Available online at www.sciencedirect.com
ScienceDirect
Journal of Multivariate Analysis 99 (2008) 1154–1176
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

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

Christian Hennig*

*Department of Statistical Science, UCL, Gower St., London WC1e 6BT, UK*

Received 5 June 2006
Available online 19 July 2007

**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 obtained by adding data points, (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 for $k$-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 1991 subject classification:* primary 62F35secondary 62H30

*Keywords:* Breakdown point; Model-based cluster analysis; Mixture model; Trimmed $k$-means; Average silhouette width; Hierarchical cluster analysis

## 1. Introduction

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

* Fax: +44 20 7383 4703.
*E-mail address:* chrish@stats.ucl.ac.uk.

0047-259X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2007.07.002

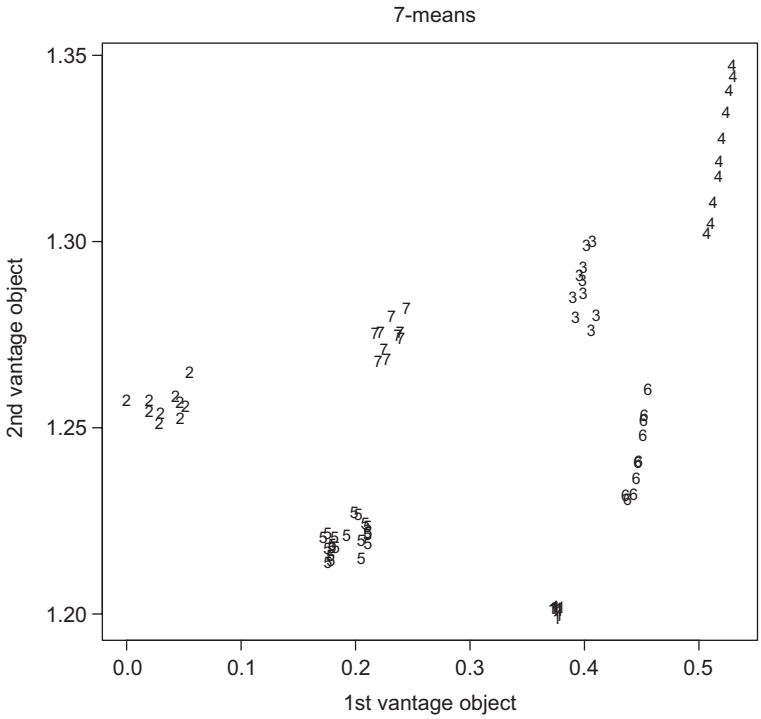


Fig. 1. First two variables of 80-images data set with 7-means clustering.

four-dimensional space by choosing four so-called "vantage objects'' and taking their distances to all objects as variables (see [23] for the full procedure). The images are from eight different scenes, and therefore there is a "true clustering'' (cluster 5 in Fig. 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 Fig. 2, and they are used to illustrate the effect of "realistic'' outliers on a clustering. The clustering in Fig. 2 is further discussed in Section 3.4.

If only one of the outliers shown as "cluster 2'' in Fig. 2 is added to the data set in Fig. 1, 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 four into two parts instead of separating the two scenes underlying cluster 5. With additional outlier, 8-means generates the clustering of Fig. 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

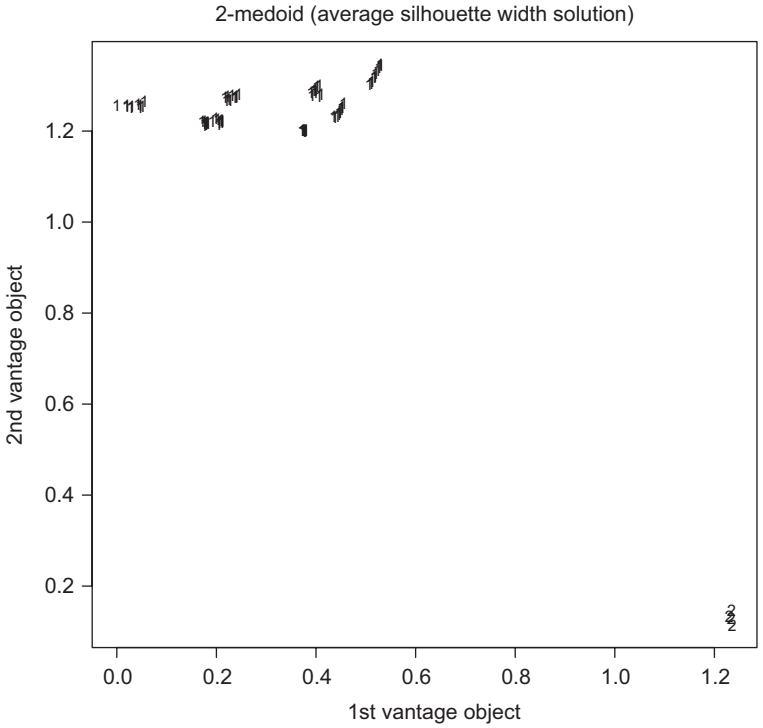


Fig. 2. Same data as in Fig. 1 with four outlying images added and average silhouette width (2-medoid) clustering.

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 ([37], a large number of references is given in [15, Chapter 7], and [34]). Recently, there are also attempts to apply key concepts of robust statistics such as the influence function [17] and the breakdown point [18,9] to certain cluster analysis methods [28,14,13,19]. 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 perturbation

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 [21].

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 [26, Chapter 2], trimmed $k$-means [7, Section 3], mixture models with and without noise and with and without estimation of the number of clusters [11,33, Section 4], single and complete linkage agglomerative clustering (Section 5). The paper is concluded with an overview of the robustness results and some discussion in Section 6.

## 2. Robustness concepts

### 2.1. The dissolution point and a dissimilarity measure between clusters

In [19], 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. [8]. 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 [19]. Note, though, that [19] only proposed the definition and gave some motivation, but no dissolution results are derived in that paper.

A sequence of mappings $E = (E_n)_{n\in\mathbb{N}}$ 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 $i \neq j \leqslant k$. 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 disjunct clustering 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 [39]. The Jaccard coefficient [25] 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 $\gamma^*$ 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 data set 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 $\gamma$ cannot be attained. See [19] for examples of data dependence of the 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

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

as a cutoff value to consider a cluster as dissolved. The definition of the Jaccard coefficient enables a simple interpretation: if $\gamma^*(C, E_n^*(\mathbf{x}_{n+g})) \leqslant \frac{1}{2}$, 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 $|E_n(\mathbf{x}_n)| \geqslant 2$. Let $\mathcal{K} \subseteq \mathbb{N}$ be the set of possible cluster numbers containing at least one element $k \geqslant 2$. Let $\mathcal{F} = \{F \text{ partition on } \mathbf{x}_n : |F| \in \mathcal{K}\}$. Then $\exists \hat{F} \in \mathcal{F} : \gamma^*(C, \hat{F}) \leqslant \frac{1}{2}$. $\frac{1}{2}$ is the smallest value for this to hold independently of $C$ and $E_n(\mathbf{x}_n)$.*

This is equivalent to [19, Lemma 3.3].

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

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

The dissolution point is defined by addition of points to the original data set here, which is not the only possibility. See Section 6 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 [19, Remark 3.5], that at least $r \geqslant 1$ clusters of $E_n(\mathbf{x}_n)$ dissolve if $|E_n(\mathbf{x}_n)| = k,\ |E_n^*(\mathbf{x}_{n+g})| = k - r$.

**Remark 2.4.** In Shi [39], 39 similarity measures between sets are compared. In [19], $\gamma_1(C, D) = \frac{2|C\cap D|}{|C|+|D|}$ 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 [16]. Many of the measures listed in Shi [39] do not fulfill these basic requirements, others are criticized by Shi for stability reasons. See [21] for a further discussion of the choice of the Jaccard coefficient.

The comparison of whole clusterings has been treated, e.g., in [37,24].

### 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.

The definition below is a bit more complicated than the intuitive description above for two reasons. The first reason is 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. Therefore, 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.

The second reason is that the idea of "strong isolation'' does not refer to an absolute value but should be defined dependent on the within-cluster distances.

**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 $n \geqslant m$ 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 $\exists E_n^*(\mathbf{x}_{n+g}) \ni D \subseteq C$.

**Remark 2.6.** It would be possible to define a weaker version of isolation robustness "of degree $\alpha$'' 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.

The dissolution point and isolation robustness are defined in order to take two different points of view. Dissolution point results derive conditions on single clusters in concrete data sets that enable robustness, while isolation robustness delivers a binary classification of clustering 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 [31] will be investigated. These versions have been proposed to robustify the $k$-means approach.

- The $k$-medoids method [27, 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 [7] optimizes the $k$ means criterion after an optimally chosen portion of $\alpha$ 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., [35]). In the present paper, the "average silhouette width'' criterion proposed for $k$-medoids (but applicable to all partitioning techniques) by Kaufman and Rousseeuw [27, Chapter 2] is considered for the $k$-medoids case. This criterion recently became very popular, see, e.g., Jörnsten [26].

**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 $n \geqslant k$ even if "$n \in \mathbb{N}$" 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

$$a(i,k) = \frac{1}{|C_j|-1}\sum_{x \in C_j} d(x_i, x), \quad b(i,k) = \min_{C_l \not\ni x_i} \frac{1}{|C_l|}\sum_{x \in C_l} d(x_i, x).$$

If $|C_j| = 1$, $s(i,k) = 0$.

For $k \geqslant 2$ (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.

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

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.

### *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 Fig. 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 $\alpha$, 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 $\alpha$ (Remark 2.6).

**Example 3.5.** An example can be constructed from the 80-images data. The left side of Fig. 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 Fig. 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 [19]. Garcia-Escudero and Gordaliza [14] 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

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

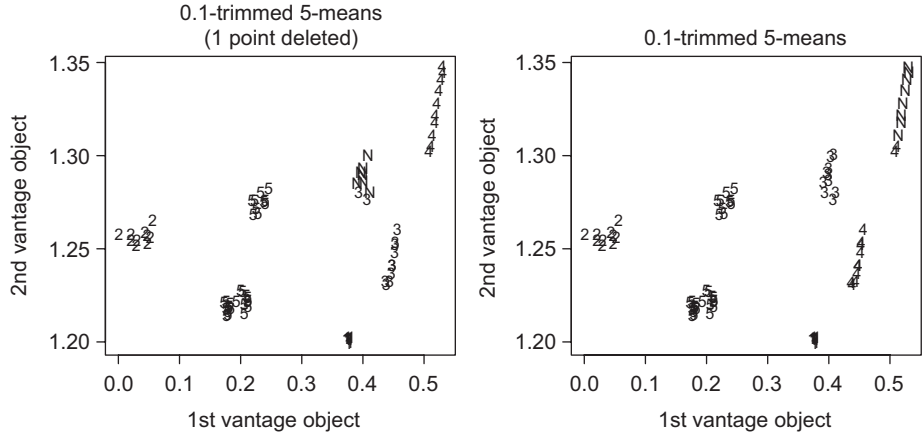


Fig. 3. Left side: 79-images data set (one image of 80-images data has been deleted, which belongs to cluster 5 of the clustering on the right side) with 0.1-trimmed 5-means clustering. Right side: same with the 80-images data set. "N'' denotes trimmed points.

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 $\alpha$-trimmed $k$-means.

**Theorem 3.6.** *Let $n - \lceil n(1-\alpha)\rceil \geqslant g \in \mathbb{N}$, $C \in E_{k,n}(\mathbf{x}_n)$ 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 $l \leqslant k$ clusters so that*

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

*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 $\Delta$ 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 $Q(\mathbf{x}_n, \mathcal{C}^*)$.

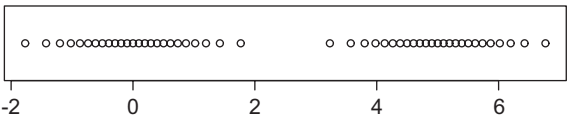


Fig. 4. "Standard'' example data set: 25 points (0,1)-NSD combined with 25 points (5,1)-NSD.

**Example 3.7.** The following definition is used to generate reproducible reference data sets:

**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 $(a, \sigma^2)$-*Normal standard data set* (NSD) with $n$ points, where $\Phi_{a,\sigma^2}$ denotes the cdf of the Normal distribution with parameters $a$, $\sigma^2$.

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 Fig. 4.

Let $k = 2,\ \alpha = 0.1$. The $\alpha$-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 five 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 five 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 five 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 $x_{n+1} \geqslant 24$ ($k$-means) or $x_{n+1} \geqslant 67$ ($k$-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 $\alpha$) seems to be rather robust and yields the solution of Fig. 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 $|C| \leqslant \frac{n}{2}$ 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 non-identical points.

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

In the data set shown in Fig. 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 $f_\theta$ is a density from some parametric family, $\sum_{j=0}^{k} \pi_j = 1,\ 0 \leqslant \pi_j \leqslant 1$ 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 [33].

Having estimated $\eta$ 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:

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

$f_{0,1}$ 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 [33,11] and the $t_\nu$-distribution, which was suggested as a more robust alternative (with $\pi_0 = 0$; [36]). Banfield and Raftery [2] 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, $\hat{u}(x) = \frac{1}{x_{\max,n} - x_{\min,n}} 1(x_{\max,n} \geqslant x \geqslant x_{\min,n})$, 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'' [33] and "mclust'' [12]. 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 $\infty$ if $\hat{a}_1 = x_1$ and $\hat{\sigma}_1 \to 0$, the parameter space has to be restricted if the likelihood is to be maximized (consistent roots of the likelihood equation may be found without restriction). Here, the restriction

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

for some pre-specified $\sigma_0$ is used (for a discussion of the choice of $\sigma_0$, see [20]). An alternative would be to assume all $\sigma_j$ to be equal.

The most frequently used methods to estimate $k$ are the information criteria AIC [1] and BIC [38]. The estimator $\hat{k}$ with BIC (AIC has about the same robustness behavior) is defined as $\hat{k} = \arg\max_k \mathrm{BIC}(k)$, where

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

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 $\eta$ for $\mathbf{x}_n$ under $k$ mixture components.

### 4.2. Robustness results

The robustness of these methods has already been investigated in Hennig [19], 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 [19, Lemma 4.1] 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 (Fig. 4) are 15.2 (classical normal mixture), about 800 ($t_3$-mixture), $3.8 \times 10^6$ ($t_1$-mixture), $3.5 \times 10^7$ (normal mixture with noise). These values depend on $\sigma_0$, 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 [20].

- The above argument does not hold if the noise component $u$ is taken as some non-zero data independent constant (improper density), because in this case the loglikelihood cannot diverge to $-\infty$, see [19, Theorem 4.11]. 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 [19, Theorems 4.13 and 4.16], 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) ($\pi_0$ *estimated or fixed* $= 0$). *Let* $g \in \mathbb{N}$, $C \in E_n(\mathbf{x}_n)$ *with* $|C| > g$. *Consider parameter vectors* $\eta^*$ *for* $1 \leqslant k^* \leqslant n$ *mixture components, so that* $\gamma^*(C, \mathcal{C}^*) \leqslant \frac{1}{2}$ *for the corresponding clustering* $\mathcal{C}^*$. *If for any such* $\eta^*$

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

*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 [19].

Unfortunately, Theorem 4.1 is not as useful as Theorem 3.6, because the optimization of (4.7) over all possible $\eta^*$ seems computationally intractable. Empirically, in the standard example data set of Fig. 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* ($\pi_0$ *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 Appendix A.

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 $\sigma_0$) 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 $\sigma_0$ 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 $\mathbf{D} = (d_{ij})_{i,j=1,\ldots,n}$, $d_{ij} = d(x_i, x_j)$, 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

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

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

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

For simplicity, it is also assumed that the non-zero distances are pairwise distinct. I will restrict considerations to the single linkage and the complete linkage method (see [15, Chapter 4], 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 $k = n-1, \ldots, 1$:

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

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

$\mathcal{C} = \bigcup_{k=1}^n \mathcal{C}_k$ 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 [15, 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 $n \geqslant k \in \mathbb{N}$, and $\mathcal{C}_{k(h)}$ with $h_{k(h)} \leqslant h$ and $h_{k(h)-1} > h$ is called the *h-level partition* for given $h \geqslant 0$.

### 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 $k \geqslant 2$, 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,*

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

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

*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*

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

The proof is given in Appendix A.

**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 Fig. 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 $C$), $d_h(C, D) = \max_{x\in D\setminus C, y\in C, d(x,y)\leqslant h} d(x, y)$ (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 $h^* = d(H)\leqslant h$) 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 $g\geqslant 2$. If $m_g \leqslant h$ and if*

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

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

$H$ may be chosen in order 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 Fig. 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 $\geqslant 13$ 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 $\Delta(E_h, \mathbf{x}_n, C_1)\geqslant \frac{2}{27}$.

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 $h\geqslant 8$, so that the result of the theorem is extremely conservative here. Fig. 5 shows data where the dissolution point bound obtained in the theorem is attained.

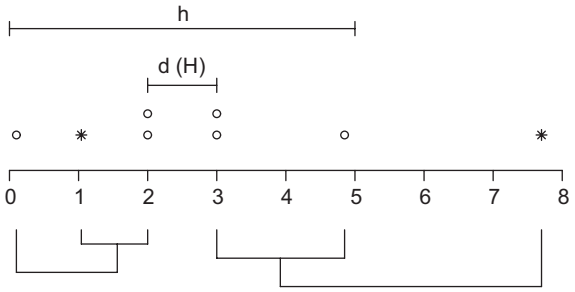


Fig. 5. Let (0.1, 2, 2, 3, 3, 4.85) (circles) form a well separated complete linkage cluster ($4.8 < h < 7.6$) in a data set. Let $H = \{2, 2, 3, 3\}$. Thus, $d(H) = 1$, $d_h(H, \mathbf{x}_n) = m_0 = 1.9$, $m_1 = 4.8$, $m_2 = 7.7$, $q_H = 0$. Therefore, $\Delta = \frac{2}{8}$. Dissolution works by adding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

## 6. 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.

Results of practical relevance are the non-robustness of the average silhouette width to estimate the number of clusters, the finding that fixing the level to cut the tree is much better than fixing the number of clusters for agglomerative hierarchical methods, and that methods to estimate the number of clusters should make it possible to classify added extreme outliers as one-point-clusters without changing the clustering of the other points, at least if such points are not trimmed or assigned to noise components.

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., [10,6]. 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 [34].

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 behavior. 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 [11]; 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 [21]). A graphical method to validate single clusters is introduced in Hennig [22]. 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 various definitions of a breakdown point [17,9]. 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 [40]. 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., [15,5,32, Section 3.5]). 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 A: Proofs

**Proof of Theorem 3.6.** Assume that $\Delta(E_k, \mathbf{x}_n, C) \leqslant \frac{g}{|C|+g}$, 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 points are changed into clusters to whose centroid they are closer than some of its former members). A contradiction of (3.6) follows from

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

$$Q(\mathbf{x}_{n+g}, \mathcal{D}) \geqslant Q(\mathbf{x}_n, \mathcal{D}^*),$$

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 non-empty different clusters $C_1 \subset \mathbf{x}_n$ and $C_2 \subset \mathbf{x}_n$, eventually $b(i, k) \leqslant \max_{x,y \in \mathbf{x}_n} d(x, y)$, where $d$ is the underlying distance, $a(i, k) \geqslant \min_{x,y \in \mathbf{x}_n} d(x, y) > 0$ 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 $\mathbf{x}_n$,

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

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

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

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, $|x-y| \geqslant i(C)$, 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 $\eta^*$ 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 $\sigma$ (namely $\max_{\sigma^* \geqslant \sigma_0} \frac{1}{\sigma^*} f_{0,1}\left(\frac{x-y}{2\sigma^*}\right)$ because $|x-y| \leqslant 2\max(|x-a|, |y-a|)$) and converges to 0 as $|x-y| \to \infty$. All proportion parameters are $\leqslant 1$, and the number of clusters is smaller or equal to $n+g$ (see [30, 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 argument guaranteeing the existence of $C \supseteq D \in E_n^*(\mathbf{x}_{n+g})$ 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 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., [4, 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 $d_{lm} \leqslant h$.

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 $q(i,g) \geqslant q^*(i,g)$,
- $q(i,g) \leqslant q^*(i,g)$ 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 $\gamma(C_i, D) \geqslant \frac{|C_i|}{|C_i|+q(i,g)}$, 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 $g \geqslant 1$ that $\max \delta_C(C_1, C_2) \leqslant m_g$, 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 $d_h(H \cap C_j, C_j) \leqslant m_{g-1}$ and the second largest possible $d_h(H \cap C_j, C_j) \leqslant m_{g-2}$. If $m_g \leqslant h$, 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.

$g = 1$: $\delta_C(C_1, C_2) \leqslant d_h(H \cap C_1, C_1) + d_h(H \cap C_2, C_2) + d(H)$, because $d$ is a metric and $d(z_1, z_2) \leqslant d(z_1, x_1) + d(z_2, x_2) + d(x_1, x_2)$ 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 $d(x_{n+1}, H) = \min_{x \in H} d(x_{n+1}, x) \leqslant d(H)$, 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 $\leqslant \max(d_h(H, \mathbf{x}_n), d(H))$ and the other one has to be $\leqslant d_h(H, \mathbf{x}_n)$.

$1 \leqslant g \to g + 1$: 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 $d(x_{n+1}, H) \leqslant d(H)$, $d(x_{n+q+1}, H) \leqslant m_q$, $q \leqslant g + 1$, 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 $x_{n+l} \in C_j$, $l \leqslant g$ or is $d_h(H, \mathbf{x}_n)$. Observe $d_h(H \cap C_j, C_j) \leqslant m_{g-1}$ 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 $m_g \leqslant h$. Observe $\gamma(C, C^*) \geqslant \frac{|H|}{|C|+q_H} > \frac{1}{2}$, because by (5.9) no more than $q_H$ points outside of $C$ can be joined with $H$. □

## References


[1] H. Akaike, A new look at the statistical identification model, IEEE Trans. Automat. Control 19 (1974) 716–723.

[2] J.D. Banfield, A.E. Raftery, Model-based Gaussian and non-Gaussian clustering, Biometrics 49 (1993) 803–821.

[4] H.-H. Bock, Automatische Klassifikation, Vandenhoeck und Ruprecht, Göttingen, 1974.

[5] G. Celeux, G. Soromenho, An entropy criterion for assessing the number of clusters in a mixture, J. Classification 13 (1996) 195–212.

[6] Z.M. Chen, J.W. Van Ness, Space-contracting, space-dilating, and positive admissible clustering algorithms, Pattern Recognition 27 (1994) 853–857.

[7] J.A. Cuesta-Albertos, A. Gordaliza, C. Matran, Trimmed $k$-means: an attempt to robustify quantizers, Annals of Statistics 25 (1997) 553–576.

[8] P.L. Davies, U. Gather, Breakdown and groups, Ann. Stat. 33 (2005) 977–1035.
[9] D.L. Donoho, P.J. Huber, The notion of breakdown point, in: P.J. Bickel, K. Doksum, J.L. Hodges Jr. (Eds.), A Festschrift for Erich L. Lehmann, Wadsworth, Belmont, CA, 1983, pp. 157–184.
[10] L. Fisher, J.W. Van Ness, Admissible clustering procedures, Biometrika 58 (1971) 91–104.
[11] C. Fraley, A.E. Raftery, How many clusters? Which clustering method? Answers via model based cluster analysis, Comput. J. 41 (1998) 578–588.
[12] C. Fraley, A.E. Raftery, Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST, J. Classification 20 (2003) 263–286.
[13] M.T. Gallegos, Clustering in the presence of outliers, in: M. Schwaiger, O. Opitz (Eds.), Exploratory Data Analysis in Empirical Research, Springer, Berlin, 2003, pp. 58–66.
[14] L.A. Garcia-Escudero, A. Gordaliza, Robustness properties of $k$ means and trimmed $k$ means, J. Amer. Statist. Assoc. 94 (1999) 956–969.
[15] A.D. Gordon, Classification, 2nd ed., Chapman and Hall, Boca Raton, FL, 1999.
[16] J.C. Gower, P. Legendre, Metric and Euclidean properties of dissimilarity coefficients, J. Classification 3 (1986) 5–48.
[17] F.R. Hampel, A general qualitative definition of robustness, Ann. Math. Statist. 42 (1971) 1887–1896.
[18] F.R. Hampel, The influence function and its role in robust estimation, J. Amer. Statist. Assoc. 69 (1974) 383–393.
[19] C. Hennig, Breakdown points for ML estimators of location-scale mixtures, Ann. Statist. 32 (2004) 1313–1340.
[20] C. Hennig, Robustness of ML estimators of location-scale mixtures, in: D. Baier, K.-D. Wernecke (Eds.), Innovations in Classification, Data Science, and Information Systems, Springer, Heidelberg, 2004, pp. 128–137.
[21] C. Hennig, A general robustness and stability theory for cluster analysis. Preprint no. 2004-07, Universität Hamburg, Fachbereich Mathematik—SPST, 2004. ⟨www.homepages.ucl.ac.uk/∼ucakche/papers/classbrd.ps⟩;
C. Hennig, Cluster-wise assessment of cluster stability, Comput. Stat. Data An. (2006), in press, doi:10.1016/j.csda.2006.11.025.
[22] C. Hennig, A method for visual cluster validation, in: C. Weihs, W. Gaul (Eds.), Classification—The Ubiquitous Challenge, Springer, Heidelberg, 2005, pp. 153–160.
[23] C. Hennig, L.J. Latecki, The choice of vantage objects for image retrieval, Pattern Recognition 36 (2003) 2187–2196.
[24] L. Hubert, P. Arabie, Comparing partitions, J. Classification 2 (1985) 193–218.
[25] P. Jaccard, Distribution de la flore alpine dans la Bassin de Dranses et dans quelques regions voisines, Bull. Soc. Vaudoise Sci. Naturelles 37 (1901) 241–272.
[26] R. Jörnsten, Clustering and classification based on the data depth, J. Multivariate Anal. 90 (2004) 67–89.
[27] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data, Wiley, New York, 1990.
[28] Y. Kharin, Robustness in Statistical Pattern Recognition, Kluwer Academic Publishers, Dordrecht, 1996.
[29] P. Legendre, L. Legendre, Numerical Ecology, second ed., Elsevier, Amsterdam, 1998.
[30] B.G. Lindsay, Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, Hayward, 1995.
[31] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.
[32] G.J. McLachlan, On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, Appl. Statist. 36 (1987) 318–324.
[33] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[34] G.W. Milligan, Clustering validation: results and implications for applied analyses, in: P. Arabie, L.J. Hubert, G. De Soete (Eds.), Clustering and Classification, World Scientific, Singapore, 1996, pp. 341–375.
[35] G.W. Milligan, M.C. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (1985) 159–179.
[36] D. Peel, G.J. McLachlan, Robust mixture modeling using the $t$ distribution, Statist. Comput. 10 (2000) 335–344.
[37] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Amer. Statist. Assoc. 66 (1971) 846–850.
[38] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[39] G.R. Shi, Multivariate data analysis in palaeoecology and palaeobiology—a review, Palaeogeography, Palaeoclimatology, Palaeoecology 105 (1993) 199–234.
[40] Y. Zuo, Some quantitative relationships between two types of finite sample breakdown point, Statist. Probab. Lett. 51 (2001) 369–375.