unstable. The addition of suitable non-outliers to the center of cluster 4 instead of the outlier added above would result in splitting up cluster 5 instead. As opposed to the situation above, it seems to be inadequate to judge this latter instability as a serious robustness problem of the clustering method, because from looking at the data alone it is rather unclear if a good clustering method should split up cluster 4, cluster 5, both, or none of them.

This illustrates that not all instabilities in cluster analysis are due to weaknesses of the clustering methods. There also exist data constellations that are unstable with respect to clustering. Some features of a clustering are expected to be more stable than others (the separation between clusters 5 and 7 is clearly more meaningful than the question if cluster 4 should be split up or not). The approach of the present paper to handle such feature-dependent instabilities is the introduction of a cluster-dependent concept of robustness. Here is an even clearer example: be there 100 one-dimensional points distributed more or less uniformly between 0 and 1, 15 points between 10 and 10.4 and 15 points between 10.6 and 11. It should be clear that a reasonably robust clustering method (estimating k, say) should assign the first 100 points to a single stable cluster, while it may depend on small variations in the data whether the remaining points are estimated as a single cluster or split up into two clusters, and no sensible clustering method can be expected to be stable in that respect.

The 80-images data set illustrates further that a proper estimation of the number of clusters, which adds clusters fitting extreme outliers, could be a key to robustness against outliers. However, not every method to estimate the number of clusters is suitable for this purpose, see Section 3.4.

The assessment of the effect of a perturbation of the data to a clustering has a long history in cluster analysis (Rand 1971, a large number of references is given in Gordon 1999, Chapter 7, and Milligan 1996). Recently, there are also attempts to apply key concepts of robust statistics such as the influence function (Hampel 1971) and the breakdown point (Hampel 1974, Donoho and Huber 1983) to certain cluster analysis methods (Kharin 1996, Garcia-Escudero and Gordaliza 1999, Gallegos 2003, Hennig 2004a). The disadvantage of the latter influence/breakdown approach is that it only applies to cluster analysis methods estimating parameters of statistical models, and the results are only comparable between methods estimating the same parameters. While being able to handle and to compare more general cluster analysis techniques, the disadvantage of the former cluster pertur-

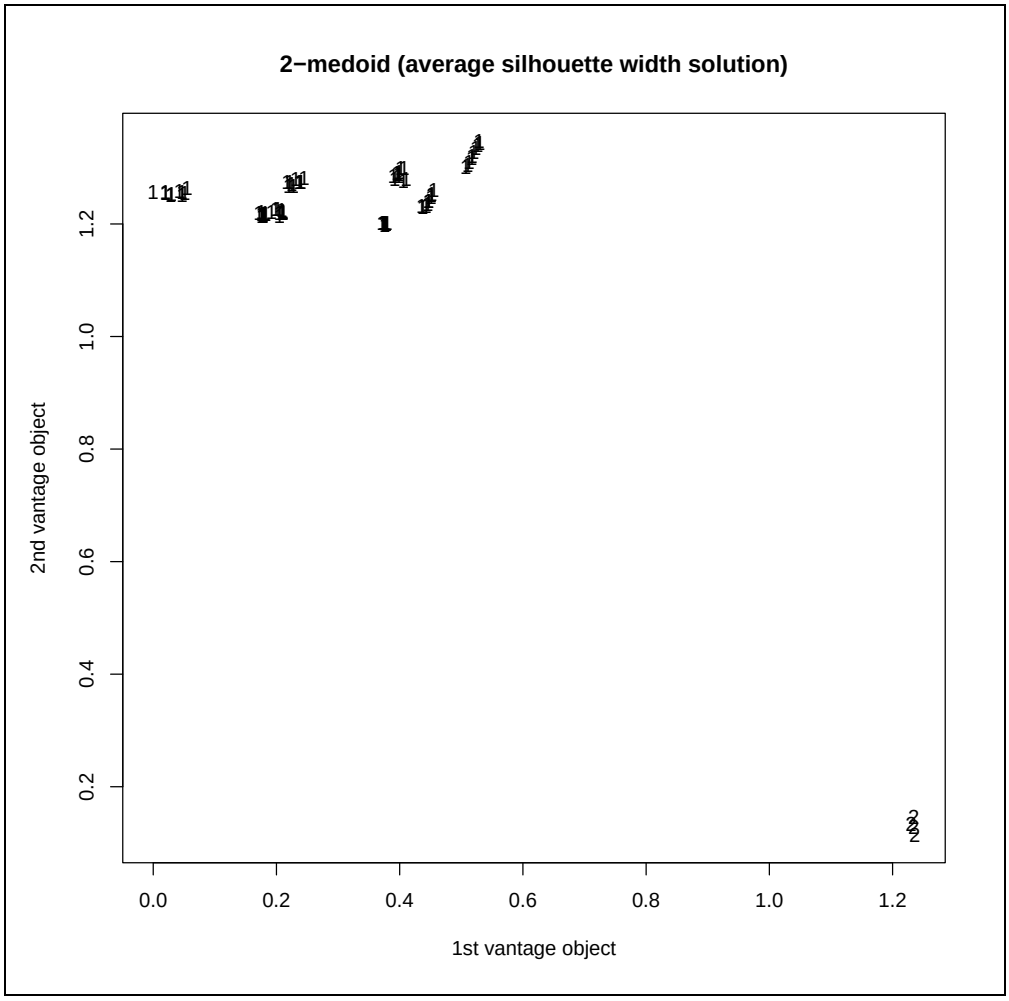

Figure 2: Same data as in Figure 1 with 4 outlying images added and average silhouette width (2-medoid) clustering

bation approach is that it consists mainly of simulation studies, and the results depend strongly on the design of these studies.

The aim of the present paper is to develop robustness concepts for cluster analysis that can be applied to a wide range of cluster analysis methods. Considerations are restricted to methods yielding disjunct clusters, but the proposed methodology can also be applied to more general cluster analysis methods (Hennig 2004c).

An important difference between the theory given here and the results published so far is that the present approach treats stability as a property of an individual cluster instead of the whole clustering. It is intuitively clear and has been demonstrated above that a single data set can contain at the same time stable and much less stable (in most cases this means: less clearly separated) clusters.

The following two concepts are introduced in Section 2:

- The "dissolution point" is an adaptation of the breakdown point concept to all individual clusters yielded by general cluster analysis methods.
- "Isolation robustness" means that a theorem of the following type can be shown: For g arbitrarily large but fixed, a cluster with a large enough isolation (minimum distance between a point inside and a point outside the cluster, depending on g) cannot be merged with points not belonging to the cluster in the original data set by addition of g points to the data set.

The concepts are applied to various cluster analysis methods, namely k-means, k-medoids with estimation of k by average silhouette width (Kaufman and Rousseeuw 1990, Chapter 2), trimmed k-means (Cuesta-Albertos, Gordaliza and Matran 1997; all in Section 3), mixture models with and without noise and with and without estimation of the number of clusters (Fraley and Raftery 1998, McLachlan and Peel 2000; Section 4), single and complete linkage agglomerative clustering (Section 5), fixed point clustering (Hennig 1997, 2002, 2005a, Hennig and Christlieb 2002; Section 6). The paper is concluded with an overview of the robustness results and some discussion in Section 7.

2 ROBUSTNESS CONCEPTS

2.1 The dissolution point and a dissimilarity measure between clusters

In Hennig (2004a), a definition of a breakdown point for a general clustering method has been proposed (though applied only to ML-estimators for location-scale mixtures), of which the definition is based on the assignments of the points to clusters and not on parameters to be estimated. This concept deviates somewhat from the traditional meaning of the term "breakdown point", since it attributes

"breakdown" to situations that are not always the worst possible ones. Furthermore, the definition is not linked to an equivariance property and it is not possible to derive a non-trivial upper bound for this definition, which may be taken as a requirement for a breakdown point definition, cf. Davies and Gather (2002). Therefore, the proposed robustness measure is called "dissolution point". It is thought to measure a kind of "breakdown" in the sense that the addition of points changes the cluster solution so strongly that the pattern of the original data can be considered as "dissolved". The definition here is a modification of that given in Hennig (2004a).

A sequence of mappings $E = (E_n)_{n\in I\!N}$ is called a general clustering method, if E_n maps a set of entities $\mathbf{x}_n = \{x_1, \ldots, x_n\}$ (this is how $\mathbf{x}_n$ is always defined throughout the paper) to a collection of subsets $\{C_1, \ldots, C_s\}$ of $\mathbf{x}_n$. Note that it is assumed that entities with different indexes can be distinguished. This means that the elements of $\mathbf{x}_n$ are interpreted as data points and that $|\mathbf{x}_n| = n$ even if, for example, for $i \neq j$, $x_i = x_j$. This could formally be achieved by writing (x_i, i) and (x_j, j) instead, but for simplicity reasons such a notation has not been chosen. Assume for the remainder of the paper that E is a disjunct cluster method (DCM), i.e., $C_i \cap C_j = \emptyset$ for $i \neq j \leq k$. Most popular DCMs yield partitions, i.e., $\bigcup_{j=1}^{k} C_j = \mathbf{x}_n$.

If E is a DCM and $\mathbf{x}_{n+g}$ is generated by adding g points to $\mathbf{x}_n$, $E_{n+g}(\mathbf{x}_{n+g})$ induces a clustering on $\mathbf{x}_n$, which is denoted by $E_n^*(\mathbf{x}_{n+g})$. Its clusters are denoted by $C_1^*, \ldots, C_{k^*}^*$. $E_n^*(\mathbf{x}_{n+g})$ is a DCM as well. k^* may be smaller than k if E produces k clusters for all n.

The definition of stability with respect to the individual clusters requires a measure for the similarity between a cluster of $E_n^*(\mathbf{x}_{n+g})$ and a cluster of $E_n(\mathbf{x}_n)$, i.e., between two subsets C and D of some finite set.

There are a lot of possible similarity measures. Such measures are used, e.g., in ecology to measure similarity of species populations of regions (Shi 1993). The Jaccard coefficient (Jaccard 1901) is presumably the most popular measure, and I suggest it for the purpose of the present paper (see Remark 2.4):

$$\gamma(C, D) = \frac{|C \cap D|}{|C \cup D|}.$$

The definition of dissolution is based on the similarity of a cluster $C \in E_n(\mathbf{x}_n)$ to its most similar cluster in $E_n^*(\mathbf{x}_{n+g})$. A similarity between C and a clustering $\hat{E}_n(\mathbf{x}_n)$ is defined by

$$\gamma^*(C, \hat{E}_n(\mathbf{x}_n)) = \max_{D\in\hat{E}_n(\mathbf{x}_n)} \gamma(C, D).$$

How small should γ^* be to say that the pattern corresponding to C in the original data is dissolved in $E_n^*(\mathbf{x}_n)$? The usual choice for a breakdown point in robust

statistics would be the worst possible value. In the present setup, this value depends on the dataset and on the clustering method. The key problem is that in a partition $E_n^*(\mathbf{x}_{n+g})$) there has to be at least one cluster that intersects with C, so that the natural minimum value 0 of γ cannot be attained. See Hennig (2004a) for examples of data dependent worst values. In general, the worst possible value may be difficult to compute, while one would judge a cluster as "broken down" or "dissolved" already in much simpler constellations of $E_n^*(\mathbf{x}_{n+g})$. I propose

$$\gamma^* \leq \frac{1}{2} = \gamma(\{x,y\},\{x\}) = \gamma(C,C_1) \text{ if } C_1 \subset C,\ |C_1| = |C|/2, \tag{2.1}$$

as a cutoff value to consider a cluster as dissolved. The definition of the Jaccard coefficient enables a simple interpretation: If $\gamma^*(C, E_n^*(\mathbf{x}_{n+g})) \leq \frac{1}{2}$, then the number of points of C and its most similar cluster in $E_n^*(\mathbf{x}_{n+g})$) for which the two clusters differ is at least as large as the number of points where they coincide.

The cutoff value $\frac{1}{2}$ can be further motivated by the following Lemma, which means that every cluster can dissolve, at least in absence of further subtle restrictions on the possible clusterings.

Lemma 2.1 *Let $E_n(\mathbf{x}_n) \ni C$ be a DCM with $|E_n(\mathbf{x}_n)| \geq 2$. Let $\mathcal{K} \subseteq I\!N$ be the set of possible cluster numbers containing at least one element $k \geq 2$. Let $\mathcal{F} = \{F \text{ partition on } \mathbf{x}_n :\ |F| \in \mathcal{K}\}$. Then $\exists \hat{F} \in \mathcal{F} :\ \gamma^*(C,\hat{F}) \leq \frac{1}{2}$. $\frac{1}{2}$ is the smallest value for this to hold.*

This is equivalent to Lemma 3.3 in Hennig (2004a).

Note that $\mathcal{F}$ is restricted here to consist of partitions, not of disjunct clusterings. The reason for this is that the claim of the Lemma would be trivial if the new clustering $\hat{F}$ would be allowed to consist of no clusters at all or to assign only very few points to clusters. The Lemma shows that dissolution is possible by new assignments of points to clusters, not only by not clustering points.

Definition 2.2 *Let $E = (E_n)_{n\in I\!N}$ be a DCM. The* **dissolution point** *of a cluster $C \in E_n(\mathbf{x}_n)$ is defined as*

$$\Delta(E,\mathbf{x}_n,C) = \min_g \left\{ \frac{g}{|C|+g} :\ \exists \mathbf{x}_{n+g} = (x_1,\ldots,x_{n+g}) :\ \gamma^*(C, E_n^*(\mathbf{x}_{n+g})) \leq \frac{1}{2} \right\}.$$

The dissolution point is defined by addition of points to the original data set here, which is not the only possibility. See Section 7 for a discussion.

Note that it would be mathematically equivalent with respect to all theory presented in this paper to define the dissolution point as the minimal g instead of $\frac{g}{|C|+g}$. I suggest $\frac{g}{|C|+g}$ because this enables comparisons between the dissolution points of different clusters and the choice of a proportion between 0 and 1 follows the tradition of the breakdown point (though there is no proof of the dissolution point to be bounded from above by $\frac{1}{2}$ under some reasonable assumptions).

Remark 2.3 *It follows from Remark 3.5 in Hennig (2004a), that at least $r \geq 1$ clusters of $E_n(\mathbf{x}_n)$ have to dissolve if $|E_n(\mathbf{x}_n)| = k,\ |E_n^*(\mathbf{x}_{n+g})| = k - r$.*

Remark 2.4 *In Shi (1993), 39 similarity measures between sets are compared. In Hennig (2004a), $\gamma_1(C, D) = \frac{2|C \cap D|}{|C|+|D|}$ has been used, which is a monotone function of the Jaccard coefficient and leads to an equivalent dissolution definition if the cutoff value $\frac{1}{2}$ is replaced by $\frac{2}{3}$. The interpretation of (2.1) seems to be most natural for the Jaccard coefficient and the cutoff value of $\frac{1}{2}$, and the Jaccard coefficient is well known and widely used (though usually for much different purposes).*

It does not depend on the number of points which are neither in C nor in D, it is symmetric and attains its minimum 0 only for disjoint sets and its maximum 1 only for equal sets. $1 - \gamma$ is a metric (Gower and Legendre 1986). Many of the measures listed in Shi (1993) do not fulfill these basic requirements, others are criticized by Shi for stability reasons. See Hennig (2004c) for a further discussion of the choice of the Jaccard coefficient.

The comparison of whole clusterings has been treated, e.g., by Rand (1971), Hubert and Arabie (1985).

2.2 Isolation robustness

In the following sections, there will be various results on dissolution points for different DCMs. While these results are informative about the nature of the methods, in most cases they do not allow a direct comparison. The concept of isolation robustness should enable such a comparison. The rough idea is that it can be seen as a minimum robustness demand on cluster analysis that an extremely well isolated cluster remains stable under the addition of points. The isolation $i(C)$ of a cluster C is defined as the minimum distance of a point of the cluster to a point not belonging to the cluster, which means that a distance structure on the data is needed. The DCMs treated in this paper, as far as they are not directly distance based, operate on the Euclidean space, so that the Euclidean distance can be used. It is further assumed that the distance measure is a metric because the idea of "isolation" is incompatible with the possibility that there may be a distance of 100 between two points and a third point can be added that has a distance of 1 to both of them.

Definition 2.5 *A DCM $E = (E_n)_{n \in \mathbb{N}}$ is called* **isolation robust***, if there exists a sequence of functions $v_m :\ \mathcal{M}_m \times \mathbb{N} \mapsto \mathbb{R},\ m \in \mathbb{N}$ (where $\mathcal{M}_m$ is the space of distance matrices between m objects permissible by the distance structure underlying the DCM) so that*
for $n \geq m$ for any data set $\mathbf{x}_n$, for given $g \in \mathbb{N}$, for any cluster $C \in E_n(\mathbf{x})$ with $|C| = m$, within-cluster distance matrix M_C and $i(C) > v_m(M_C, g)$ and for any data set $\mathbf{x}_{n+g}$, where g points are added to $\mathbf{x}_n$ the following statement holds:
For all $D \in E_n^*(\mathbf{x}_{n+g}) :\ D \subseteq C$ or $D \subseteq \mathbf{x}_n \setminus C$ and $\exists E_n^*(\mathbf{x}_{n+g}) \ni D \subseteq C$.

Note that a well isolated cluster may be unstable not because of robustness problems with the DCM, but because of internal inhomogeneity. Isolation robustness addresses only robustness of a good separation, not robustness of a large homogeneity. Under a sensible DCM, it is always possible to construct data in which a rather inhomogeneous cluster is split up in more than one part under addition of a single point. The definition allows C to be split up and prevents only that parts of C are joined with parts of $\mathbf{x}_n \setminus C$ in the same cluster.

Remark 2.6 *It would be possible to define a weaker version of isolation robustness "of degree α" by demanding the existence of $v_m(M_C, g)$ only for $g < \alpha m$. With such a definition, it would not be necessary that for a large enough isolation the definition above holds for arbitrarily large g, which may be even larger than n. However, the following theory will show that isolation robustness is either violated already for $g = 1$ or it holds for arbitrarily large g, thus $\alpha = \infty$, for any of the discussed methods.*

3 VARIATIONS ON k-MEANS

3.1 Definition of methods

In the following subsection, dissolution and isolation robustness of some versions of the k-means clustering method (MacQueen 1967) will be investigated. These versions have been proposed to robustify the k-means approach.

- The k-medoids method (Kaufman and Rousseeuw 1990, Chapter 2), which uses (in its default form) the L_1-norm instead of the squared L_2-norm and uses optimally chosen cluster members instead of means as cluster centers. Thus, it can also be applied to data that come as distance matrix (the distances being not necessarily L_1-norms) and is a modification of k-medians.

- The trimmed k-means method (Cuesta-Albertos, Gordaliza and Matran 1997) optimizes the k means criterion after an optimally chosen portion of α of the data has been left out.

- The number of clusters k is often treated as fixed. It is also possible to estimate this number. Many criteria have been proposed to do this (see, e.g., Milligan and Cooper 1985). In the present paper, the "average silhouette width" criterion proposed for k-medoids (but applicable to all partitioning techniques) by Kaufman and Rousseeuw (1990, Chapter 2) is considered for the k-medoids case. This criterion recently became very popular, see, e.g., Jörnsten (2004).

Definition 3.1 *The* k**-means clustering** *of* $\mathbf{x}_n$ *is defined by*

$$E_n(\mathbf{x}_n) = \underset{\{C_1,\ldots,C_k\} \text{ partition of } \mathbf{x}_n}{\arg\min} \sum_{i=1}^{n} \min_j \|x_i - \bar{x}_j\|_2^2, \tag{3.1}$$

where $\bar{x}_j = \frac{1}{|C_j|}\sum_{x_i \in C_j} x_i$ *and* $\|\bullet\|_p$ *denotes the* L_p*-norm.*

(For ease of notation, assume $n \geq k$ even if "$n \in I\!N$" is written.)

Definition 3.2 *The* k**-medoids clustering** *of* $\mathbf{x}_n$ *is defined by*

$$E_n(\mathbf{x}_n) = \underset{\{C_1,\ldots,C_k\} \text{ partition of } \mathbf{x}_n,\ \tilde{x}_1 \in C_1,\ldots,\tilde{x}_k \in C_k}{\arg\min} \sum_{i=1}^{n} \min_j \|x_i - \tilde{x}_j\|_1. \tag{3.2}$$

Definition 3.3 *The* α**-trimmed** k**-means clustering** *of* $\mathbf{x}_n$ *is defined by*

$$E_n(\mathbf{x}_n) = \underset{\{C_1,\ldots,C_k\} \text{ partition of } \mathbf{y} \subset \mathbf{x}_n,\ |\mathbf{y}| = \lceil n(1-\alpha) \rceil}{\arg\min} \sum_{i=1}^{n} 1(x_i \in \mathbf{y}) \min_j \|x_i - \bar{x}_j\|_2^2, \tag{3.3}$$

where $\lceil z \rceil$ *is the smallest integer larger or equal to* z *and* $1(\bullet)$ *denotes the indicator function.*

Definition 3.4 *For* $x_i \in \mathbf{x}_n$ *with underlying distance measure* d*, a clustering* $E_{k,n}(\mathbf{x}_n) = \{C_1, \ldots, C_k\}$ *and* $x_i \in C_j$*,* $s(i,k) = \frac{b(i,k)-a(i,k)}{\max(a(i,k),b(i,k))}$ *is called* **silhouette width** *of point* x_i*, where*

$$a(i,k) = \frac{1}{|C_j| - 1} \sum_{x \in C_j} d(x_i, x), \ \ b(i,k) = \min_{x_i \notin C_l} \frac{1}{|C_l|} \sum_{x \in C_l} d(x_i, x).$$

If $|C_j| = 1$*,* $s(i,k) = 0$*.*
For $k \geq 2$ *(it is not possible to estimate* $k = 1$ *with this method; the method may be accompanied with a test detecting the presence of any clustering), let* E_k *be a partitioning method with* $|E_{k,n}(\mathbf{x}_n)| = k$ *for all data sets.*

$$E_n(\mathbf{x}_n) = E_{\hat{k},n}(\mathbf{x}_n) \text{ with } \hat{k} = \underset{k \in \{2,\ldots,n\}}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} s(i,k)$$

is called **average silhouette width-clustering** *corresponding to the partitioning method* E_k*.*

Maximizing the average silhouette width means that, on average, the distance of the points to their neighboring clusters is large compared to the distance to their own clusters, so that an optimal solution can be expected to yield homogeneous clusters (which is easier for large k), but so that neighboring clusters are far away from each other (which is not possible with k too large). The average silhouette width in the given form assumes partitions and is therefore not applicable to trimmed k-means.

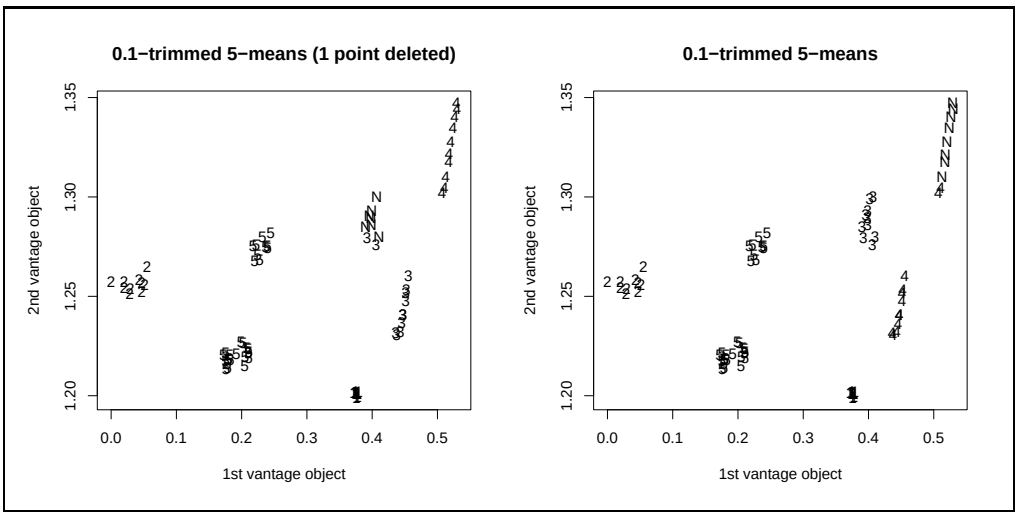

Figure 3: Left side: 79-images data set (one image of 80-images data has been deleted, which belongs to cluster 5 of the clustering on the right side) with 0.1-trimmed 5-means clustering. Right side: same with the 80-images data set. "N" denotes trimmed points.

3.2 General robustness problems with fixed k

With fixed k, all robustness results for the versions of k-means defined above (and for most other reasonable clustering methods) depend on the structure of the whole data set. A characterization of dissolution robustness in terms of an individual cluster and its isolation is impossible. Therefore, all these methods are not isolation robust. Here are the reasons:

- For k-means and k-medoids, consider a sequence of single outliers x_{n+1} to be added to the data set $\mathbf{x}_n$ so that $\min_{x\in\mathbf{x}_n} \|x_{n+1} - x\|_1 \to \infty$ (then, of course, also the L_2-distance converges to infinity). If x_{n+1} is grouped together in the same cluster with points of $\mathbf{x}_n$, the target criterion converges to infinity. If, for the clustering $E_{k,n+1}(\mathbf{x}_{n+1})$, $D = \{x_{n+1}\}$ is chosen as the first cluster and $\mathbf{x}_n$ is partitioned into $k-1$ clusters, the target criterion is bounded from above. Therefore, if the outlier is extreme enough, the best solution is to partition $\mathbf{x}_n$ into $k-1$ clusters, which means that at least one of the original clusters has to be dissolved because of Remark 2.3. If all clusters are strongly isolated, points of at least two of them will be merged into the same cluster (this happens to cluster 3 in Figure 1). Isolation robustness is impossible.

- For trimmed k-means, single extreme outliers can be trimmed. However, isolation robustness is still not possible, because for an arbitrarily strongly isolated cluster C, a constellation with $k+1$ groups of points (including C)

with very similar structure and isolation with the following properties can always be constructed: In the k-clusters solution of the resulting data set, C is a cluster and there is one cluster D corresponding to two others of the $k+1$ groups (the two groups are joined or, depending on α, one of them is as a whole or partly trimmed). If a single point is added close to the mean of one of the groups corresponding to D, then the two groups corresponding to D yield two new clusters and C is joined with another group or trimmed (or some of its points are joined and some are trimmed) instead. Thus, trimmed k-means is unstable if k is not well chosen. This violates even isolation robustness of degree α (Remark 2.6).

Example 3.5 *An example can be constructed from the 80-images data. The left side of Figure 3 shows a 0.1-trimmed 5-means solution for 79 of the 80 points. The solution for all 80 points is shown on the right side (the point that has been left out on the left side belongs to cluster 3). In this solution, some members of the well-separated former cluster 4 are joined with a part of the former cluster 3 and the other former members of cluster 4 are trimmed. Similar things would happen if the separation between all "natural groups" in the data (the clusters shown in Figure 1, say) would be uniformly increased. The separation of the former cluster 4 does not prevent parts of it from being joined with points very far away by adding a single point.*

These arguments hold for more general clustering methods with fixed k, and it has been presumed (and shown for mixture models) that the estimation of k is crucial for robustness in cluster analysis (Hennig 2004a). Garcia-Escudero and Gordaliza (1999) have already shown the non-robustness of k-means and k-medians. They show that trimmed k-means is often breakdown robust (breakdown defined in terms of the estimated means), but that the robustness is data dependent (see also Example 3.9). In fact, while trimmed k-means are not isolation robust, a useful dissolution robustness result can be derived.

3.3 Trimmed k-means, fixed k

For a given data set $\mathbf{x}_n$ and a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$, which is a partition of some $\mathbf{y}(\mathcal{C}) \subseteq \mathbf{x}_n$ (interpreted as non-exhaustive clustering on $\mathbf{x}_n$), let

$$Q(\mathbf{x}_n, \mathcal{C}) = \sum_{i=1}^{n} 1(x_i \in \mathbf{y}(\mathcal{C})) \min_{j \in \{1,\ldots,k\}} \|x_i - \bar{x}_j\|_2^2.$$

Let $B_n(\mathcal{C}) = \mathbf{x}_n \setminus \mathbf{y}(\mathcal{C})$ be the set of the trimmed points. Let $E_k = (E_{k,n})_{n \in \mathbb{N}}$ be the α-trimmed k-means.

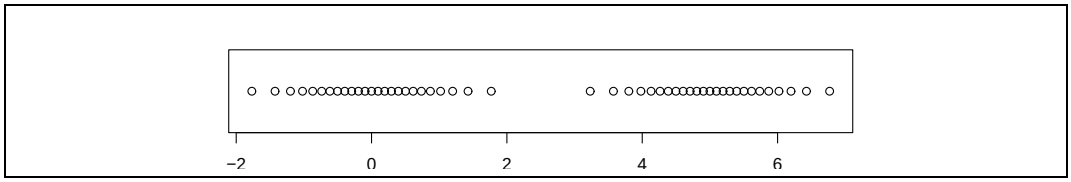

Figure 4: "Standard" example dataset: 25 points (0,1)-NSD combined with 25 points (5,1)-NSD

Theorem 3.6 *Let $n-\lceil n(1-\alpha)\rceil \geq g \in \mathbb{N}$, $C \in E_{k,n}(\mathbf{x}_n)$ with $|C| > g$. Consider partitions $\mathcal{C}^*$ of subsets $\mathbf{y}(\mathcal{C}^*) \subset \mathbf{x}_n$ with $|\mathbf{y}(\mathcal{C}^*)| = \lceil (n+g)(1-\alpha)\rceil - g$ into $l \leq k$ clusters so that*

$$\gamma^*(C, \mathcal{C}^*) \leq \tfrac{1}{2}, \tag{3.4}$$

there exist l possible centroids so that $\mathcal{C}^*$ assigns every point of $\mathbf{y}(\mathcal{C}^*)$ to the closest centroid and all points of $\mathbf{y}(\mathcal{C}^*)$ are closer to their closest centroid than any point of $\mathbf{x}_n \setminus \mathbf{y}(\mathcal{C}^*)$ is close to any of the centroids. (3.5)

If for any such $\mathcal{C}^*$

$$\min_{y_1,\ldots,y_g \in B_n(E_{k,n}(\mathbf{x}_n))} \sum_{i=1}^{g} \min_j \|y_i - \bar{x}_j\|_2^2 < Q(\mathbf{x}_n, \mathcal{C}^*) - Q(\mathbf{x}_n, E_{k,n}(\mathbf{x}_n)), \tag{3.6}$$

then $\Delta(E_k, \mathbf{x}_n, C) > \frac{g}{|C|+g}$.

The proof is given in the appendix. Note that (3.5) means that $\mathcal{C}^*$ can occur as an induced clustering of a clustering on some $\mathbf{x}_{n+g}$.

The theorem says that the cluster C cannot be dissolved by adding g points, if there are g points among the originally trimmed points that are fitted well enough by the original clusters. Dissolution point theorems are useful if they enable the computation of the dissolution point of a given cluster in a given data set without being forced to find the worst g points to be added. The computation of Δ according to Theorem 3.6 may be difficult, as (3.6) requires to be evaluated for all possible partitions $\mathcal{C}^*$. However, in simple situations it is easy to guess how to minimize $Q(\mathbf{x}_n, \mathcal{C}^*)$.

Example 3.7 *The following definition is used to generate reproducible reference datasets:*

Definition 3.8 *$\Phi^{-1}_{a,\sigma^2}(\frac{1}{n+1}), \ldots, \Phi^{-1}_{a,\sigma^2}(\frac{n}{n+1})$ is called a (a, σ^2)-***Normal standard dataset*** *(NSD) with n points, where Φ_{a,σ^2} denotes the cdf of the Normal distribution with parameters a, σ^2.*

I will use a data set consisting of two NSDs with 25 points each, with $(a, \sigma^2) = (0,1),\ (5,1)$, *respectively, as standard example, to which the robustness results are to be applied, see Figure 4.*

Let $k = 2,\ \alpha = 0.1$. *The* α*-trimmed* k*-means is obtained by trimming the four extreme points of the two NSDs and one further point of the four extreme points of the remaining data. There are two resulting clusters corresponding to the remaining points of the two NSDs, one with 22 (let this be the* C *of interest) and one with 23 points,* $Q(\mathbf{x}_n, E_{k,n}(\mathbf{x}_n)) = 24.79$,

$$\min_{y_1,\ldots,y_g \in B_n(E_{k,n}(\mathbf{x}_n))} \sum_{i=1}^{g} \min_j \|y_i - \bar{x}_j\|_2^2 = 14.75.$$

For $g = 6$, C *can be dissolved because only 5 points are trimmed and one extreme outlier can remain, which has to be fitted by its own cluster, compare Section 3.2. Let therefore* $g = 5$. *How can* $Q(\mathbf{x}_n, \mathcal{C}^*)$ *be minimized over partitions of the 45 points of* $\mathbf{y}(E_{k,n}(\mathbf{x}_n))$ *that dissolve* C*? Because of (3.5), the clusters of* $\mathcal{C}^*$ *have to be topologically connected. The two obvious possibilities to do this are to take a subcluster of* C *with 11 points, trim the 5 points at one side of the NSD of which* C *is a subset, and join the remaining points with the other NSD, which leads to* $Q(\mathbf{x}_n, \mathcal{C}^*) = 131.14$, *or to form a cluster with 44 points containing* C, *trim the 5 most extreme points on the opposite side and take the second cluster to fit the remaining single point, which even yields* $Q(\mathbf{x}_n, \mathcal{C}^*) = 259.38$. *Thus, (3.6) is fulfilled and* $\Delta(E_2, \mathbf{x}_n, C) = \frac{6}{28}$. *For* k*-means and* k*-medoids, for* C_1, C_2 *being the original clusters with 25 points each,* $\Delta(E_2, \mathbf{x}_n, C_j) = \frac{1}{21},\ j = 1, 2$, *because if* $x_{n+1} \geq 24$ *(*k*-means) or* $x_{n+1} \geq 67$ *(*k*-medoids) is added, the two original clusters are merged.*

Example 3.9 *For the 8-images data, 0.1-trimmed 7-means (and also trimmed 7-means with other choices of* α*) seems to be rather robust and yields the solution of Figure 1 with some points of cluster 4 being trimmed. A small enough number of added outliers is trimmed and does no further harm than reducing the number of trimmed points of cluster 4.*

The separations between the clusters seem to be different enough that "isolation dissolution" as in Example 3.5 could not be constructed for 0.1-trimmed 6-means by leaving out only one point.

3.4 Average silhouette width

In Section 3.2 it has been presumed that the robustness problems of k-means and k-medoids are mainly caused by the fixed number of clusters k. Unfortunately, the average silhouette width method to estimate k does not yield a better robustness behavior. The following theorem shows that if a single extreme enough outlier is added to a data set, the average silhouette width clustering consists of only two

clusters one of which consists of only the outlier. Therefore, no isolation robustness is possible and the dissolution point of any cluster C with $|C| \leq \frac{n}{2}$ is the smallest possible value $\frac{1}{|C|+1}$.

Theorem 3.10 *Let $\mathbf{x}_{n+1} = \mathbf{x}_n \cup \{x_{n+1}\}$, where $\mathbf{x}_n$ is a fixed data set with n pairwise different points. If x_{n+1} large enough,*

$$E_{n+1}(\mathbf{x}_{n+1}) = \{\mathbf{x}_n, \{x_{n+1}\}\},$$

where $(E_n)_{n\in\mathbb{N}}$ is the average silhouette width clustering corresponding to k-means or k-medoids.

The assumption that the points of $\mathbf{x}_n$ are pairwise different is not crucial. It can be seen from the proof that the given clustering will be preferred to any clustering with $k < n$ but large including at least one cluster that contains two nonidentical points.

Example 3.11 *In the standard example data set (Figure 4), the necessary size of an outlier so that the average silhouette width clustering joins the two original clusters by estimating 2 clusters, one of which consists only of the outlier, is 67.*

In the data set shown in Figure 2, four outliers have been added to the 80-images data. This results in only two clusters as shown, so that all original clusters are dissolved. Up to three of the shown outliers make up a new cluster and leave the original clustering unchanged.

4 MIXTURE MODELS

4.1 Definition of methods

In cluster analysis based on mixture models (including a model for "noise"-points), the data is assumed to be generated i.i.d. by a distribution of the form

$$f_\eta(x) = \sum_{j=1}^{k} \pi_j f_{\theta_j}(x) + \pi_0 u(x), \tag{4.1}$$

where f_θ is a density from some parametric family, $\sum_{j=0}^{k} \pi_j = 1,\ 0 \leq \pi_j \leq 1$ for $j = 0, \ldots, k,\ \eta = (k, \pi_0, \ldots, \pi_k, \theta_1, \ldots, \theta_k)$. u models points not belonging to any cluster ("noise component"). The "classical" mixture model assumes $\pi_0 = 0$. For literature on models like these and more structured models (mixtures of regressions etc.), see McLachlan and Peel (2000).

Having estimated η by $\hat{\eta} = (\hat{k}, \hat{\pi}_0, \ldots, \hat{\pi}_k, \hat{\theta}_1, \ldots, \hat{\theta}_k)$, and, if necessary, u by $\hat{u}$ (it may be assumed that $\hat{k}$ is constant or $\hat{\pi}_0 = 0$), a clustering on $\mathbf{x}_n$ can be generated

by $E_n(\mathbf{x}_n) = \{C_1, \ldots, C_{\hat{k}}\}$. For $j = 1, \ldots, \hat{k}$:

$$C_j = \left\{ x \in \mathbf{x}_n : \ \hat{\pi}_j f_{\hat{\theta}_j}(x) > \hat{\pi}_0 \hat{u}(x), \ j = \arg\max_l \hat{\pi}_l f_{\hat{\theta}_l}(x) \right\}, \tag{4.2}$$

given a rule to break ties in the $\hat{\pi}_j f_{\hat{\theta}_j}(x)$. For simplicity reasons, in the present paper one-dimensional data and mixture models of the following form are considered:

$$f_\eta(x) = \sum_{j=1}^{k} \pi_j f_{a_j,\sigma_j}(x) + \pi_0 u(x), \text{where } f_{a,\sigma}(x) = \frac{1}{\sigma} f_{0,1}\left(\frac{x-a}{\sigma}\right), \tag{4.3}$$

$f_{0,1}$ being continuous, symmetrical about 0, monotonically decreasing on $[0, \infty]$, larger than 0 on $\mathbb{R}$. Of particular interest is the standard normal distribution, which is often used in cluster analysis (McLachlan and Peel 2000, Fraley and Raftery 1998) and the t_ν-distribution, which was suggested as a more robust alternative (with $\pi_0 = 0$; Peel and McLachlan 2000). Banfield and Raftery (1993) suggested robustification of the classical normal mixture by including a noise component where u is taken to be the uniform distribution over the convex hull of the data, i.e., for one-dimensional data, $\hat{u}(x) = \frac{1}{x_{max,n} - x_{min,n}} 1(x_{max,n} \geq x \geq x_{min,n})$, where $x_{max,n}$ and $x_{min,n}$ are the maximum and the minimum of $\mathbf{x}_n$. Basic robustness properties will carry over to the multivariate case.

For fixed $\hat{k} = k$, $\hat{\eta}$ can be estimated by maximum likelihood, which is implemented by means of the EM-algorithm in the software packages "EMMIX" (McLachlan and Peel 2000) and "mclust" (Fraley and Raftery 2003). Because the loglikelihood

$$L_{n,k}(\eta, \mathbf{x}_n) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j f_{a_j,\sigma_j}(x_i) + \frac{\pi_0}{x_{max,n} - x_{min,n}} \right) \tag{4.4}$$

converges to ∞ if $\hat{a}_1 = x_1$ and $\hat{\sigma}_1 \to 0$, the parameter space has to be restricted. Here, the restriction

$$\sigma_j \geq \sigma_0 > 0 \tag{4.5}$$

for some pre-specified σ_0 is used (for a discussion of the choice of σ_0, see Hennig 2004b). An alternative would be to assume all σ_j to be equal.

The most frequently used methods to estimate k are the information criteria AIC (Akaike 1974) and BIC (Schwarz 1978). The estimator $\hat{k}$ with BIC (AIC has about the same robustness behavior) is defined as $\hat{k} = \arg\max_k \text{BIC}(k)$, where

$$\text{BIC}(k) = 2L_{n,k}(\hat{\eta}_{n,k}, \mathbf{x}_n) - q(k) \log n, \tag{4.6}$$

where $q(k)$ denotes the number of free parameters, i.e., $q(k) = 3k - 1$ for the classical mixture and $q(k) = 3k$ with noise component, and $\hat{\eta}_{n,k}$ denotes the ML-estimator of η for $\mathbf{x}_n$ under k mixture components.

4.2 Robustness results

The robustness of these methods has already been investigated in Hennig (2004a), where parameter breakdown points have been considered and immediate consequences of these results for dissolution points have been outlined. Here is a summary of these results for fixed k:

- For fixed k, the situation for all considered methods is similar to Section 3.2. If a single point $x_{n+1} \to \infty$ is added to $\mathbf{x}_n$, then it follows from Lemma 4.1 of Hennig (2004a) that eventually $\{x_{n+1}\}$ is a cluster. The necessary sizes of an outlier to dissolve the original $k = 2$ clusters by merging in the standard example data set (Figure 4) are 15.2 (classical normal mixture), about 800 (t_3-mixture), 3.8×10^6 (t_1-mixture), 3.5×10^7 (normal mixture with noise). These values depend on σ_0, which was chosen as 0.025. Note that the clusterings of the t-mixtures and the noise component approach are somewhat robust even under the addition of more outliers, as long as they are not all at the same point (Hennig 2004b).
- The above argument does not hold if the noise component u is taken as some nonzero data independent constant (improper density), because in this case the loglikelihood cannot diverge to $-\infty$, see Theorem 4.11 of Hennig (2004a). The same discussion as given for trimmed k-means in Section 3.2 applies. Unfortunately, dissolution results will be less tractable than Theorem 3.6, because such results will be similarly difficult to evaluate (and more difficult to derive) than those given in Theorem 4.1 below.

If extreme outliers are added under estimated k, the BIC will enlarge the number of mixture components to fit the outliers, as opposed to the average silhouette width. However, while it can be shown that the parameter estimators of the original mixture components are prevented from diverging to infinity (Theorems 4.13 and 4.16 of Hennig 2004a), cluster dissolution is still possible by adding points that change the local clustering structure. A corresponding theorem is easily derived:

Theorem 4.1 *For a data set $\mathbf{x}_n$, let $\hat{k}$ be a maximizer of the BIC, and let $E = (E_n)_{n\in\mathbb{N}}$ be the corresponding maximum likelihood method according to (4.4) (π_0 estimated or fixed $= 0$). Let $g \in \mathbb{N}$, $C \in E_n(\mathbf{x}_n)$ with $|C| > g$. Consider parameter vectors η^* for $1 \leq k^* \leq n$ mixture components, so that $\gamma^*(C, \mathcal{C}^*) \leq \frac{1}{2}$ for the corresponding clustering $\mathcal{C}^*$. If for any such η^**

$$\left[L_{n,\hat{k}}(\hat{\eta}_{n,\hat{k}}, \mathbf{x}_n) - L_{n,k^*}(\eta^*, \mathbf{x}_n) - \frac{1}{2}(5g + 3\hat{k} - 3k^* + 2n)\log(n+g) + n\log n\right] > 0, \tag{4.7}$$

then $\Delta(E, \mathbf{x}_n, C) > \frac{g}{|C|+g}$.

The proof is completely analogous to the proof of Theorem 4.13 (Theorem 4.16 with noise component) in Hennig (2004a).

Unfortunately, Theorem 4.1 is not as useful as Theorem 3.6, because the optimization of (4.7) over all possible η^* seems computationally intractable. Empirically, in the standard example data set of Figure 4, the addition of 12 (normal mixture with and without noise component) or 13 points (t_1-mixture) between the two original clusters yield $\hat{k} = 1$ and therefore dissolution of both clusters.

The isolation robustness result is more expressive.

Theorem 4.2 *Let E be a clustering method defined by maximizing (4.4) for given k and the BIC over k (π_0 estimated or fixed $= 0$). Then E is isolation robust. (The corresponding function v_m does only depend on g, but not on M_C.)*

The proof is given in the Appendix.

The fact that v_m does not depend on the distance matrix within C in this case is a consequence of the missing invariance property. If a clustering would not change under multiplication of the data with a constant, the required isolation for robustness should not be constant but depend on some spread measure of C. Invariance is violated by (4.5), and multiplying a data set with an extremely large factor (depending on σ_0) would result in a clustering where $\hat{k}$ would equal the number of pairwise distinct points in the data. This is irrelevant in practice, and clusterings can be considered as "practically invariant" under linear transformations, unless σ_0 is chosen far too small.

5 AGGLOMERATIVE HIERARCHICAL METHODS

5.1 Definition of methods

Most agglomerative hierarchical methods assume that the objects of a data set $\mathbf{x}_n$ are characterized by an $n \times n$ distance matrix $\mathbf{D} = (d_{ij})_{i,j=1,...,n}$, $d_{ij} = d(x_i, x_j)$, the distance between x_i and x_j. In the present paper, d is assumed to be a metric, and it is assumed that the definition of d_{ij} does not depend on the presence or absence of other points in the data set. Furthermore, it is assumed that the underlying object space $\mathcal{O} \supset \mathbf{x}_n$ is rich enough that

$$\forall x \in \mathbf{x}_n,\ d^* \in I\!R^+\ \exists y \in \mathcal{O}: \qquad d(x,y) = d^*, \tag{5.1}$$

$$\forall x, y \in \mathcal{O}, \qquad I\!R^+ \ni d^* < d(x,y)\ \exists z \in \mathcal{O}:$$
$$d(x,z) = d^*,\ d(y,z) = d(x,y) - d^*. \tag{5.2}$$

These assumptions ensure that the possible "locations" of points to be added to $\mathbf{x}_n$ are not too restricted.

For simplicity, it is also assumed that the nonzero distances are pairwise distinct. I will restrict considerations to the single linkage and the complete linkage method

(see Chapter 4 of Gordon 1999, for references). The isolation robustness results will carry over to compromises between these two methods such as average linkage.

Definition 5.1 *Let $\delta : \mathcal{P}(\mathbf{x}_n) \times \mathcal{P}(\mathbf{x}_n) \mapsto \mathbb{R}_0^+$ be a dissimilarity measure between data subsets. Let $\mathcal{C}_n = \{\{x\} : x \in \mathbf{x}_n\}$, $h_n = 0$. For $k = n-1, \ldots, 1$:*

$$(A_k, B_k) = \underset{A,B \in \mathcal{C}_{k+1}}{\arg\min}\ \delta(A, B),\ h_k = \delta(A_k, B_k), \tag{5.3}$$

$$\mathcal{C}_k = \{A_k \cup B_k\} \cup \mathcal{C}_{k+1} \setminus \{A_k, B_k\}. \tag{5.4}$$

$\mathcal{C} = \bigcup_{k=1}^n \mathcal{C}_k$ is called

a) Single linkage hierarchy, *if $\delta(A, B) = \delta_S(A, B) = \min_{x_i \in A, x_j \in B} d_{ij}$,*

b) Complete linkage hierarchy, *if $\delta(A, B) = \delta_C(A, B) = \max_{x_i \in A, x_j \in B} d_{ij}$,*

There are two simple methods to obtain a partition from a hierarchy. The first one is to cut the hierarchy at a prespecified number of clusters k, the second one is to cut the hierarchy at a given distance level h (the reader is referred to Gordon 1999, Section 3.5, for more sophisticated methods to estimate the number of clusters).

Definition 5.2 *Given a hierarchy $\mathcal{C} = \bigcup_{k=1}^n \mathcal{C}_k$ on $\mathbf{x}_n$ defined as in Definition 5.1 equipped with a monotonically decreasing sequence of level number $h_1, \ldots, h_n$, see (5.3), $\mathcal{C}_k$ is called the* **k-number partition** *for given $n \geq k \in \mathbb{N}$, and $\mathcal{C}_{k(h)}$ with $h_{k(h)} \leq h$ and $h_{k(h)-1} > h$ is called the* **h-level partition** *for given $h \geq 0$.*

5.2 Robustness results

While the k-number and the h-level partition are similarly simple, their robustness properties are different. The discussion in Section 3.2 applies to the k-number partition (not only of single and complete linkage clustering, but also of all other agglomerative methods that I know). An extreme enough outlier x_{n+1} always forms a cluster on its own, as long as $k \geq 2$, because $\delta(\{x_{n+1}\}, A)$ can be driven to infinity for all $A \subset \mathbf{x}_n$.

The h-level partition (denoted $E_h = (E_{h,n})_{n \in \mathbb{N}}$ in the following) is more stable. Let h be fixed, $E_{h,n}(\mathbf{x}_n) = \mathcal{C}_{k(h)} = \{C_1, \ldots, C_{k(h)}\}$.

Here are the results for single linkage. For two clusters C_i, C_j, $i, j = 1, \ldots, k(h)$, let $g_{(i,j)} = \lceil \delta_S(C_i, C_j)/h \rceil$. If C_i and C_j were the only clusters, this would be the number of additional points needed to join C_i and C_j. For given C_i, $g \in \mathbb{N}$, let $q(i, g)$ be the maximum number of points of $\mathbf{x}_n$ which are not members of C_i, but can be joined with C_i if g points are added to $\mathbf{x}_n$.

Theorem 5.3 *Given the notation above, where E_h is the h-level partition of the single linkage hierarchy,*

$$\gamma(C_i, E_n^*(\mathbf{x}_{n+g}) \geq \frac{|C_i|}{|C_i| + q(i,g)}, \tag{5.5}$$

$$\frac{|C_i|}{|C_i| + q(i,g)} > \frac{1}{2} \Rightarrow \Delta(E_h, \mathbf{x}_n, C_i) > \frac{g}{|C_i| + g}. \tag{5.6}$$

Further,

$$q(i,g) = \max_{\{C_{j_1},\ldots,C_{j_l}\} \in \mathcal{D}_g(C_i)} \sum_{m=1}^{l} |C_{j_m}| - |C_i|, \tag{5.7}$$

where $\mathcal{D}_g(C_i)$ denotes the set of all "g-reachable cluster trees", i.e., subsets S of $\mathcal{C}_{k(h)}$ with the following properties:

- $C_i \in S$,
- *there exists $Q \subseteq \{(C_{j_1}, C_{j_2}),\ C_{j_1} \neq C_{j_2} \in S\}$ so that the graph with the members of S as vertices and Q as the set of edges is a tree, i.e., a connected graph without circles, and*

$$\sum_{(C_{j_1}, C_{j_2}) \in Q} g_{(j_1,j_2)} \leq g. \tag{5.8}$$

The proof is given in the Appendix.

Corollary 5.4 *The h-level partition of the single linkage hierarchy is isolation robust.*

This follows because for given g and $i(C_i) = \min_{j \neq i} \delta_S(C_i, C_j)$ large enough, $g_{(i,j)}$ for any $j \neq i$ is larger than g and $q(i,g) = 0$.

Example 5.5 *The isolation of the two clusters corresponding to the NSDs in the standard example data set of Figure 4 is 1.462 and the largest within-cluster distance is 0.343. The 2-number partition would join the two original clusters if a single point at 8.23 (original data maximum plus isolation) is added. For the h-level partition, h could be chosen between 0.343 and 1.462 to generate two clusters. If $h > 0.713$ (half of the isolation), $\Delta(E_h, \mathbf{x}_n, C_j) = \frac{1}{26}$. If $h = 0.344$, $\Delta(E_h, \mathbf{x}_n, C_j) = \frac{4}{29}$. While the stability depends on h chosen favorably with respect to the data, the theory does not allow h to be chosen data-dependent. The main problem with the h-level approach is that h has to be chosen by use of background knowledge, and such knowledge does not always exist. Furthermore, the h-level clusterings are not invariant with respect to multiplying all distances with a constant.*

The h-level partition of complete linkage is trivially isolation robust, because under complete linkage no two points with a distance larger than h can be in the same cluster. Therefore, if $i(C) > h$, no point of C can be together in the same cluster with a point that has not been in C before the addition of g points with g arbitrarily large.

Contrary to single linkage, complete linkage h-level clusters can be split by addition of points. Therefore it is more difficult to prevent dissolution. The following theorem gives a condition which prevents dissolution. Let $E_h = (E_{h,n})_{n\in\mathbb{N}}$ be the h-level partition of complete linkage, $d(C)$ be the diameter of a set C (maximum distance within C), $d_h(C,D) = \max_{x\in D\setminus C,\ y\in C,\ d(x,y)\leq h} d(x,y)$ (the maximum over $\emptyset$ is 0).

Theorem 5.6 *For a given $C \in E_{h,n}(\mathbf{x}_n)$, let $H \subset C$ be a subcluster of C (i.e., a member of $E_{h^*,n}(\mathbf{x}_n)$ with $h^* = d(H) \leq h$) with $|H| > \frac{|C|}{2}$. Define*

$$m_0 = \max(d(H), d_h(H,\mathbf{x}_n)), \quad m_1 = d(H) + d_h(H,\mathbf{x}_n) + m_0,$$
$$m_g = d(H) + m_{g-1} + m_{g-2}$$

for $g \geq 2$. If $m_g \leq h$ and if

$$q_H = |\{y \in \mathbf{x}_n \setminus C : \min_{x\in H} d(x,y) \leq h\}| < 2|H| - |C|, \tag{5.9}$$

then $\Delta(E_h, \mathbf{x}_n, C) > \frac{g}{|C|+g}$. H may be chosen to minimize m_g.

According to this theorem, $d_h(H,\mathbf{x}_n)$ has to be much smaller than h to enable good dissolution robustness. This can happen if C is strongly isolated and its diameter is much smaller than h. However, the proof of the theorem deals with a very specific worst-case situation, and it will be very conservative for lots of data sets. This can be seen in the following example. A better result under additional restrictions may be possible.

Example 5.7 *The 2-number partition would join the two original clusters in the data set of Figure 4 if a single point at about 11.8 is added. For the h-level partition, h could be chosen between 3.54 and 8.53 to generate two clusters. Theorem 5.6 does not yield a better lower bound than $\frac{1}{26}$ for the dissolution point of one of the clusters, the (0,1)-NSD, say. The only subcluster with ≥ 13 points is $H = \{x_{11}\dots x_{25}\}$, $d(H) = 1.96$. Even for $h = 3.54$, there are points in the (5,1)-NSD which are closer than h to all points of H, and $d_h(H,\mathbf{x}_n) = 3.54$. In fact, $d_h(H,\mathbf{x}_n) > \frac{h}{2}$ for any h between 3.54 and 8.53, enforcing $m_1 > h$. The theorem does not apply until $h = 9.05$ and the second cluster is chosen as an (11.7,1)-NSD, in which case $q_H = 4$ and $m_1 = 9.04$, thus $\Delta(E_h, \mathbf{x}_n, C_1) \geq \frac{2}{27}$.*

However, the worst case scenario of the proof of Theorem 5.6 is impossible here and in fact I have not been able to dissolve one of the two clusters by adding any $g < |C|$ points unless $h \geq 8$, so that the result of the theorem is extremely conservative here. Figure 5 shows data where the dissolution point bound obtained in the theorem is attained.

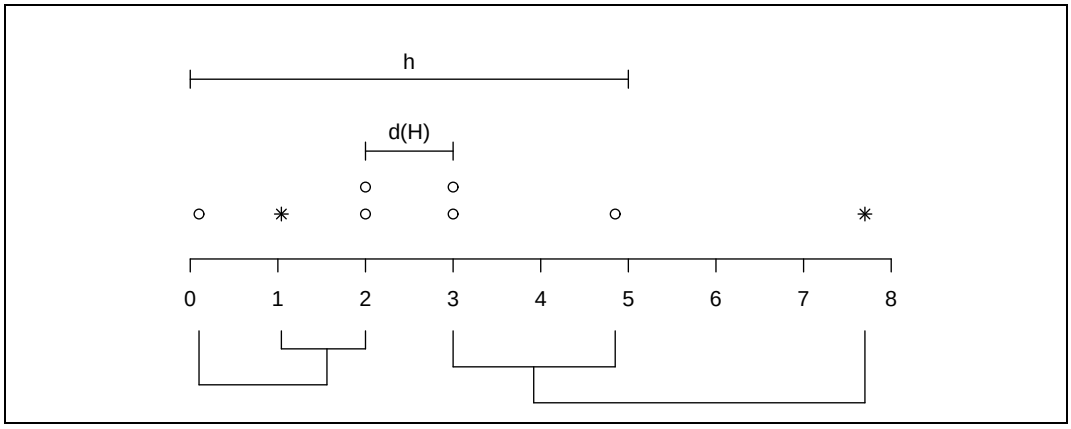

Figure 5: Let $(0.1, 2, 2, 3, 3, 4.85)$ (circles) form a well separated complete linkage cluster ($4.8 < h < 7.6$) in a data set. Let $H = \{2, 2, 3, 3\}$. Thus, $d(H) = 1$, $d_h(H, \mathbf{x}_n) = m_0 = 1.9$, $m_1 = 4.8$, $m_2 = 7.7$, $q_H = 0$. Therefore, $\Delta = \frac{2}{8}$. Dissolution works by adding points at 1.04 and 7.7 (stars). The resulting complete linkage clusters are shown below the x-axis.

6 FIXED POINT CLUSTERS

Note: This section is not part of the version of this paper which I submitted for journal publication. It may be published later elsewhere.

6.1 Definition of fixed point clusters

Fixed point cluster (FPC) analysis has been introduced by Hennig (1997). FPC analysis has been applied to clusterwise linear regression (Hennig, 1997, 2002, 2003) and normal-shaped clusters of p-dimensional data (Hennig and Christlieb, 2002) based on the Mahalanobis distance. The latter is introduced here.

The basic idea of FPC analysis is that a cluster can be formalized as a data subset, which is homogeneous in the sense that it does not contain any outlier, and which is well separated from the rest of the data meaning that all other points are outliers with respect to the cluster. That is, the FPC concept is a local cluster concept: It does not assume a cluster structure or some parametric model for the whole dataset. It is based only on the cluster candidate itself and its relation to its surroundings.

In order to define FPCs, a definition of an outlier with respect to a data subset is needed. The definition should be based only on a parametric model for the non-outliers (reference model), but not for the outliers. That is, if the Gaussian family is taken as reference model, the whole dataset is treated as if it came from

a contamination mixture

$$(1-\epsilon)N_p(a,\mathbf{\Sigma})+\epsilon P^*, \quad 0\leq\epsilon<1, \tag{6.1}$$

where p is the number of variables, $N_p(a,\mathbf{\Sigma})$ denotes the p-dimensional Gaussian distribution with mean vector a and covariance matrix $\mathbf{\Sigma}$, and P^* is assumed to generate points well separated from the core area of $N_p(a,\mathbf{\Sigma})$. The principle to define the outliers is taken from Becker and Gather (1999). They define α-outliers as points that lie in a region with low density such that the probability of the so-called outlier region is α under the reference distribution. α has to be small in order to match the impression of outlyingness. For the $N_p(a,\mathbf{\Sigma})$-distribution, the α-outlier region is

$$\{x:\ (x-a)^t\mathbf{\Sigma}^{-1}(x-a)>\chi^2_{p;1-\alpha}\},$$

$\chi^2_{p;1-\alpha}$ denoting the $1-\alpha$-quantile of the χ^2-distribution with p degrees of freedom. In a concrete situation, a and $\mathbf{\Sigma}$ are not known, and they have to be estimated. This is done for Mahalanobis FPCs by the sample mean and the maximum likelihood covariance matrix. (Note that these estimators are non-robust, but they are reasonable if they are only applied to the non-outliers.)

A dataset $\mathbf{x}_n$ consists of p-dimensional points. Data subsets are represented by an indicator vector $w\in\{0,1\}^n$. Let $\mathbf{x}_n(w)$ be the set with only the points x_i, for which $w_i=1$, and $n(w)=\sum_{i=1}^n w_i$. Let $m(w)=\frac{1}{n(w)}\sum_{w_i=1}x_i$ the mean vector and $\mathbf{S}(w)=\frac{1}{n(w)}\sum_{w_i=1}(x_i-m(w))(x_i-m(w))'$ the ML covariance matrix estimator for the points indicated by w.

The set of outliers from $\mathbf{x}_n$ with respect to a data subset $\mathbf{x}_n(w)$ is

$$\{x:\ (x-m(w))'\mathbf{S}(w)^{-1}(x-m(w))>\chi^2_{p;1-\alpha}\}.$$

That is, a point is defined as an outlier w.r.t $\mathbf{x}_n(w)$, if its Mahalanobis distance to the estimated parameters of $\mathbf{x}_n(w)$ is large.

An FPC is defined as a data subset which is exactly the set of non-outliers w.r.t. itself:

Definition 6.1 *A data subset $\mathbf{x}_n(w)$ of $\mathbf{x}_n$ is called* **Mahalanobis fixed point cluster** *of level α, if for $i=1,\ldots,n$:*

$$w=\left(1\left[(x_i-m(w))'\mathbf{S}(w)^{-1}(x_i-m(w))\leq\chi^2_{p;1-\alpha}\right]\right)_{i=1,\ldots,n}. \tag{6.2}$$

If $\mathbf{S}(w)^{-1}$ does not exist, the Moore-Penrose inverse is taken instead on the supporting hyperplane of the corresponding degenerated normal distribution, and $w_i=0$ for all other points.

For combinatorial reasons it is impossible to check (6.2) for all w. But FPCs can be found by a fixed point algorithm defined by

$$w^{k+1}=\left(1\left[(x_i-m(w^k))'\mathbf{S}(w^k)^{-1}(x_i-m(w^k))\leq\chi^2_{p;1-\alpha}\right]\right)_{i=1,\ldots,n}. \tag{6.3}$$

This algorithm is shown to converge toward an FPC in a finite number of steps if $\chi^2_{p;1-\alpha} > p$ (which is always fulfilled for $\alpha < 0.25$, i.e., for all reasonable choices of α) in Hennig and Christlieb (2002). Note that the proof requires the use of the ML-estimator for the covariance matrix, i.e., division is by $n(w)$ instead of $n(w) - 1$ for the UMVU estimator.

The problem here is the choice of reasonable starting configurations w_0. While, according to this definition, there are many very small FPCs, which are not very meaningful (e.g., all sets of p or fewer points are FPCs), an FPC analysis aims at finding all substantial FPCs, where "substantial" means all FPCs corresponding to well separated, not too small data subsets which give rise to an adequate description of the data by a model of the form (6.1). For clusterwise regression, this problem is discussed in depth in Hennig (2002) along with an implementation, which is included in the add-on package "fpc" for the statistical software system R. In the same package, there is also an implementation of Mahalanobis FPCs. There, the following method to generate initial subsets is applied:

For every point of the dataset, one initial configuration is chosen, so that there are n runs of the algorithm (6.3). For every point, the p nearest points in terms of the Mahalanobis distance w.r.t. $\mathbf{S}(1, \ldots, 1)$ are added, so that there are $p + 1$ points. Because such configurations often lead to too small clusters, the initial configuration is enlarged to contain n_{start} points. To obtain the $p + 2$nd to the n_{start}th point, the covariance matrix of the current configuration is computed (new for every added point) and the nearest point in terms of the new Mahalanobis distance is added.

$n_{start} = 20 + 4p$ is chosen as the default size of initial configurations in package "fpc". This is reasonable for fairly large datasets, but should be smaller for small datasets. Experience shows that the effective minimum size of FPCs that can be found by this method is not much smaller than n_{start}. The default choice for α is 0.99; $\alpha = 0.95$ produces in most cases more FPCs, but these are often too small, compare Example 6.5. A simulation study comparing different choices for the parameters of the algorithm and a discussion of a fuzzy version of fixed point clusters can be found in Hennig (2005b).

Note that Mahalanobis FPCs are invariant under linear transformations, i.e., for any invertible matrix $\mathbf{A}$ and vector $b \in I\!R^p$, and any FPC $\mathbf{x}_n(w)$, $\mathbf{y}_n(w)$ is also an FPC for $\mathbf{y}_n = \{\mathbf{A}x_i + b \ : \ x_i \in \mathbf{x}_n\}$ and vice versa.

6.2 Robustness results for fixed point clusters

To derive a lower bound for the dissolution point of a fixed point cluster, the case $p = 1$ is considered for the sake of simplicity. This is a special case of both Mahalanobis and clusterwise linear regression FPCs.

$E_n(\mathbf{x}_n)$ is taken as the collection of all data subsets fulfilling (6.2). $E_n(\mathbf{x}_n)$ is not a partition, because FPCs may overlap and not all points necessarily belong to any

FPC.

FPCs are robust against gross outliers in the sense that

an FPC $\mathbf{x}(w)$is invariant against any change, especially addition of points, outside its domain $\{(x-m(w))'\mathbf{S}(w)^{-1}(x-m(w))\leq c\}$, (6.4)

$c=\chi^2_{p;1-\alpha}$, because such changes simply do not affect its definition. However, FPCs can be affected by points added inside their domain, which is, for $p=1$,

$$D(w)=[m(w)-s(w)\sqrt{c},m(w)+s(w)\sqrt{c}],\ s(w)=\sqrt{S(w)}.$$

The aim of the following theory is to characterize a situation in which an FPC is stable under addition of points. The key condition is the separateness of the FPC, i.e., the number of points in its surrounding (which is bounded by k_2 in (6.7)) and the number of points belonging to it but close to its border (which is bounded by k_1 in (6.6)). The derived conditions for robustness (in the sense of a lower bound on the dissolution point) are somewhat stronger than presumably needed, but the theory reflects that that the key ingredient for stability of an FPC is to have few points close to the border (inside and outside).

In the following, $\mathbf{x}_n(w)$ denotes a Mahalanobis FPC in $\mathbf{x}_n$.

Let $S_{gk}(w)$ be the set containing the vectors $(m_{+g},s^2_{+g},m_{-k},s^2_{-k})$ with the following property:

Property $A(g,k,\mathbf{x}_n(w))$**:** Interpret $\mathbf{x}_n(w)$ as an FPC in itself, i.e., $\mathbf{y}_{\tilde{n}}=\mathbf{x}_n(w)=\mathbf{y}_{\tilde{n}}(1,\ldots,1)$ $(\tilde{n}=n(w))$.
$(m_{+g},s^2_{+g},m_{-k},s^2_{-k})$ possess the Property $A(g,k,\mathbf{x}_n(w))$, if it is possible to add points $y_{\tilde{n}+1},\ldots,y_{\tilde{n}+g}$ to $\mathbf{y}_{\tilde{n}}$, so that if the algorithm (6.3) started from the original FPC $\mathbf{y}_{\tilde{n}}$ is run on the dataset $\mathbf{y}_{\tilde{n}+g}=\mathbf{y}_{\tilde{n}}\cup\{y_{\tilde{n}+1},\ldots,y_{\tilde{n}+g}\}$ and converges to a new FPC $\mathbf{y}_{\tilde{n}+g}(w^*)$, m_{+g} and s^2_{+g} are the values of the mean and variance of the points $\{y_{\tilde{n}+1},\ldots,y_{\tilde{n}+g}\}\cap\mathbf{y}_{\tilde{n}+g}(w^*)$, and m_{-k} and s^2_{-k} are the values of the mean and variance of the points lost in the algorithm, i.e., $\mathbf{y}_{\tilde{n}}\setminus\mathbf{y}_{\tilde{n}+g}(w^*)$, where it is assumed that $|\mathbf{y}_{\tilde{n}}\setminus\mathbf{y}_{\tilde{n}+g}(w^*)|\leq k$. Mean and variance of 0 points are taken to be 0. Note that always $(0,0,0,0)\in S_{gk}(w)$, because of (6.4) and the added points can be chosen outside the domain of $\mathbf{y}_{\tilde{n}}$.

In the proof of Theorem 6.2 it will be shown that an upper bound of the domain of $\mathbf{y}_{\tilde{n}+g}(w^*)$ in the situation of Property $A(g,k,\mathbf{x}_n(w))$ (assuming $m(w)=0,\ s(w)=1$) is

$$\mathbf{x}_{max}(g,k,m_{+g},s^2_{+g},m_{-k},s^2_{-k})=$$
$$\frac{n_gm_{+g}-km_{-k}}{n_1}+\sqrt{c\left(\frac{n(w)+n_gs^2_{+g}-ks^2_{-k}}{n_1}+\frac{c_1m^2_{+g}+c_2m^2_{-k}+c_3m_{+g}m_{-k}}{n_1^2}\right)}, \quad (6.5)$$

where $n_g = |\{w_j^* = 1 : \ j \in \{n+1, \ldots, n+g\}\}|$ is the number of points added during the algorithm,

$$n_1 = n(w) + n_g - k, \ c_1 = (n(w) - k)n_g, \ c_2 = -(n(w) + n_g)k, \ c_3 = 2kn_g.$$

Further define for $g, k \geq 0$

$$\begin{aligned} x_{maxmax}(g,k) &= \max_{(m_{+g},s^2_{+g},m_{-k},s^2_{-k})\in S_{gk}(w)} \mathbf{x}_{max}(g,k,m_{+g},s^2_{+g},m_{-k},s^2_{-k}), \\ x_{maxmin}(g,k) &= \min_{(m_{+g},s^2_{+g},m_{-k},s^2_{-k})\in S_{gk}(w)} \mathbf{x}_{max}(g,k,m_{+g},s^2_{+g},m_{-k},s^2_{-k}). \end{aligned}$$

Note that $x_{maxmin}(g,k) \leq \sqrt{c} \leq x_{maxmax}(g,k)$, because $(0,0,0,0) \in S_{gk}(w)$. $x_{maxmax}(g,k)$ is nondecreasing in g, because points can always be added far away that they do not affect the FPC, and therefore a maximum for smaller g can always be attained for larger g. By analogy, $x_{maxmin}(g,k)$ is non-increasing.

Theorem 6.2 *Let $\mathbf{x}_n(w)$ be an FPC in $\mathbf{x}_n$. Let $\mathbf{x}_{n+g} = \{x_1, \ldots, x_{n+g}\}$. If $\exists k_1, k_2$ with*

$$\begin{aligned} k_1 \ \leq \ & |\mathbf{x}_n \cap \\ & ([m(w) - s(w)x_{maxmax}(g+k_1,k_2), m(w) - s(w)\sqrt{c}] \cup \\ & [m(w) + s(w)\sqrt{c}, m(w) + s(w)x_{maxmax}(g+k_1,k_2)])|, \qquad (6.6) \\ k_2 \ \leq \ & |\mathbf{x}_n \cap \\ & ([m(w) - s(w)\sqrt{c}, m(w) - s(w)x_{maxmin}(g+k_1,k_2)] \cup \\ & [m(w) + s(w)x_{maxmin}(g+k_1,k_2), m(w) + s(w)\sqrt{c}])|, \qquad (6.7) \end{aligned}$$

then

$$\gamma^*(\mathbf{x}(w), E_n^*(\mathbf{x}_{n+g})) \geq \frac{n(w) - k_2}{n(w) + k_1}. \qquad (6.8)$$

If $\frac{n(w)-k_2}{n(w)+k_1} > \frac{1}{2}$, then $\boldsymbol{\Delta}(\mathbf{x}_n(w), \mathbf{x}_n) > \frac{g}{n(w)+g}$.

The proof is given in the appendix. k_1 is the maximum number of points in $\mathbf{x}_n$ outside the FPC $\mathbf{x}_n(w)$ that can be added during the algorithm, k_2 is the maximum number of points inside the FPC $\mathbf{x}_n(w)$ that can be lost during the algorithm due to changes caused by the g new points.

Theorem 6.2 shows the structure of the conditions needed for stability, but in the given form it is not obvious how strong these conditions are (and even not if they are possible to fulfill) for a concrete dataset. It is difficult to evaluate $x_{maxmax}(g+k_1,k_2)$ and $x_{maxmin}(g+k_1,k_2)$ and the conditions (6.6) and (6.7), where k_1 and k_2 also appear on the right hand sides. The following Lemma will give somewhat conservative bounds for $x_{maxmax}(g+k_1,k_2)$ and $x_{maxmin}(g+k_1,k_2)$ which can be evaluated more easily. The conditions (6.6) and (6.7) can then be checked for any given g, k_1 and k_2.

Lemma 6.3 *For $g \geq 0, 0 \leq k < n(w)$:*

$$\begin{aligned} x_{maxmax}(g,k) &\leq x^*_{maxmax}(g,k,m^*_{+g}), && (6.9)\\ x_{maxmin}(g,k) &\geq x^*_{maxmin}(g,k,m^*_{-k}), && (6.10) \end{aligned}$$

where for $0 \leq k < n(w)$

$$\begin{aligned} x^*_{maxmax}(0,k,m_{+g}) &= \sqrt{c},\ \textit{for } g > 0:\\ x^*_{maxmax}(g,k,m_{+g}) &= \frac{gm_{+g} + k\sqrt{c}}{n_1} + \sqrt{c\left(\frac{n(w) + g(a_{max}(g)^2 - m^2_{+g})}{n_1} + \frac{c_1 m^2_{+g}}{n_1^2}\right)},\\ x^*_{maxmin}(g,k,m_{-k}) &= \frac{-gm^*_{+g} - km_{-k}}{n_1} + \\ &\quad \sqrt{c\left(\frac{n(w) - k(c - m^2_{-k})}{n_1} + \frac{c_2 m^2_{-k} - c_3 m_{-k} m^*_{+g}}{n_1^2}\right)},\\ a_{max}(g) &= x^*_{maxmax}(g-1,k,m^*_{+(g-1)}),\\ m^*_{+g} &= \frac{1}{g}\sum_{i=1}^{g} a_{max}(i),\\ m^*_{-k} &= \operatorname*{arg\,min}_{m_{-k} \in [0,\sqrt{c}]} x^*_{maxmin}(g,k,m_{-k}), \end{aligned}$$

The proof is given in the appendix. For the minimization needed to obtain m^*_{-k}, the zeros of the derivative of $x^*_{maxmin}(g,k,m_{-k})$ are the zeros of $tm^2_{-k} + um_{-k} + v$ where

$$\begin{aligned} t &= k^3 + \frac{c^2}{n_1^2} - n_1 ck^2 + 2k^2 c(n(w) + g) - \frac{k^2(n(w)+g)^2 c}{n_1},\\ u &= -\frac{2kgm^*_{+g}}{n_1} + 2k^2 gcm^*_{+g} - \frac{2k^2(n(w)+g)gcm^*_{+g}}{n_1},\\ v &= k^2 n(w) - k^3 c + \frac{(n(w)-k)g(m^*_{+g})^2}{n_1} - \frac{k^2 g^2 c(m^*_{+g})^2}{n_1}. \qquad (6.11) \end{aligned}$$

Isolation robustness of type (a) is not adequate for FPC analysis, because every set of identical points forms an FPC. But a consequence of the theory above is that FPC analysis is isolation robust of type (b).

Theorem 6.4 *FPC analysis is isolation robust of type (b) under the following condition on an FPC $C = \mathbf{x}_n(w)$ with $i(C) > v_m(M_C, g)$:*

$$\begin{aligned} \exists k_2 :\ \frac{|C| - k_2}{|C|} &> \frac{1}{2},\\ k_2 &\leq |T(C) \cap ([-s(w)\sqrt{c}, -s(w)x_{maxmin}(g,k_2)] \cup \\ &\quad [s(w)x_{maxmin}(g,k_2), s(w)\sqrt{c}])|, \qquad (6.12) \end{aligned}$$

where $T(C) = \mathbf{x}_n(w) - m(w)$ is C transformed to mean 0.

Example 6.5 *Consider the standard example dataset of 50 points, see Figure 4. For* $\alpha = 0.99$*, the computation scheme outlined in Section 6.1 finds two FPCs, namely the two NSDs, for* n_{start} *down to 4. Let* $\mathbf{x}_n(w)$ *be the (0,1)-NSD,* $m(w) = 0, s(w)^2 = 0.788$*,* $D(w) = [-2.287, 2.287]$*. The largest point is 1.769, the second largest one is 1.426, the smallest point in the data not belonging to* $\mathbf{x}(w)$ *is 3.231, the second smallest one is 3.574. If* $g = 1$ *point is added,* $s(w)x^*_{maxmax}(1, 0, m^*_{+1}) = 2.600, s(w)x^*_{maxmin}(1, 0, m^*_{-0}) = 2.154$*. Thus, (6.6) and (6.7) hold for* $k_1 = k_2 = 0$*. The same holds for* $g = 2$*:* $s(w)x^*_{maxmax}(2, 0, m^*_{+2}) = 3.000, s(w)x^*_{maxmin}(2, 0, m^*_{-0}) = 2.019$*. For* $g = 3$*:* $s(w)x^*_{maxmax}(3, 0, m^*_{+3}) = 3.528, s(w)x^*_{maxmin}(3, 0, m^*_{-0}) = 1.879$*. This means that (6.6) does not hold for* $k_1 = 0$*, because the smallest point of the (5,1)-NSD would be included into the corresponding FPC.* $g = 3$ *and* $k_1 = 1$ *in Theorem 6.2 correspond to* $g = 4$ *in Lemma 6.3. For* $g = 4$*:* $s(w)x^*_{maxmax}(4, 0, m^*_{+4}) = 4.250, s(w)x^*_{maxmin}(4, 0, m^*_{-0}) = 1.729$*. This means that for* $g = 3$*, neither* $k_1 = 1$*, nor* $k_2 = 0$ *works, and in fact an iteration of (6.3) with added points* $2.286, 2.597, 2.929$ *leads to dissolution, namely to an FPC containing all 50 points of the dataset. Thus,* $\Delta(E_n, \mathbf{x}_n, \mathbf{x}_n(w)) = \frac{3}{28}$*.*

For $\alpha = 0.95$*, there is also an FPC* $\mathbf{x}_n(w_{0.95})$ *corresponding to the (0,1)-NSD, but it only includes 23 points, the two most extreme points on the left and on the right are left out. According to the theory, this FPC is not dissolved by being joined with the (5,1)-NSD, but by implosion. For* $g = 1$*,* $s(w_{0.95})x^*_{maxmax}(1, 0, m^*_{+1}) = 1.643, s(w_{0.95})x^*_{maxmin}(1, 0, m^*_{-0}) = 1.405$*. This means that the points* $-1.426, 1.426$ *can be lost.* $s(w_{0.95})x^*_{maxmax}(1, 2, m^*_{+1}) = 1.855, s(w_{0.95})x^*_{maxmin}(1, 2, m^*_{-2}) = 0.988$*, which indicates that* k_2 *is still too small for (6.7) to hold. Nothing better can be shown than* $\Delta(E_{n,0.05}, \mathbf{x}_n, \mathbf{x}_n(w_{0.05})) \geq \frac{1}{24}$*. However, here the conservativity of the dissolution bound matters (the worst case of the mean and the variance of the two left out points used in the computation of* $x^*_{maxmin}(1, 2, m^*_{-2})$ *cannot be reached at the same time in this example) and dissolution by addition of one (or even two) points seems to be impossible.*

7 SUMMARY AND DISCUSSION

The aim of this paper was to provide a stability theory for cluster analysis that can be applied to general methods for disjoint clustering. Here is a summary of the results concerning the different clustering methods:

- All examined methods with a fixed number of clusters and without trimming (k-means, k-medoids, normal or t-mixture with fixed k, k-number partitions of agglomerative hierarchical clusterings) can be spoiled by adding a single outlier.
- The same holds for the average silhouette width estimation of k and for the mixture model with noise, if the density of the noise component is taken as the uniform density on the convex hull of the data. However, in the latter

case, the outlier(s) that have to be added to the data set to spoil the original clustering have to be extremely and presumably unrealistically large (the same holds for t_1-mixtures).

- Trimmed k-means and the normal mixture with fixed k and a data-independent noise density are not isolation robust (which seems to matter mainly if k is misspecified), but well enough separated clusters in data sets with not too many outliers and a well specified number of clusters are robust against dissolution with these methods.

- Normal and t-mixture models with k estimated by the BIC or the AIC and the h-level partitions of single and complete linkage are isolation robust. In practice, well enough separated clusters will be robust against dissolution with these methods.

In spite of the generality of the definitions given in the present paper, a general quality ranking of the methods by means of the results is not justified. For example, the dissolution result for h-level complete linkage is weak, but seemingly more conservative than the results for other methods. The trimmed k-means is not isolation robust but outperforms at least the isolation robust h-level single linkage in the one-dimensional standard example data set as well as some isolation robust methods in other data sets I have seen. This, however, requires a suitable choice of k. While the theoretical results given in the present paper do not indicate a robustness advantage of complete linkage over single linkage, it seems that such an advantage exists in practice, because isolated clusters can be "chained" by single linkage usually under addition of much fewer points than by complete linkage. More sensible definitions could be needed to capture such differences.

Robustness and stability are not the only requirements of a good clustering. For example, there are many data sets where the density of points is high in a central area of the data space, which might be significantly clustered (though the clusters are not strongly isolated), but the density of points becomes much lower toward the borders of the data region. If single linkage (be it the k-number or the h-level partition) is applied to such data, the solution is often one very large cluster containing all central points and a lot of clusters containing only one or two more or less outlying points. This general structure is then very robust against addition or removal of points (only the exact composition of the outlier clusters changes), but it is not very useful. The most interesting patterns are not revealed. Therefore, the robustness properties should be regarded as one of a number of desirable features for a cluster analysis method. In the literature, lists of such desirable features have been investigated for a long time to assess the quality of different methods, see, e.g., Fisher and van Ness (1971), Chen and van Ness (1994). Differences between cluster stability in concrete data sets and theoretical properties of the methods with respect to idealized situations have already been noted by Milligan (1996).

The mixture models and the agglomerative hierarchical methods have also been applied to the 80-images data set. Single and complete linkage showed the expected robustness behaviour. The addition of a single outlier dissolved a well separated cluster using the k-number partitions, while the h-level partitions with properly chosen h were reasonable and robust.

The add-on package `mclust` for the statistical software R (`www.R-project.org`) for normal mixture modeling with BIC, however, ended up with suboptimal and non-robust solutions because of computational problems. These were seemingly caused by the occurrence of non-invertible covariance matrices during the iterations of the EM-algorithm (the software is described in Fraley and Raftery 2003; other implementations of normal mixtures seem to be sensitive to problems of this kind as well). This illustrates that the practical stability of a clustering algorithm may deviate seriously from the theoretical robustness of the underlying global optimization problem.

Concerning the practical relevance of the results, I have to admit that it was very difficult to find a real data set illustrating at least the most interesting theoretical results given in the present paper. The reason is that the results concern well separated clusters (not only isolation robustness, but also the assumptions of the dissolution point theorems are connected to good separation), while most cluster analysis methods yield at least some not well separated and often very unstable clusters in most real data sets. Therefore, the robustness theory should be complemented with methods to assess the stability of single clusters in a concrete clustering. A publication on using the Jaccard similarity for this task is in preparation (see Hennig 2004c). A graphical method to validate single clusters is introduced in Hennig (2005a). A choice of a cluster analysis method for a particular application has always to depend on the data set and on the aim of the study.

In the robustness literature there are verious definitions of a breakdown point (Hampel 1971, Donoho and Huber 1983). In particular, breakdown (and dissolution) can be defined via addition and replacement of points (deletion is usually not considered, because replacement is clearly stronger). In many situations, addition and replacement are equivalent, see Zuo (2001). Unfortunately, this is not the case in cluster analysis. As a simple example, consider two extremely well separated homogeneous clusters, one with 100 and the other with 10 points. The number of points to be added to lead the smaller cluster into dissolution can be arbitrarily large if an isolation robust method is used. Under replacement, the 10 points of the smaller cluster have simply to be taken into the domain of the other cluster. For single linkage, it is impossible to split a cluster by addition, but it would be possible by replacement. Therefore it would be interesting if replacement based definitions would reveal similar characteristics of the methods. The addition approach has been taken here for the sake of simplicity.

An interesting result is the role of outliers in cluster robustness. Outliers are extremely dangerous for some methods with fixed k, but are completely harmless for mixtures with BIC-estimated k and h-level partitions. It is interesting if this holds for other methods to estimate k (see, e.g., Section 3.5 of Gordon 1999, Celeux and Soromenho 1996, McLachlan 1987). With fixed k, trimmed k-means can handle a moderate number of outliers (unless k is ill-specified) and t-mixtures and normal mixtures with noise are only sensitive to such extreme outliers that they can be easily discarded by a routine inspection of the data (less extreme outliers may be dangerous if there are some of them at the same point). Local instability, caused by points between the clusters or at their borders, seems often to be the more difficult robustness problem in cluster analysis.

APPENDIX: PROOFS

Proof of Theorem 3.6: Assume that $\Delta(E_k, \mathbf{x}_n, C) \leq \frac{g}{|C|+g}$, i.e., it is possible to add g points to $\mathbf{x}_n$ so that C is dissolved. Let $x_{n+1}, \ldots, x_{n+g}$ be the corresponding points, $\mathbf{x}_{n+g}$ be the resulting data set, $\mathcal{D} = E_{k,n+g}(\mathbf{x}_{n+g})$, $\mathcal{D}^* = E^*_{k,n}(\mathbf{x}_{n+g})$.

Let $\mathcal{F}$ be a clustering on $\mathbf{x}_{n+g}$, which is defined as follows: Take the original clusters $C_1, \ldots, C_k$ and add the g minimizing points $y_1, \ldots, y_g \in B_n(E_{k,n}(\mathbf{x}_n))$ of $\sum_{i=1}^g \min_j \|y_i - \bar{x}_j\|_2^2$ to their corresponding clusters C_j. Trim

$$B_{n+g}(\mathcal{F}) = B_n(E_{k,n}(\mathbf{x}_n)) \cup \{x_{n+1}, \ldots, x_{n+g}\} \setminus \{y_1, \ldots, y_g\}.$$

Because maximal g points have been added to any C_j, C with $|C| > g$ is not dissolved in the induced clustering, which equals $\mathcal{F}$, because all added g points have been trimmed. But C is assumed to dissolve. Therefore,

$$Q(\mathbf{x}_{n+g}, \mathcal{D}) < Q(\mathbf{x}_{n+g}, \mathcal{F}).$$

Because $\mathcal{D}$ is a trimmed k-means clustering, $\mathcal{D}^*$ fulfills (3.5), where the centroids are the cluster means of $\mathcal{D}$ (otherwise $\mathcal{D}$ could be improved by changing assignments so that points are assigned to the cluster with the closest centroid and trimmed point are changed into clusters to whose centroid they are closer than some of its former members). A contradiction of (3.6) follows from

$$\begin{aligned} Q(\mathbf{x}_{n+g}, \mathcal{F}) &= \min_{y_1,\ldots,y_g \in B_n(E_{k,n}(\mathbf{x}_n))} \sum_{i=1}^{g} \min_j \|y_i - \bar{x}_j\|_2^2 + Q(\mathbf{x}_n, E_{k,n}(\mathbf{x}_n)), \\ Q(\mathbf{x}_{n+g}, \mathcal{D}) &\geq Q(\mathbf{x}_n, \mathcal{D}^*), \end{aligned}$$

because all summands of $Q(\mathbf{x}_n, \mathcal{D}^*)$ also appear in $Q(\mathbf{x}_{n+g}, \mathcal{D})$.

Proof of Theorem 3.10: Consider $x_{n+1} \to \infty$. For $\mathcal{D} = \{\mathbf{x}_n, \{x_{n+1}\}\}$ get $s(n+1, 2) = 0$ and $s(i, 2) \to 1$ for $i = 1, \ldots, n$, because $a(i, 2)$ does not change while $b(i, 2) \to \infty$. Thus, $\frac{1}{n+1} \sum_{i=1}^{n+1} s(i, k) \to \frac{n}{n+1}$.

Because of the arguments given in Section 3.2, $\{x_{n+1}\}$ will be contained eventually in the optimal clustering for any k. For any partition in which there are nonempty different clusters $C_1 \subset \mathbf{x}_n$ and $C_2 \subset \mathbf{x}_n$, eventually $b(i,k) \leq \max_{x,y\in\mathbf{x}_n} d(x,y)$, where d is the underlying distance, $a(i,k) \geq \min_{x,y\in\mathbf{x}_n} d(x,y) > 0$ as long as x_i does not form a cluster in itself, and therefore there exists a constant c so that $\frac{1}{n+1}\sum_{i=1}^{n+1} s(i,k) < c < \frac{n}{n+1}$. For large enough x_{n+1}, this is worse than $\mathcal{D}$, and therefore $\mathcal{D}$ is the average silhouette width clustering.

Proof of Theorem 4.2: First consider the case without noise component, i.e., $\pi_0 = 0$. Let $C \in E_n(\mathbf{x}_n)$ with isolation $i(C)$, $|E_n(\mathbf{x}_n)| = \hat{k}$. Let $f_{max} = \frac{1}{\sigma_0} f_{0,1}(0)$. Under addition of g points to $\mathbf{x}_n$,

$$BIC(n+g) \geq 2\sum_{i=1}^{n+g} \log\left(\sum_{j=1}^{n+g} \frac{1}{n+g} f_{max}\right) - (3(n+g)-1)\log(n+g). \quad \text{(A.1)}$$

The latter can be attained by fitting $\mathbf{x}_{n+g}$ with the following $n+g$ mixture components:

$$a_j = x_j,\ \sigma_j = \sigma_0,\ \pi_j = \frac{1}{n+g},\ j = 1,\ldots,n+g.$$

If this would be the solution maximizing the BIC, there would be no violation of isolation robustness, because every point would form a cluster, so that there would be no cluster in $E_n^*(\mathbf{x}_{n+g})$ joining points of C and of $\mathbf{x}_n \setminus C$.

Suppose that there exists $D \in E_n^*(\mathbf{x}_{n+g})$ so that neither $D \subseteq C$ nor $D \subseteq \mathbf{x}_n \setminus C$, i.e., $\exists x, y \in \mathbf{x}_n :\ x \in C \cap D,\ y \in (\mathbf{x}_n \setminus C) \cap D$. Thus, $|x-y| \geq i(C)$, and there exists a mixture component l in $\eta^* = \hat{\eta}_{n+g,\hat{k}^*}$ ($\hat{k}^*$ maximizing the BIC for $\mathbf{x}_{n+g}$; the components of η^* being denoted by $\pi_j^*, a_j^*, \sigma_j^*$) so that

$$l = \arg\max_j \pi_j^* f_{a_j^*,\sigma_j^*}(x) = \arg\max_j \pi_j^* f_{a_j^*,\sigma_j^*}(y).$$

By choosing $i(C)$ large enough, at least one of the $f_{a_l^*,\sigma_l^*}(z),\ z = x, y$ can be made arbitrarily small, and therefore $\sum_{j=1}^{\hat{k}^*} \pi_j^* f_{a_j^*,\sigma_j^*}(z)$ and even $L_{n+g,\hat{k}^*}(\eta^*, \mathbf{x}_{n+g})$ can be made arbitrarily small as well. Hence, $i(C)$ can be made so large that $2L_{n+g,\hat{k}^*}(\eta^*, \mathbf{x}_{n+g}) - 3(\hat{k}^*-1)\log(n+g)$ is smaller than the lower bound in (A.1), which contradicts the existence of $D \in E_n^*(\mathbf{x}_{n+g})$ joining points of C and $\mathbf{x}_n \setminus C$. Since $E_n^*(\mathbf{x}_{n+g})$ is a partition, it must contain C or a subset of C.

There exists an upper bound on $\min(f_{a,\sigma}(x), f_{a,\sigma}(y))$, which is independent of a and σ (namely $\max_{\sigma^*\geq\sigma_0} \frac{1}{\sigma^*} f_{0,1}\left(\frac{x-y}{2\sigma^*}\right)$ because $|x-y| \leq 2\max(|x-a|, |y-a|)$) and converges to 0 as $|x-y| \to \infty$. All proportion parameters are ≤ 1, and the number of clusters is smaller or equal to $n+g$ (see Lindsay 1995, p.22). (A.1) is independent of $\mathbf{x}_n$ and C, and therefore the above argument holds for large enough $i(C)$ uniformly over all $\mathbf{x}_n$ and C for given n. This proves isolation robustness.

If a noise component is added, $E_n^*(\mathbf{x}_{n+g})$ is not necessarily a partition, so that the last argument does no longer hold. The former arguments are not affected by

introduction of the noise component. It remains to show that $E_n^*(\mathbf{x}_{n+g})$ contains C or a subset of C, which means that not all members of C are assigned to the noise component. But by choosing $i(C)$ large enough, $\frac{1}{x_{max}-x_{min}}$ becomes arbitrarily small, and assigning even a single point of C to the noise component again can make the loglikelihood arbitrarily small in contradiction to (A.1) with $\pi_0^* = 0$.

Proof of Theorem 5.3: It is well known (see, e.g., Bock 1974, p. 389) that the single linkage $h-$level clusters are the connectivity components of the graph $G(\mathbf{x}_n)$ where all members of the data set are the vertices and there is an edge between x_l and x_m whenever $d_{lm} \leq h$.

Since it is not possible to reduce connectivity components by adding points, $\exists D \in E_n^*(\mathbf{x}_{n+g}) : \ C_i \subseteq D$. Let $q^*(i,g)$ be the right side of (5.7). $q(i,g) = q^*(i,g)$ holds because

- two clusters C_j and C_l can always be linked by adding $g_{(j,l)}$ equidistant points between the points x_j and x_l with $d_{jl} = \delta_S(C_j, C_l)$ because of (5.2); $\sum_{m=1}^{l} |C_{j_m}| - |C_i|$ points can be joined by adding g points if $\{C_{j_1}, \ldots, C_{j_l}\} \in \mathcal{D}_g(C_i)$ because of (5.8), therefore $q(i,g) \geq q^*(i,g)$,
- $q(i,g) \leq q^*(i,g)$ because for all $x, y \in D$ there must be a path P between x and y in $G(\mathbf{x}_{n+g})$, and the minimum set of clusters from $E_{h,n}(\mathbf{x}_n)$ needed to cover $P \cap \mathbf{x}_n$, i.e. the path without the g added points, can obviously be joined by these g points, fulfills (5.8) and is therefore a member of $\mathcal{D}_g(C_i)$.

Get $\gamma(C_i, D) \geq \frac{|C_i|}{|C_i|+q(i,g)}$, therefore (5.5). (5.6) follows directly.

Proof of Theorem 5.6: Suppose that in the induced clustering $E_n^*(\mathbf{x}_{n+g})$ the points of H are not in the same cluster. It will be shown by complete induction over $g \geq 1$ that $\max \delta_C(C_1, C_2) \leq m_g$, where the maximum is taken over $C_1, C_2 \in E_{n+g}(\mathbf{x}_{n+g})$ with $C_1 \cap H \neq \emptyset, \ C_2 \cap H \neq \emptyset$ and that furthermore for such $C_j, \ j = 1, 2$, the largest possible $d_h(H \cap C_j, C_j) \leq m_{g-1}$ and the second largest possible $d_h(H \cap C_j, C_j) \leq m_{g-2}$. If $m_g \leq h$, the clusters C_1, C_2 would be joined in the h-level partition, because all distinct clusters must have distances larger or equal to h from each other.
$g = 1$: $\delta_C(C_1, C_2) \leq d_h(H \cap C_1, C_1) + d_h(H \cap C_2, C_2) + d(H)$, because d is a metric and $d(z_1, z_2) \leq d(z_1, x_1) + d(z_2, x_2) + d(x_1, x_2)$ for $z_1 \in C_1, \ z_2 \in C_2, \ x_1 \in C_1 \cap H, \ x_2 \in C_2 \cap H$. Observe $d(x_{n+1}, H) = \min_{x \in H} d(x_{n+1}, x) \leq d(H)$, because otherwise the points of H would be joined as in the original data set at the level $d(H)$, before x_{n+1} can change anything about H. Points $x \in H$ not being in the same cluster as x_{n+1} can only be joined with $y \in \mathbf{x}_n$ if $d(x, y) < d_h(H, \mathbf{x}_n)$. Thus, one of the $d_h(H \cap C_j, C_j), \ j = 1, 2$ (namely where $x_{n+1} \in C_j$) has to be $\leq \max(d_h(H, \mathbf{x}_n), d(H))$ and the other one has to be $\leq d_h(H, \mathbf{x}_n)$.
$1 \leq g \to g+1$: Order the points $x_{n+1}, x_{n+2}, \ldots$ so that the smaller $d(x_{n+j}, H)$, the smaller the index. Observe still $d(x_{n+1}, H) \leq d(H)$, $d(x_{n+q+1}, H) \leq m_q, \ q \leq g+1$,

the latter because otherwise all clusters containing points of H obtained after addition of x_{n+q} are joined before x_{n+q+1} can affect them. Thus, for $g+1$ added points, m_g is the largest possible value for $d_h(H \cap C_j, C_j),\ j=1,2$, and it can only be reached if x_{n+g+1} is a member of the corresponding cluster. The largest possible $d_h(H \cap C_j, C_j),\ j=1,2$ for $x_{n+g+1} \notin C_j$ can be attained by either one of $x_{n+l} \in C_j,\ l \leq g$ or is $d_h(H, \mathbf{x}_n)$. Observe $d_h(H \cap C_j, C_j) \leq m_{g-1}$ for all these possibilities. This finishes the induction.
This means that all points of H are in the same induced h-level cluster C^* if g points are added and $m_g \leq h$. Observe $\gamma(C, C^*) \geq \frac{|H|}{|C|+q_H} > \frac{1}{2}$, because by (5.9), no more than q_H points outside of C can be joined with H.

Proof of Theorem 6.2: Because of the invariance of FPCs and the equivariance of their domain under linear transformations, assume w.l.o.g. $m(w)=0,\ s(w)=1$.

First it is shown that $\mathbf{x}_{max}(g,k,m_{+g},s^2_{+g},m_{-k},s^2_{-k})$ as defined in (6.5) is the upper border of the domain of $\mathbf{y}_{\tilde{n}+g}(w^*)$ in the situation of Property $A(g,k,\mathbf{x}_n(w))$, i.e.,

$$\mathbf{x}_{max}(g,k,m_{+g},s^2_{+g},m_{-k},s^2_{-k}) = m(w^*) + \sqrt{c}s(w^*). \tag{A.2}$$

Assume, w.l.o.g., that the k points to be lost during the algorithm are $y_1,\ldots,y_k$ and the n_g added points are $y_{\tilde{n}+1},\ldots,y_{\tilde{n}+n_g}$, thus $\mathbf{y}_{\tilde{n}+g}(w^*) = \{y_{k+1},\ldots,y_{\tilde{n}+n_g}\}$, $|\mathbf{y}_{\tilde{n}+g}(w^*)| = n_1$. Now, by straightforward arithmetic:

$$\begin{aligned}
m(w^*) &= \frac{n(w)m(w)+n_gm_{+g}-km_{-k}}{n_1} = \frac{n_gm_{+g}-km_{-k}}{n_1},\\
s(w^*)^2 &= \frac{1}{n_1}\left(\sum_{i=1}^{\tilde{n}}(y_i-m(w^*))^2 + \sum_{w_i^*=1,w_i=0}(y_i-m(w^*))^2 - \sum_{i=1}^{k}(y_i-m(w^*))^2\right)\\
&= \frac{1}{n_1}\left(\sum_{i=1}^{\tilde{n}}\left(y_i-\frac{n_gm_{+g}-km_{-k}}{n_1}\right)^2\right.\\
&\quad + \sum_{w_i^*=1,w_i=0}\left(y_i-m_{+g}+\frac{(n(w)-k)m_{+g}+km_{-k}}{n_1}\right)^2\\
&\quad \left.- \sum_{i=1}^{k}\left(y_i-m_{-k}+\frac{(n(w)+n_g)m_{-k}-n_gm_{+g}}{n_1}\right)^2\right)\\
&= \frac{n(w)s(w)^2+n_gs^2_{+g}-ks^2_{-k}}{n_1}\\
&\quad +\frac{1}{n_1^3}[(n(w)n_g^2+n_g(n(w)-k)^2-kn_g^2)m^2_{+g}\\
&\quad +(n(w)k^2+n_gk^2-(n(w)+n_g)^2k)m^2_{-k}\\
&\quad +(2kn_g(n(w)-k)-2n(w)kn_g+2kn_g(n(w)+n_g))m_{+g}m_{-k}]
\end{aligned}$$

$\Rightarrow$ (A.2).

(6.8) remains to be shown (the bound on Δ then follows directly from Definition 2.2). From (A.2), get that in the situation of Property $A(g, k, \mathbf{x}_n(w))$, the algorithm (6.3), which is known to converge, will always generate FPCs in the new dataset $\mathbf{y}_{\tilde{n}+g}$ with a domain $[x^-, x^+]$, where

$$x^- \in [-x_{maxmax}(g,k), -x_{maxmin}(g,k)],\ x^+ \in [x_{maxmin}(g,k), x_{maxmax}(g,k)], \tag{A.3}$$

if started from $\mathbf{x}_n(w)$. Note that, because of (6.4), the situation that $\mathbf{x}_n(w) \subset \mathbf{x}_n$ is FPC, g points are added to $\mathbf{x}_n$, k_1 further points of $\mathbf{x}_n \setminus \mathbf{x}_n(w)$ are included in the FPC and k_2 points from $\mathbf{x}_n(w)$ are excluded during the algorithm (6.3) is equivalent to the situation of property $A(g + k_1, k_2)$. Compared to $\mathbf{x}(w)$, if g points are added to the dataset and no more than k_1 points lie in $[-x_{maxmax}(g + k_1, k_2), -\sqrt{c}] \cup [\sqrt{c}, x_{maxmax}(g + k_1, k_2)]$, no more than $g + k_1$ points can be added to the original FPC $\mathbf{x}_n(w)$. Only the points of $\mathbf{x}_n(w)$ in $[-\sqrt{c}, -x_{maxmin}(g + k_1, k_2)] \cup [x_{maxmin}(g + k_1, k_2), \sqrt{c}]$ can be lost. Under (6.7), these are no more than k_2 points, and under (6.6), no more than k_1 points of $\mathbf{x}_n$ can be added. The resulting FPC $\mathbf{x}_{n+g}(w^*)$ has in common with the original one at least $n(w) - k_2$ points, and $|\mathbf{x}_n(w) \cup (\mathbf{x}_n \cap \mathbf{x}_{n+g}(w^*))| \leq n(w) + k_1$, which proves (6.8). □

The following proposition is needed to show Lemma 6.3:

Proposition 7.1 *Assume $k < n(w)$. Let $\mathbf{y} = \{x_{n+1}, \ldots, x_{n+g}\}$. In the situation of Property $A(g, k, \mathbf{x}_n(w))$, $m_{+g} \leq m^*_{+g}$ and $\max(\mathbf{y} \cap \mathbf{x}_{n+g}(w^*)) \leq a_{max}(g)$.*

Proof by induction over g.
$g = 1$: $x_{n+1} \leq \sqrt{c}$ is necessary because otherwise the original FPC would not change under (6.3).
$g > 1$: suppose that the proposition holds for all $h < g$, but not for g. There are two potential violations of the proposition, namely $m_{+g} > m^*_{+g}$ and $\max(\mathbf{y} \cap \mathbf{x}_{n+g}(w^*)) > a_{max}(g)$. The latter is not possible, because in previous iterations of (6.3), only $h < g$ points of $\mathbf{y}$ could have been included, and because the proposition holds for h, no point larger than $x^*_{maxmax}(g-1, k, m^*_{+(g-1)})$ can be reached by (6.3). Thus, $m_{+g} > m^*_{+g}$. Let w.l.o.g. $x_{n+1} \leq \ldots \leq x_{n+g}$. There must be $h < g$ so that $x_{n+h} > a_{max}(h)$. But then the same argument as above excludes that x_{n+h} can be reached by (6.3). Thus, $m_{+g} \leq m^*_{+g}$, which proves the proposition. □

Proof of Lemma 6.3: Proof of (6.9): Observe that $\mathbf{x}_{max}(g, k, m_{+g}, s^2_{+g}, m_{-k}, s^2_{-k})$ is enlarged by setting $s^2_{-k} = 0$, $n_g = g$ and by maximizing s^2_{+g}. $s^2_{+g} \leq a_{max}(g)^2 - m^2_{+g}$ because of Proposition 7.1. Because $x_{maxmax}(g, k) \geq \sqrt{c}$, if $\mathbf{x}_{max}(g, k, m_{+g}, s^2_{+g}, m_{-k}, s^2_{-k})$ is maximized in a concrete situation, the points to be left out of $\mathbf{x}_n(w)$ must be the smallest points of $\mathbf{x}_n(w)$. Thus, $-\sqrt{c} \leq m_{-k} \leq m(w) = 0$.

Further, $c_2 \leq 0, c_3 \geq 0$. To enlarge $\mathbf{x}_{max}(g, k, m_{+g}, s^2_{+g}, m_{-k}, s^2_{-k})$, replace the term $-km_{-k}$ in (6.5) by $k\sqrt{c}$, $c_2 m^2_{-k}$ by 0 and $c_3 m_{+g} m_{-k}$ by 0 (if $m_{+g} < 0$ then

$m_{-k} = 0$, because in that case m_{+g} would enlarge the domain of the FPC in both directions and $\mathbf{x}_{n+g}(w^*) \supseteq \mathbf{x}_n(w)$). By this, obtain $x^*_{maxmax}(g, k, m_{+g})$, which is maximized by the maximum possible m_{+g}, namely m^*_{+g} according to Proposition 7.1.

Proof of (6.10): To reduce $\mathbf{x}_{max}(g, k, m_{+g}, s^2_{+g}, m_{-k}, s^2_{-k})$, set $s^2_{+g} = 0$ and observe $s^2_{-k} \leq (c - m^2_{-k})$. The minimizing m_{-k} can be assumed to be positive (if it would be negative, $-m_{-k}$ would yield an even smaller $\mathbf{x}_{max}(g, k, m_{+g}, s^2_{+g}, m_{-k}, s^2_{-k})$). $c_1 \geq 0$, and therefore $n_g m_{+g}$ can be replaced by $-g m^*_{+g}$, $c_1 m^2_{+g}$ can be replaced by 0, and $c_3 m_{-k} m_{+g}$ can be replaced by $-c_3 m_{-k} m^*_{+g}$. This yields (6.10).

Proof of Theorem 6.4: Note first that for one-dimensional data, $T(C)$ can be reconstructed from the distance matrix M_C. If $i(C) > s(w)\mathbf{x}_{maxmax}(g, k_2)$, there are no points in the transformed data set $\mathbf{x}_n - m(w)$ that lie in

$$([-s(w)\mathbf{x}_{maxmax}(g, k_2), -s(w)\sqrt{c}] \cup [s(w)\sqrt{c}, s(w)\mathbf{x}_{maxmax}(g, k_2)]),$$

and it follows from Theorem 6.2 that

$$\exists D \in E^*_n(\mathbf{x}_{n+g}) : \; D \subseteq C, \; \gamma(C, D) \geq \frac{|C| - k_2}{|C|} > \frac{1}{2}.$$

$v_m(M_C, g) = s(w)\mathbf{x}_{maxmax}(g, k_2)$ is finite by Lemma 6.3 and depends on C and $\mathbf{x}_n$ only through $s(w)$ and k_2, which can be determined from M_C.

References

Akaike, H. (1974), "A new look at the statistical identification model", *IEEE Transactions on Automatic Control* 19, pp. 716-723

Banfield, J. D. and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics* 49, pp. 803-821.

Becker, C. and Gather, U. (1999), "The masking breakdown point of multivariate outlier identification rules". *Journal of the American Statistical Association*, 94, pp. 947-955.

Bock, H.-H. (1974), *Automatische Klassifikation.* Vandenhoeck und Ruprecht, Göttingen.

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

Chen, Z. M. and Van Ness, J. W. (1994), "Space-contracting, space-dilating, and positive admissible clustering algorithms", *Pattern Recognition*, 27, pp. 853-857.

Cuesta-Albertos, J. A., Gordaliza, A. and Matran, C. (1997), "Trimmed k-means: An Attempt to Robustify Quantizers", *Annals of Statistics*, 25, 553-576.

Davies, P. L. and Gather, U. (2002), *Breakdown and Groups,* Technical Report 57/2002, SFB 475, Universität Dortmund, `http://wwwstat.mathematik.uni-essen.de/davies/brkdown220902.ps.gz`. To appear in *Annals of Statistics*.

Donoho, D. L. and Huber, P. J. (1983), "The notion of Breakdown point", in Bickel, P. J., Doksum, K. and Hodges jr., J. L. (Eds.): *A Festschrift for Erich L. Lehmann,* Wadsworth, Belmont, CA, pp. 157-184.

Fisher, L. and Van Ness, J. W. (1971), "Admissible Clustering Procedures", *Biometrika*, 58, 91-104.

Fraley, C., and Raftery, A. E. (1998), "How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis", *Computer Journal*, 41, pp. 578-588.

Fraley, C., and Raftery, A. E. (2003), "Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST", *Journal of Classification*, 20, pp. 263-286.

Gallegos, M. T. (2003), Clustering in the Presence of Outliers, in Schwaiger, M. and Opitz, O. (eds.): *Exploratory Data Analysis in Empirical Research,* Springer, Berlin, pp. 58-66.

Garcia-Escudero, L. A., and Gordaliza, A. (1999), "Robustness Properties of k Means and Trimmed k Means", *Journal of the American Statistical Association,* 94, pp. 956-969.

Gordon, A. D. (1999), *Classification* (2nd ed.).Chapman and Hall, Boca Raton.

Gower, J. C. and Legendre, P. (1986), "Metric and Euclidean properties of dissimilarity coefficients", *Journal of Classification,* 3, pp. 5-48.

Hampel, F. R. (1971), "A General Qualitative Definition of Robustness, *Annals of Mathematical Statistics",* 42, pp. 1887-1896.

Hampel, F. R. (1974), "The Influence Function and Its Role in Robust Estimation", *Journal of the American Statistical Association,* 69, pp. 383-393.

Hennig, C. (1997), Fixed Point Clusters and their relation to stochastic models, in Klar, R. and Opitz, O. (eds.): *Classification and knowledge organization,* Springer, Berlin, pp. 20-28.

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

Hennig, C. (2003), Clusters, outliers, and regression: fixed point clusters, *Journal of Multivariate Analysis,* 86, pp. 183-212.

Hennig, C. (2004a), "Breakdown points for ML estimators of location-scale mixtures", *Annals of Statistics,* 32, pp. 1313-1340.

Hennig, C. (2004b), "Robustness of ML Estimators of Location-Scale Mixtures", in Baier, D. and Wernecke, K.-D. (eds.): *Innovations in Classification, Data Science, and Information Systems.* Springer, Heidelberg, pp. 128-137.

Hennig, C. (2004c), *A general robustness and stability theory for cluster analysis.* Preprint no. 2004-07, Universität Hamburg, Fachbereich Mathematik - SPST `www.homepages.ucl.ac.uk/~ucakche/papers/classbrd.ps`

Hennig, C. (2005a), "A method for visual cluster validation", in Weihs, C. and Gaul, W. (eds.): *Classification - The Ubiquitous Challenge.* Springer, Heidelberg, pp. 153-160.

Hennig, C. (2005b), "Fuzzy and Crisp Mahalanobis Fixed Point Clusters", in Baier, D., Decker, R., and Schmidt-Thieme, L. (eds.): *Data Analysis and Decision Support.* Springer, Heidelberg 2005, pp. 47-56.

Hennig, C. and Latecki, L. J. (2003) "The choice of vantage objects for image retrieval", *Pattern Recognition,* 36, pp. 2187-2196.

Hennig, C. and Christlieb, N. (2002), Validating visual clusters in large datasets: fixed point clusters of spectral features, *Computational Statistics and Data Analysis,* 40, pp. 723-739.

Hubert, L. and Arabie, P. (1985), "Comparing Partitions", *Journal of Classification* 2, pp. 193-218.

Jörnsten, R. (2004), "Clustering and classification based on the data depth", *Journal of Multivariate Analysis* 90, pp. 67-89.

Kaufman, L. and Rousseeuw, P. J. (1990), *Finding Groups in Data.* Wiley, New York.

Kharin, Y. (1996), *Robustness in Statistical Pattern Recognition*, Kluwer Academic Publishers, Dordrecht.

Jaccard, P. (1901), "Distribution de la flore alpine dans la Bassin de Dranses et dans quelques regions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles",* 37, pp. 241-272.

Legendre, P. and Legendre, L. (1998), *Numerical ecology* (2nd ed.). Elsevier, Amsterdam.

Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Hayward.

MacQueen, J. (1967), "Some methods for classification and analysis of multivariate observations", *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 281-297.

McLachlan, G. J. (1987), On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, *Applied Statistics,* 36, pp. 318-324.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, Wiley, New York.

Milligan, G. W. (1996), "Clustering validation: results and implications for applied analyses". In: Arabie, P., Hubert, L. J. and De Soete, G. (eds.): *Clustering and Classification*, World Scientific, Singapore, pp. 341-375.

Milligan, G. W. and Cooper, M. C. (1985), "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, 50, pp. 159-179.

Peel, D. and McLachlan, G. J. (2000), "Robust mixture modeling using the t distribution", *Statistics and Computing,* 10, pp. 335-344.

Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association,* 66, pp. 846-850.

Schwarz, G. (1978), "Estimating the dimension of a model", *Annals of Statistics* 6, pp. 461-464.

Shi, G. R. (1993), "Multivariate data analysis in palaeoecology and palaeobiology - a review", *Palaeogeography, Palaeoclimatology, Palaeoecology,* 105, pp. 199-234.

Zuo, Y. (2001), "Some quantitative relationships between two types of finite sample breakdown point", *Statistics and Probability Letters* 51, pp. 369-375.