Dissolution and isolation robustness of fixed point clusters

C. Hennig[1]

Dept. of Statistical Science,
University College London,
Gower St, London WC1E 6BT, United Kingdom

Abstract. The concepts of a dissolution point (which is an adaptation of the breakdown point concept to cluster analysis) and isolation robustness are introduced for general clustering methods, generating possibly overlapping clusterings. Robustness theorems for fixed point clusters (Hennig (2002, 2003, 2005)) are shown.

1 Introduction

Stability and robustness are important issues in cluster analysis. The addition of single outliers to the dataset can change the outcome of several cluster analysis methods completely. In Hennig (2007), a general theory for robustness and stability in cluster analysis has been introduced and discussed in depth. Several examples for lacking robustness in cluster analysis have been given, and the robustness of a wealth of cluster analysis methods including single and complete linkage, k-means, trimmed k-means and normal mixtures with fixed and estimated number of clusters has been investigated.

The present paper applies the theory given there to fixed point clusters (FPC; Hennig (2002, 2003, 2005), Hennig and Christlieb (2002)).

In Section 2, the two robustness concepts dissolution point and isolation robustness are introduced. In Section 3, after fixed point clustering has been introduced, results about the dissolution point and isolation robustness of fixed point clusters are given, which are proven in Section 4.

2 Robustness concepts

2.1 The dissolution point and a dissimilarity measure between clusters

In Hennig (2007), a definition of a breakdown point for a general clustering method has been proposed, of which the definition is based on the assignments of the points to clusters and not on parameters to be estimated. The breakdown point is a classical robustness concept, which measures the minimal amount of contamination of the dataset that suffices to drive an estimator arbitrarily far away from its initial value. There are various breakdown

point definitions in the literature. The present paper makes use of the principle of the sample addition breakdown point (Donoho and Huber (1983)), in which contamination is defined by adding points to the dataset. The concept introduced here deviates somewhat from the traditional meaning of the term "breakdown point", since it attributes "breakdown" to situations, which are not always the worst possible ones. Therefore, the proposed robustness measure is called "dissolution point" in the present paper, even though it is thought to measure a "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".

A sequence of mappings $E = (E_n)_{n\in\mathbb{N}}$ is called a general clustering method (GCM), 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 $\{C_1, \ldots, C_k\}$ 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 $i \neq j$, $x_i = x_j$. k is not assumed to be fixed, and, as opposed to Hennig (2007), clusters are allowed to overlap.

If E is a GCM 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 $C_1^*, \ldots, C_{k^*}^*$ (these clusters have the form $C_i \cap \mathbf{x}_n,\ C_i \in E_{n+g}(\mathbf{x}_{n+g})$, but the notation doesn't necessarily imply $C_i \cap \mathbf{x}_n = C_i^* \forall i$). If E is a partitioning method, $E_n^*(\mathbf{x}_{n+g})$ is a partition as well. k^* may be smaller than k even if E produces k clusters for all n.

It is essential to observe that different clusters of the same clustering on the same data may have a different stability. Thus, it makes sense to define stability with respect to the individual clusters. This 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. For the definition of the dissolution point, the Jaccard similarity between sets is proposed:

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

A similarity between C and a clustering $\hat{E}_n(\mathbf{x}_n)$ is defined by

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

Definition 1. Let $E = (E_n)_{n\in\mathbb{N}}$ be a GCM. 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 choice of the cutoff value $\frac{1}{2}$ is motivated in Hennig (2006).

2.2 Isolation robustness

Results on dissolution points are informative about the methods, but they do not necessarily allow a direct comparison of different methods, because they usually need method-specific assumptions. The concept of isolation robustness is thought to enable such a comparison. The rough idea is that it can be seen as a minimum robustness requirement of 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 definition of isolation robustness in Hennig (2007) applies to partitioning methods only. Fixed point clustering is not a partitioning method, and therefore a new definition, which I call "of type II", is presented here. Let $\mathcal{M}_m$ be the space of distance matrices between m objects permissible by the distance structure underlying the GCM. Call a cluster $C \in E_n(\mathbf{x})$ "fulfilling the isolation condition for a function v_m" (v_m is assumed to map $\mathcal{M}_m \times I\!N \mapsto I\!R$) if $|C| = m$, M_C is its within-cluster distance matrix, $i(C) > v_m(M_C, g)$, and an additional homogeneity condition is fulfilled, which may depend on the GCM.

Definition 2. A GCM $E = (E_n)_{n \in I\!N}$ is called **isolation robust** of type II, if there exists a sequence of functions $v_m,\ m \in I\!N$, such that
for $n \geq m$ for any dataset $\mathbf{x}_n$, for given $g \in I\!N$, for any cluster $C \in E_n(\mathbf{x})$ fulfilling the isolation condition for v_m, and for any dataset $\mathbf{x}_{n+g}$, where g points are added to $\mathbf{x}_n$, there exists $D \in E_n^*(\mathbf{x}_{n+g}) :\ D \subseteq C,\ \gamma(C, D) > \frac{1}{2}$.

The reason that the isolation condition requires a further unspecified homogeneity condition is that extreme isolation of clusters should prevent by no means that a cluster is split up into several smaller clusters by the addition of points. This is not considered by the isolation robustness concept, and therefore a further condition on C has to ensure that this does not happen.

3 Fixed point clusters

3.1 Definition of fixed point clusters

FPC analysis has been introduced as a method for overlapping clustering for clusterwise linear regression (Hennig (2002, 2003)) and normal-shaped clusters of p-dimensional data (Hennig (2005), Hennig and Christlieb (2002)), which should be robust against outliers.

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 a debfinition of outliers with respect to the cluster candidate itself.

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

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

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.

Note that it is not assumed that the whole dataset can be partitioned into clusters of this kind, and therefore this does not necessarily introduce a parametric model for the whole dataset.

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

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

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 (the degrees of freedom of the χ^2-distribution may be adapted in this case).

For combinatorial reasons it is impossible to check (2) for all w. But FPCs can be found by a fixed point algorithm defined by

$$w^{k+1} = \left(1\left[(x_i - m(w^k))'\mathbf{S}(w^k)^{-1}(x_i - m(w^k)) \leq \chi^2_{p;1-\alpha}\right]\right)_{i=1,\ldots,n}. \quad (3)$$

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).

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 (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 (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 $p+$ 2nd 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 1.

Note that Mahalanobis FPCs are invariant under linear transformations.

3.2 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 (intercept only).

$E_n(\mathbf{x}_n)$ is taken as the collection of all data subsets fulfilling (2). $E_n(\mathbf{x}_n)$ is not necessarily 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 $\{(x-m(w))'\mathbf{S}(w)^{-1}(x-m(w)) \leq c\},\ c=\chi^2_{p;1-\alpha}$, (4)

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 (7)) and the number of points belonging to it but close to its border (which is bounded by k_1 in (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 on itself as complete dataset, i.e., on $\mathbf{y}_{\tilde{n}}=\mathbf{x}_n(w)=\mathbf{y}_{\tilde{n}}(1,\ldots,1)$ $(\tilde{n}=n(w))$.
$(m_{+g}, s^2_{+g}, m_{-k}, s^2_{-k})$ has the Property $A(g,k,\mathbf{x}_n(w))$ if there exist points $y_{\tilde{n}+1},\ldots,y_{\tilde{n}+g}$ such that, if the algorithm (3) 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 started from the initial dataset $\mathbf{y}_{\tilde{n}}$, it converges to a new FPC $\mathbf{y}_{\tilde{n}+g}(w^*)$ such that 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., $\mathbf{y}_{\tilde{n}}\setminus\mathbf{y}_{\tilde{n}+g}(w^*)$ (implying $|\mathbf{y}_{\tilde{n}}\setminus\mathbf{y}_{\tilde{n}+g}(w^*)|\leq k$). Mean and variance of 0 points are taken to be 0. Note that always $(0,0,0,0)\in S_{gk}(w)$, because of (4) and the added points can be chosen outside the domain of $\mathbf{y}_{\tilde{n}}$.

In the proof of Theorem 1 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

$$\mathbf{x}_{max}(g,k,m_{+g},s^2_{+g},m_{-k},s^2_{-k})=$$
$$\frac{n_g m_{+g}-k m_{-k}}{n_1}+\sqrt{c\left(\frac{n(w)+n_g s^2_{+g}-k s^2_{-k}}{n_1}+\frac{c_1 m^2_{+g}+c_2 m^2_{-k}+c_3 m_{+g} m_{-k}}{n_1^2}\right)}, \quad (5)$$

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 $g, k \geq 0$

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

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 nonincreasing.

Theorem 1. *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*

$$\begin{aligned} k_1 \leq |\mathbf{x}_n \cap & \\ & ([m(w) - s(w)x_{maxmax}(g + k_1, k_2), m(w) - s(w)\sqrt{c}] \cup \\ & [m(w) + s(w)\sqrt{c}, m(w) + s(w)x_{maxmax}(g + k_1, k_2)])|, \qquad (6) \\ k_2 \leq |\mathbf{x}_n \cap & \\ & ([m(w) - s(w)\sqrt{c}, m(w) - s(w)x_{maxmin}(g + k_1, k_2)] \cup \\ & [m(w) + s(w)x_{maxmin}(g + k_1, k_2), m(w) + s(w)\sqrt{c}])|, \qquad (7) \end{aligned}$$

then

$$\gamma^*(\mathbf{x}(w), E_n^*(\mathbf{x}_{n+g})) \geq \frac{n(w) - k_2}{n(w) + k_1}. \qquad (8)$$

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

The proof is given in Section 4. 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 1 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) and (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 and will be applied in Example 1. The conditions (6) and (7) can then be checked for any given g, k_1 and k_2.

Lemma 1. *For $g \geq 0, 0 \leq k < n(w)$:*

$$x_{maxmax}(g,k) \leq x^*_{maxmax}(g,k,m^*_{+g}), \tag{9}$$
$$x_{maxmin}(g,k) \geq x^*_{maxmin}(g,k,m^*_{-k}), \tag{10}$$

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

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

The proof is given in Section 4. 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

$$t = k^3 + \frac{c^2}{n_1^2} - n_1 ck^2 + 2k^2 c(n(w) + g) - \frac{k^2(n(w)+g)^2 c}{n_1},$$
$$u = -\frac{2kgm^*_{+g}}{n_1} + 2k^2 gcm^*_{+g} - \frac{2k^2(n(w)+g)gcm^*_{+g}}{n_1},$$
$$v = k^2 n(w) - k^3 c + \frac{(n(w)-k)g(m^*_{+g})^2}{n_1} - \frac{k^2 g^2 c(m^*_{+g})^2}{n_1}. \tag{11}$$

Theorem 2. *FPC analysis is isolation robust of type II under the following condition on an FPC $C = \mathbf{x}_n(w)$ with $i(C) > v_m(M_C, g)$:*

$$\exists k_2 : \frac{|C| - k_2}{|C|} > \frac{1}{2},$$
$$k_2 \leq |T(C) \cap ([-s(w)\sqrt{c}, -s(w)x_{maxmin}(g,k_2)] \cup [s(w)x_{maxmin}(g,k_2), s(w)\sqrt{c}])|, \tag{12}$$

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

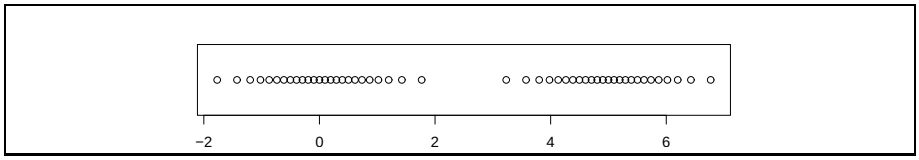

Fig. 1. Example dataset

Example 1. The dataset shown in Figure 1 consists of two datasets of the form $\Phi^{-1}_{a,\sigma^2}(\frac{1}{n+1}),\ldots,\Phi^{-1}_{a,\sigma^2}(\frac{n}{n+1})$ with $n = 25$, $(a,\sigma^2) = (0,1)$, $(5,1)$, respectively. For $\alpha = 0.99$, the computation scheme outlined in Section 3.1 finds two FPCs, namely the two separated initial datasets, for n_{start} down to 4. Let $\mathbf{x}_n(w)$ be the 25 points generated with $(a,\sigma^2) = (0,1)$, $m(w) = 0, s(w)^2 = 0.788$, $D(w) = [-2.287, 2.287]$. 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) and (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) does not hold for $k_1 = 0$, because the smallest point belonging to $(a,\sigma^2) = (5,1)$ would be included into the corresponding FPC. $g = 3$ and $k_1 = 1$ in Theorem 1 correspond to $g = 4$ in Lemma 1. 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 (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 $(a,\sigma^2) = (0,1)$, 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 points corresponding to $(a,\sigma^2) = (5,1)$, 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 (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.

4 Proofs

Proof of Theorem 1: 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 (5) is the upper bound of the domain of $\mathbf{y}_{\tilde{n}+g}(w^*)$ in the situation of Property $A(g, k, \mathbf{x}_n(w))$, i.e.,

$$\mathbf{x}_{max}(g, k, m_{+g}, s^2_{+g}, m_{-k}, s^2_{-k}) = m(w^*) + \sqrt{c}s(w^*). \tag{13}$$

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:

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

This proves (13).

It remains to prove (8), then the bound on Δ follows directly from Definition 1. From (13), get that in the situation of Property $A(g, k, \mathbf{x}_n(w))$, the algorithm (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

$$x^- \in [-x_{maxmax}(g,k), -x_{maxmin}(g,k)],\ x^+ \in [x_{maxmin}(g,k), x_{maxmax}(g,k)], \tag{14}$$

if started from $\mathbf{x}_n(w)$. Note that, because of (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 (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 (7), these are no more than k_2 points, and under (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 (8).

The following proposition is needed to show Lemma 1:

Proposition 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 (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 (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 (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 (3). Thus, $m_{+g} \leq m^*_{+g}$, which proves the proposition.

Proof of Lemma 1: Proof of (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 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 (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 1.

Proof of (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 $-gm^*_{+g}$, $c_1 m^2_{+g}$ can be replaced by 0, and $c_3 m_{-k} m_{+g}$ can be replaced by $-c_3 m_{-k} m^*_{+g}$. This yields (10).

Proof of Theorem 2: 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 dataset $\mathbf{x}_n - m(w)$ 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 1 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 1 and depends on C and $\mathbf{x}_n$ only through $s(w)$ and k_2, which can be determined from M_C.

References

BECKER, C. and GATHER, U. (1999): The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association, 94, 947–955.*

DONOHO, D. L. and HUBER, P. J. (1983): The Notion of Breakdown Point. In: P. J. Bickel, K. Doksum, and J. L. Hodges jr., J. L. (Eds.): *A Festschrift for Erich L. Lehmann*, Wadsworth, Belmont, CA, 157-184.

HENNIG, C. (2002): Fixed Point Clusters for Linear Regression: Computation and Comparison. *Journal of Classification 19, 249–276.*

HENNIG, C. (2003): Clusters, Outliers, and Regression: Fixed Point Clusters. *Journal of Multivariate Analysis, 86, 183–212.*

HENNIG, C. (2005): Fuzzy and Crisp Mahalanobis Fixed Point Clusters. In: D. Baier, R. Decker, and L. Schmidt-Thieme, L. (Eds.): *Data Analysis and Decision Support.* Springer, Heidelberg, 47-56.

HENNIG, C. (2007): Dissolution Point and Isolation Robustness: Robustness Criteria for General Cluster Analysis Methods. *Journal of Multivariate Analysis* (in press), DOI: 10.1016/j.jmva.2007.07.002.

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