A Robust Method for Clustering Analysis

María Teresa Gallegos
Fakultät für Mathematik und Informatik
Universität Passau
D-94030 Passau
Germany

Abstract: We develop a robust clustering method which unites Rousseeuw's minimum covariance determinant method and the determinant criterion of clustering analysis.

Key Words: Cluster Analysis, multivariate data, outliers, robust methods, within–groups covariance matrix, minimum covariance determinant estimator, determinant criterion

AMS 1991 Subject Classification: Primary 62J30; secondary 62F35

1 Introduction

This paper deals at the same time with two well–known problems of multivariate statistics: cluster analysis and robust estimation for multivariate distributions.

The task of the first one is the partition of a set of n realizations of a mixture which has N components in N disjoint subsets or clusters. The elements of the N clusters are then used in order to estimate the parameters of the components of the mixture. The aim of the second one is the detection and/or accommodation of contaminations or spurious elements among a set n of observations which are assumed to come from a (parametric) statistical population. Their previous discard and/or accommodation make possible a clean estimation of the parameters of the generating statistical law.

If the basic statistical population is a mixture consisting of only *one* component there exist many robust methods. For a good overview see [2], Chapter 7. In order to *detect* contaminations in the case where the mixture consists of *more than one* component we extend one of these procedures: Rousseeuw's [17] minimum covariance determinant (MCD). Here, one seeks for a subset k of the n observations whose covariance matrix has lowest determinant. The robust estimate of the mean vector respect. covariance matrix is the sample mean respect. the sample covariance of this subset. With respect to other robust methods such as the M–estimators of Campbell (1980) the MCD estimator has a high breakdown point: A value of $(d+1)^{-1}$ (d is the dimension) against the $\frac{n-k+1}{n}$ of MCD. We recall that the breakdown value of an estimator is the minimum proportion of contamination that an estimator can bear before it succumbs.

On the topic of decomposition of mixtures in absence of outliers many mathematical techniques are already developed. Some of them are Bayesian algorithms for estimating the parameters, cf. Mardia et al. (1979), Chapter 13.2, or the EM algorithm studied in Csiszár–Tusnády (1981). It is also possible to apply clustering algorithms; for an overview see Bock (1996). In this communication and in order to deal with decomposition of mixtures in the presence of outliers we extend the so-called determinant criterion of cluster analysis, cf. [4], Cap.4, §16(a). This method looks for a partition of the set of observations whose pooled within–groups sums of squares and products (SSP) matrix has least determinant.

To our knowledge not much work has been done on the topic of detection of outliers for mixture models. It seems that Kahr [10], and Kahrin and Zhuk [11], [12], and [13] are the pioneers on this research area. Since the method we propose here differs radically from the ones proposed by these authors, we do not describe their work in more detail here, and we refer the interested reader to the literature indicated above.

After giving some general notations and preliminary results in the Subsections 1.1 and 1.2 we formalize in 2.1 the problem described above by specifying the way to follow in order to detect contaminations among observations coming from a mixture. The chosen procedure unites in a suitable way the MCD and the determinant criterion already mentioned above. We look for a subset of k observations out of the whole set of n observations and for a partition of these k points in N clusters whose pooled within–groups sums of squares and products (SSP) matrix has least determinant. Extending the terminology and notation of Rousseeuw we call any such solution a minimum covariance determinant clustering estimator and we denote it by MCDC. In Subsection 2.2, we give a probabilistic formulation of the problem using the mixture model due to Tukey–Huber, cf. [9], together with a suitable extension of the classic clustering setup. We further show that if first, the different components of the mixture are normally distributed with equal covariance matrices and if second, a condition which demands that the mixtures and the outlier population must be sufficiently separated is satisfied then the maximum likelihood estimate for the given statistical model coincides with our MCDC estimator.

The calculation of a MCDC requires the computation of a subset of size k out of the n observations and the subsequent partition of this subset in N clusters. Hence, for most real data sets, algorithms which compute the MCDC are unfeasible and, therefore, algorithms which calculate an approximation to the MCDC are desirable. For his MCD, Rousseeuw [17] has developed such an algorithm. In order to extend this method for our purposes we first develop some technical results in Chapter 3. Although the purpose of these statements is the final derivation of Theorem 4.2 in Chapter 4, much of the given results are of general interest.

The common denominator of all of them is the derivation of sufficient conditions on the traces of two positive definite matrices such that the determinant of one of them is smaller than the corresponding one of the other. With these theoretical tools at hand we are capable to generalize the building block of Rousseeuw's algorithm in Chapter 4 obtaining in this way the basic step of our procedure: the CC–step. Given an arbitrary configuration, a CC–step iteration yields another configuration with lower or equal value of the determinant of the corresponding SSP. The totality of the algorithm is made up of the generation of several initial configurations and, for each one of these, the subsequent iteration of CC–steps until the objective function reaches a stable value. At its end the algorithm picks out that configuration with least SSP determinant. This is the approximation to the MCDC. A detailed description of this algorithm together with some simulation results which show its goodness are given in Chapter 6.

Recall that the breakdown value of an estimator is an indicator of its robustness. Estimators with zero breakdown value are not robust. Hence, this paper would be somehow incomplete if it lacked a word about this topic. Chapter 5 deals with the computation of the breakdown value of the MCDC estimator when the size k of the subset we are looking for equals $k = \lfloor \frac{n+dN+1}{2} \rfloor$ (the MCD estimator, $N = 1$, has its highest possible breakdown value at this size k). It turns out that the breakdown point is positive and that it decreases as the number of clusters increases. The first one of these two facts pleads for the robustness of the algorithm.

Finally, in Chapter 7, we consider the situation where, in addition, each cluster is constrained to have no more than a certain number of observations and we develop an analogous estimator to the MCDC which take into consideration these restrictions.

1.1 General notation and results

Let $\mathbb{N} = \{0, 1, 2, \cdots\}$ be the set of natural numbers and $\mathbb{N}_> = \{1, 2, \cdots\}$ be the set of natural numbers strictly greater than 0. Let $d \in \mathbb{N}_>$. We denote by $\mathbf{I}_d$ the identity matrix from dimension d. Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times d}$ be two matrices. The following definitions or assertions are well–known.

(a) Let $\mathbf{A}$ be symmetric. We write $\mathbf{A} > 0$ if $\mathbf{A}$ is positive definite and $\mathbf{A} \geq 0$ if $\mathbf{A}$ is positive semi–definite.

(b) We denote by trace $\mathbf{A}$ the trace of the matrix $\mathbf{A}$ and by det $\mathbf{A}$ its determinant.

(c) We denote a $d \times d$ diagonal matrix with diagonal entries $d_i, 1 \leq i \leq d$, by $\mathrm{diag}(d_i)_{1 \leq i \leq d}$.

(d) Let $\mathbf{A}$ be symmetric. By the spectral decomposition theorem or Jordan decomposition theorem, the matrix $\mathbf{A}$ can be written as

(1) $\quad \mathbf{A} = \Gamma \Lambda \Gamma,$

where $\mathbf{\Lambda} \in \mathbb{R}^{d \times d} = \mathrm{diag}(\lambda_i)_{1 \leq i \leq d}$ is a diagonal matrix with the ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ of $\mathbf{A}$, and Γ is an orthogonal matrix whose columns are standardized eigenvectors. If $\mathbf{A} > 0$ then define the rational powers

(2) $\quad \mathbf{A}^{r/s} = \mathbf{\Gamma} \mathbf{\Lambda}^{r/s} \mathbf{\Gamma}, \quad \text{where} \quad \mathbf{\Lambda}^{r/s} = \mathrm{diag}(\lambda_i^{r/s})_{1 \leq i \leq d},$

for $s \in \overline{N}$ and $r \in \mathbb{R}$. If $\mathbf{A} > 0$ then (2) holds if the exponents are restricted to be nonnegative.

(e) Let $\mathbf{A} \geq 0$. We have det $\mathbf{A} = \prod_{k=1}^{d} \lambda_k$ and trace $\mathbf{A} = \sum_{k=1}^{d} \lambda_k$. Moreover,

(3) $\quad \det \mathbf{A} \leq (\frac{1}{d} \operatorname{trace} \mathbf{A})^d.$

If $\mathbf{A} > 0$ then we have equality in (3) if and only if all the eigenvalues of this matrix are equal.

(f) Let $\mathbf{A} > 0$ and $\mathbf{B}$ symmetric. We have $\operatorname{trace} \mathbf{A}^{-1}\mathbf{B} = \operatorname{trace} \mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}$ and $\det \mathbf{A}^{-1}\mathbf{B} = \det \mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}}$.

(g) The symbol $N_d(\boldsymbol{\mu}, \mathbf{\Sigma})$ denotes the d–dimensional multinormal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$.

We next give some specific (and well–known) definitions and results concerning cluster analysis which will be used in the sequel.

1.2 Preliminaries

Let $C, D \subseteq \mathbb{R}^d$ be two nonempty, finite sets. Let $\mathcal{C} = \{C_1, C_2, \ldots, C_N\}$ be a partition of the set C in N clusters. (It may be possible that $C_i = \emptyset$ for some $i \in 1..N$). We will often identify $\mathcal{C}$ with the integral interval $1..N$.

(a) Let

$$(4) \quad \mathbf{W}(D) := \sum_{\mathbf{x} \in D} \mathbf{x}\mathbf{x}^t$$

be the matrix of sums and products of the set D.

(b) For all $\mathbf{x} \in C$, let $b_{\mathcal{C}}(\mathbf{x})$ $(\in 1..N)$ be the cluster where $\mathbf{x}$ belongs.

(c) For all $i \in 1..N$ such that $\#C_i > 0$ let

$$(5) \quad \mathbf{m}_{\mathcal{C}}(i) := \frac{1}{\#C_i} \sum_{\mathbf{x} \in C_i} \mathbf{x} \quad (\in \mathbb{R}^d)$$

be the sample mean vector of the i–th cluster and let

$$(6) \quad \mathbf{V}_{\mathcal{C}}(i) := \frac{1}{\#C_i} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(i))(\mathbf{x} - \mathbf{m}_{\mathcal{C}}(i))^t \ (\in \mathbb{R}^{d \times d})$$

be the sample covariance matrix of the i–th cluster.

(d) For all $\mathbf{x} \in C$, let $\mathbf{m}_{\mathcal{C}}(\mathbf{x})$ be the mean of the cluster where $\mathbf{x}$ belongs.

(e) For $\mathbf{m}_1, \dots, \mathbf{m}_N \in \mathbb{R}^d$ let

$$(7) \quad \mathbf{W}(\mathcal{C}, \mathbf{m}_1, \dots, \mathbf{m}_N) := \sum_{i=1}^{N} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \ (\in \mathbb{R}^{d \times d})$$

be the pooled within–groups sums of squares and products (SSP) matrix of the partition $\mathcal{C}$ about $\mathbf{m}_1, \dots, \mathbf{m}_N$. We denote this matrix also by $\mathbf{W}(\mathcal{C}, \mathbf{m})$. Note that $\mathbf{W}(\mathcal{C}, \mathbf{m}_{\mathcal{C}})$ is the pooled within–groups sums of squares and products (SSP) matrix of the partition $\mathcal{C}$. We denote this matrix simply by $\mathbf{W}(\mathcal{C})$.

(f) We have

$$(8) \quad \mathbf{W}(\mathcal{C}) = \sum_{\mathbf{x} \in C} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(\mathbf{x}))^t \mathbf{W}(\mathcal{C})^{-1} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(\mathbf{x})),$$

and

$$(9) \quad \sum_{i=1}^{N} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(i))^t \mathbf{W}(\mathcal{C})^{-1} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(i)) = d.$$

The Equality (8) is plain. In order to see the validity of (9) note that, by Definition (7) we have

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{\mathbf{x}\in C_i}(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i))^t\mathbf{W}(\mathcal{C})^{-1}(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i))\\
=\ &\sum_{i=1}^{N}\sum_{\mathbf{x}\in C_i}\mathrm{trace}(\underbrace{\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i))^t}\,\underbrace{\mathbf{W}(\mathcal{C})^{-1}(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i))}\\
=\ &\mathrm{trace}\,\mathbf{W}(\mathcal{C})^{-1}\sum_{i=1}^{N}\sum_{\mathbf{x}\in C_i}(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i))(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i))^t\\
=\ &\mathrm{trace}\,\mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{C})\\
=\ &\mathrm{trace}\,\mathbf{I}_d = d.
\end{aligned}$$

(g) For all $\mathbf{x}\in\mathbb{R}^d$ and all $i\in 1..N$, define the relative distances

$$(10)\quad d_{\mathcal{C}}(\mathbf{x},i) := \sqrt{(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i))^t\mathbf{W}(\mathcal{C})^{-1}(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i))}.$$

(h) Let $\mathbf{m}_1,\dots,\mathbf{m}_N,\mathbf{c}_1,\dots,\mathbf{c}_N\in\mathbb{R}^d$. We have

$$\begin{aligned}
(11)\qquad &\mathbf{W}(\mathcal{C},\mathbf{c})-\mathbf{W}(\mathcal{C},\mathbf{m})\\
=\ &\sum_{i=1}^{N}\#C_i\left(\mathbf{m}_{\mathcal{C}}(i)(\mathbf{m}_i-\mathbf{c}_i)^t+(\mathbf{m}_i-\mathbf{c}_i)\mathbf{m}_{\mathcal{C}}(i)^t+\mathbf{c}_i\mathbf{c}_i^t-\mathbf{m}_i\mathbf{m}_i^t\right).
\end{aligned}$$

Particularly, for $\mathbf{c}=\mathbf{m}_{\mathcal{C}}$ we obtain

$$(12)\quad \mathbf{W}(\mathcal{C},\mathbf{m}) = \mathbf{W}(\mathcal{C})+\sum_{i=1}^{N}\#C_i(\mathbf{m}_i-\mathbf{m}_{\mathcal{C}}(i))(\mathbf{m}_i-\mathbf{m}_{\mathcal{C}}(i))^t.$$

(i) Finally, let us denote by $\mathcal{P}(C)$ the set of all partitions of the set C in N clusters, and by $\tilde{\mathcal{P}}(C)$ the set of all partitions of the set C in N nonempty clusters.

2 The Problem

2.1 Background

Let $n\in\mathbb{N}_>$. Let $S=\{\mathbf{x}_1,\mathbf{x}_2,\dots,\mathbf{x}_n\}\subseteq\mathbb{R}^d$ be a set of n observations. Let $k\in 1..n$ and let $N\in 1..k$ $(n\geq d+1)$.

The problem (which we denote by (P)) is to find a *subset* C of S with k elements and a *partition* $\mathcal{C}$ of the set C in N clusters such that the pooled within–groups sums of squares and products matrix (SSP) of the partition $\mathbf{W}(\mathcal{C})$ has the lowest determinant, i.e.,

$$\text{(P)} \qquad \min\ \det \mathbf{W}(\mathcal{C}),$$

where this minimum ranges over all pairs $(C, \mathcal{C}) \in \bigcup_{D \in \binom{S}{k}} \{D\} \times \mathcal{P}(D)$. This question unites two problems of statistics: Setting $N = 1$ we obtain Rousseeuw's minimum covariance determinant (MCD) (1998) which is a highly robust estimator of multivariate location and scatter. Its aim is to find k observations (out of n) whose covariance matrix has the lowest determinant. Setting $k = n$ we obtain the well–known problem of multivariate cluster analysis whose aim is to find a partition $\mathcal{S}$ of the observations S in N clusters such that SSP matrix has the lowest determinant. Hence, we are seeking for a robust clustering method.

Following the terminology introduced by Rousseeuw (1998) we call the estimator associated with (P) the *minimum covariance determinant clustering* estimator and denote it by MCDC (the additional C with respect to Rousseeuw's MCD stands for clustering). Therefore, for $N = 1$ the MCDC is Rousseeuw's MCD, and for $k = n$ the MCDC is the determinant criterion of cluster analysis, cf. [4], Cap.4, §16(a). The MCDC estimate of location is defined by the sample mean of the different clusters, whereas the MCDC estimate of scatter is given by the covariance matrix of the different clusters.

Finally, note that (P) may be written as

$$(13) \quad \min_{C \in \binom{S}{k}} \ \min_{\mathcal{C} \in \mathcal{P}(C)} \ \det \mathbf{W}(\mathcal{C}).$$

Moreover, for an arbitrary $C \in \binom{S}{k}$ the minimization problem $\min_{\mathcal{C} \in \mathcal{P}(C)} \ \det \mathbf{W}(\mathcal{C})$ is the determinant criterion for the set of observations C.

The following paragraph extends the usual statistical clustering setup, cf. [15], Section 13.2, to the situation (P).

2.2 A Probabilistic Formulation

Let $f_j, g : (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)) \to \mathbb{R}$, $j \in 1..N$, be probability density functions on $\mathbb{R}^d$. Let $(p_j)_{j\in 1..N}$ be a family of numbers in $]0,1[$ such that $\sum_{j=1}^n p_j \le 1$. Set $q := 1 - \sum_{j=1}^n p_j$. Furthermore, let $X_1, X_2, \ldots, X_n : (\Omega, P) \to \mathbb{R}^d$ be n independent random variables such that

$$(14)\quad X_i \sim \sum_{j=1}^{N} p_j f_j + qg, \qquad i \in 1..n.$$

That is, the statistical population from which the observation $\mathbf{x}_i$, $1 \le i \le n$, arises has the Tukey–Huber form, cf. [9], where q is the probability of appearance of an outlier. It is well known that, if $q = 0$ then multivariate cluster analysis deals with estimation of the prior probabilities p_j, $1 \le j \le n$, and the parameters of the densities f_j, $j \in 1..N$. We now extend the probabilistic formulation of this traditional setup in order to deal with the presence of a contaminating population g, i.e., q is assumed to be strictly positive. For this sake, let $\mathbf{x} := \mathbf{x}_1, \ldots, \mathbf{x}_n$ be n independent realizations of the random variables (14). We assume that k of them arise from one of the subpopulations f_j, $j \in 1..N$, and the remaining $n-k$ come from g. Therefore, the parameter set is $\Theta := \bigcup_{C \in \binom{S}{k}} \{C\} \times \mathcal{P}(C)$, where the element $(C, \mathcal{C})$ defines the subset $C = \{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}\}$ of those k observations belonging to one of the subpopulations f_j, $j \in 1..N$, and $\mathcal{C}$ divides this set C in N clusters C_j, $j \in 1..N$, where the j–th cluster consists of the observations coming from the subpopulation f_j. Hence, each observation $\mathbf{x}_i \in S \setminus C$ comes from the subpopulation g. The likelihood function is

$$(15)\quad L_{\mathbf{x}}(C, \mathcal{C}) = \left[\prod_{j=1}^{N} \prod_{\mathbf{x} \in C_j} f_j(\mathbf{x}) \right] \left[\prod_{\mathbf{x}_i \in S \setminus C} g(\mathbf{x}_i) \right],$$

and each element $(C, \mathcal{C}) \in \operatorname{argmax}_{(C,\mathcal{C})\in\Theta} L_{\mathbf{x}}(C, \mathcal{C})$ is a maximum likelihood estimator of the unknown quantities.

2.3 Proposition (Statistical justification of MCDC)

Let us assume that for all $j \in 1..N$, $f_j \sim N_d(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$, where the mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ and the *common* covariance matrix $\boldsymbol{\Sigma} > 0$ are *unknown*. For each realization $\mathbf{x}$ for which there exists an element $(C, \mathcal{C}) \in \Theta$ such that the implication

$$(16)\quad (C, \mathcal{C}) \in \operatorname{argmax}_{(D,\mathcal{D})\in\Theta} \prod_{j=1}^{N} \prod_{\mathbf{x} \in D_j} f_j(\mathbf{x}\,; \mathbf{m}_{\mathcal{D}}(j), \mathbf{V}_{\mathcal{D}}(j)) \;\Rightarrow\; S \setminus C \in \operatorname{argmax}_{D \in \binom{S}{(n-k)}} \prod_{\mathbf{x}_i \in D} g(\mathbf{x}_i),$$

holds the maximum likelihood estimator is the grouping which minimizes $\det \mathbf{W}(\mathcal{C})$.

Proof. The condition (16) together with (15) and standard operations, cf. [15], pp. 362-3, show that

$$ML(\mathbf{x}) = \mathrm{argmax}_{(C,\mathcal{C})\in\Theta} \prod_{j=1}^{N} \prod_{\mathbf{x}\in C_j} f_j(\mathbf{x}\ ; \mathbf{m}_{\mathcal{C}}(j), \mathbf{V}_{\mathcal{C}}(j)) = \mathrm{argmin}_{(C,\mathcal{C})\in\Theta} \det \mathbf{W}(\mathcal{C}).$$

This is the claim. □

2.4 Remarks

(a) Clearly, the Condition (16) is satisfied if the statistical population f_j, $1 \le j \le n$, and the outlier population g are well separated. If this condition is not met then the given assertion may not be true. Let, for example, $n = 3$, $d = 1$, $N = 1$, and $k = 2$. Moreover, let $x > 0$ be a real number, let $f \sim N_1(0, 1)$, and let $g = \epsilon\delta_x + \epsilon\delta_{-x} + (1 - 2\epsilon)\delta_0$ for a small $\epsilon > 0$. Furthermore, let the observations be given by $x_1 = x$, $x_2 = -x$, and $x_3 = 0$. Clearly, the maximum likelihood estimator is $C = \{-x, x\}$. But the solution of the optimization problem on the left hand side of (16) is $C = \{0, x\}$ or $C = \{0, -x\}$.

(b) Clearly, for each realization $\mathbf{x}$, we have that if $(C, \mathcal{C}) \in ML(\mathbf{x})$ then $\mathcal{C}$ must be a solution of the usual clustering setup with observations in the set C.

In the next section we derive some technical tools which will be used in Section 4 in order to develop algorithms which compute points $(C, \mathcal{C}) \in \Theta$ with small SSP–determinant. Moreover, some of these results are of interest on their own.

3 Technical results

Consider the Preliminaries 1.2. The next lemma is the main building block for our subsequent results. Surprisingly, we have not found it in the literature. We must remark, however, that similar results exist in the technical tools of cluster analysis.

3.1 Lemma

Let $\mathbf{A} > 0$ and $\mathbf{B} \ge 0$.

(a) If

$$\text{(17)}\quad \operatorname{trace}\ \mathbf{A}^{-1}\mathbf{B} \underset{(\leq)}{<} d \qquad \text{or, equivalently,} \qquad \operatorname{trace}\ \mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}} \underset{(\leq)}{<} d$$

then $\det \mathbf{B} \underset{(\leq)}{<} \det \mathbf{A}$.

(b) Let trace $\mathbf{A}^{-1}\mathbf{B} \leq d$. Then $\det \mathbf{A} = \det \mathbf{B}$ if and only if $\mathbf{A} = \mathbf{B}$.

Proof. We prove first (a). Clearly, $\mathbf{A}^{-\frac{1}{2}}\mathbf{B}\mathbf{A}^{-\frac{1}{2}} \geq 0$. Hence, by (3) together with (17), we have

$$(\det \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2})^{1/d} \leq \frac{1}{d} \operatorname{trace} \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} \leq 1.$$

Therefore,

$$1 \geq \det \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} = \det \mathbf{A}^{-1}\mathbf{B} = \frac{\det \mathbf{B}}{\det \mathbf{A}}.$$

This is the first claim. In order to prove the "if" of our second assertion note that

$$\det \mathbf{A} = \det \mathbf{B} \Leftrightarrow \det \mathbf{A}^{-1}\mathbf{B} = \det \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} = 1 \text{ and } \mathbf{B} > 0.$$

Therefore, by (3) applied to the matrix $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} (> 0)$ together with the given hypothesis we obtain first $1 = (\det \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2})^{\frac{1}{d}} \leq \frac{1}{d} \operatorname{trace} \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} \leq 1$. Therefore, we have equality in (3) and all the eigenvalues of the matrix $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$ are equal. This assertion clearly implies the claim. □

We now investigate some consequences of the result above. Clearly, if we set $\mathbf{A} = \mathbf{I}_d$ in 3.1(b) then we obtain a *sufficient* condition that a positive semi–definite matrix $\mathbf{B} \geq 0$ must satisfy in order to be the identity matrix. Another conclusion follows from the observation that 3.1(a) together with the left to right implication in (b) may be written as a suitable optimization problem. We formalize these two facts in the following Corollary.

3.2 Corollary

(a) Let $\mathbf{B} \geq 0$. If trace $\mathbf{B} \leq d$ then $\det \mathbf{B} = 1$ implies $\mathbf{B} = \mathbf{I}_d$.

(b) Let $\mathbf{A} > 0$. The optimization problem

$$\max\ \det \mathbf{B} \quad \text{constrained on}$$

$$\mathbf{B} \geq 0 \ \text{ and } \ \operatorname{trace} \mathbf{A}^{-1}\mathbf{B} \leq d$$

has as unique solution the matrix $\mathbf{A}$.

Let $D \subseteq \mathbb{R}^d$ be nonempty. Letting now in 3.1(a) $\mathbf{B}$ be the matrix of sums and products of the set D, cf. (4), we obtain the following

3.3 Corollary

Let $\mathbf{A} > 0$. Let $D \subseteq \mathbb{R}^d$ be a nonempty set.

(a) If

$$(18)\quad \sum_{\mathbf{x}\in D} \mathbf{x}^t\mathbf{A}^{-1}\mathbf{x} \underset{(\leq)}{<} d$$

then

$$(19)\quad \det \mathbf{W}(D) \underset{(\leq)}{<} \det \mathbf{A}.$$

(b) If in (18) we have $'' \leq''$ then equality holds in (19) if and only if $\mathbf{W}(D) = \mathbf{A}$.

Proof. Let in Lemma 3.1 $\mathbf{B} = \mathbf{W}(D)$. By assumption $\mathbf{A} > 0$ and, clearly, $\mathbf{W}(D) \geq 0$. Therefore,

$$\text{trace }\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} = \sum_{\mathbf{x}\in D} \text{trace } \underbrace{\mathbf{A}^{-1/2}\mathbf{x}}_{\mathbf{A}}\underbrace{\mathbf{x}^t\mathbf{A}^{-1/2}}_{\mathbf{B}} = \sum_{\mathbf{x}\in D} \text{trace } \underbrace{\mathbf{x}^t\mathbf{A}^{-1}\mathbf{x}}_{\text{number}} = \sum_{\mathbf{x}\in D} \mathbf{x}^t\mathbf{A}^{-1}\mathbf{x} \leq d,$$

where the last inequality follows from (18). The claim follows now as a direct application of this lemma. $\square$

Let now $(C, \mathcal{C})$, $\mathbf{W}(\mathcal{C})$, and $\mathbf{m}_\mathcal{C}$ as in the Preliminaries 1.2. We use Lemma 3.1 in order to establish some optimal properties of the SSP matrix of the partition $\mathcal{C}$ compared with the corresponding SSP matrix about N arbitrary points in $\mathbb{R}^d$.

3.4 Corollary

Let $\mathbf{m}_1, \ldots, \mathbf{m}_N \in \mathbb{R}^d$. The following assertions hold.

(a) If $\mathbf{W}(\mathcal{C}) > 0$ then $\mathbf{W}(\mathcal{C}, \mathbf{m}) > 0$.

(b) Let $\mathbf{W}(\mathcal{C}) > 0$. Let $\mathbf{c}_1, \ldots, \mathbf{c}_N \in \mathbb{R}^d$ such that

$$(20)\quad \sum_{i=1}^{N} \#C_i \left(2\mathbf{m}_\mathcal{C}(i)^t\mathbf{W}(\mathcal{C},\mathbf{m})^{-1}(\mathbf{m}_i - \mathbf{c}_i) + \mathbf{c}_i^t\mathbf{W}(\mathcal{C},\mathbf{m})^{-1}\mathbf{c}_i - \mathbf{m}_i^t\mathbf{W}(\mathcal{C},\mathbf{m})^{-1}\mathbf{m}_i\right) \underset{(\leq)}{<} 0.$$

Then

(21) $\det \mathbf{W}(\mathcal{C}, \mathbf{c}) \underset{(\leq)}{<} \det \mathbf{W}(\mathcal{C}, \mathbf{m})$.

Here, we have equality if and only $\mathbf{W}(\mathcal{C}, \mathbf{c}) = \mathbf{W}(\mathcal{C}, \mathbf{m})$.

(c) Furthermore, we have

(22) $\det \mathbf{W}(\mathcal{C}) \leq \det \mathbf{W}(\mathcal{C}, \mathbf{m})$.

We have equality if and only if $\mathbf{W}(\mathcal{C}) = \mathbf{W}(\mathcal{C}, \mathbf{m})$ or, equivalently, if $\mathbf{m}_{\mathcal{C}}(i) = \mathbf{m}(i)$ for all $i \in 1..N$ such that $\#C_i > 0$.

Proof. The Assertion (a) follows from (12) together with the hypothesis $\mathbf{W}(\mathcal{C}) > 0$. In order to prove (b), let in Lemma 3.1 $\mathbf{A} = \mathbf{W}(\mathcal{C}, \mathbf{m})$ and $\mathbf{B} = \mathbf{W}(\mathcal{C}, \mathbf{c})$. By Part (a) we have $\mathbf{A}, \mathbf{B} > 0$. Hence,

$$\begin{aligned}
& \operatorname{trace}\ \mathbf{A}^{-1/2}(\mathbf{B} - \mathbf{A})\mathbf{A}^{-1/2} \\
= & \sum_{i=1}^{N} \#C_i \Big(\operatorname{trace}\ \mathbf{W}(\mathcal{C}, \mathbf{m})^{-1/2} \mathbf{m}_{\mathcal{C}}(i)(\mathbf{m}_i - \mathbf{c}_i)^t \mathbf{W}(\mathcal{C}, \mathbf{m})^{-1/2} \\
& \qquad + \operatorname{trace}\ \mathbf{W}(\mathcal{C}, \mathbf{m})^{-1/2}(\mathbf{m}_i - \mathbf{c}_i)\mathbf{m}_{\mathcal{C}}(i)^t \mathbf{W}(\mathcal{C}, \mathbf{m})^{-1/2} \\
& \qquad + \operatorname{trace}\ \mathbf{W}(\mathcal{C}, \mathbf{m})^{-1/2}(\mathbf{c}_i \mathbf{c}_i^t - \mathbf{m}_i \mathbf{m}_i^t)\mathbf{W}(\mathcal{C}, \mathbf{m})^{-1/2} \Big) \\
= & \sum_{i=1}^{N} \#C_i \left(2\mathbf{m}_{\mathcal{C}}(i)^t \mathbf{W}(\mathcal{C}, \mathbf{m})^{-1}(\mathbf{m}_i - \mathbf{c}_i) + \mathbf{c}_i^t \mathbf{W}(\mathcal{C}, \mathbf{m})^{-1}\mathbf{c}_i - \mathbf{m}_i^t \mathbf{W}(\mathcal{C}, \mathbf{m})^{-1}\mathbf{m}_i \right) \\
< & \ 0.
\end{aligned}$$

Here the last claim follows from (20). Plainly, this last inequality implies trace $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} < d$. Hence, the hypothesis of Lemma 3.1 is satisfied and the asserted inequality follows as a direct application of it.

We finally prove Part (c). If $\det \mathbf{W}(\mathcal{C}) = 0$ then the assertion is trivially satisfied. Suppose now that $\det \mathbf{W}(\mathcal{C}) > 0$. The proof of the first two claims is analogous to the one given above. However, this time one must take into account (12). In order to prove the last claim note that, by (12), the matrix $\mathbf{W}(\mathcal{C})^{-\frac{1}{2}}(\mathbf{W}(\mathcal{C}, \mathbf{m}) - \mathbf{W}(\mathcal{C}))\mathbf{W}(\mathcal{C})^{-\frac{1}{2}}$ is positive semi–definite. Therefore, the trace of it equals 0 if and only if $\mathbf{W}(\mathcal{C}, \mathbf{m}) - \mathbf{W}(\mathcal{C}) = 0$. Assertion which clearly implies the claim. $\square$.

The statement 3.4(c) may be written as follows: Let $\mathbf{W}(\mathcal{C}) > 0$. Each tuple $(\mathbf{m}_1, \ldots, \mathbf{m}_N) \in \mathbb{R}^{dN}$ satisfying $\mathbf{m}_i = \mathbf{m}_\mathcal{C}(i)$ for all $i \in 1..N$ such that $\#C_i > 0$ is a solution of the optimization problem

$$(23) \qquad \min \ \det \mathbf{W}(\mathcal{C}, \mathbf{m}) \quad \text{constrained on} \\ \mathbf{m}_1, \ldots, \mathbf{m}_N \in \mathbb{R}^{dN}.$$

For $N = 1$ this result is well known, cf. [15], Exercises and Complements 1.4.2.

With these results at hand we are now able to prove the main result of this section. It gives a sufficient condition on the configurations $(D, \mathcal{D})$ and $(C, \mathcal{C})$ which assures the reduction of the determinant of the SSP matrix of the first one against the determinant of the SSP matrix of the second one. We need first a definition.

3.5 Definition

Let $\emptyset \neq D \subseteq \mathbb{R}^d$. Associated with each $\mathbf{x} \in D$ let there be a cluster $c_\mathbf{x} \in \mathcal{C}$. Note that we have not necessarily $\mathbf{x} \in c_\mathbf{x}$. Let

$$(24) \quad D_i := \{\mathbf{x} \in D | c_\mathbf{x} = i\}$$

for all $i \in 1..N$ and $\mathcal{D} := \{D_1, \ldots, D_N\}$. Note that the partition $\mathcal{D}$ depends on $\mathcal{C}$.

3.6 Theorem

Consider the Definition 3.5. If

$$(25) \quad \sum_{\mathbf{x} \in D} (\mathbf{x} - \mathbf{m}_\mathcal{C}(c_\mathbf{x}))^t \mathbf{W}(\mathcal{C})^{-1} (\mathbf{x} - \mathbf{m}_\mathcal{C}(c_\mathbf{x})) \underset{(\leq)}{<} d$$

then

$$(26) \quad \det\ \mathbf{W}(\mathcal{D}) \underset{(\leq)}{<} \det\ \mathbf{W}(\mathcal{C}).$$

Moreover, we have equality in (26) if and only if $\mathbf{m}_\mathcal{D}(i) = \mathbf{m}_\mathcal{C}(i)$ for all $i \in 1..N$ such that $\#D_i > 0$ and $\mathbf{W}(\mathcal{D}) = \mathbf{W}(\mathcal{C})$.

Proof. If det $\mathbf{W}(\mathcal{D}) = 0$ then there is nothing to prove. Otherwise, let in Corollary 3.3 $\mathbf{x} := \mathbf{x} - \mathbf{m}_{\mathcal{C}}(c_{\mathbf{x}})$ for all $\mathbf{x} \in D$ and $\mathbf{A} = \mathbf{W}(\mathcal{C})$. We obtain

$$(27) \quad \mathbf{W}(D) = \sum_{\mathbf{x} \in D} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(c_{\mathbf{x}}))(\mathbf{x} - \mathbf{m}_{\mathcal{C}}(c_{\mathbf{x}}))^t = \mathbf{W}(\mathcal{D}, \mathbf{m}_{\mathcal{C}}(1), \ldots, \mathbf{m}_{\mathcal{C}}(N)),$$

where the last equality follows from the definition of D_i, $i \in 1..N$, cf. (24), and $\mathcal{D}$. The Condition (25) is now the Hypothesis (18). Therefore,

$$(28) \quad \det \mathbf{W}(\mathcal{D}, \mathbf{m}_{\mathcal{C}}),$$

with equality if and only if $\mathbf{W}(\mathcal{D}, \mathbf{m}_{\mathcal{C}}) = \mathbf{W}(\mathcal{C})$. Let now in Corollary 3.4 $\mathcal{C} := \mathcal{D}$ and $\mathbf{m}_i := \mathbf{m}_{\mathcal{C}}(i)$ for all $i \in 1..N$. Part (c) implies

$$(29) \quad \det \ \mathbf{W}(\mathcal{D}) \leq \det \ \mathbf{W}(\mathcal{D}, \mathbf{m}_{\mathcal{C}}).$$

Moreover, we have equality (29) if and only if $\mathbf{W}(\mathcal{D}) = \mathbf{W}(\mathcal{D}, \mathbf{m}_{\mathcal{C}})$ or, equivalently, if $\mathbf{m}_{\mathcal{C}}(i) = m_{\mathcal{D}}(i)$ for all $i \in 1..N$ such that $|D_i| > 0$. Therefore, by (28), (29) together with assertions following (29) we have

$$\det \mathbf{W}(\mathcal{D}) \leq \det \mathbf{W}(\mathcal{C}).$$

Here, we have equality if and only if $\mathbf{m}_{\mathcal{C}}(i) = \mathbf{m}_{\mathcal{D}}(i)$ for all $i \in 1..N$ such that $|D_i| > 0$ and $\mathbf{W}(\mathcal{D}) = \mathbf{W}(\mathcal{C})$. □

Therefore, if a set $\{(\mathbf{x}, c_{\mathbf{x}}) | \mathbf{x} \in D, c_{\mathbf{x}} \in 1..N\}$ satisfying (25) is available then it is possible to "construct" a new configuration $(D, \mathcal{D})$ such that its SSP has a lower determinant, cf. 3.5.

Consider now the situation where a configuration $(E, \mathcal{E})$, $E \subseteq \mathbb{R}^d$ and $\mathcal{E} \in \mathcal{P}(E)$, is given. We want to know if the determinant of its SSP is lower than the one corresponding to the configuration $(C, \mathcal{C})$. The following corollary gives a sufficient conditions for this fact to be true.

3.7 Corollary

Let $\emptyset \neq E \subseteq \mathbb{R}^d$. Let $\mathcal{E} = \{E_1, E_2, \ldots, E_N\}$ be a partition of E in N clusters.

(a) If

(30) $\operatorname{trace} \mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{E},\mathbf{m}_{\mathcal{C}}) \underset{(\leq)}{<} d$ or, equivalently,

(31) $\displaystyle\sum_{i=1}^{N}\sum_{x\in E_i}(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i))^t\mathbf{W}(\mathcal{C})^{-1}(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(i)) \underset{(\leq)}{<} d$

then $\det \mathbf{W}(\mathcal{E}) \underset{(\leq)}{<} \det \mathbf{W}(\mathcal{C})$. Equality holds if and only if $\mathbf{m}_{\mathcal{E}}(i) = \mathbf{m}_{\mathcal{C}}(i)$ for all $i \in 1..N$ such that $|E_i| > 0$ and $\mathbf{W}(\mathcal{E}) = \mathbf{W}(\mathcal{C})$.

(b) Let $\mathbf{W}(\mathcal{E}) > 0$. Each solution of the optimization problem

(32) $\min \ \det \mathbf{W}(\mathcal{C})$ constrained on

$$(\mathbf{m}_{\mathcal{C}},\mathbf{W}(\mathcal{C})) \in \mathbb{R}^{dN}\times\mathbb{R}^{d\times d} \text{ such that } \mathbf{W}(\mathcal{C}) > 0 \text{ and } \operatorname{trace} \mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{E},\mathbf{m}_{\mathcal{C}}) \leq d$$

satisfies $\mathbf{W}(\mathcal{C}) = \mathbf{W}(\mathcal{E})$ and $\mathbf{m}_{\mathcal{C}}(i) = \mathbf{m}_{\mathcal{E}}(i)$ for all $i \in 1..N$ such that $\#E_i > 0$. Moreover, if $\#E_i > 0$ for all $i \in 1..N$ then the unique solution of (32) is $(\mathbf{m}_{\mathcal{E}},\mathbf{W}(\mathcal{E}))$.

Proof. In order to prove Part (a), let in Definition 3.5 $D := E$ and $c_{\mathbf{x}} := b_{\mathcal{E}}(\mathbf{x})$, cf. 1.2(b), $\mathbf{x} \in E$. Clearly, (31) is the hypothesis (25) of Theorem 3.6 for the situation considered and the claim follows as a direct application of this theorem. The Assertion (b) is a direct consequence of Part (a). □

3.8 Remarks

(a) By Corollary 3.2, the conditions (i) $\operatorname{trace} \mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{D}) \leq d$ and (ii) $\det \mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{D}) = 1$ assure equality between $\mathbf{W}(\mathcal{D})$ and $\mathbf{W}(\mathcal{C})$. However, these conditions do not imply equality between the mean vectors $\mathbf{m}_{\mathcal{C}}$ and $\mathbf{m}_{\mathcal{D}}$. Let, for example, $d = 1$, $N = 1$, $C = \{-1, 1\}$ and $D = \{1, 3\}$. The Corollary 3.7 gives us the (additional) condition to be satisfied in order to have (also) equality between the mean vectors, namely (iii) $\operatorname{trace} \mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{D}, m_{\mathcal{C}}) \leq d$ and (iv) $\det \mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{D}) = 1$. It is easy to see that (iii) implies (i). Indeed, using (12), we obtain

$$\begin{aligned}
& \operatorname{trace} \mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{D},\mathbf{m}_{\mathcal{C}}) - \operatorname{trace} \mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{D}) \\
= & \sum_{i=1}^{N} \#D_i(\mathbf{m}_{\mathcal{C}}(i)-\mathbf{m}_{\mathcal{D}}(i))^t\mathbf{W}(\mathcal{C})^{-1}(\mathbf{m}_{\mathcal{C}}(i)-\mathbf{m}_{\mathcal{D}}(i)) \\
\geq & \ 0.
\end{aligned}$$

(b) Note that the optimization problem (32) may be written as

$$\text{(33)} \qquad \min \ \det \mathbf{W}(\mathcal{C}) \ \text{ constrained on}$$
$$(C, \mathcal{C}) \ \text{ such that } \ C \subseteq \mathbb{R}^d, \mathcal{C} \in \mathcal{P}(C), \mathbf{W}(\mathcal{C}) > 0 \text{ and trace}\, \mathbf{W}(\mathcal{C})^{-1}\mathbf{W}(\mathcal{E}, \mathbf{m}_{\mathcal{C}}) \leq d.$$

(c) For $N = 1$ the Statement 3.7 (b) follows from [8], Theorem on p. 50.

4 The CC–step as an extension of the C–step

Consider the Background 2.1. In this section, we apply the theory of the previous section in order to develop some algorithms which compute "good" approximations for the MCDC. For his MCD, Rousseeuw [17] has developed such a method. Here, we generalize this algorithm in the situation (P). All of our methods have as (theoretical) principle the idea suggested by Inequality (25) (previous adaptation to the situation considered). Namely, the use of the trace in order to obtain information about the determinant. It is worthwhile to remark (again) that this key idea is a generalization of the one invented by Rousseeuw.

The following paragraphs fit Definition 3.5 and Theorem 3.6 for our purposes.

4.1 Definition

Let $\emptyset \neq C_{\text{new}} \subseteq S$ such that $\# C_{\text{new}} = k$. Associated with each $\mathbf{x} \in C_{\text{new}}$ let there be a cluster $c_{\mathbf{x}} \in \mathcal{C}$. For all $i \in 1..N$, define

$$\text{(34)} \quad C_{\text{new},i} := \{\mathbf{x} \in C_{\text{new}} | c_{\mathbf{x}} = i\}.$$

Let further $\mathcal{C}_{\text{new}} := \{C_{\text{new},1}, \ldots, C_{\text{new},N}\}$. Finally, for all $i \in 1..N$ such that $C_{\text{new},i} = \emptyset$ define $\mathbf{m}_{\mathcal{C}_{\text{new}}}(i) := \mathbf{m}_{\mathcal{C}}(i)$.

4.2 Theorem

Let $\emptyset \neq M \subseteq S$ such that $\# M = k$. Let $c_{\mathbf{x}} \in \mathcal{C}$ be as in 3.5 with $D = M$. If

$$\text{(35)} \quad \sum_{\mathbf{x} \in M} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(c_{\mathbf{x}}))^t \mathbf{W}(\mathcal{C})^{-1} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(c_{\mathbf{x}})) \underset{(\leq)}{<} \sum_{\mathbf{x} \in C} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(\mathbf{x}))^t \mathbf{W}(\mathcal{C})^{-1} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(\mathbf{x}))$$

then, with the definitions $C_{\text{new}} := M$ and 4.1, we have

$$(36) \quad \det\ \mathbf{W}(\mathcal{C}_{\text{new}}) \underset{(\leq)}{<} \det\ \mathbf{W}(\mathcal{C}).$$

Moreover, we have equality in (36) if and only if $\mathbf{m}_{\mathcal{C}_{\text{new}}} = \mathbf{m}_{\mathcal{C}}$ and $\mathbf{W}(\mathcal{C}_{\text{new}}) = \mathbf{W}(\mathcal{C})$.

Proof. Recalling the equality

$$\sum_{\mathbf{x}\in C}(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(\mathbf{x}))^t\mathbf{W}(\mathcal{C})^{-1}(\mathbf{x}-\mathbf{m}_{\mathcal{C}}(\mathbf{x})) = d,$$

cf. 1.2(f), the claim follows as a direct application of the Theorem 3.6 together with the convention $\mathbf{m}_{\mathcal{C}_{\text{new}}}(i) := \mathbf{m}_{\mathcal{C}}(i)$ for all $i \in 1..N$ such that $C_{\text{new},i} = \emptyset$. □

This result is the basis of the following building block for our algorithms: Starting from a configuration $(C, \mathcal{C})$, look for a subset M of S with k observations and for a family of clusters $(c_{\mathbf{x}})_{\mathbf{x}\in M} \subseteq \mathcal{C}$ such that the corresponding sum of the (squared) Mahalanobis distances is smaller than the current one. Equalityt (35) says that this new configuration is a better approximation for the MCDC. Let us call this building block the SCC–step. In order to formalize it, we next define the matrix of (squared) distances (with respect to the partition $\mathcal{C}$) as

$$(37) \quad \mathbf{D}_{\mathcal{C}} := (d_{\mathcal{C}}(\mathbf{x}, j)^2)_{\mathbf{x}\in S, j\in 1..N} \in \mathbb{R}^{n\times N},$$

where $d_{\mathcal{C}}(\mathbf{x}, j)$ is defined in (10).

4.3 The SCC–step

Let a configuration $(C, \mathcal{C})$ be given.

(i) Compute the matrix of distances $\mathbf{D}_{\mathcal{C}}$;

(ii) Find a subset C_{new} with k elements and a family of clusters $(c_{\mathbf{x}})_{\mathbf{x}\in C_{\text{new}}} \subseteq \mathcal{C}$ such that (35) is satisfied;

(iii) Calculate the new clusters $C_{\text{new},i}$, $i \in 1..N$, and the new partition $\mathcal{C}_{\text{new}}$ according to (34);

(iv) Compute $(m_{\text{new}}, \mathbf{W}(\mathcal{C}_{\text{new}}))$ according to 4.1.

Note that the SCC–step does not specify how to choose the subset C_{new} nor the family $(c_{\mathbf{x}})_{\mathbf{x}\in C_{\text{new}}}$. The following paragraphs give a way to follow in order to obtain such elements. As expected, the resulting SCC–step is the extension of Rousseeuw's C–step, cf. [17], Section 3. We need first some preliminaries.

4.4 Definition

For each $\mathbf{x}_i \in S$, let c_i be an element of the set $\operatorname{argmin}_{j\in 1..N} d_{\mathcal{C}}(\mathbf{x}_i, j)$. Moreover, let $\pi : 1..n \to 1..n$ be a permutation of the observations such that

(38) $d_{\mathcal{C}}(\mathbf{x}_{\pi(1)}, c_{\pi(1)}) \le d_{\mathcal{C}}(\mathbf{x}_{\pi(2)}, c_{\pi(2)}) \le \ldots \le d_{\mathcal{C}}(\mathbf{x}_{\pi(n)}, c_{\pi(n)})$.

For the sake of brevity let us denote by $\mathcal{C}_{\text{per}}$ the set of all such permutations. The following result is a direct application of Theorem 4.2 and the definition above.

4.5 Corollary

Let, in Theorem 4.2, $C_{\text{new}} := \{\mathbf{x}_{\pi(1)}, \mathbf{x}_{\pi(2)}, \ldots, \mathbf{x}_{\pi(k)}\}$. Then

(39) $\det\ \mathbf{W}(\mathcal{C}_{\text{new}}) \le \det\ \mathbf{W}(\mathcal{C})$

with equality if and only if $\mathbf{m}_{\mathcal{C}_{\text{new}}} = \mathbf{m}_{\mathcal{C}}$ and $\mathbf{W}(\mathcal{C}_{\text{new}}) = \mathbf{W}(\mathcal{C})$. In the case $N = 1$ Corollary 4.5 is Theorem 1 of [17].

Following Rousseeuw's terminology we refer to the construction in Corollary 4.5 as a *CC–step*. (Again, the additional C with respect to Rousseeuw's C–step stands for clustering). In algorithmic terms, the CC–step can be described as follows:

4.6 The CC–step

Let a configuration $(C, \mathcal{C})$ be given.

(i) Compute the matrix of distances $\mathbf{D}_{\mathcal{C}}$.

(ii) For each $\mathbf{x} \in S$, determine the set $\operatorname{argmin}_{j\in 1..N} d_{\mathcal{C}}(\mathbf{x}, j)$.

(iii) (α) Compute the set of all possible configurations $(M, \mathcal{M})$ such that $M = \{\mathbf{x}_{\pi(1)}, \mathbf{x}_{\pi(2)}, \ldots, \mathbf{x}_{\pi(k)}\}$ for an element $\pi \in \mathcal{C}_{\text{per}}$ and $\mathcal{M}$ is defined by a family $(c_{\mathbf{x}})_{\mathbf{x}\in M} \in \prod_{\mathbf{x}\in M} \operatorname{argmin}_{j\in 1..N} d_{\mathcal{C}}(\mathbf{x}, j)$, cf. (34).

(β) From this set choose a configuration $(C_{\rm new}, \mathcal{C}_{\rm new})$ with least SSP–determinant.

(iv) Compute $(m_{\rm new}, \mathbf{W}(\mathcal{C}_{\rm new}))$ according to 4.1.

4.7 Remarks

(a) For $N = 1$, the CC–step is the C–step, whereas for $k = n$ the CC–step is (essentially) the algorithm given in [4], Cap. 4, pp. 190–1.

(b) Consider the set defined in 4.6(iii)(α). Two different elements in this set may define configurations with unequal determinants. We illustrate this fact with two examples.

Example 1. Let $d = 2$, $n = 8$, $\mathbf{x}_i := (\cos(i\frac{\Pi}{4}), \sin(i\frac{\Pi}{4}))$, $i \in 0..7$, $N = 1$, and $k = 4$. Moreover, let $C := \{\mathbf{x}_0, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_6\}$. Therefore, $\mathbf{m}_{\mathcal{C}} = (0, 0)$ and $\mathbf{W}(\mathcal{C}) = 2\mathbf{I}_d$. The matrix of distances $\mathbf{D}_{\mathcal{C}}$ is constant with all entries equal to $\frac{1}{2}$. Hence, each subset of S with 4 elements is an admissible set $C_{\rm new}$. But the MCD's are of the form $\{\mathbf{x}_{(i+j) \mod 7} | j \in 0..3\}$ for an element $i \in 0..7$.

Example 2. Let $d = 1$, $S = \{-2, -1, 0, 1, 2\}$, $N = 2$, and $k = 4$. Moreover, let $C_1 = \{-1, 1\}$ and $C_2 = \{-2, 2\}$. Therefore, $m_{\mathcal{C}} = (0, 0)$, $\mathbf{W}(\mathcal{C}) = 10$. Clearly, $M_1 = \{0, 1, -1, 2\}$ and $M_2 = \{-2, 0, 1, -1\}$ are admissible candidates for the new set. Moreover, for all $\mathbf{x} \in S$, we have $\operatorname{argmin}_{j\in 1..2} d_{\mathcal{C}}(\mathbf{x}, j) = 1..2$. But among all these possible configurations the MCDC's are $\{\{-2, -1\}, \{0, 1\}\}$ and $\{\{-1, 0\}, \{1, 2\}\}$.

(c) If the set given in 4.6(iii)(α) is a singleton then the accomplishment of task (β) is trivial. Otherwise, due to the remark above a configuration with least determinant must be computed explicitly.

(d) (Other SCC–steps). There are many other possibilities for the generation of SCC–steps. Let, for example, in Definition 4.4, $j \in 1..k$ and set $M_j := \{\mathbf{x}_{\pi(1)}, \mathbf{x}_{\pi(2)}, \ldots, \mathbf{x}_{\pi(j)}\}$ for a permutation $\pi \in C_{\rm per}$. The corresponding set of clusters is given as in 4.4. From the set $C \setminus M_j$, choose $k - j$ observations. The associated cluster for each of these observations is the current one. Let finally in substep 4.6(iii) $C_{\rm new}$ be the union of M_j with these chosen observations. Clearly, this set satisfies (35).

Now, starting from an initial configuration $(C_0, \mathcal{C}_0)$ and iterating CC-steps, we obtain a sequence of configurations $((C_n, \mathcal{C}_n))_{n\geq 0}$ such that $\det \mathbf{W}(\mathcal{C}_{n+1}) \leq \det \mathbf{W}(\mathcal{C}_n)$ for all $n \in \mathbb{N}$.

Since the sequence $(\det \mathbf{W}(\mathcal{C}_n))_{n\in\mathbb{N}}$ converges and there are only a finite number of possible configurations, this iterative process must terminate in a finite number of steps with either $\det \mathbf{W}(\mathcal{C}_n) = 0$ or $\det \mathbf{W}(\mathcal{C}_{n+1}) = \det \mathbf{W}(\mathcal{C}_n)$. If the first case occurs then the corresponding configuration is an optimum. Otherwise, by Theorem 4.2, we have $\mathbf{m}_{\mathcal{C}_{n+1}} = \mathbf{m}_{\mathcal{C}_n}$ and $\mathbf{W}(\mathcal{C}_{n+1}) = \mathbf{W}(\mathcal{C}_n)$ and, therefore, $\mathbf{D}_{\mathcal{C}_{n+1}} = \mathbf{D}_{\mathcal{C}_n}$. Moreover, by the very definition of substep 4.6(iii), all the configurations defined by the set given in (β) have the same determinant. Hence, by 4.2, each of these configurations has $\mathbf{m}_{\mathcal{C}_n}$ as the vector of sample means and $\mathbf{W}(\mathcal{C}_n)$ as the SSP matrix. Therefore, a new iteration on $\mathcal{C}_n$ does not reduce the objective function. Although this is not sufficient for $\mathcal{C}_n$ to be a global minimum, it is a necessary condition. Different may be the situation if we take in 4.6(iii)(β) *any* configuration: Consider the Example 1 above and take as initial configuration $C_0 = \{\mathbf{x}_0, \mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_6\}$. If we take as C_1 the admissible configuration $\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_7\}$ then the equality $\det \mathbf{W}(C_1) = \det \mathbf{W}(C_0)$ together with the second claim of Theorem 4.2 implies $\mathbf{W}(C_1) = \mathbf{W}(C_0)$. But, one more iteration with $C_2 = \{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ as admissible set yields a MCD. Note that the transition from C_0 to C_1 is a SCC–step. Therefore, the example above concretises the difference between an algorithm that successively iterates CC–steps and has as stop criterion $\det \mathbf{W}(\mathcal{C}_n) = 0$ or $\det \mathbf{W}(\mathcal{C}_{n+1}) = \det \mathbf{W}(\mathcal{C}_n)$ and an algorithm using SCC–steps and the same stop criterion. Namely, the first one yields a configuration which satisfies a necessary condition for a configuration to be a global minimum, whereas the second one may or may not terminate with such a configuration. Note that one may subsequently iterate both SCC–steps and CC–steps. Although the next example shows that iterating SCC–steps may yield a "better" configuration than the use of CC–steps one after another, the discussion given above justifies our concentrating on the latter algorithm.

4.8 Examples

Let $(C_{\text{new}}, \mathcal{C}_{\text{new}})$ respect. $(D_{\text{new}}, \mathcal{D}_{\text{new}})$ be the result of the application of the CC-step respect. a SCC–step to the configuration $(C, \mathcal{C})$. Then by the definitions of a CC–step and a SCC–step the inequalities

$$(40)\quad \sum_{\mathbf{x}\in C_{\text{new}}} (\mathbf{x} - \mathbf{m}_{\mathcal{C}}(c_{\mathbf{x}}))^t \mathbf{W}(\mathcal{C})^{-1}(\mathbf{x} - \mathbf{m}_{\mathcal{C}}(c_{\mathbf{x}})) \le \sum_{\mathbf{y}\in D_{\text{new}}} (\mathbf{y} - \mathbf{m}_{\mathcal{C}}(\mathbf{y}))^t \mathbf{W}(\mathcal{C})^{-1}(\mathbf{y} - \mathbf{m}_{\mathcal{C}}(\mathbf{y})) \le 1$$

hold.

(a) However, (40) does not (always) imply $\det \mathbf{W}(\mathcal{C}_{\text{new}}) \le \det \mathbf{W}(\mathcal{D}_{\text{new}})$. Consider the following situation: Let $d = 1$, $N = 1$, $k = 2$, and $S = \{0, 3, 4, -x, x\}$, where $x \geq 6$. Moreover, let $C = \{-x, x\}$. Therefore, $\mathbf{m}_{\mathcal{C}} = 0$ and $\mathbf{W}(\mathcal{C}) = 2x^2$. The solution of the CC–step is $C_{\text{new}} = \{0, 3\}$ and a solution of a SCC–step is $D_{\text{new}} = \{3, 4\}$. Here, we have $\mathbf{W}(\mathcal{C}_{\text{new}}) = \frac{9}{2} > \frac{1}{2} = \mathbf{W}(\mathcal{D}_{\text{new}})$. Moreover, D_{new} is the MCD for the situation considered. Hence, here a SCC–step is better than the CC–step.

(b) A situation where (40) does imply $\det \mathbf{W}(\mathcal{C}_{\text{new}}) \leq \det \mathbf{W}(\mathcal{D}_{\text{new}})$ is the following: Let d, N, k as in (a), $S = \{0, 1, 4, -x, x\}, x \geq 6$, and $C = \{-x, x\}$. Here, the application of the CC–step yields $C_{\text{new}} = \{0, 1\}$, which is the MCD.

The following corollary extends Corollary 1 of [17] to "robust" clustering and Theorem 16.1 of [4] to clustering under the presence of outliers.

4.9 Corollary

Let $(C^\star, \mathcal{C}^\star)$ be a MCDC for the situation 2.1. Then

(a) each cluster $C_i^\star$, $i \in 1..N$, is separated from $S \setminus C^\star$ by an ellipsoid; and

(b) two different clusters $C_j^\star$ and $C_h^\star$ are separated by the hyperplane h_{ij} given by

$$h_{jh}(\mathbf{y}) := \left[\mathbf{y} - \frac{1}{2}(\mathbf{m}_{\mathcal{C}^\star}(j) + \mathbf{m}_{\mathcal{C}^\star}(h))\right]^t \mathbf{W}(\mathcal{C}^\star)^{-1}(\mathbf{m}_{\mathcal{C}^\star}(j) - \mathbf{m}_{\mathcal{C}^\star}(h)) = 0.$$

The observations in the cluster $C_j^\star$ are those which satisfy $h_{jh}(\mathbf{y}) > 0$.

Proof. We first prove (a). Clearly, the application of the CC–step to the configuration $(C^\star, \mathcal{C}^\star)$ results in the same configuration. Hence, the set $C^\star$ may be written as

$$C^\star = \{\mathbf{x}_{\pi^\star(1)}, \mathbf{x}_{\pi^\star(2)}, \ldots, \mathbf{x}_{\pi^\star(k)}\},$$

where $\pi^\star : 1..n \to 1..n$ is a permutation satisfying 4.6(iii)(β). Let now $j \in 1..N$ be arbitrary. Therefore, all the observations $\mathbf{x}_i \in C_j^\star$ satisfy $d_{\mathcal{C}^\star}(\mathbf{x}_i, \mathbf{m}_{\mathcal{C}^\star}(j))^2 \leq d_{\mathcal{C}^\star}(\mathbf{x}_{\pi^\star(k)}, c^\star_{\pi^\star(k)})^2 =: m$, whereas all $\mathbf{x}_i \notin C^\star$ satisfy $d_{\mathcal{C}^\star}(\mathbf{x}_i, \mathbf{m}_{\mathcal{C}^\star}(j))^2 \geq m$. Take the ellipsoid $E_j := \{\mathbf{x} \in \mathbb{R}^d | (\mathbf{x} - \mathbf{m}_{\mathcal{C}^\star}(j))^t \mathbf{W}_{\mathcal{C}^\star}^{-1}(\mathbf{x} - \mathbf{m}_{\mathcal{C}^\star}(j)) \leq m\}$. Then $C_j^\star \subseteq E_j$ and $S \setminus C^\star$ is contained in the closure of $\complement E_j$. The assertion (b) follows from the remark at the end of 2.1 together with the Theorem 16.1 of [4]. $\square$

5 An essay: On the breakdown value of the MCDC

In this paragraph we compute the breakdown value of the MCDC estimates of locations and scatters of the different components of the mixture (14) when $k = \lfloor \frac{n+dN+1}{2} \rfloor$. We extend in this way some results of [14]. Loosely speaking, the finite–sample version of the breakdown value of an estimator measures the minimum fraction of outliers that can *completely* spoil the estimate. Consider then the Paragraphs 2.1 and 2.2 and let us denote by $\mathbf{m}_n(S) := (\mathbf{m}_{n,1}(S), \ldots, \mathbf{m}_{n,N}(S)) \in \mathbb{R}^{dN}$ an estimate of locations, by $\mathbf{V}_n(S) = (\mathbf{V}_{n,1}(S), \ldots, \mathbf{V}_{n,N}(S)) \in (\mathbb{R}^{d\times d})^N$ an estimate of scatters, and by $\mathbf{W}_n(S) \in (\mathbb{R}^{d\times d})^N$ an estimate of covariance of the mixture (14) based on the set of observations S. We say that $\mathbf{m}_n(S)$ is *translation invariant* respect. *affine equivariant* if for all $1 \le j \le N$ the estimate of location $\mathbf{m}_{n,j}(S)$ for the j-th component of the mixture is translation invariant respect. affine equivariant, cf. [14], Def. 2.2. We say that $\mathbf{V}_n(S)$ is *affine equivariant* if each $\mathbf{V}_{n,j}(S)$, $1 \le j \le N$, is affine equivariant, cf. [14], Def. 2.2.

The natural extension of the breakdown point of the estimate of a vector of locations given in [7] or [14], Def. 2.2, is

$$(41)\quad \epsilon_N(\mathbf{m}_n(S)) := \min_{1\le m\le n} \left\{ \frac{m}{n} \mid \sup_O \min_{1\le j\le N} ||\mathbf{m}_{n,j}(S) - \mathbf{m}_{n,j}(O)|| = \infty \right\},$$

where the supremum is taken over all possible sets O of n observations obtained by replacing m points of S by arbitrary points in $\mathbb{R}^d$. Furthermore, the natural extension of the breakdown point of an estimate vector of scatters given in [7] or [14], Section 2, is given by

$$(42)\quad \epsilon_N(\mathbf{V}_n(S)) = \min_{1\le m\le n} \left\{ \frac{m}{n} \mid \sup_O \min_{1\le j\le N} d_M(\mathbf{V}_{n,j}(S) - \mathbf{V}_{n,j}(O)) = \infty \right\},$$

where $d_M(\mathbf{A}, \mathbf{B}) := \max\{|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_d(\mathbf{A})^{-1} - \lambda_d(\mathbf{B})^{-1}|\}$ $(\infty - \infty := 0)$.

Note that these definitions support the notion of a breakdown value given above, namely that in order to corrupt a vector estimate one must break down *all* of its components.

The breakdown value of $\mathbf{W}_n(S)$ is defined by, cf. [7] or [14], Section 2,

$$(43)\quad \epsilon_N(\mathbf{W}_n(S)) = \min_{1\le m\le n} \left\{ \frac{m}{n} \mid \sup_O d_M(\mathbf{W}_N(S), \mathbf{W}_N(O)) = \infty \right\}.$$

5.1 Remarks

Let $1 \le j \le N$. Let

$$(44)\quad \epsilon_1(\mathbf{m}_{n,j}(S)) := \min_{1 \le m \le n} \left\{ \frac{m}{n} \mid \sup_O ||\mathbf{m}_{n,j}(S) - \mathbf{m}_{n,j}(O)|| = \infty \right\}$$

be the breakdown value of the estimate of location $\mathbf{m}_{n,j}(S)$ for the j–th component of the mixture. Let

$$(45)\quad \epsilon_1(\mathbf{V}_{n,j}(S)) := \min_{1 \le m \le n} \left\{ \frac{m}{n} \mid \sup_O d_M(\mathbf{V}_{n,j}(S), \mathbf{V}_{n,j}(O)) = \infty \right\}$$

be the breakdown value of the estimate of scatter $\mathbf{V}_{n,j}(S)$ for the j–th component of the mixture.

The quantities $\min_{1 \le j \le N} \epsilon_1(\mathbf{m}_{n,j}(S))$ and $\max_{1 \le j \le N} \epsilon_1(\mathbf{m}_{n,j}(S))$ are alternative definitions of breakdown values for $\mathbf{m}_n$, whereas $\min_{1 \le j \le N} \epsilon_1(\mathbf{V}_{n,j}(S))$ and $\max_{1 \le j \le N} \epsilon_1(\mathbf{V}_{n,j}(S))$ are alternative definitions of breakdown values for $\mathbf{V}_n$. Clearly,

$$(46)\quad \min_{1 \le j \le N} \epsilon_1(\mathbf{m}_{n,j}(S)) \le \max_{1 \le j \le N} \epsilon_1(\mathbf{m}_{n,j}(S)) \le \epsilon_N(\mathbf{m}_n(S))$$

and

$$(47)\quad \min_{1 \le j \le N} \epsilon_1(\mathbf{V}_{n,j}(S)) \le \max_{1 \le j \le N} \epsilon_1(\mathbf{V}_{n,j}(S)) \le \epsilon_N(\mathbf{V}_n(S)).$$

For any $\mathbf{v} \in \mathbb{R}^d$ and any nonsingular $d \times d$ matrix $\mathbf{A}$, let $\mathbf{A}S + \mathbf{v} := \{\mathbf{A}\mathbf{x}_1 + \mathbf{v}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{v}\}$.

The next result is a generalization of [14], Lemma 2.1. It follows as a direct application of this latter result together with the componentwise definitions of translation invariance and affine equivariance given above.

5.2 Lemma

(a) If $\mathbf{m}_n(S)$ is translation invariant, then for any $\mathbf{v} \in \mathbb{R}^d$ we have $\epsilon(\mathbf{m}_n(S+\mathbf{v})) = \epsilon(\mathbf{m}_n(S))$ and for any $1 \le j \le N$ we have that $\epsilon_1(\mathbf{m}_{n,j}(S + \mathbf{v})) = \epsilon_1(\mathbf{m}_{n,j}(S))$.

(b) If $\mathbf{m}_n(S)$ is affine equivariant, then for any $\mathbf{v} \in \mathbb{R}^d$ and for any nonsingular $d \times d$ matrix $\mathbf{A}$ we have that $\epsilon(\mathbf{m}_n(\mathbf{A}S + \mathbf{v})) = \epsilon(\mathbf{m}_n(S))$ and $\epsilon_1(\mathbf{m}_{n,j}(AS + \mathbf{v})) = \epsilon_1(\mathbf{m}_{n,j}(S))$ for all $1 \le j \le N$.

(c) If $\mathbf{V}_n(S)$ is affine equivariant, then for any $\mathbf{v} \in \mathbb{R}^d$ and for any nonsingular $d \times d$ matrix $\mathbf{A}$ we have that $\epsilon(\mathbf{V}_n(\mathbf{A}S + \mathbf{v})) = \epsilon(\mathbf{V}_n(S))$ and $\epsilon_1(\mathbf{V}_{n,j}(AS + \mathbf{v})) = \epsilon_1(\mathbf{V}_{n,j}(S))$ for all $1 \leq j \leq N$.

5.3 Theorem

When $\mathbf{m}_n(S) \in \mathbb{R}^{d\times d}$ is translation invariant, then $\min_{1\leq j\leq N} \epsilon_N(\mathbf{m}_{n,j}(S)) \leq \lfloor \frac{n+1}{2} \rfloor / n$.

Proof. Note that

$$\min_{1\leq j\leq N} \epsilon_1(\mathbf{m}_{n,j}(S)) = \min_{1\leq m\leq n} \left\{ \frac{m}{n} \mid \sup_O ||\mathbf{m}_n(S) - \mathbf{m}_n(O)|| = \infty \right\}, \tag{48}$$

where the supremum is taken over all possible sets O of n observations obtained by replacing m points of S by arbitrary points in $\mathbb{R}^d$. The claim follows now as a direct application of [14], Theorem 2.1. □

In the sequel we assume that

$$\mathbf{W}_n(S) = \sum_{j=1}^{N} n_j \mathbf{V}_{n,j}(S) \in \mathbb{R}^{d\times d}. \tag{49}$$

The following remarks establish some connections between the breakdown values and the determinants of $\mathbf{V}_n$ and $\mathbf{W}_n$.

5.4 Remarks

(a) Let $d = 1$. We have

$$\min_{1\leq j\leq N} \epsilon_1(\mathbf{V}_{n,j}(S)) \leq \epsilon_N(\mathbf{W}_n(S)) \leq \epsilon_N(\mathbf{V}_n(S)). \tag{50}$$

Indeed, let $\epsilon_N(\mathbf{W}_n(S)) = \frac{m}{n}$. Let O be an arbitrary set of observations obtained by replacing $m-1$ arbitrary points of S by points in $\mathbb{R}$. We have $0 < \mathbf{W}_n(O) < K$ for a suitable constant $K > 0$. By (49), there exists a $1 \leq l \leq N$ such that $\mathbf{V}_{n,l}(O) > 0$, and for all $1 \leq j \leq N$ we have $\mathbf{V}_{n,j}(O) < K_j$ for suitable constants $K_1, \ldots, K_N > 0$. These two last assertions imply that $0 < \mathbf{V}_{n,l}(O) < \infty$, and hence $\min_{1\leq j\leq N} d_M(\mathbf{V}_{n,l}(S), \mathbf{V}_{n,l}(O)) < \infty$. The right hand side of (50) follows now from the Definition (42). Let now O be a set of n observations obtained by replacing m points of S by *adequate* points in $\mathbb{R}$ and such that $\mathbf{W}_n(O) = 0$ or $\mathbf{W}_n(O) = \infty$. By (49), either $\mathbf{V}_{n,l}(O) = \infty$ for some l, $1 \le l \le N$, or $\mathbf{V}_{n,j}(O) = 0$ for all j, $1 \le j \le N$. Consequently, $\mathbf{V}_{n,l}(O) = \infty$ or $\mathbf{V}_{n,l}(O) = 0$ or, equivalently, $\min_{1\le j\le N} \epsilon_1(\mathbf{V}_{n,j}(S)) \le \frac{m}{n}$. This is the left hand side of (50).

Let O be a set of n observations.

(b) For all $1 \le j \le N$ such that $\mathbf{V}_{n,j} > 0$, we have

$$(51)\quad \det \mathbf{W}_n(O) = \det \mathbf{V}_{n,j}(O) \det(n_j \mathbf{I}_d + \sum_{\substack{1\le l\le N\\ l\neq j}} n_l \mathbf{V}_{n,j}^{-1/2} \mathbf{V}_{n,l} \mathbf{V}_{n,j}^{-1/2}).$$

Let now

$$\kappa_j := \max\{\lambda | \lambda \text{ eigenvalue of } n_j \mathbf{I}_d + \sum_{\substack{1\le l\le N\\ l\neq j}} n_l \mathbf{V}_{n,j}^{-1/2} \mathbf{V}_{n,l} \mathbf{V}_{n,j}^{-1/2}\}.$$

Since the eigenvalues of the matrix $n_j \mathbf{I}_d + \sum_{\substack{1\le l\le N\\ l\neq j}} n_l \mathbf{V}_{n,j}^{-1/2} \mathbf{V}_{n,l} \mathbf{V}_{n,j}^{-1/2}$ are bounded above by $n_j + \kappa_j$, and below by n_j, we obtain

$$(52)\quad n_j^d \det \mathbf{V}_{n,j}(O) \le \det \mathbf{W}_n(O) \le (n_j + \kappa_j)^d \det \mathbf{V}_{n,j}(O).$$

(c) From (52) follows that if $\det \mathbf{W}_n(O) < \infty$ then $\det \mathbf{V}_{n,j}(O) < \infty$ for all $1 \le j \le N$, and if $\det \mathbf{W}_n(O) = 0$ then $\det \mathbf{V}_{n,j}(O) = 0$ for all $1 \le j \le N$.

(d) The claim $\det \mathbf{W}_n(O) = \infty$ does not imply $\det \mathbf{V}_{n,j}(O) = \infty$ for some j, $1 \le j \le N$. As an example, let $d = N = 2$, and let

$$\mathbf{V}_{n,1}(O) = \begin{pmatrix} 2K & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{V}_{n,2}(O) = \begin{pmatrix} 0 & 0 \\ 0 & 2K \end{pmatrix},$$

for an arbitrary $K > 0$. Similarly, the claim $\det \mathbf{W}_n(O) > 0$ does not imply $\det \mathbf{V}_{n,j}(O) > 0$ for a $1 \le j \le N$. As an example, let $d = N = 2$, and let

$$\mathbf{V}_{n,1}(O) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{V}_{n,2}(O) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Let us denote by $\mathbf{m}_n^\star(S) := (\mathbf{m}_{n,1}^\star(S), \ldots, \mathbf{m}_{n,N}^\star(S))$ the estimate of locations defined by the MCDC, and by $\mathbf{V}_n^\star(S) := (\mathbf{V}_{n,1}^\star(S), \ldots, \mathbf{V}_{n,N}^\star(S))$ respect. $\mathbf{W}_n^\star(S)$ the estimate of scatters respect. the estimate of covariance defined by the MCDC. Clearly, $\mathbf{m}_n^\star(S)$ is translation and affine equivariant, whereas $\mathbf{V}_n^\star(S)$ and $\mathbf{W}_n^\star(S)$ is affine equivariant.

For the proof of the following theorem we assume that

($\star$) the covariance matrix of each set having at least $d+1$ elements is positive definite.

The following result is an extension of [14], Theorem 3.1.

5.5 Theorem

Let $k = \lfloor \frac{n+dN+1}{2} \rfloor$.

(a) If $d = 1$ then the breakdown value of the MCDC estimate of locations is $\lfloor \frac{n+N}{2} \rfloor / n$.

For the next two assertions assume $n \geq (d+2)N - 1$ if $n - dN + 1$ is even, and assume $n \geq (d+2)N$ otherwise.

(b) If $d = 1$ then the breakdown value of the MCDC estimate of scatters and covariance is $\lfloor \frac{n-dN+1}{2} \rfloor / n$.

(c) If $d \geq 2$ then the breakdown value of the MCDC estimate of locations, scatters, and covariance is $\lfloor \frac{n-dN+1}{2} \rfloor / n$.

Proof. We first prove (a). Let $d = 1$ and let $N \geq 2$. Clearly, if we replace $\lfloor \frac{n+N}{2} \rfloor$ original observations then each subset of k elements of the new set contains at least N new observations. Set $M := \max\{|\mathbf{x}_i| \mid i \in 1..n\}$. Let the corrupted observations be all different and of type $M \cdot 10^{lK}$ for an $l \in 1..\lfloor \frac{n+N}{2} \rfloor$ and a constant $K > 0$ big enough. Let us denote by O the contaminated set. It is easy to see that the MCDC configuration of the set of observations O consists of $N-1$ singletons $\{M \cdot 10^{lK}\}$ for an $l \in 2..\lfloor \frac{n+N}{2} \rfloor$ and one cluster containing all the original observations in O together with the contamination $M \cdot 10^K$. Clearly, $||\mathbf{m}_{n,j}^\star(O)|| \to \infty$ as $K \to \infty$, $j \in 1..N$. Consequently, the breakdown value of $\mathbf{m}_{n,j}^\star(O)$ is at most $\lfloor \frac{n+N}{2} \rfloor / n$. Let us now show that it is at least $\lfloor \frac{n+N}{2} \rfloor / n$. For this sake, let us denote by O an arbitrary corrupted set obtained by replacing $\lfloor \frac{n+N}{2} \rfloor - 1$ or less original observations by points in $\mathbb{R}$. Then each subset of O with k observations contains at least one original observation. Consequently at least one cluster, say the first one, of the MCDC configuration for the set of observations O contains original observations. Now, O contains

$$n + 1 - \left\lfloor \frac{n+N}{2} \right\rfloor \geq \left\lfloor \frac{n+N}{2} \right\rfloor = k$$

original observations. Therefore, by the definition of the MCDC we have

$$\det \mathbf{W}_n^\star(O) \leq \det \mathbf{W} < \infty,$$

where $\mathbf{W}$ is the SSP matrix of an arbitrary configuration of a set D having k original observations of the set O. Furthermore, each $\det \mathbf{V}_{n,j}^\star(O)$ is proportional to the length of the corresponding cluster. Consequently, the length of each cluster is bounded above. If possible, let $||\mathbf{m}_{n,1}^\star(O)||$ be arbitrary large. A moment of reflection shows that this fact implies that the length of the first cluster also goes to infinity. This is a contradiction.

We now prove simultaneously (b) and (c). We first prove that $\epsilon_N(\mathbf{m}_n^\star(S))$, $\epsilon_N(\mathbf{V}_n^\star(S))$ and $\epsilon_N(\mathbf{W}_n^\star(S))$ are at most $\lfloor\frac{n-dN+1}{2}\rfloor$ if $d \geq 2$, and that $\epsilon_N(\mathbf{V}_n^\star(S))$ and $\epsilon_N(\mathbf{W}_n^\star(S))$ are at most $\lfloor\frac{n-N+1}{2}\rfloor$ if $d = 1$. Choose N disjoint subsets of S with d elements each. Denote by H_j, $1 \leq j \leq N$, the $d-1$–dimensional hyperplane generated by each of these subsets. Replace $\lfloor\frac{n-dN+1}{2}\rfloor$ observations of S by points on each of the H_j's (the given assumption makes this fact possible). Let us denote by O_1 the set containing the N disjoint subsets and the $\lfloor\frac{n-dN+1}{2}\rfloor$ replacements. O_1 contains $dN + \lfloor\frac{n-dN+1}{2}\rfloor = \lfloor\frac{n+dN+1}{2}\rfloor$ points. Let us now enlarge the set O_1 to a set O by adding arbitrary $n - \lfloor\frac{n+dN+1}{2}\rfloor$ observations contained in the set $S \setminus O_1$. The condition $(\star)$ assures that no point in the set $O \setminus O_1$ lies on one of the H_j's. Therefore, there exists a MCDC configuration $(C^\star, \mathcal{C}^\star)$ for the set of observations O, $k = \lfloor\frac{n+dN+1}{2}\rfloor$ and N satisfying $\det \mathbf{W}_n^\star(O) = 0$ and $C_j^\star \subseteq H_j$, $1 \leq j \leq N$. By 5.4(c) (or the second fact above) we have $\det \mathbf{V}_{n,j}^\star(O) = 0$ for all $1 \leq j \leq N$. Hence, $\epsilon_N(\mathbf{W}_{n,j}^\star(S)), \epsilon_N(\mathbf{V}_{n,j}^\star(S)) \leq \lfloor\frac{n-dN+1}{2}\rfloor$ for all $1 \leq j \leq N$ and $\epsilon_N(\mathbf{V}_n^\star(S)) \leq \lfloor\frac{n-dN+1}{2}\rfloor$. Send now the contaminations in the H_j's in a suitable way to infinity. Since each $C_j^\star$, $1 \leq j \leq N$, contains at least one contamination, all the mean vectors $\mathbf{m}_{n,j}^\star(O)$s become infinitely large and, therefore, $\epsilon_1(\mathbf{m}_{n,j}(S)) \leq \lfloor\frac{n-dN+1}{2}\rfloor$, $1 \leq j \leq N$ and $\epsilon_N(\mathbf{m}_n^\star(S)) \leq \lfloor\frac{n-dN+1}{2}\rfloor$.

We next prove that $\epsilon_N(\mathbf{m}_n^\star(S))$, $\epsilon_N(\mathbf{V}_n^\star(S))$ and $\epsilon_N(\mathbf{W}_n^\star(S))$ are at least $\lfloor\frac{n-dN+1}{2}\rfloor$ if $d \geq 2$, and that $\epsilon_N(\mathbf{V}_n^\star(S))$ and $\epsilon_N(\mathbf{W}_n^\star(S))$ are at least $\lfloor\frac{n-N+1}{2}\rfloor$ if $d = 1$. Indeed, let us now replace $\lfloor\frac{n-dN+1}{2}\rfloor - 1$ or less observations contained in S by arbitrary points in $\mathbb{R}^d$ and denote by O the new set. Then each subset of O having $\lfloor\frac{n+dN+1}{2}\rfloor$ points contains at least

$$\left\lfloor\frac{n+dN+1}{2}\right\rfloor - \left\lfloor\frac{n-dN+1}{2}\right\rfloor + 1 \geq dN + 1$$

original observations. Therefore, at least one cluster, say the first one, contains $d+1$ original observations. Therefore, by the Condition $(\star)$, we have $\mathbf{V}^\star_{n,1}(O) > 0$, and by 5.4(c) we have also $\mathbf{W}^\star_n(O) > 0$. Note that $|O \cap S| \geq \lfloor \frac{n+dN+1}{2} \rfloor$. Therefore, if we denote by $\mathbf{W} \in \mathbb{R}^{d\times d}$ the SSP matrix defined by $\lfloor \frac{n+dN+1}{2} \rfloor$ points in this set and an arbitrary configuration $\mathcal{C}$ then by the definition of the MCDC for the set of observations O, $k = \lfloor \frac{n+dN+1}{2} \rfloor$ and N we have

(53) $0 < \det \mathbf{W}^\star_n(O) \leq \det \mathbf{W} < \infty,$

i.e., $\mathbf{W}^\star_n(S)$ does not break down. Now, by 5.4(c), we have also $\det \mathbf{V}^\star_{n,1}(O) < \infty$ for all $1 \leq j \leq N$. We have shown that $0 < \det \mathbf{V}^\star_{n,1}(O) < \infty$ and consequently $\mathbf{V}^\star_n(S)$ does not break down. Let now in [14], Lemma 3.1, $p := d$, $\mathbf{v}_1, \ldots, \mathbf{v}_{d+1} \in \mathbb{R}^d$ be those original observations which belong to the cluster $C^\star_1$, and let

$$E(\mathbf{t}, \mathbf{C}) := \{\mathbf{x} \in \mathbb{R}^d | (\mathbf{x} - \mathbf{m}^\star_{n,1}(O))^t \mathbf{W}^\star_n(O)^{-1} (\mathbf{x} - \mathbf{m}^\star_{n,1}(O)) \leq \max_{\mathbf{y} \in C^\star_1} d_{\mathcal{C}^\star}(\mathbf{y}, \mathbf{m}_{\mathcal{C}^\star}(1))^2\},$$

where $d_{\mathcal{C}^\star}$ is given by (10). By the discussion above the $\mathbf{v}_i$'s span a nonempty simplex in $\mathbb{R}^d$. By this lemma we have that for each $V > 0$ there exists a constant $M > 0$, which depends only on the $\mathbf{v}_i$'s such that $||\mathbf{m}^\star_{n,1}(O)|| > M$ implies $\det E(\mathbf{t}, \mathbf{C}) > V$. But, by the right hand side of (53), this cannot be the case, and we conclude that $||\mathbf{m}^\star_{n,1}(O)||$ is bounded above, i.e., $\mathbf{m}^\star_{n,1}(O)$ does not break down and $\mathbf{m}^\star_n(O)$ does not either. □

6 The algorithm

In this section, we generalize the algorithm given in [17], Section 4, to the situation (P). Concisely:

Generate many initial configurations $(C, \mathcal{C})$ and iterate CC–steps until convergence is reached. Take the limit configuration with the least determinant.

6.1 On the initial configuration

In order to create an initial subset C_0, we refer the reader to Paragraph 4.1 of [17]. For the generation of an initial configuration $\mathcal{C}_0$ of the set C_0 consider the following possibilities:

(a) choose at random a subset $C_0 \in \binom{S}{k}$. Independently, draw at random a subset of N elements of C_0 say $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ and put the observation $\mathbf{y}_i$ in the cluster $C_{0,i}$. For each of the $k - N$ remaining points of this initial set select at random a cluster $C_{0,i}$, $i \in 1..N$. Clearly, this procedure is equivalent to drawing at random an element of the set $\tilde{\mathcal{P}}(C_0)$.

(b) choose randomly a subset of S with $\max\{d+1, N+1\}$ elements. Following the procedure (a), construct a partition of this set in N clusters. If the corresponding SSP matrix has zero determinant then select at random an observation from the rest and add it to a cluster selected at random. Repeat this step until the SSP of the current partition $(D, \mathcal{D})$ has a strictly positive determinant. Compute the matrix of distances $\mathbf{D}_{\mathcal{D}}$. For each $\mathbf{x}_i \in S$, let c_i be an element of the set $\operatorname{argmin}_{j \in 1..N} d_{\mathcal{D}}(\mathbf{x}_i, j)$. Moreover, let $\pi : 1..n \to 1..n$ be a permutation of the observations such that

$$d_{\mathcal{D}}(\mathbf{x}_{\pi(1)}, c_{\pi(1)}) \leq d_{\mathcal{D}}(\mathbf{x}_{\pi(2)}, c_{\pi(2)}) \leq \ldots \leq d_{\mathcal{D}}(\mathbf{x}_{\pi(n)}, c_{\pi(n)}).$$

Put $C_0 := \{\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(k)}\}$ and $C_{0,j} := \{\mathbf{x}_i \in C_0 | c_i = j\}$, $j \in 1..N$.

6.2 Experimental results

A data set consisting of $n = 450$ observations in two dimensions was generated. The first 200 of them came from the population $N((-3,0), \mathbf{I}_d)$ and the next 200 came from the population $N((3,0), \mathbf{I}_d)$. The remaining 50 observations (or contaminations) came from a normal population with mean vector $(0, -5)$ and covariance matrix $36\mathbf{I}_d$. The Figure 1 shows the generated data set.

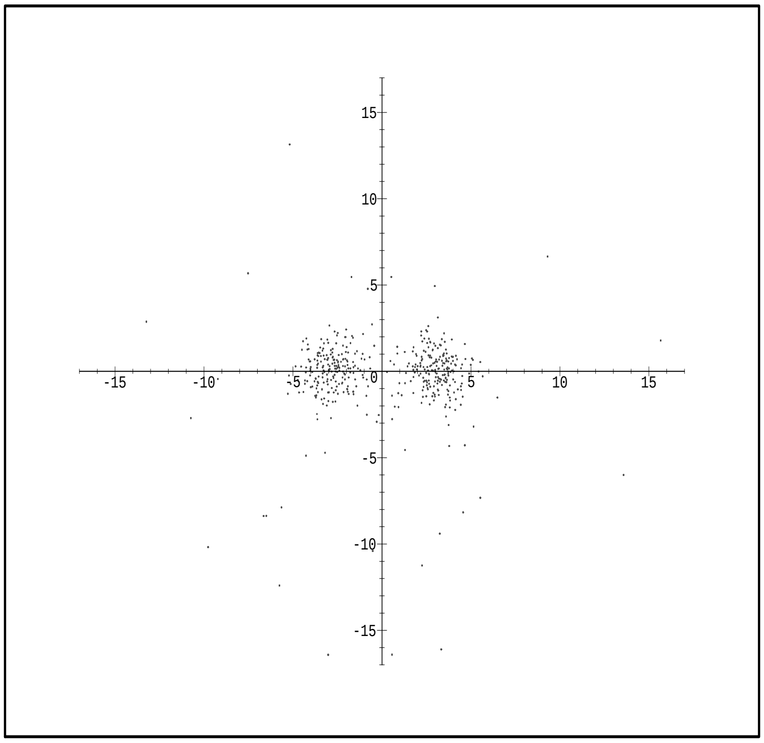

Figure 1. The simulated data set: 415 observations coming from the mixture $\frac{4}{9}N((-3,0),\mathbf{I}_d)+\frac{4}{9}N((3,0),\mathbf{I}_d)+\frac{1}{9}N((0,-5),36\mathbf{I}_d)$.

We now apply our algorithm in order to find a MCDC (or at least a good approximation for it) of size $k=400$. Moreover, we use the method 6.1(a) in order to construct 415 initial configurations. The Figure 2 shows the configuration with the least determinant among the 415 trials. Thus, we are looking for the configuration $\mathcal{C}=\{1..200,201..400\}$ or, more precisely, for robust estimators of the prior probabilities, the mean vectors, and the covariance matrices of the first two components of the mixture given above. The (sample) mean vectors of the generated clusters are $\widehat{\mathbf{m}}_1=(-2.8,0.12)$ and $\widehat{\mathbf{m}}_2=(3.0,0.08)$, whereas the (sample) covariance matrices are

$$\widehat{\mathbf{V}}_1=\begin{pmatrix} 0.94 & 0.082 \\ & 1.093 \end{pmatrix} \quad \text{and} \quad \widehat{\mathbf{V}}_2=\begin{pmatrix} 0.88 & -0.07 \\ & 1.06 \end{pmatrix}.$$

This yields the SSP matrix

$$\widehat{\mathbf{W}}=\begin{pmatrix} 364.82 & 2.78 \\ & 434.03 \end{pmatrix}$$

with determinant $\det \widehat{\mathbf{W}} = 158334$.

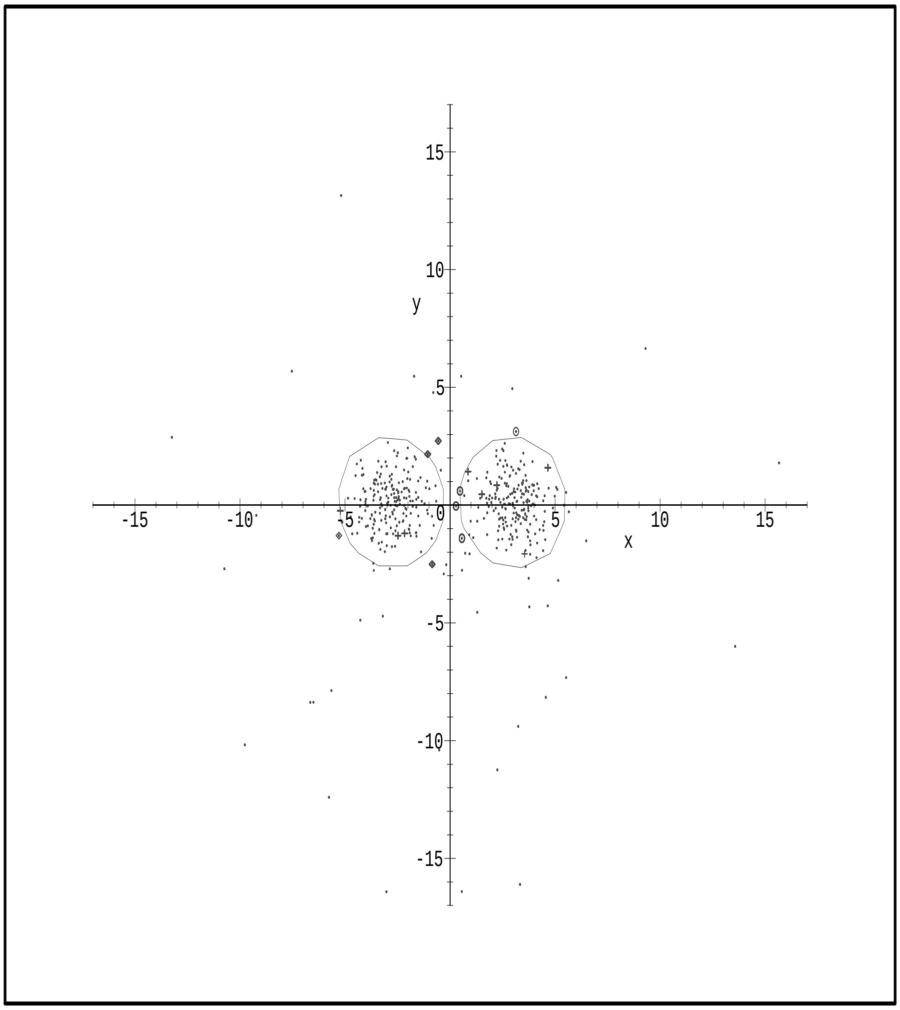

Figure 2. The final configuration. The diamonds and circles are those observations which come from the left and right populations respectively and which the algorithm classifies as outliers. The crosses are those outliers which the algorithm classifies as noncontaminations. The ellipsoids are those given in Corollary 4.9(a).

The value of the objective function is $\det \mathbf{W}_{\mathrm{MCDC}} = 142574$. Note that it is lower than the one corresponding to the "true" configuration. The (approximated) MCDC mean vectors are $\mathbf{m}_{1,\mathrm{MCDC}} = (-2.8, 0.1)$ and $\mathbf{m}_{2,\mathrm{MCDC}} = (3.0, 0.08)$, whereas the (approximated) MCDC covariance matrices are

$$\mathbf{V}_{1,\mathrm{MCDC}} = \begin{pmatrix} 0.886 & 0.04 \\ & 1.03 \end{pmatrix} \quad \text{and} \quad \mathbf{V}_{2,\mathrm{MCDC}} = \begin{pmatrix} 0.83 & -0.096 \\ & 1.04 \end{pmatrix}.$$

Finally, the (approximated) MCDC prior probabilities are $p_{1,\mathrm{MCDC}} = \frac{199}{400}$ and $p_{2,\mathrm{MCDC}} = \frac{201}{400}$. A comparison of these values with the simulated ones pleads for the goodness of the algorithm.

The descriptive statistics in Table 1 provide some insight into the algorithm.

# of CC–steps	# of trials	min. value of det $\mathbf{W}$	max. value of det $\mathbf{W}$	mean value	standard deviation
3	142	142574	142656	142601	26
4	62	142596	478652	153456	59372
5	45	142596	477643	164935	83461
6	42	142596	485294	224973	144961
7	26	142603	485294	246165	155284
8	27	142603	485294	292722	167811
9	13	142603	489542	221384	143775
10	10	142603	478135	309817	167172
11	16	142656	481193	310710	168058
12..20	32	142603	489774	290754	167965

Table 1. Summary statistics for the data of the algorithm.

More precisely, let $\mathbf{x} := \mathbf{x}_1, \ldots, \mathbf{x}_n$ be a realization of the mixture (14). The construction method specified by 6.1(a) implies that the random initial configuration $IC(\mathbf{x}) : (\Omega, P) \to \bigcup_{C \in \binom{S}{k}} \{C\} \times \tilde{\mathcal{P}}(C)$ is uniformly distributed. Clearly, the random number of CC–steps until convergence is a function of $IC(\mathbf{x})$ and we denote it by $L(\mathbf{x}) : (\Omega, P) \to 0..l(\mathbf{x})$. Usually, it is not possible to compute the maximal length $l(\mathbf{x})$. In our concrete case, the 415 trials make possible to estimate it among other parameters like expectation, variance and mode. The Figure 3 shows the (count) histogram of the data given in the Table 1.

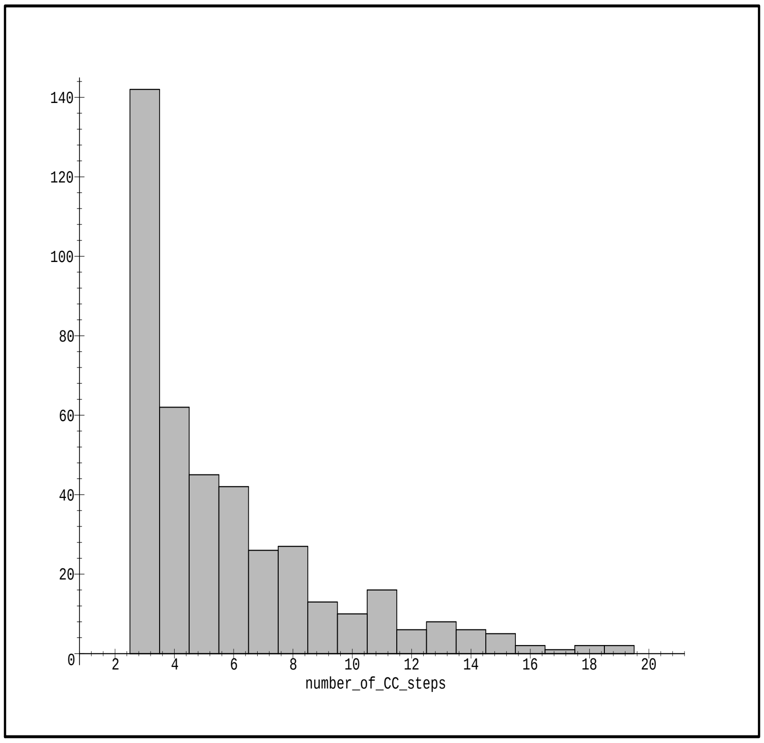

Figure 3. Count histogram of the random variable $L(\mathbf{x})$ for the realization $\mathbf{x}$.

Therefore, the estimate for $l(\mathbf{x})$ is 20, whereas the estimate for the mode is 3. The estimate for the expectation is 6.7 and the estimate for the standard deviation is 3.38. Moreover, a look at the fourth column of Table 1 shows that the value of the objective function tends to increase as the number of required CC–steps for convergence increases. This fact motivates to run the algorithm until the third CC–step, only. More or less the same behavior was observed in [17], Remark 4.2.

7 The CC–step under additional constraints

Assume now that, in addition to the conditions stated by (P), the j–th cluster, $j \in 1..N$, can not contain more than b_j observations, where $\sum_{j=1}^{N} b_j \geq n$. In this section we adapt the CC–step to these constraints, and we write the corresponding optimization problem as a combinatorial one. Moreover, we show that its solution is determined by the solution of a linear optimization problem known by the name of *Hitchock problem*, cf. [16], which is motivated by the following situation: we have s sources of some commodity, each with a supply of a_i units, $i \in 1..s$, and t terminals, each of which has a demand of b_j units, $j \in 1..t$. Furthermore, we know that the cost of sending the commodity form source i to terminal j is c_{ij}. How do we satisfy the demand at minimum cost? The next definition formalizes this situation.

7.1 Definition: The Hitchcock problem

Let $s, t \geq 1$. For all $i \in 1..s$ and $j \in 1..t$, let $a_i \geq 0$, $b_j \geq 0$, and $c_{ij} \geq 0$. Furthermore, let $\sum_{i=1}^{s} a_i \leq \sum_{j=1}^{t} b_j$. An instance of the *Hitchcock problem* is the following LP with variables f_{ij}, $i \in 1..s$, $j \in 1..t$:

$$\text{(HP)} \qquad \min \sum_{i,j} c_{ij} f_{ij} \quad \text{constrained on}$$

$$\begin{cases} \sum_{j=1}^{t} f_{ij} = a_i, & i \in 1..s, \\ \sum_{i=1}^{s} f_{ij} \leq b_j, & j \in 1..t, \\ f_{ij} \in \{0, 1\}. \end{cases}$$

7.2 The optimization problem defined by the constrained CC–step as an instance of the Hitchcock problem

Let a configuration $(C, \mathcal{C})$ be given. In Definition 7.1 let s be the number n of observations and t be the number N of clusters. Moreover, for all $i \in 1..n$, and for all $j \in 1..N$, let $a_i := 1$ and $c_{ij} := d_{\mathcal{C}}(\mathbf{x}_i, j)^2$. The optimization problem

$$(54) \quad \min \sum_{i,j} d_{\mathcal{C}}(\mathbf{x}_i, j)^2 f_{ij} \quad \text{constrained on}$$

$$\begin{cases} \sum_{j=1}^{N} f_{ij} = 1, & i \in 1..n, \\ \sum_{i=1}^{n} f_{ij} \leq b_j, & j \in 1..N, \\ f_{ij} \in \{0, 1\}. \end{cases}$$

with variables f_{ij}, $i \in 1..n$, $j \in 1..N$, is an instance of the Hitchcock problem. Moreover, each of its solutions defines an assignment $(a_{\mathcal{C}}(i))_{i \in 1..n}$ given by $a_{\mathcal{C}}(i) = j \;\Leftrightarrow\; f_{ij} = 1$, $i \in 1..n, j \in 1..N$, which has minimum cost among all admissible ones. Let us denote by $A_{\mathcal{C}}$ the set of all solutions.

The following definition is analogous to Section 4.4.

7.3 Definition

For each element $a_{\mathcal{C}} \in A_{\mathcal{C}}$, let $\pi : 1..n \to 1..n$ be a permutation of the observations such that

$$(55) \quad d_{\mathcal{C}}(\mathbf{x}_{\pi(1)}, a_{\mathcal{C}}(\pi(1))) \leq d_{\mathcal{C}}(\mathbf{x}_{\pi(2)}, a_{\mathcal{C}}(\pi(2))) \leq \ldots \leq d_{\mathcal{C}}(\mathbf{x}_{\pi(n)}, a_{\mathcal{C}}(\pi(n))).$$

Let us denote by $\mathcal{C}_{\text{per}}(a_{\mathcal{C}})$ the set of all such permutations. With these tools at hand, we are now able to write the CC–step for the situation considered. We denote it by CCC–step.

7.4 The CCC–step

Let a configuration $(C, \mathcal{C})$ be given.

(i) Compute the matrix of distances $\mathbf{D}_{\mathcal{C}}$.

(ii) Compute the set $A_{\mathcal{C}}$ of solutions of (54).

(iii) (α) Compute the set of all possible configurations $(M, \mathcal{M})$ such that $M = \{\mathbf{x}_{\pi(1)}, \mathbf{x}_{\pi(2)}, \ldots, \mathbf{x}_{\pi(k)}\}$, where $\pi \in \mathcal{C}_{\text{per}}(a_{\mathcal{C}})$ for an element $a_{\mathcal{C}} \in A_{\mathcal{C}}$ and $\mathcal{M}$ is defined by the family $(a_{\mathcal{C}}(\pi(i))_{i \in 1..k}$, cf. (34).

(β) From this set choose a configuration $(C_{\text{new}}, \mathcal{C}_{\text{new}})$ with smallest SSP–determinant

(iv) Compute $(m_{\text{new}}, \mathbf{W}(\mathcal{C}_{\text{new}}))$ according to 4.1.

Acknowledgment. I thank Professor Dr. Gunter Ritter for pointing out the topic to me and for reading the manuscript.

References

[1] Barnett, V.D. 1976. The ordering of multivariate data. J. Roy. Statist. Soc. A, 139, 318–355.

[2] Barnett, V. and T. Lewis. 1994. Outliers in statistical data, 3rd Edition, Wiley, Chichester, UK.

[3] Bock, H.H. 1989. Probabilistic aspects in cluster analysis. In: O.Opitz, Ed., Conceptual and Numerical Analysis of Data. Springer–Verlag., Berlin.

[4] Bock, H.H. 1974. Automatische Klassifikation. Goettingen Vandenhoeck & Ruprecht.

[5] Campbell, N. A. 1980. Robust procedures in multivariate analysis I: Robust covariance estimation. Applied statistics, **29**, 231–237.

[6] Csiszár, I. and G. Tusnády. 1981. Alternating Minimization and the EM–Algorithm. Colloquium on Information Theory, Budapest.

[7] Donoho, D.L. and P.J. Huber. 1983. The notion of a breakdown point. In A Festschrift for Erich L. Lehmann. P.J. Bickel, K.A. Doksum and J .L. Hodges, Jr., eds. 157–184. Wadsworth, Belmont, Calif.

[8] Grűbel, R. 1988. A Minimal Characterization of the Covariance Matrix. Metrika, Volume 35, 1988, pp. 49–52.

[9] Huber, P.J. 1981. Robust Statistics. John Wiley and Sons, New York.

[10] Kharin, Yu.S. 1996. Robustness in Statistical Pattern Recognition. Kluwer Academic Publishers, Dordrecht/Boston/London.

[11] Kharin, Yu.S. and E.E. Zhuk. 1993. Asymptotics robustness in cluster analysis for the case of Tukey–Huber distortions. In: O.Opitz, B. Lausen, R. Klar, Ed., Information and Classification: Concepts, Methods and Applications. Springer–Verlag, New York

[12] Kharin, Yu.S. and E.E. Zhuk. 1994. Robustness in statistical pattern recognition under "contaminations" of training samples. In: Proc. of the 12th IAPR Int. Conf. on Pattern Recognition. Vol **2**. IEEE Comp. Press, Washington.

[13] Kharin, Yu.S. and E.E. Zhuk. 1998. Filtering of multivariate samples for clustering under "outliers". Pattern Recognition Letters, 19(12), pp. 1077–1085.

[14] Lopuhaä, H.P. and P. Rousseeuw. 1991. Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariances Matrices. The Annals of Statistics, 19, 229–248.

[15] Mardia, K.V., J.T. Kent and J.M. Bibby. 1979. Multivariate Analysis. Academic Press, London, New York, Toronto, Sydney, San Francisco.

[16] Papadimitriou, C. and K. Steiglitz. 1998. Combinatorial Optimization. Dover Publ., Mineola, New York.

[17] Rousseeuw, Peter J. and K. Van Driessen. 1999. A Fast algorithm for the Minimum Covariance Determinant Estimator. Technometrics, 41, 212–223.