Adv Data Anal Classif (2009) 3:135–167
DOI 10.1007/s11634-009-0044-9

REGULAR ARTICLE

Trimming algorithms for clustering contaminated grouped data and their robustness

María Teresa Gallegos · Gunter Ritter

Received: 23 May 2009 / Revised: 8 August 2009 / Accepted: 16 August 2009 /
Published online: 2 September 2009

Abstract We establish an affine equivariant, constrained heteroscedastic model and criterion with trimming for clustering contaminated, grouped data. We show existence of the maximum likelihood estimator, propose a method for determining an appropriate constraint, and design a strategy for finding reasonable clusterings. We finally compute breakdown points of the estimated parameters thereby showing asymptotic robustness of the method.

Keywords Statistical clustering · Robust clustering · Trimming algorithm · Breakdown points · Heteroscedasticity · HDBT ratio

Mathematics Subject Classification (2000) Primary 62H30; Secondary 62F35

1 Introduction

1.1 Background

Statistical clustering methods start from a statistical model of the data deriving from it, in general by the maximum likelihood or maximum a posteriori paradigms, a cluster criterion to be optimized. Various problems, expected and unexpected ones, are encountered on this way. First, the criteria do not possess maxima in general so that special precautions have to be taken. Second, the criteria possess so-called "local maxima" and "minimum distance partitions" (MDP's), some of them reasonable solutions

M. T. Gallegos
Institute for Data Analysis, Salzweg, Germany

G. Ritter (✉)
Fakultät für Informatik und Mathematik, Universität Passau, Passau, Germany
e-mail: ritter@fim.uni-passau.de

but others containing spurious, undesirable clusters. Third, optimization of the criteria is not easy. Fourth, the clustering method obtained may not be robust in the sense that deviations from the model may grossly falsify the result.

Solutions to some of these problems are available. Hathaway (1985), following a proposal by Dennis (1981), Beale and Thompson (oral communications), investigated constraints on the scale parameters $v_1, \ldots, v_g$ of a univariate normal mixture of g components showing that they mitigated or even avoided some of the problems. Hathaway also indicated how to extend the constraints to d dimensions. We call them the *HDBT constraints*. Denoting the covariance matrices of the g (multivariate) components by $V_1, \ldots, V_g$, they may be written

$$V_j \succeq cV_\ell, \quad 1 \leq j, \ell \leq g, \tag{1}$$

for some constant $c > 0$. The symbol $\succeq$ stands for the positive semidefinite or Löwner ordering on the space of symmetric matrices and the constant c is necessarily bounded above by 1 so that $0 < c \leq 1$. The constraints are affine equivariant and mean that the covariance matrices V_j must not be too different in size and shape. They are a generalization of homoscedasticity, i.e., equality of all covariance matrices, which they contain as the special case $c = 1$. We also define the *HDBT ratio* of a g-tuple $\mathbf{V} = (V_1, \ldots, V_g)$ of positive-definite matrices as the maximum c for which the constraints (1) hold. It is easy to see that it can be computed as

$$r_{\text{HDBT}}(\mathbf{V}) = \max\{c \mid V_j \succeq c\, V_\ell \quad \text{for all} \quad j, \ell\} = \min_{j,\ell,k} \lambda_k \left(V_\ell^{-1/2} V_j V_\ell^{-1/2} \right), \tag{2}$$

where $\lambda_1(A), \ldots, \lambda_d(A)$ denote the d eigenvalues of a symmetric d by d matrix A. The HDBT ratio of a clustering is the HDBT ratio of its scatter matrices. Hathaway showed in the univariate context that, besides guaranteeing the maximum likelihood estimate and its consistency, the HDBT constraints removed many undesirable local optima. In the clustering context, Pollard (1981) proves (for the homoscedastic, spherical normal model) that the optimal solution is consistent in a certain sense. This means that the global maximum is the favorite solution if the data set is large. But the asymptotic nature of this result must not be overlooked. If the data set is small or of medium size then experience shows that the optimal solution may again be undesirable. We will show here that the HDBT constraints are of benefit also in the clustering context.

Outliers, i.e., observations discordant with the posited populations, are known to severely hamper the performance of statistical methods, see Barnett and Lewis (1994); Ritter and Gallegos (1997); Becker and Gather (1999). Clustering algorithms deemed to be robust actually break down under the influence of a single gross outlier, see García-Escudero and Gordaliza (1999). Nevertheless, there are nowadays some robust trimming methods based on classification models. Cuesta-Albertos et al. (1997) and García-Escudero and Gordaliza (1999) proposed a trimmed extension of the k-means algorithm conjecturing on the basis of empirical studies that its breakdown point applied to "well-structured" data sets could be large. Gallegos and Ritter (2005) undertook a mathematical analysis of a trimmed homoscedastic classification model obtaining among other things a high asymptotic breakdown point of the covariance

matrices. The mean values turn out to be more fragile but we were able to show that their maximum likelihood estimates, too, are robust in the presence of well-separated data sets. The majority of data sets is neither spherical nor homoscedastic and it is desirable to extend these methods and results to the general heteroscedastic case. However, it is well known that homoscedasticity cannot be dispensed with without additional cost since the very existence of a maximum likelihood or maximum a posteriori estimate already poses a problem. Moreover, one cannot expect robustness if clusters with arbitrarily different covariance matrices are allowed.

To our knowledge, the first heteroscedastic, normal classification model with full covariance structure and trimming is Rocke and Woodruff (1999) MINO. Besides trimming they used also constraints on the cluster sizes n_j, $1 \leq j \leq g$, in order to enforce the existence of maximum likelihood estimates. The constraints $n_j \geq d+1$ protect scatter matrices against singularity if the data are in general position. Gallegos and Ritter (2009) extended their method to maximum a posteriori estimation and showed that their algorithm leads to a standard problem from combinatorial optimization, λ-*assignment*, a special transportation problem. Despite trimming, these methods do not act robustly on all data sets. García-Escudero et al. (2008) present a constrained heteroscedastic trimming algorithm relaxing the requirements on sphericity in García-Escudero and Gordaliza (1999) and of equality of shapes in Gallegos and Ritter (2005). They also prove convergence of the parameter estimates as the size of the data set tends to infinity. The limit is given by the parameters obtained from the related cluster criterion for the underlying mixture if they are unique. However, their constraints lack affine equivariance. Here, we propose and analyse a robust, affine equivariant, heteroscedastic, full normal classification model. Specializations to normal submodels such as the diagonal or spherical are immediate and left to the interested reader.

1.2 Outline

In Sect. 2, we first use again the statistical clustering model with "spurious" outliers presented in Gallegos and Ritter (2005, 2009) in order to derive a heteroscedastic clustering criterion with trimming. Its maximum exists provided that some constraints are introduced. In the normal case, contrary to Rocke and Woodruff (1999) and Gallegos and Ritter (2009), we apply here the HDBT constraints (1) on the covariance matrices obtaining a trimmed, heteroscedastic, affine equivariant cluster criterion, the *Trimmed Determinant Criterion* (TDC). It is the extension of the homonymous criterion appearing in Gallegos and Ritter (2005) to the heteroscedastic case. We propose and substantiate an iterative and alternating reduction step for finding MDP's w.r.t. the posterior density. It consists of three successive steps: maximum likelihood estimation of parameters, maximum a posteriori classification, and trimming.

Of course, the minimizer of the TDC depends on the constant c in (1). The space of possible solutions increases as c decreases. However, if c is chosen too small, the optimal clustering turns out to be undesirable in many cases of real and synthetic data sets, see Sect. 5. Although it provides optimal fit of estimated populations and clusters

it may be unbalanced in the sense that its HDBT ratio is excessively small. In most applications, cluster balance turns out to be an important asset of a credible solution. Since the solution with the best *fit* often lacks sufficient *balance* we need a trade-off between the two and solutions which combine a large posterior density with a large HDBT ratio are more promising. This means that we are facing a problem of *biobjective optimization*. Making a compromise by optimizing the target function under a fixed constraint c is not advisable for two reasons. First it introduces a parameter in the algorithm that must be known a priori. What is more, the optimal solution under the HDBT constraints is hard to find, at least in the multivariate case. The crux is the estimation step. In Sect. 2.4, we propose instead a heuristic method based on a plot of the posterior density versus the HDBT ratio of MDP's or local optima for finding reasonable clusterings together with a constant c.

The aim of a trimming algorithm is robustness. We show here that, as an additional benefit besides existence of solutions and balance, the HDBT constraints render the estimates obtained from the TDC *robust*. Mutatis mutandis, the properties of the homoscedastic case, Gallegos and Ritter (2005), remain valid if the HDBT constraints are used instead. Constraints serving a similar purpose can be designed for statistical models other than normality. The method first uses the number of clusters and the number of discarded elements as fixed parameters. In Sect. 2.5, we comment on their choice.

In Sects. 3 and 4, we offer a theoretical robustness analysis of the TDC estimates showing first that the estimates of the covariance matrices are indeed robust under the HDBT constraints. The same cannot be said about the location parameters if arbitrary data sets are allowed, Sect. 4. However, the question of their robustness has an affirmative answer for data sets that possess a certain *separation property*. The larger the constraint c is the more robust the method turns out to be. These results are obtained from a mathematical analysis of breakdown points.

Thus, the consideration of HDBT ratio and constraints serves five purposes: it guarantees a solution, it reduces local optima, it avoids spurious clusters, it adds robustness, and it is a key to feasible solutions. In the final Sect. 5, we report on our experience with two numerical data sets.

1.3 Notation

The n *elements* or *objects* to be clustered are numbered $1, \ldots, n$. Associated with them are n *observations* or *data points* $x_1, \ldots, x_n$ in a sample space E which we collect in the data set $D = \{x_1, \ldots, x_n\}$. Given natural numbers $g \geq 2$ and $r \leq n$, a *solution* of the trimming and clustering problem is given by an r-element subset $R \subseteq \{1, \ldots, n\}$ and a partition of R in g groups or clusters $C_1, \ldots, C_g$. It is most easily specified by an array $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_n)$ of *labels* $\ell_i, 0 \leq \ell_i \leq g$, which is *admissible* in the sense that exactly $n - r$ *labels* ℓ_i are 0. If $\ell_i \geq 1$ then object i is *retained* and assigned to class ℓ_i. If $\ell_i = 0$ then object i is *discarded*, i.e., not assigned to a class. (If $\boldsymbol{\ell}$ is a "meaningful" assignment then discarded objects may be regarded as "outliers" and retained elements as "regular.") The g clusters defined by the assignment $\boldsymbol{\ell}$ are written $C_j(\boldsymbol{\ell}) = \{i \mid \ell_i = j\}, 1 \leq j \leq g$.

Their cardinalities are $n_j = n_j(\boldsymbol{\ell}) = |C_j(\boldsymbol{\ell})|$ and we have $\sum_{j=1}^{g} n_j = r$. We denote the set of all admissible assignments by Λ_r and allow one or more clusters to be empty.

We also consider g distributional models (*classes*) on E with class-specific parameters $\gamma_j \in \Gamma_j$ and density functions $f_{\gamma_j}, 1 \leq j \leq g$. The joint parameter $\gamma = (\gamma_1, \ldots, \gamma_g)$ is contained in some subspace $\Gamma \subseteq \Gamma_1 \times \cdots \times \Gamma_g$ of the product. Given an assignment $\boldsymbol{\ell} \in \Lambda_r$ with retained objects R and a parameter $\gamma \in \Gamma$, we abbreviate $f[R \mid \boldsymbol{\ell}, \gamma] = \prod_{j=1}^{g} \prod_{i \in C_j(\boldsymbol{\ell})} f_{\gamma_j}(x_i)$. We call it the *trimmed likelihood function*. If $E = \mathbb{R}^d$ and if the model is normal then $\gamma_j = (m_j, V_j)$ with the location parameters $m_j \in \mathbb{R}^d$ and the covariance matrices $V_j \in \mathrm{PD}(d)$, the cone of symmetric, positive-definite d by d matrices. We gather $\mathbf{m} = (m_1, \ldots, m_g)$ and $\mathbf{V} = (V_1, \ldots, V_g)$. We will often need the positive semidefinite or Löwner ordering $\preceq$ on $\mathrm{PD}(d)$.

Estimates of the parameters γ_j, m_j, and V_j w.r.t. an assignment $\boldsymbol{\ell}$ are denoted by $\gamma_j(\boldsymbol{\ell}), m_j(\boldsymbol{\ell})$, and $V_j(\boldsymbol{\ell})$, respectively. We also abbreviate $\gamma(\boldsymbol{\ell}) = (\gamma_1(\boldsymbol{\ell}), \ldots, \gamma_g(\boldsymbol{\ell}))$, $\mathbf{m}(\boldsymbol{\ell}) = (m_1(\boldsymbol{\ell}), \ldots, m_g(\boldsymbol{\ell}))$, $\mathbf{V}(\boldsymbol{\ell}) = (V_1(\boldsymbol{\ell}), \ldots, V_g(\boldsymbol{\ell}))$. A bar as in $\overline{x}$ denotes a sample mean and the letters W and S indicate (pooled) SSP matrices and scatter matrices, respectively. The precise meaning becomes clear from various additional specifications as subscripts or in parentheses. E.g., $\overline{x}_T = \frac{1}{|T|} \sum_{i \in T} x_i$ is the sample mean of a non-empty subset $T \subseteq \{1, \ldots, n\}$, $W_T = \sum_{i \in T} (x_i - \overline{x}_T)(x_i - \overline{x}_T)^{\mathrm{T}}$ ($S_T = \frac{1}{|T|} W_T$) is its SSP matrix (scatter matrix), and $W(\boldsymbol{\ell}) = \sum_{j=1}^{g} W_{C_j(\boldsymbol{\ell})}$ ($S(\boldsymbol{\ell}) = \frac{1}{r} W(\boldsymbol{\ell})$) is the pooled SSP matrix (pooled scatter matrix) of the retained elements w.r.t. $\boldsymbol{\ell}$. Likewise, $\overline{x}_j(\boldsymbol{\ell}) = \overline{x}_{C_j(\boldsymbol{\ell})}$, $W_j(\boldsymbol{\ell}) = W_{C_j(\boldsymbol{\ell})}$, and $S_j(\boldsymbol{\ell}) = S_{C_j(\boldsymbol{\ell})}$. Sample means and SSP and scatter matrices of empty clusters are put to zero.

The entropy of a probability vector $(p_1, \ldots, p_g)$ is $\mathrm{H}(p_1, \ldots, p_g) = -\sum_{j=1}^{g} p_j \ln p_j$. Finally, $\boldsymbol{\ell}^*$ denotes an optimal assignment and a $*$ indicates parameter estimates w.r.t. $\boldsymbol{\ell}^*$. E.g., $m_j^* = m_j(\boldsymbol{\ell}^*)$ is the estimated mean of its jth cluster.

2 Statistical model, criteria and algorithm

Gallegos and Ritter (2005, 2009) established parametric classification models with trimming for data with so-called "spurious" outliers for a data set D of n observations in some sample space E as explained in Sect. 1.3. At least $r \leq n$ of the data are regular, i.e., independent draws from the g class-specific densities $f_{\gamma_1}, \ldots, f_{\gamma_g}$, each, $(\gamma_1, \ldots, \gamma_g) \in \Gamma$. The remaining $n - r$ observations may, but do not have to be gross outliers. Besides the population parameters, their assignment to the g classes and the number of occurrences of each class are unknown. Therefore, the number of classes, the number of outliers in the data set, the outliers themselves, the class assignments, and the population parameters are subject to estimation. Applying the ideas presented in these papers to the present context of constrained parameters $\Gamma \subseteq \Gamma_1 \times \cdots \times \Gamma_g$, we obtain a trimmed a posteriori density, i.e., the a posteriori probability w.r.t. the assignment and the likelihood function w.r.t. the parameters γ_j. We use it here as our starting point referring the interested reader to the communications cited above for the details.

2.1 The trimmed a posteriori cluster criterion

The *trimmed a posteriori log-density* for (ℓ, γ) in the setup just described is

$$\begin{aligned} &- r\mathrm{H}\left(\frac{n_1(\ell)}{r}, \ldots, \frac{n_g(\ell)}{r}\right) + \ln f\left[R \mid \ell, \gamma\right] \\ &\quad = -r\mathrm{H}\left(\frac{n_1(\ell)}{r}, \ldots, \frac{n_g(\ell)}{r}\right) + \sum_{j=1}^{g} \sum_{i:\ell_i=j} \ln f_{\gamma_j}(x_i). \end{aligned}$$

It implies the *trimmed maximum a posteriori cluster criterion*

$$\begin{aligned} &- r\mathrm{H}\left(\frac{n_1(\ell)}{r}, \ldots, \frac{n_g(\ell)}{r}\right) + \ln f[R \mid \ell, \gamma(\ell)] \\ &\quad = -r\mathrm{H}\left(\frac{n_1(\ell)}{r}, \ldots, \frac{n_g(\ell)}{r}\right) + \max_{\gamma \in \Gamma} \sum_{j=1}^{g} \sum_{i:\ell_i=j} \ln f_{\gamma_j}(x_i) \end{aligned} \tag{3}$$

to be maximized w.r.t. all admissible assignments ℓ. The use of the entropy H of the cluster proportions $n_j(\ell)/r$ goes back to Symons (1981) and accounts for unequal cluster sizes. It distinguishes the maximum a posteriori from the maximum likelihood estimator.

It must be noted that the maximum w.r.t. $\gamma \in \Gamma$ required in criterion (3) does not exist in general for all Γ and all $\ell \in \Lambda_r$. If $\Gamma = \Gamma_1 \times \cdots \times \Gamma_g$, i.e., if the parameters γ_j may be chosen freely in the factors Γ_j then the maximum, if it exists, and the sum over j commute so that the double sum reduces to

$$\sum_{j=1}^{g} \max_{\gamma \in \Gamma_j} \sum_{i:\ell_i=j} \ln f_{\gamma}(x_i) = \sum_{j=1}^{g} \sum_{i:\ell_i=j} \ln f_{\gamma_j(\ell)}(x_i). \tag{4}$$

Sometimes, the maximum likelihood estimate $\gamma_j(\ell)$ w.r.t. $C_j(\ell)$ appearing here does not exist, e.g. in a normal model if $C_j(\ell)$ is too small. The problem may be circumvented in various ways. A first is restricting Γ (or parts of it) to a compact subset (together with continuity of the likelihoods $\gamma \mapsto f_\gamma(x)$). This has the effect that the estimator looses equivariance. A second way requires that each cluster should contain sufficiently many data points together with an assumption on their locations such as "general position" (affine independence of any $d+1$ elements) in the normal case, see Rocke and Woodruff (1999) and Gallegos and Ritter (2009). If the data are in general position and if we allow only assignments ℓ with cluster sizes $n_j(\ell) \geq b$ for some lower bound $b \geq d+1$ then the maximum of criterion (3) exists with *free* parameters and (4) shows that, up to a constant, the criterion reduces to minimization of

$$2r\mathrm{H}\left(\frac{n_1(\ell)}{r}, \ldots, \frac{n_g(\ell)}{r}\right) + \sum_{j=1}^{g} n_j(\ell) \cdot \ln \det S_j(\ell); \tag{5}$$

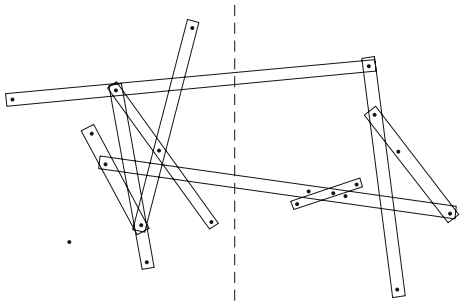

Fig. 1 A synthetic data set with two clusters of ten points, each, randomly sampled from the normal distributions N_{-2e_1,I_2} and N_{2e_1,I_2}, respectively (separated by the *dashed line*). There are no outliers. Shown are nine almost collinear "spurious clusters." The partitions defined by them all mask the genuine partition in two clusters, their negative log-posteriors (5) falling below its value 65.96. However, the HDBT ratio (2) of the genuine partition is 1/1.69 whereas the largest of the spurious ones shown is 1/2757 (the cluster of five points). The optimal unconstrained solution uses the uppermost horizontal cluster and has a negative log-posterior (5) of 60.95 but an HDBT ratio of 1/66 244

here $S_j(\boldsymbol{\ell})$ is the scatter matrix of cluster j w.r.t. $\boldsymbol{\ell}$. (In the outlier-free context, see also Symons (1981), criterion (11). The SSP matrix appearing there must be replaced with the scatter matrix which was plainly intended.) In this case, the estimates of means and covariance matrices are the sample means and scatter matrices of the optimal clusters. However, the sizes or shapes of the estimated covariance matrices may sometimes be too different to be credible, cf. Fig. 1, so that the lower bound b has to be properly chosen. We will follow here a third way using the HDBT constraints (1) on the covariance matrices.

2.2 The normal case: Trimmed Determinant Criterion

We next specialize criterion (3) to the general normal case with parameters $\gamma_j = (m_j, V_j)$, $1 \le j \le g$, under the HDBT constraints. Letting

$$\mathcal{V}_c = \{\mathbf{V} = (V_1, \ldots, V_g) \mid V_j \succ 0,\ V_j \succeq cV_\ell \quad \text{for all } j, \ell,\ 1 \le j, \ell \le g\}, \quad 0 < c \le 1,$$

we show first that the HDBT constraints guarantee the existence of the maximum w.r.t. the population parameters γ in criterion (3). We need an analytic lemma.

Lemma 1 *If the data D are in general position and if $r \ge gd + 1$ then, for any assignment $\boldsymbol{\ell} \in \Lambda_r$ (some clusters may be empty), the minimum of*

$$\sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \left[\ln \det V_j + \operatorname{tr}\left(V_j^{-1} S_j(\boldsymbol{\ell}) \right) \right]$$

w.r.t. $\mathbf{V} \in \mathcal{V}_c$ exists for any $0 < c \le 1$.

Proof The HDBT constraints imply $\det V_j \geq \det(cV_\ell)$ and $V_j^{-1} \preceq cV_\ell^{-1}$. Hence, we have for any $1 \leq \ell \leq g$

$$\sum_{j=1}^{g} n_j \left[\ln\det V_j + \operatorname{tr}\left(V_j^{-1} S_j(\boldsymbol{\ell})\right)\right] \geq \sum_{j=1}^{g} n_j \left[\ln\det(cV_\ell) + \operatorname{tr}\left(cV_\ell^{-1} S_j(\boldsymbol{\ell})\right)\right]$$
$$= r \ln\det(cV_\ell) + c \operatorname{tr}\left(V_\ell^{-1} W(\boldsymbol{\ell})\right),$$

where $W(\boldsymbol{\ell})$ is the pooled SSP matrix specified by $\boldsymbol{\ell}$. By assumption there is some cluster, say ℓ, of size $n_\ell(\boldsymbol{\ell}) \geq d+1$. By general position, its SSP matrix is positive definite so that $W(\boldsymbol{\ell}) \geq \varepsilon I_d$ with some constant $\varepsilon > 0$ that depends only on the data. Hence

$$\sum_{j=1}^{g} n_j \left[\ln\det V_j + \operatorname{tr}\left(V_j^{-1} S_j(\boldsymbol{\ell})\right)\right] \geq r \ln\det(cV_\ell) + \varepsilon c \operatorname{tr} V_\ell^{-1}.$$

As $\mathbf{V}$ approaches the boundary of $\mathcal{V}_c$, i.e., as some V_j approaches the boundary of $\mathrm{PD}(d)$, again by the HDBT constraints, so does V_ℓ. It is well known that this implies that the right, and hence the left side of the above inequality tends to ∞. This proves the claim. □

Now standard normal estimation theory shows that, for any admissible assignment $\boldsymbol{\ell}$, the partial maximizer w.r.t. the means m_j in (3) [here, $\gamma_j = (m_j, V_j)$] depends only on $C_j(\boldsymbol{\ell})$ and is given by the sample means of the clusters defined by $\boldsymbol{\ell}$,

$$m_j(\boldsymbol{\ell}) = \begin{cases} \overline{x}_j(\boldsymbol{\ell}), & \text{if } C_j(\boldsymbol{\ell}) \neq \emptyset, \\ \text{arbitrary, e.g. } 0, & \text{otherwise,} \end{cases} \quad 1 \leq j \leq g. \tag{6}$$

Omitting the entropy term, the partial maximum w.r.t. the location parameter $\mathbf{m}(\boldsymbol{\ell})$ is

$$\text{const} - \frac{1}{2} \sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \left[\ln\det V_j + \operatorname{tr}\left(V_j^{-1} S_j(\boldsymbol{\ell})\right)\right].$$

According to Lemma 1, this expression attains its maximum w.r.t. $\mathbf{V} \in \mathcal{V}_c$, i.e., under the HDBT constraints for any $0 < c \leq 1$. Summing up, after a change of sign, the (HDBT constrained) trimmed maximum a posteriori cluster criterion (3) becomes in the normal case the (heteroscedastic) Trimmed Determinant Criterion

$$r \cdot \mathrm{H}\left(\frac{n_1(\boldsymbol{\ell})}{r}, \ldots, \frac{n_g(\boldsymbol{\ell})}{r}\right) + \min_{\mathbf{V} \in \mathcal{V}_c} \frac{1}{2} \sum_{j=1}^{g} n_j(\boldsymbol{\ell}) \left[\ln\det V_j + \operatorname{tr}\left(V_j^{-1} S_j(\boldsymbol{\ell})\right)\right]. \tag{TDC}$$

It is to be minimized w.r.t. all $\boldsymbol{\ell} \in \Lambda_r$ and it contains the scatter matrices $S_j(\boldsymbol{\ell})$ of $C_j(\boldsymbol{\ell})$. Finally, we denote the minimizing assignment by $\boldsymbol{\ell}^*$, $R^* = \{i \mid \ell_i \neq 0\}$ is the set of regular elements w.r.t. $\boldsymbol{\ell}^*$, and the partition of R^* associated with $\boldsymbol{\ell}^*$ is

$(C_1^*, \ldots, C_g^*)$. The optimal assignment $\boldsymbol{\ell}^*$ induces estimates m_j^* and V_j^* of the location and scale parameters m_j and V_j which we call the TDC *parameter estimates*. They are $m_j^* = m_j(\boldsymbol{\ell}^*)$ as in (6), and the minimizers w.r.t. $\mathbf{V} \in \mathcal{V}_c$ appearing in the TDC, where $\boldsymbol{\ell}^*$ is inserted for $\boldsymbol{\ell}$.

There are only few cases where the minimizing parameters V_j for given $\boldsymbol{\ell}$ are known to us in closed form. One is the unconstrained model where they are the scatter matrices if clusters are large enough. If the scatter matrices happen to satisfy the constraints then they are the solutions also in the constrained case. Another is the homoscedastic case, $c = 1$, where the common estimate of the V_j's is the pooled scatter matrix $S(\boldsymbol{\ell})$, see Sect. 1.3; up to an additive constant, the TDC reduces to

$$r\left\{\mathrm{H}\left(\frac{n_1(\boldsymbol{\ell})}{r}, \ldots, \frac{n_g(\boldsymbol{\ell})}{r}\right) + \frac{1}{2}\ln\det S(\boldsymbol{\ell})\right\}. \tag{7}$$

Without the entropy term, this is the criterion of the same name derived in Gallegos and Ritter (2005). Finally, a univariate case is treated in Proposition 1.

The optimal clustering may contain empty clusters, an indication that the number of clusters g has been chosen too large. E.g., if a data set is a clear sample from a single univariate normal population then the optimal partition in two clusters will leave one cluster empty. A simple example is $n = r = 4$, $D = \{0, 3, 4, 7\}$, and $c = 1$. Some values of the criterion (7) are

$$\begin{cases} 3.66516, & \text{for the partition } \{D, \emptyset\}, \\ 3.79572, & \text{for the partition } \{\{0, 3, 4\}, \{7\}\}, \\ 4.39445, & \text{for the partition } \{\{0, 3\}, \{4, 7\}\}. \end{cases}$$

The remaining partitions need not to be considered, either by symmetry or since they cannot be optimal. Hence the method returns a single nonempty cluster. Empty clusters become less likely as c decreases.

2.3 Minimum distance partitions and optimization

Several strategies for optimizing the TDC are available, among them local descent methods on a suitably defined graph structure on Λ_r and alternating methods of type k-means. An apparent disadvantage of these methods is their getting stuck in suboptimal solutions such as local minima or MDP's. A closer analysis of the situation shows however that particular suboptimal solutions often deserve more attention than the absolute optimum of the criterion itself. It is therefore interesting to generate local optima and MDP's.

We propose next an alternating method of type k-means for producing MDP's of criterion (3). We begin by rewriting the trimmed posterior density in a different form:

$$-r\mathrm{H}\left(\frac{n_1(\boldsymbol{\ell})}{r},\ldots,\frac{n_g(\boldsymbol{\ell})}{r}\right)+\sum_{j=1}^{g}\sum_{i:\ell_i=j}\ln f_{\gamma_j}(x_i)=\sum_{i:\ell_i\neq 0}\left(\ln\frac{n_{\ell_i}}{r}+\ln f_{\gamma_{\ell_i}}(x_i)\right)$$
$$=\sum_{i:\ell_i\neq 0}u_{i,\ell_i}$$

with $u_{i,j}=\ln\frac{n_j}{r}+\ln f_{\gamma_j}(x_i)$, the logarithm of the posterior probability of j for x_i. For given parameters γ_j, this sum is maximized w.r.t. $\boldsymbol{\ell}$ by assigning each object i according to the maximum a posteriori discriminant rule and by discarding the $n-r$ observations with the overall smallest posterior probabilities. Given a labelling $\boldsymbol{\ell}$, the unconstrained maximum in criterion (3) chooses as γ the (unconstrained) maximum likelihood estimate for the retained observations. As a consequence the following strategy improves the criterion (3) starting from an initial admissible labelling $\boldsymbol{\ell}$. It extends the k-means algorithm and its generalization by Schroeder (1976) to trimming. We first keep the parameters g and r fixed.

Multipoint reduction step

// Input: An admissible labelling $\boldsymbol{\ell}$;
// Output: An admissible labelling $\boldsymbol{\ell}_{\text{new}}$ with larger value of the criterion *or* the response "fail."
Estimation: *if* some cluster $C_j(\boldsymbol{\ell})$ does not allow maximum likelihood estimation of its parameters, respond "fail;"
else update each γ_j with the maximum likelihood estimate for $C_j(\boldsymbol{\ell})$ (no constraints);
Classification: assign each of the n objects i to the cluster j with maximum posterior probability $u_{i,j}$ to obtain a labelling $\boldsymbol{\ell}'$;
Trimming: discard the $n-r$ objects i with smallest values u_{i,ℓ'_i} from $\boldsymbol{\ell}'$ to obtain $\boldsymbol{\ell}_{\text{new}}$;

In the Classification step *all* misfits are removed, hence the name "multipoint." (Other schemes are possible, e.g., the "single-point" reduction step which removes just one misfit. They all improve the criterion.) In the Trimming step the r observations which best fit in their clusters are retained. Note that either step may leave one or more clusters empty. Iteration of the three steps will eventually result in a stationary configuration since there is only a finite number of labellings and since the criterion continually improves. The solution attained at convergence is self-consistent [or a (*free*) *minimum distance partition*] in the sense that clustering and parameters generate each other.

The preceding reduction step disregards constraints (e.g. HDBT if the TDC is considered) in the Estimation step. We claim that MDP's at the boundary of the constraints do not deserve much interest. A solution at the boundary depends on the precise value of c (see Proposition 1). But there is no precise value since it is unknown. Moreover, even if c were known, the Estimation step would require the maximum likelihood estimate w.r.t. $\mathcal{V}_c$. We do not know of a practicable analytical solution to the associated constrained optimization problem in Euclidean space $\mathbb{R}^d$ for $d\geq 2$ and numerical

methods such as gradient descent would lead to inefficient overall algorithms. An exception are free MDP's that happen to satisfy the constraints—they are automatically *constrained* MDP's. The constant c must be estimated together with the assignment and the other parameters. In Sect. 2.4, we will propose a method based on free MDP's or free local optima.

We can say more in the univariate case. Given $\boldsymbol{\ell}$, we denote the sample variance of cluster j by s_j and $w_j = n_j s_j$. Our next proposition deals with arbitrary g and covers the general constrained case if $g = 2$.

Proposition 1 *Let $d = 1$, let $g \geq 2$, let $r \geq g+1$, and let $0 < c \leq 1$. Let $\boldsymbol{\ell}$ be such that the sample variances s_j satisfy $s_2 > 0$ and $cs_\ell \leq s_j \leq s_\ell/c$ for all $3 \leq j \leq g,\ \ell < j$.*[1] *(In other words, the sample variances satisfy the HDBT constraints except, possibly, for the pair s_1, s_2.) Then partial minimization of the TDC w.r.t. $\mathbf{V} = (v_1, \ldots, v_g) \in \mathcal{V}_c$ is solved by*

$$\begin{cases} v_1(\boldsymbol{\ell}) = s_1,\ v_2(\boldsymbol{\ell}) = s_2, & \text{if } cs_1 \leq s_2 \leq s_1/c, \\ v_1(\boldsymbol{\ell}) = \frac{w_1+w_2/c}{n_1+n_2},\ v_2(\boldsymbol{\ell}) = \frac{cw_1+w_2}{n_1+n_2}, & \text{if } s_2 < cs_1, \\ v_1(\boldsymbol{\ell}) = \frac{w_1+cw_2}{n_1+n_2},\ v_2(\boldsymbol{\ell}) = \frac{w_1/c+w_2}{n_1+n_2}, & \text{if } s_1 < cs_2, \end{cases}$$

and $v_j(\boldsymbol{\ell}) = s_j,\ 3 \leq j \leq g$.

Proof Let us abbreviate $h_j(v) = n_j(\ln v + \frac{s_j}{v})$. In the present case, the partial minimum of the TDC w.r.t. $(v_1, \ldots, v_g)$ can be rewritten in the form (omitting the entropy term)

$$\begin{aligned} h &:= \min_{\substack{v_1>0 \\ cv_\ell \leq v_j \leq v_\ell/c,\ \ell<j}} \sum_{j=1}^{g} h_j(v_j) \\ &= \min_{v_1>0} \left\{ h_1(v_1) + \min_{cv_1 \leq v_2 \leq v_1/c} \left\{ h_2(v_2) + \min_{\substack{cv_\ell \leq v_j \leq v_\ell/c \\ \ell<j,\ j\geq 3}} \sum_{j\geq 3} h_j(v_j) \right\} \right\} \\ &\geq \min_{v_1>0} \left\{ h_1(v_1) + \min_{cv_1 \leq v_2 \leq v_1/c} \left\{ h_2(v_2) + \sum_{j\geq 3} \min_{v>0} h_j(v) \right\} \right\} \\ &= \min_{v_1>0} \left\{ h_1(v_1) + \min_{cv_1 \leq v_2 \leq v_1/c} h_2(v_2) \right\} + \sum_{j\geq 3} h_j(s_j), \end{aligned}$$

[1] This presupposes that the clusters $2, \ldots, g$ contain at least two different elements, each.

since h_j, $j \geq 3$, assumes its unconstrained minimum at $v = s_j$ (> 0). The constrained minimizer of $h_2(v_2)$ w.r.t. v_2 is

$$\widetilde{v}_2(v_1) = \begin{cases} s_2, & cs_2 < v_1 < s_2/c, \\ cv_1, & v_1 \geq s_2/c, \\ v_1/c, & v_1 \leq cs_2, \end{cases}$$

and we have thus shown

$$h \geq \min_{v_1 > 0} \{h_1(v_1) + h_2(\widetilde{v}_2(v_1))\} + \sum_{j \geq 3} h_j(s_j). \tag{8}$$

The function $v_1 \mapsto h_2(\widetilde{v}_2(v_1))$ is differentiable, monotone decreasing in $]0, cs_2]$, constant in $[cs_2, s_2/c]$, and monotone increasing in $[s_2/c, \infty[$. It follows that the sum $v_1 \mapsto h_1(v_1) + h_2(\widetilde{v}_2(v_1))$ has a minimum which is attained in the interval where the minimum of the unimodal function $h_1(v_1)$ is located. The minimizer of the lower bound (8) turns out to be the value $v_1(\ell)$ given in the proposition.

We have, thus, shown that the target function is nowhere less than its value at the parameters stated in the proposition. The proof will be finished if we show that these parameters satisfy the HDBT constraints. This is true by assumption for all pairs (j, ℓ), $j, \ell \geq 3$, and was ensured for the pair $(1, 2)$. The remaining pairs $(1, j)$, $(2, j)$, $j \geq 3$, follow from elementary estimates based on the constraints assumed for (s_1, s_j) and (s_2, s_j). The condition $r \geq g + 1$ ensures that the minimum w.r.t. $v_1 > 0$ exists so that $v_j(\ell) > 0$ for all j. □

2.4 Overall algorithm and choice of the constant c

Iteration of (unconstrained) multipoint reduction steps strives for labellings with large values of criterion (3) (or small values of the TDC). If the "fail" signal does not occur then the iteration stalls at some unconstrained MDP for the reasons stated before. However, it does not necessarily represent an interesting solution so that the process has to be replicated, for possibly many different, randomly or expediently chosen, initial assignments or parameters. The number of replications needed depends on the data set and on the initial assignments.

Two different outcomes of this algorithm are possible. It may happen that all replications return the signal "fail." This occurs typically if the data set contains very small clusters or if the number of clusters, g, has been chosen too large. E.g., if one attempts to group a d-dimensional data set of less than $g(d + 1)$ elements in g clusters (normal model, $r = n$) then "fail" signals, only, are returned. In this case, the parameters g and/or r must be adapted. Reducing r discards very small clusters. Moreover, clusters large enough to allow estimation of their parameters can be enforced by putting lower bounds on cluster sizes in the reduction step if r is large enough, cf. Gallegos and Ritter (2009).

Otherwise, we obtain unconstrained MDP's and we have to decide which one to use. The optimum of the criterion does not guarantee a reasonable clustering as experience

shows and a solution close to the desired one cannot be estimated without a further assumption. In most normal cases, we are interested in solutions that combine a large value of the criterion with a large HDBT ratio. Of course, this is not a law. Rescaling Fig. 1 in such a way that the five-point cluster becomes spherical, we obtain an oblong, vertical data set which contains the quintuple as a region of concentration. This might suggest a partition in two clusters with five and 15 elements. But we contend that this is not the point of view to be taken in general. The criterion measures how well the estimated populations *fit* their clusters. Declaring the HDBT ratio of a solution a measure of its *balance*, we postulate that, in general, it is good fit *combined* with high balance that characterizes a feasible solution. Since it occurs only rarely that the best fitting solution enjoys high (but not the highest) balance, this leads to a *biobjective optimization problem* which calls for a compromise.

Here is a simple heuristic method that finds a well-fitting, balanced clustering together with a constant c: Generate a large number of (unconstrained) MDP's and display their HDBT ratios versus the values of their criteria in a negative double-logarithmic plot as shown in Figs. 3 and 5 in Sect. 5. The convex hull of all MDP's will usually show a knee at its left lower part. The extreme point at the knee determines the favorite solution and c. Often, the MDP's are supported from below by an almost horizontal line segment and this MDP is found close to its left end. It is not unusual that it has an HDBT ratio of a hundredth or less. The plot provides also some guidance about the number of replications needed. Run the algorithm until its convex hull has stabilized.

The method may also be applied with local optima instead of MDP's.

2.5 Choice of the parameters g and r

Criteria and reduction step (or steepest descent) depend on two parameters, the number of clusters g and the number of retained elements r. So far we have designed a tool that allows us to establish interesting clusterings for arbitrary but fixed pairs (g, r). This is a substantial reduction of the complexity of the data analytic problem but the task of reducing the number of pairs, maybe even to one, remains. For obvious reasons, r should be chosen no larger than and close to the (unknown) number of regular elements in the data set. We give some guidelines for the selection of g and r.

Recently, Neykov et al. (2007) proposed a simple method that estimates both parameters at a time, the *trimmed BIC*. They establish a table of BIC values indexed by g and r proposing to use the parameter values where the minima w.r.t. g stabilize. There are many other methods and we compile first some known methods for estimating the number of classes of uncontaminated data sets.

2.5.1 The number of classes of uncontaminated data

For the number of clusters there are essentially three approaches, cf. Milligan and Cooper (1985) and Gordon (1999), *cluster validation*, the so-called *elbow criterion*, and *model selection criteria*. Cluster validation assesses the quality of a partition and may be divided in two branches: tests and validity measures. The classical test, due to

Wolfe (1970), is a likelihood ratio test for the hypothesis of k clusters against $(k-1)$ clusters. Bock (1985) discusses some significance tests for distinguishing between the hypothesis of a homogeneous population versus the alternative of heterogeneity. Chen et al. (2004) propose a modified likelihood ratio test for a mixture of two components versus $g \geq 3$. Also normality tests may sometimes be beneficial in this respect, see the comprehensive review by Mecklin and Mundfrom (2004). Validity measures are functionals of partitions and usually measure between cluster separation and within cluster cohesion (or "compactness"); see, e.g., Bezdek et al. (1999). In the case of (almost) spherical models, the total within-clusters sum of squared distances about the centroids is used as a measure of cohesion and the total between-clusters sum of squared distances for separation; cf. Milligan and Cooper (1985) and the abridged presentation of their work by Gordon (1999). The elbow criterion identifies the number of clusters as the location where the decrease of some cluster criterion flattens markedly. For a refinement of this method see Tibshirani et al. (2001).

Comparing the maximum likelihoods or a posteriori densities between solutions for different numbers of classes does not make sense since each additional class allows better fit so that these values increase with g. A *model selection criterion* counteracts this tendency by subtracting a penalty term that increases with g from the maximum of the log-likelihood or from the posterior log-density. Schwarz (1978) proposed his popular Bayesian Information Criterion (BIC) for exponential families. In the uncontaminated case, its penalty term is $\frac{q}{2} \cdot \ln n$, q being the total dimension of the parametric model. There is some practical evidence that supports BIC as a means for estimating the number of clusters of *mixture models*, too; see the discussion in McLachlan and Peel (2000), Ch. 6. Moreover, Kéribin (2000) described a family of penalty terms, among them BIC, which *asymptotically* as $n \to \infty$ neither over- nor underestimate the correct number of components of a mixture model $\sum_{j=1}^{g} \pi_j f_{\gamma_j}$ if the class-conditional populations satisfy certain regularity conditions and the parameters certain constraints. Her interesting result is applicable, e.g., to Gaussian families if the mean values are bounded and if the covariance matrices are bounded below in the Löwner ordering by a positive multiple of the identity matrix. In the case of a mixture, $q = q(g)$ equals $g - 1$ (for the mixing rates) plus the total number of (real) parameters of the g components.

We propose BIC with this value of q also for our clustering model if there is sufficient separation. For a justification, we compare the maximum a posteriori density (3) in an outlier-free context, $r = n$, with the maximum likelihood of the related g-class mixture model under suitable constraints as in Kéribin's theorem. Let $\boldsymbol{\ell}^*$ be the optimal maximum a posteriori assignment and let π^* and γ^* be the optimal mixing rates and population parameters of the mixture model. For any g, the optimal value of criterion (3) is no larger than that of the mixture model:

$$
\begin{aligned}
&-n\mathrm{H}\left(\frac{n_1(\boldsymbol{\ell}^*)}{n}, \ldots, \frac{n_g(\boldsymbol{\ell}^*)}{n}\right) + \sum_{j=1}^{g} \sum_{i:\ell_i^*=j} \ln f_{\gamma_j(\boldsymbol{\ell}^*)}(x_i) \\
&= \sum_i \left\{ \ln \frac{n_{\ell_i^*}(\boldsymbol{\ell}^*)}{n} + \ln f_{\gamma_{\ell_i^*}(\boldsymbol{\ell}^*)}(x_i) \right\}
\end{aligned}
$$

$$
\begin{aligned}
&= \ln \prod_i \frac{n_{\ell_i^*}(\boldsymbol{\ell}^*)}{n} f_{\gamma_{\ell_i^*}(\boldsymbol{\ell}^*)}(x_i) \le \ln \prod_i \sum_j \frac{n_j(\boldsymbol{\ell}^*)}{n} f_{\gamma_j(\boldsymbol{\ell}^*)}(x_i) \\
&\le \max_{\pi,\gamma} \ln \prod_i \sum_j \pi_j f_{\gamma_j}(x_i) \\
&= \ln \prod_i \sum_j \pi_j^* f_{\gamma_j^*}(x_i).
\end{aligned}
$$

On the other hand, if the data set is well separated in g clusters then $f_{\gamma_j^*}(x_i) \ll f_{\gamma_{\ell_i^*}^*}(x_i)$ for all $j \neq \ell_i^*$, $1 \le i \le n$, $f_{\gamma_{\ell_i^*}^*}(x_i) \approx f_{\gamma_{\ell_i^*}(\boldsymbol{\ell}^*)}(x_i)$, and $\pi_j^* \approx \frac{n_j(\boldsymbol{\ell}^*)}{n}$, $1 \le j \le g$, so that, for this g, the third and the last terms of the above chain are close and both ends almost meet. This reasoning supports BIC as a penalty term also for maximum a posteriori partitioning in the case of large data sets and good separation.

2.5.2 *The number of outliers*

An approach to estimating the number of clusters can be combined with a test for estimating the number of outliers. In a first step, establish a table of the optimal clusterings for all (reasonable) numbers of clusters, g, and all numbers of discarded elements, $n - r$. It is, of course, sufficient to perform the procedure with a lacunary set of values $n - r$. Next, reduce the number of possible solutions by validating them w.r.t. absence of outliers with a multiple testing procedure. Tests for *goodness of fit* of the regular densities and the clusters $C_j(\boldsymbol{\ell}^*)$, $1 \le j \le g$, *normality tests*, see Mecklin and Mundfrom (2004) extensive survey article, and methods for *outlier detection* or *identification*, see Becker and Gather (1999), are available for this task. If g admits an acceptable pair $(g, n-r)$, keep the one with maximum r (=: r_g) as a candidate. After having run through all values of g, at most one pair is left in each line of the table so that the complexity of the problem is again substantially reduced. It remains to choose the favorite g. Since the estimated numbers r_g of regular observations depend on g, the numbers of objects have to be normalized, e.g. to n. By consistency of parameter estimation, cf. Gallegos and Ritter (2009), Theorems 1 and 2, the value of the maximum a posteriori criterion (3) increases approximately linearly with the number r, asymptotically, at least if there is sufficient separation. Therefore, we propose to combine the TDC estimates with the following *corrected BIC* in order to estimate the number of clusters:

$$
\operatorname*{argmax}_g \left\{ -n\,\mathrm{H}\left(\frac{n_1(\boldsymbol{\ell}^*)}{r_g}, \ldots, \frac{n_g(\boldsymbol{\ell}^*)}{r_g}\right) + \frac{n}{r_g} \sum_{j=1}^{g} \sum_{\ell_i^*=j} \ln f_{\gamma_j(\boldsymbol{\ell}^*)}(x_i) - \frac{q(g)}{2} \ln n \right\}. \tag{9}
$$

Experience with various data sets has shown the effectiveness of this method, see again Gallegos and Ritter (2009).

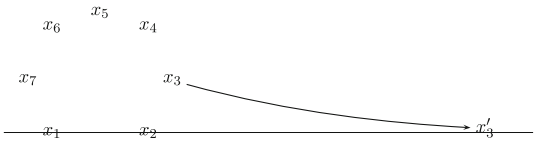

Fig. 2 Non-robustness of criterion (5) with full normal covariance matrices in the *free* heteroscedastic case with minimum cluster size 3. Data point x_3 is replaced with some x_3' close to the abscissa and far away. Minimization of the criterion (5) discards x_7 generating the clustering $\{\{x_1, x_2, x_3'\}, \{x_4, x_5, x_6\}\}$

3 Robustness

Although criterion and algorithm involve trimming, neither the estimates of the means nor those of the covariance matrices would be robust without HDBT constraints. In fact, no matter how r is chosen they would break down under the influence of a single outlier. An example is provided by the data set consisting of seven points $x_1, \ldots, x_7$ shown in Fig. 2. We use criterion (5) to subdivide it in two groups of minimum cluster size 3 and $r = 6$, i.e., we discard one object. There are two equivalent optimal clusterings, $\{\{x_2, x_3, x_4\}, \{x_5, x_6, x_7\}\}$, x_1 discarded, and $\{\{x_3, x_4, x_5\}, \{x_6, x_7, x_1\}\}$, x_2 discarded. We now replace x_3 with a distant outlier x_3' close to the abscissa, say $x_3' = (a, a^{-2})$ for large a. Although we discard one point, the criterion does not choose the "right" one, x_3'. In fact, x_3' creates together with x_1 and x_2 a cluster with a small determinant of its scatter matrix which determines the optimal clustering. This turns out to be $\{\{x_1, x_2, x_3'\}, \{x_4, x_5, x_6\}\}$, x_7 discarded. As a consequence, neither do mean and largest eigenvalue of the scatter matrix of the slim cluster remain bounded as $a \to \infty$ nor does the smallest eigenvalue remain bounded away from zero.

We show in this and the following section that the HDBT constraints do not only guarantee existence of a solution but also robustness of the TDC. Our main results are the following:

(i) If r is large enough then the TDC estimates of the covariance matrices resist $n - r + g - 1$ arbitrary replacements. On the other hand they break down under $n - r + g$ suitable replacements, see Theorem 1;
(ii) there exists a data set such that the TDC estimate of at least one mean (i.e., a sample mean) breaks down with two suitable replacements no matter how many objects we discard, see Theorem 2;
(iii) if the data set bears a clear structure of g clusters and if r is large enough and properly chosen then the TDC estimates of all means resist $n - r$ arbitrary replacements. On the other hand, it is possible to break down one mean with $n - r + 1$ suitable replacements, see Theorem 3.

3.1 Breakdown values

The finite-sample breakdown value of an estimator, Hodges (1967) and Donoho and Huber (1983), measures the minimum fraction of gross outliers that can *completely* spoil the estimate. Two types of breakdown points are customary, the *addition* and the *replacement* breakdown point. The former refers to the addition of $n - r$ outliers to

a data set of r regular observations and the latter to $n - r$ replacements in a data set of n regular observations. The former is technically simpler since the set of regular observations is *fixed*, but one needs two estimators, one for r data and one for n. By contrast, the latter considers all $\binom{n}{r}$ possible replacements of $n - r$ observations but needs only one estimator for n objects. For this reason, we deal with replacements only.

Let $\delta : \mathcal{A} \to \Theta$ be an estimator on its natural domain of definition $\mathcal{A} \subseteq E^n$ of admissible data sets of size n (e.g., "general position" in case of the m.l.e. under normal assumptions). Given $m \leq n$, we say that $M \in \mathcal{A}$ is an *m-modification* of $D \in \mathcal{A}$ if it arises from D by modifying at most m observations of D in an (admissible but otherwise) arbitrary way. An estimator δ "breaks down with D under m replacements" if the set

$$\{\delta(M) \mid M \text{ is } m\text{-modification of } D\} \subseteq \Theta$$

is not relatively compact in Θ. Of course, there is no breakdown if Θ is compact. The *individual* breakdown point for the data set D is the number

$$\beta(\delta, D) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} \mid \delta \text{ breaks down with } D \text{ under } m \text{ replacements} \right\}.$$

It is the minimal fraction of replacements in D that may cause δ to break down. The individual breakdown point is not an interesting concept per se since it depends on a single data set. It tells the statistician how many gross outliers the data set M under his or her study may contain without causing excessive damage if the imaginary "clean" data set that should have been observed were D. Now let $\mathcal{K} \subseteq \mathcal{A}$ be some subclass of admissible data sets. The *restricted* breakdown point of δ w.r.t. $\mathcal{K}$, cf. Gallegos and Ritter (2005), is

$$\beta(\delta, \mathcal{K}) = \min_{D \in \mathcal{K}} \beta(\delta, D).$$

The restricted breakdown point depends only on δ and the subclass $\mathcal{K}$. It provides information about the robustness of δ if the hypothetic "clean" data set D that should have been observed instead of the contaminated data set M had been a member of $\mathcal{K}$. Finally, we call Donoho and Huber's breakdown point the *universal* breakdown point

$$\beta(\delta) = \beta(\delta, \mathcal{A}).$$

It depends solely on the estimator. The restricted breakdown value may be seen as a relaxed version of it. We have the chain of inequalities

$$\beta(\delta) \leq \beta(\delta, \mathcal{K}) \leq \beta(\delta, D), \quad D \in \mathcal{K}.$$

We deal here with breakdown points of the estimates of the parameters $m_j \in \mathbb{R}^d$ and $V_j \in \mathrm{PD}(d)$ obtained from (minimizing) the TDC w.r.t. $\boldsymbol{\ell}$, $\mathbf{m}$, and $\mathbf{V}$. The relatively compact subsets of the parameter space $\mathbb{R}^d$ of the means are the bounded

subsets of $\mathbb{R}^d$. A subset of PD(d) is relatively compact if the eigenvalues of its elements are bounded and bounded away from zero. This is equivalent to saying that the subset is bounded above and below by positive definite matrices in the Löwner ordering $\preceq$.

We first show that the minimum of the TDC, with any $0 < c \leq 1$, provides an asymptotically robust estimate of the covariance matrices V_j and compute the universal breakdown point. We need a lemma. It exploits the pooled SSP matrix $W(\boldsymbol{\ell})$ of an admissible assignment $\boldsymbol{\ell}$.

Lemma 2 *Let* $\mathbf{V} = (V_1, \ldots, V_g) \in \mathcal{V}_c$, *let* $\mathbf{m} = (m_1, \ldots, m_g) \in \mathbb{R}^{gd}$, *let* $\boldsymbol{\ell}$ *be an admissible labelling, and let* R *be the set of retained objects w.r.t.* $\boldsymbol{\ell}$. *We have for all* $\ell, 1 \leq \ell \leq g$,

$$2 \ln f[R \mid \boldsymbol{\ell}, \mathbf{m}, \mathbf{V}] \leq -r \ln \det(2\pi c V_\ell) - c \operatorname{tr}\left(W(\boldsymbol{\ell}) V_\ell^{-1}\right).$$

Proof By the HDBT constraints, we have

$$\begin{aligned}
2 \ln f[R \mid \boldsymbol{\ell}, \mathbf{m}, \mathbf{V}] &= -\sum_{1 \leq j \leq g} \left\{ n_j(\boldsymbol{\ell}) \ln \det(2\pi V_j) + \sum_{i \in C_j(\boldsymbol{\ell})} (x_i - m_j)^{\mathrm{T}} V_j^{-1} (x_i - m_j) \right\} \\
&\leq -\sum_{1 \leq j \leq g} \left\{ n_j(\boldsymbol{\ell}) \ln \det(2\pi c V_\ell) + c \operatorname{tr} \sum_{i \in C_j(\boldsymbol{\ell})} (x_i - m_j)(x_i - m_j)^{\mathrm{T}} V_\ell^{-1} \right\} \\
&\leq -r \ln \det(2\pi c V_\ell) - c \operatorname{tr} \sum_{1 \leq j \leq g} \sum_{i \in C_j(\boldsymbol{\ell})} \left(x_i - \overline{x}_j(\boldsymbol{\ell})\right) \left(x_i - \overline{x}_j(\boldsymbol{\ell})\right)^{\mathrm{T}} V_\ell^{-1} \\
&= -r \ln \det(2\pi c V_\ell) - c \operatorname{tr}\left(W(\boldsymbol{\ell}) V_\ell^{-1}\right).
\end{aligned}$$

The following theorem deals with the universal breakdown point of the TDC estimates of the covariance matrices. □

Theorem 1 *Let the data* D *be in general position and assume* $r \geq gd + 1$.

(a) *If* $2r \geq n + g(d+1)$ *then the TDC estimates of the covariance matrices remain in a compact subset of* PD(d) *that depends only on the original data set* D *as at most* $n - r + g - 1$ *data points of* D *are replaced in an arbitrary way.*

(b) *It is possible to replace* $n - r + g$ *elements of* D *in such a way that the largest eigenvalue of the TDC estimate of some covariance matrix* (*and hence of all covariance matrices*) *exceeds any given number.*

(c) *If* $2r \geq n + g(d+1)$ *then* $\beta_{\mathrm{var}}(n, r, g) = \frac{n-r+g}{n}$.

Proof (a) We first note that, no matter what the admissibly modified data set M is, the constrained maximum posterior density and, hence, the constrained maximum likelihood $f[R^* \mid \boldsymbol{\ell}^*, \mathbf{m}^*, \mathbf{V}^*]$ remains bounded below by a strictly positive constant that depends only on the original data set D. To this end, we compare the optimal solution with one that is constrained irrespective of the constant c. Indeed, let $\boldsymbol{\ell}$ be the labelling that assigns the remaining $r - g + 1$ original points to the first cluster C_1

and $g-1$ replacements y_j to one-point clusters $C_j = \{y_j\}$, $2 \le j \le g$. Moreover, let $\mathbf{m} = (0, y_2, \ldots, y_g)$, and let $V_j = I_d$ for all $1 \le j \le g$. By optimality, we have

$$
\begin{aligned}
&- r\mathrm{H}\left(\frac{n_1(\boldsymbol{\ell}^*)}{r}, \ldots, \frac{n_g(\boldsymbol{\ell}^*)}{r}\right) + \ln f\left[R^* \mid \boldsymbol{\ell}^*, \mathbf{m}(\boldsymbol{\ell}^*), \mathbf{V}(\boldsymbol{\ell}^*)\right] \\
&\quad \ge -r\mathrm{H}\left(\frac{r-g+1}{r}, \frac{1}{r}, \ldots, \frac{1}{r}\right) + \ln f[R \mid \boldsymbol{\ell}, \mathbf{m}, I_d] = \text{const} - \frac{1}{2}\sum_{\ell_i=1} \|x_i\|^2.
\end{aligned}
$$

The right side of this expression does not depend on the replacements.

Now, by assumption, we replace at most $n-r+g-1 \le r-(gd+1)$ (≥ 0) data points of D so that, for any assignment, at least one cluster contains at least $d+1$ original points $T \subseteq D$. This is in particular true for an optimal assignment $\boldsymbol{\ell}^*$. By general position, it follows $W(\boldsymbol{\ell}^*) \succeq W_T \succeq \varepsilon I_d$ for some $\varepsilon > 0$. Lemma 2 and the initial remark imply

$$
-r \ln\det\left(2\pi c V_1^*\right) - c\,\mathrm{tr}\left(W(\boldsymbol{\ell}^*)V_1^{*-1}\right) \ge 2\ln f\left[R^* \mid \boldsymbol{\ell}^*, \mathbf{m}^*, \mathbf{V}^*\right] \ge \text{const} > -\infty.
$$

Now, it is well known that the set of matrices V_1^* for which the left side is bounded below is a compact subset of $\mathrm{PD}(d)$. The HDBT constraints finally imply that the associated set of g-tuples $(V_1^*, \ldots, V_g^*)$ is a compact subset of $\mathrm{PD}(d)^d$. This proves Claim (a).

(b) Modify D by $n-r+g$ replacements at a large distance from each other and from all original data points to obtain M. Each r-element subset of M contains at least g replacements. Moreover, there is a cluster C of size at least two that contains at least one replacement. Indeed, if no cluster contains two replacements then each cluster contains at least one and, by $r \ge gd+1$, one of them contains another element. Now, let C_ℓ be such a cluster, let $y \in C_\ell$ be a replacement, and let $x \in C_\ell,\ x \ne y$. We have

$$
\begin{aligned}
W_\ell(\boldsymbol{\ell}) &\ge \left\{\left(x - \frac{x+y}{2}\right)\left(x - \frac{x+y}{2}\right)^{\mathrm{T}} + \left(y - \frac{x+y}{2}\right)\left(y - \frac{x+y}{2}\right)^{\mathrm{T}}\right\} \\
&= \frac{1}{2}(y-x)(y-x)^{\mathrm{T}}.
\end{aligned}
$$

Now let $\left(\boldsymbol{\ell}^*, (m_j^*)_j, (V_j^*)_j\right)$ be optimal parameters of the TDC. Comparing them with the inferior parameters $\left(\boldsymbol{\ell}^*, (m_j^*), (2V_j^*)\right)$ and noting that the entropy terms coincide, we infer

$$
\begin{aligned}
0 &\le \sum_j n_j(\boldsymbol{\ell}^*)\left\{\ln\det 2V_j^* + \mathrm{tr}\left((2V_j^*)^{-1}S_j(\boldsymbol{\ell}^*)\right) - \left[\ln\det V_j^* + \mathrm{tr}\left(V_j^{*-1}S_j(\boldsymbol{\ell}^*)\right)\right]\right\} \\
&= \sum_j n_j(\boldsymbol{\ell}^*)\left\{d\ln 2 - \frac{1}{2}\mathrm{tr}\left(V_j^{*-1}S_j(\boldsymbol{\ell}^*)\right)\right\} \le dr\ln 2 - \frac{1}{2}\mathrm{tr}\left(V_\ell^{*-1}W_\ell(\boldsymbol{\ell}^*)\right) \\
&\le dr\ln 2 - \frac{1}{4}(y-x)^{\mathrm{T}}V_\ell^{*-1}(y-x).
\end{aligned}
$$

The resulting inequality $(y-x)^{\mathrm{T}} V_\ell^{*-1}(y-x) \le 4dr \ln 2$ implies that at least one eigenvalue of V_ℓ^* exceeds any positive number as the distance between x and y is chosen large enough.

Claim (c) follows from (a) and (b). □

It is interesting to remark that the TDC estimates of the covariance matrices withstand $g-1$ more outliers than there are discarded elements, $n-r$. Outliers that are spread out may be assigned to one-point clusters and outliers located close together may form a cluster of their own. In each case the optimal assignment does not completely destroy the estimates.

The asymptotic breakdown point of an estimator is its limit as $n \to \infty$.

Corollary 1 *If $r = \lfloor \alpha n \rfloor$ for some $\alpha > 1/2$ then the universal asymptotic breakdown point of the TDC estimates of the covariance matrices is $1-\alpha$.*

As noted after Lemma 1, the TDC estimates of the means are the sample means defined by the optimal assignment. Contrary to the covariance matrices their universal breakdown point is low. In order to show this, we need a lemma and denote (univariate) scatter values and sums of squares by the letters s and w, respectively.

Lemma 3 *Let $F \cup \{z_1, \ldots, z_{g-2}\} \cup \{y_1, y_2\} \subseteq \mathbb{R}$ be a data set of r pairwise distinct elements. If $w_{\{y_1,y_2\}} \le \frac{2c}{r-2} w_F$ then the constrained normal m.l.e.'s $v_j(\boldsymbol{\ell})$ of the variances v_j for the partition $\boldsymbol{\ell} = \{F, \{z_1\}, \ldots, \{z_{g-2}\}, \{y_1, y_2\}\}$ are*

$$v_1(\boldsymbol{\ell}) = \frac{w_F + w_{\{y_1,y_2\}}/c}{r} \quad \text{and} \quad v_j(\boldsymbol{\ell}) = c\, v_1(\boldsymbol{\ell}), \quad 2 \le j \le g.$$

Proof Putting $s_1 = s_F$ and $s_g = s_{\{y_1,y_2\}}$, the TDC requires minimizing the expression

$$h(v_1, \ldots, v_g) := n_1 \left(\ln v_1 + \frac{s_1}{v_1} \right) + \sum_{2 \le j \le g-1} \ln v_j + 2 \left(\ln v_g + \frac{s_g}{v_g} \right)$$

w.r.t. $(v_1, \ldots, v_g) \in \mathcal{V}_c$. We start with the minimum of h on the larger set $\mathcal{V}_c' = \{(v_1, \ldots, v_g) \in \mathbb{R}^g_> \mid c v_1 \le v_j \le v_1/c,\ 2 \le j \le g\} \supseteq \mathcal{V}_c$. Since $\min_{c v_1 \le v_j \le v_1/c} \ln v_j = \ln c v_1$, dynamic optimization shows

$$\begin{aligned}
&\min_{\mathbf{v} \in \mathcal{V}_c'} h(v_1, \ldots, v_g) \\
&\quad = \min_{c v_1 \le v_g \le v_1/c} \left\{ n_1 \left(\ln v_1 + \frac{s_1}{v_1} \right) + \sum_{2 \le j \le g-1} \min_{c v_1 \le v_j \le v_1/c} \ln v_j + 2 \left(\ln v_g + \frac{s_g}{v_g} \right) \right\} \\
&\quad = (g-2) \ln c + \min_{c v_1 \le v_g \le v_1/c} \left\{ \left((r-2) \ln v_1 + \frac{w_1}{v_1} \right) + \left(2 \ln v_g + \frac{w_g}{v_g} \right) \right\}.
\end{aligned}$$

This is a virtual two-cluster problem. The second line of the three cases in Proposition 1 shows that, under the assumption $\frac{w_g}{2} \le c \frac{w_1}{r-2}$ stated in the lemma, its solution is indeed given by the values claimed for $v_1(\boldsymbol{\ell})$ and $v_g(\boldsymbol{\ell})$. Finally, the vector

$(v_1(\boldsymbol{\ell}), cv_1(\boldsymbol{\ell}) \ldots, cv_1(\boldsymbol{\ell}))$ is even located in $\mathcal{V}_c$ so that it is the minimum w.r.t. the smaller parameter set, too. □

Our next theorem deals with the universal breakdown point of the TDC estimates of the means.

Theorem 2 *Let the data D be in general position and let $g \geq 2$.*

(a) *If $n \geq r + 1$ and $r \geq gd + 2$ then the TDC estimates of all means remain bounded by a constant that depends only on the data set D as* one *observation is arbitrarily replaced. (In the case of ties the solution is returned that has the largest discarded element.)*

(b) *Under the standard assumption $r \geq gd + 1$, there is a data set such that the TDC estimate of one sample mean breaks down as* two *particular observations are suitably replaced.*

(c) *Under the assumptions of (a) we have $\beta_{\text{mean}}(n, r, g) = \frac{2}{n}$.*

Proof (a) We show by contradiction that an optimal assignment $\boldsymbol{\ell}^*$ discards a remote replacement. Thus, assume that the replacement y lies in cluster ℓ. The cluster must contain a second (original) element x since, by the convention, y would otherwise be swapped with a discarded original element without change of the TDC. Now, by the assumption $r \geq gd + 2$, the retained data points contain at least $gd + 1$ original elements so that one cluster has at least $d + 1$ of them. Whether this is cluster ℓ or not, general position of D and this remark imply $\det W(\boldsymbol{\ell}^*) \to \infty$ as $\|y\| \to \infty$. We now use Lemma 2 which says that

$$2 \ln f\left[R^* \mid \boldsymbol{\ell}^*, \mathbf{m}^*, \mathbf{V}^*\right] \leq -r \ln \det(2\pi c V_\ell^*) - c \operatorname{tr}\left(W(\boldsymbol{\ell}^*) V_\ell^{*-1}\right).$$

It is well known that, given a positive-definite matrix W, the minimum of the function $V \mapsto \ln\det V + \operatorname{tr} W V^{-1}$ is $\ln\det W + d$. Hence, the right side of the inequality tends to $-\infty$ as $\|y\| \to \infty$ and so does the left side. On the other hand, by the assumption $r < n$, there exists an assignment $\boldsymbol{\ell}$ such that $y \notin R$. Optimality of $\boldsymbol{\ell}^*, \mathbf{m}^*, \mathbf{V}^*$ implies

$$\begin{aligned} -r\mathrm{H}\left(\frac{n_1(\boldsymbol{\ell}^*)}{r}, \ldots, \frac{n_g(\boldsymbol{\ell}^*)}{r}\right) + \ln f\left[R^* \mid \boldsymbol{\ell}^*, \mathbf{m}^*, \mathbf{V}^*\right] \\ \geq -r\mathrm{H}\left(\frac{n_1(\boldsymbol{\ell})}{r}, \ldots, \frac{n_g(\boldsymbol{\ell})}{r}\right) + \ln f[R \mid \boldsymbol{\ell}, \mathbf{0}, I_d]. \end{aligned}$$

Since the entropies are bounded, this means that $\ln f[R^* \mid \boldsymbol{\ell}^*, \mathbf{m}^*, \mathbf{V}^*]$ has a finite lower bound that does not depend on y, a contradiction to what was found before.

(b) A proof in the multivariate case requires a subtle construction of a data set. It must secure that the optimal solution retains at least one outlier. As a main hurdle one has to avoid point patterns that are almost degenerate and mask the desired solution just as in Fig. 1. A construction for the case $c = 1$ appears in Gallegos and Ritter (2005). For the sake of illustration, we treat here general c confining ourselves to the univariate case. Since Claim (b) is plainly true if $r \geq n - 1$, we assume $r \leq n - 2$ and proceed in three steps.

(α) Construction of the modified data set M:
Let x_i, $1 \le i \le r-g$, be strictly increasing and put $F = \{x_1, \ldots, x_{r-g}\}$, let $K > 0$, and choose $z_1 < z_2 < \cdots < z_{n-r+g-2}$ such that

(i) $z_1 - x_{r-g} \ge K$ and $z_{\ell+1} - z_\ell \ge K$ for all $1 \le \ell < n-r+g-2$.

Let $0 < \varepsilon \le \sqrt{c\frac{w_F}{r-2}}$, let $y > z_{n-r+g-2}+\varepsilon$, define the replacements $y_{1,2} = y \pm \varepsilon$, and put $M = \{x_1, \ldots, x_{r-g}, z_1, \ldots, z_{n-r+g-2}, y_1, y_2\}$. Plainly, M is in general position.
Let $\widetilde{\boldsymbol{\ell}}$ be the assignment associated with the clustering $\{F, \{z_1\}, \ldots, \{z_{g-2}\}, \{y_1, y_2\}\}$ ($z_{g-1}, \ldots, z_{n-r+g-2}$ discarded).

(β) The maximum a posteriori density for $\widetilde{\boldsymbol{\ell}}$ does not depend on K and y:
Since $w_{\{y_1,y_2\}} = 2\varepsilon^2 \le \frac{2c}{r-2} w_F$, Lemma 3 shows $v_1(\widetilde{\boldsymbol{\ell}}) = \frac{w_F + w_{\{y_1,y_2\}}/c}{r}$ and $v_2(\widetilde{\boldsymbol{\ell}}) = \cdots = v_g(\widetilde{\boldsymbol{\ell}}) = c v_1(\widetilde{\boldsymbol{\ell}})$. Twice the logarithm of the corresponding posterior density equals

$$2\left((r-g)\ln\left(\frac{r-g}{r}\right) + 2\ln\left(\frac{2}{r}\right)\right) - r\ln v_1(\widetilde{\boldsymbol{\ell}}) - g\ln c - r(1+\ln 2\pi).$$

(γ) If K is large enough then no assignment $\boldsymbol{\ell}$ of r points from the set $F \cup \{z_1, \ldots, z_{n-r+g-2}\}$ is optimal:
By $r \le n-2$, the set contains at least r elements. Since $|F| = r-g$ and since $r > g$, any such assignment $\boldsymbol{\ell}$ creates a cluster $C_\ell(\boldsymbol{\ell})$ which contains some z_k and some other point. From (i), it follows

$$w(\boldsymbol{\ell}) \ge w_{C_\ell(\boldsymbol{\ell})} \underset{K\to\infty}{\longrightarrow} \infty. \tag{10}$$

By Lemma 2, twice its log-likelihood is bounded above by

$$-r\ln(2\pi c\, v_j(\boldsymbol{\ell})) - c\frac{w(\boldsymbol{\ell})}{v_j(\boldsymbol{\ell})} \le -r\left(\ln 2\pi c^2/r + \ln w(\boldsymbol{\ell}) + 1\right) \underset{K\to\infty}{\longrightarrow} -\infty, \quad 1 \le j \le g;$$

here we have used the maximum of the left side as a function of the TDC estimate $v_j(\boldsymbol{\ell})$ and (10). The claim follows from (β) since there are only finitely many $\boldsymbol{\ell}$'s.

Finally, choose K as in (γ). The optimal solution retains at least one y_h causing at least one sample mean to break down as $y \to \infty$. This proves Part (b) in the special case and Part (c) follows from (a) and (b). □

As a consequence, the asymptotic universal breakdown value of the TDC estimates of the means is zero. More cannot be expected. The reason is that the universal breakdown point makes a statement on any data set for any g, even if these two do not fit together. On the other hand, García-Escudero and Gordaliza (1999), carried out experiments with trimmed k-means observing that the means of a clear cluster structure are hard to break down with the algorithm. We offer next an analysis of this phenomenon in the present situation.

4 Restricted breakdown point of the TDC estimates of the means

Dealing with the homoscedastic case, we computed in Gallegos and Ritter (2005) the restricted breakdown point of the TDC estimates of the means w.r.t. a class of data sets with a certain *separation property* thus defining what we mean by a "clear cluster structure." The separation property defined there is not satisfied for large data sets so that asymptotic robustness does not follow. Besides carrying over the theory to the heteroscedastic case we will also remove this weakness here.

The proof of the main result of this section, Theorem 3, depends on lemmas which we first state and prove. Let $\mathcal{P} = \{P_1, \ldots, P_g\}$ be a partition of D and let $\emptyset \neq T \subseteq D$. The partition $T \cap \mathcal{P} = \{T \cap P_1, \ldots, T \cap P_g\}$ is the *trace* of $\mathcal{P}$ in T. Let $g' \geq 1$ be a natural number and let $\mathcal{T} = (T_1, \ldots, T_{g'})$ be some partition of T. The *common refinement* of $\mathcal{T}$ and $\mathcal{P}$ is denoted by $\mathcal{T} \sqcap \mathcal{P} = \{T_k \cap P_j \mid k \leq g', j \leq g\}$, a partition of T (some clusters may be empty). The pooled SSP matrix of T w.r.t. some partition $\mathcal{T}$ is defined by

$$W_{\mathcal{T}} = \sum_{j \leq g'} W_{T_j}.$$

The following proposition states a basic condition which implies robustness of the TDC estimates of the means.

Proposition 2 *Let the data D be in general position, let $g \geq 2$ and $gd + 1 < r < n$, and let q be an integer such that $\max\{2r - n, gd + 1\} \leq q < r$. Assume that D possesses a partition $\mathcal{P}$ in g clusters such that, for all $T \subseteq D, q \leq |T| < r$, and all partitions $\mathcal{T}$ of T in $g - 1$ clusters (some clusters may be empty), the pooled SSP matrix satisfies*

$$\det W_{\mathcal{T}} \geq g^2 \max_{R \in \binom{D}{r}, R \supseteq T} \det\left(\frac{1}{c^2} W_{R \cap \mathcal{P}}\right). \tag{11}$$

Then the individual breakdown point of the TDC estimates of the means satisfies

$$\beta_{\text{mean}}(n, g, r, D) \geq \frac{1}{n}(r - q + 1).$$

Proof Let M be any admissible data set obtained from D by modifying at most $r - q$ elements and let $(R^*, \ell^*, (m_j^*)_{j=1}^g, (V_j^*)_{j=1}^g)$ be a TDC estimate for M. We will show that its sample means m_j^* are bounded by a number that depends solely on the original data D. Our proof proceeds in several steps.

(α) The matrices V_j^* are bounded above and below by positive-definite matrices that depend only on D, not on the replacements:

Let R_j^* be the jth cluster generated by ℓ^*. Since $|R^*| = r$, $R^* = \bigcup_{j=1}^g R_j^*$ has at least $q \geq gd + 1$ original observations so that some R_j^* contains at least $d + 1$ original observations. The proof now finishes as that of Theorem 1 (a).

(β) If R_j^* contains some original observation, then m_j^* is bounded by a number that depends only on D:

By (α), $\operatorname{tr} W(\boldsymbol{\ell}^*)$ remains bounded above by a constant which depends solely on the original data D. Now, let $x \in R_j^* \cap D$. We have $W(\boldsymbol{\ell}^*) \succeq (x - m_j^*)(x - m_j^*)^{\mathrm{T}}$ and, hence, $\|x - m_j^*\|^2 \leq \operatorname{tr} W(\boldsymbol{\ell}^*)$ and the claim follows.

(γ) If R_j^* contains some replacement then $\|m_j^*\| \to \infty$ as the replacement tends to ∞:

This is proved like (β) where x is now the replacement.

From (β) and (γ) it follows: as the replacements tend to ∞ then, in the long run, each R_j^*, $1 \leq j \leq g$, consists solely of original observations or solely of modifications. We next put $c_{d,r} = -\frac{dr}{2}(1 + \ln 2\pi)$ and state:

(δ) $-r\mathrm{H}\left(\frac{n_1^*}{r}, \ldots, \frac{n_g^*}{r}\right) + \ln f\left[R^* \mid \boldsymbol{\ell}^*, \mathbf{m}^*, \mathbf{V}^*\right] < c_{d,r} - dr \ln c - \frac{r}{2} \ln \det \frac{W(\boldsymbol{\ell}^*)}{r}$,

whenever $0 < n_j^* < r$ for some j:

On account of Lemma 2 and of the assumption, the left side is strictly bounded above by

$$-dr \ln c - \frac{dr}{2} \ln 2\pi - \frac{1}{2}\left[r \ln \det(V_1^*/c) + \operatorname{tr}\left(W(\boldsymbol{\ell}^*)(V_1^*/c)^{-1}\right)\right].$$

Part (α) and normal estimation theory now show that the function $A \mapsto r \ln \det(A/c) + \operatorname{tr}\left(W(\boldsymbol{\ell}^*)(A/c)^{-1}\right)$, $A \succeq 0$, attains its minimum value $r\left[\ln \det\left(\frac{W(\boldsymbol{\ell}^*)}{r}\right) + d\right]$ at $\frac{cW(\boldsymbol{\ell}^*)}{r}$ and the claim follows.

(ϵ) R^* contains no modification with a sufficiently large norm:

Assume on the contrary that R^* contains a large replacement. In view of the remark right after (γ), some cluster, say R_g^*, consists solely of replacements. Note that $r > |R^* \cap D| \geq q$. Let $T = R^* \cap D$ and let $\mathcal{T} = \{R_1^* \cap D, \ldots, R_{g-1}^* \cap D\}$. From Steiner's formula we have the relation $W(\boldsymbol{\ell}^*) \succeq W_{\mathcal{T}}$ between the pooled SSP matrices and hypothesis (11) implies

$$\det W(\boldsymbol{\ell}^*) \geq \det W_{\mathcal{T}} \geq g^2 \max_{R \in \binom{D}{r}, R \supseteq T} \det\left(\frac{1}{c^2} W_{R \cap \mathcal{P}}\right).$$

Hence,

$$2d \ln c + \ln \det \frac{W(\boldsymbol{\ell}^*)}{r} \geq 2 \ln g + \max_{R \in \binom{D}{r}, R \supseteq T} \ln \det \frac{1}{r} W_{R \cap \mathcal{P}}. \tag{12}$$

Now, denoting the assignment associated with $R \cap \mathcal{P}$ by $\boldsymbol{\ell}_{R\cap\mathcal{P}}$, and writing $\mathbf{m}_{R\cap\mathcal{P}} = (m_{R\cap P_1}, \ldots, m_{R\cap P_g})$ and $S_{R\cap\mathcal{P}} = \frac{1}{r} W_{R\cap\mathcal{P}}$, the pooled scatter matrix, we have

$$
\begin{aligned}
& r \ln g + \min_{R\in\binom{M\cap D}{r}} - \ln f\left[R \mid \boldsymbol{\ell}_{R\cap\mathcal{P}}, \mathbf{m}_{R\cap\mathcal{P}}, S_{R\cap\mathcal{P}}\right] \\
&= -c_{d,r} + r \ln g + \frac{r}{2} \min_{R\in\binom{M\cap D}{r}} \ln\det S_{R\cap\mathcal{P}} \\
&\le -c_{d,r} + r \ln g + \frac{r}{2} \min_{T\subseteq R\in\binom{M\cap D}{r}} \ln\det S_{R\cap\mathcal{P}} \\
&\le -c_{d,r} + r \ln g + \frac{r}{2} \max_{T\subseteq R\in\binom{D}{r}} \ln\det S_{R\cap\mathcal{P}} \\
&\le -c_{d,r} + dr \ln c + \frac{r}{2} \ln\det V(\boldsymbol{\ell}^*) \\
&< r\mathrm{H}\left(\frac{n_1^*}{r}, \ldots, \frac{n_g^*}{r}\right) - \ln f\left[R^* \mid \boldsymbol{\ell}^*, \mathbf{m}^*, \mathbf{V}^*\right], \qquad (13)
\end{aligned}
$$

where the last two inequalities follow from (12) and (δ), respectively. Note that Part (δ) is applicable since $R^* \cap D \neq \emptyset$ implies $n_j^* > 0$ for some $j < g$ and since $n_g^* > 0$ as well. The last expression above is the minimum of the TDC. It is no larger than its value at the clustering $R \cap \mathcal{P}$ with the parameters $\mathbf{m}_{R\cap\mathcal{P}}$ and $S_{R\cap\mathcal{P}}$ for all $R \in \binom{M\cap D}{r}$. By an elementary property of the entropy, the latter value is no larger than the first line of (13). This contradiction proves Claim (ϵ).

Finally, Part (β) shows that all sample means m_j^* remain bounded by a number that depends only on D. This proves the proposition. □

In the remainder of this section, we show that the hypothesis of Proposition 2 actually states a separation property. We need more notation. Let $g \ge 2$. Given an integer $u \ge 1$ and a real number ϱ, $0 < \varrho < 1$, we define the number

$$
q_{u,\varrho} = \max\left\{2r - n, (g-1)gd + 1, \frac{n-u}{1-\varrho}\right\}.
$$

If $n > r > (g-1)gd + 1$ and $u \ge n - (1-\varrho)(r-1)$ then $q = \lceil q_{u,\varrho} \rceil$ satisfies the assumption in Proposition 2.

Let $\mathcal{P}$, T, and $\mathcal{T}$ be as in Proposition 2. Our next, combinatorial, lemma gives conditions that secure the existence of sufficiently many elements of T in each class P_j and a large intersection $T_k \cap P_j$ for some pair (k, j).

Lemma 4 *Let $\mathcal{P} = \{P_1, \ldots, P_g\}$ be a partition of D in $g \ge 2$ clusters of size $\ge u$, let $T \subseteq D$ such that $q_{u,\varrho} \le |T| < r$, and let $\mathcal{T} = \{T_1, \ldots, T_{g-1}\}$ be a partition of T (some T_k's may be empty). Then*:

(a) *For all j, we have $|T \cap P_j| \geq \varrho|T|$.*
(b) *At least one T_k contains elements of two different P_j's.*
(c) *There are clusters T_k and P_j such that $|T_k \cap P_j| \geq \frac{q_{u,\varrho}}{(g-1)g}$ $(> d)$.*

Proof (a) Assume on the contrary that $|T \cap P_j| < \varrho|T|$ for some j. From $D \supseteq T \cup P_j$ we infer

$$\begin{aligned} n &\geq |T| + |P_j| - |T \cap P_j| > |T| + u - \varrho|T| = u + (1-\varrho)|T| \\ &\geq u + (1-\varrho)q_{u,\varrho} \geq u + n - u \end{aligned}$$

by definition of $q_{u,\varrho}$, a contradiction.

(b) Since $\varrho|T| > 0$ and since there are more P_j's than T_k's, this follows from the pigeon hole principle with (a).

(c) The observations in T are spread over the $(g-1)g$ disjoint sets of the form $T_k \cap P_j$. If (c) did not hold, we would have $|T| < q_{u,\varrho}$, contradicting one of the assumptions. □

The theorem on the breakdown point of the TDC estimates of the means presented in this section applies to a class of clustered data sets with a certain separation property which we now present. We put

$$\kappa_\varrho = \begin{cases} (1-\varrho)\varrho, & g = 2, \\ \varrho/2, & g \geq 3. \end{cases}$$

Definition (Separation property) Let $u \in \mathbb{N}$ such that $1 \leq u \leq n/g$ and let $0 < \varrho < 1$. We denote by $\mathcal{L}_{u,\varrho,c}$ the system of all d-dimensional admissible data sets D of size n which have the following *separation property*:

D possesses a partition $\mathcal{P}$ in g subsets of size at least u such that, for all subsets $T \subseteq D$, $q_{u,\varrho} \leq |T| < r$ and for all partitions $\mathcal{T} = \{T_1, \ldots, T_{g-1}\}$ of T in $g-1$ clusters, we have

$$\begin{aligned} &1 + \kappa_\varrho \cdot \min_{\substack{k,\, j \neq \ell\,: \\ T_k \cap P_h \neq \emptyset,\ h = j,\, \ell}} \left(\overline{x}_{T_k \cap P_j} - \overline{x}_{T_k \cap P_\ell}\right)^{\mathrm{T}} \left(\frac{W_{\mathcal{T} \sqcap \mathcal{P}}}{|T|}\right)^{-1} \left(\overline{x}_{T_k \cap P_j} - \overline{x}_{T_k \cap P_\ell}\right) \\ &\geq g^2 \frac{\max_{R \in \binom{D}{r},\, R \supseteq T} \det \frac{1}{c^2} W_{R \cap \mathcal{P}}}{\det W_{\mathcal{T} \sqcap \mathcal{P}}}. \end{aligned} \tag{14}$$

According to Lemma 4 (b), the minimum extends over at least one triple (k, j, ℓ), $j \neq \ell$, and by Lemma 4 (c), the pooled scatter matrix $S_{\mathcal{T} \sqcap \mathcal{P}}$ is bounded below by a positive-definite matrix which depends only on D. Condition (14) is affine equivariant. We require the minimum of the Mahalanobis distances of the submeans $\overline{x}_{T_k \cap P_j}$ and $\overline{x}_{T_k \cap P_\ell}$ of P_j and P_ℓ appearing on its left-hand side to be large. Thus, condition (14) means that the partition $\mathcal{P}$ subdivides the data set in well-separated clusters, it is the "natural" partition of D. The set $\mathcal{L}_{u,\varrho,c}$ increases with decreasing u and with increasing $\varrho \leq 1/2$.

We show next that any data set D in $\mathcal{L}_{u,\varrho,c}$ satisfies the hypotheses of Proposition 2.

Lemma 5 *Let $g \geq 2$, let $n > r > (g-1)gd+1$, let $u \in \mathbb{N}$ and $0 < \varrho < 1$ satisfy $n-(1-\varrho)(r-1) \leq u \leq n/g$. Let $D \in \mathcal{L}_{u,\varrho,c}$, let $T \subseteq D$ be such that $q_{u,\varrho} \leq |T| < r$, and let $\mathcal{T} = \{T_1, \ldots, T_{g-1}\}$ be a partition of T (some T_k's may be empty). We have*

$$\det W_{\mathcal{T}} \geq g^2 \max_{R \in \binom{D}{r}, R \supseteq T} \det \frac{1}{c^2} W_{R \cap \mathcal{P}}.$$

Proof An application of Gallegos and Ritter (2005), Lemma A.3, to each T_k, $1 \leq k < g$, with partition $\{T_k \cap P_1, \ldots, T_k \cap P_g\}$, $1 \leq j \leq g$, shows first

$$\begin{aligned} W_{\mathcal{T}} &= \sum_{k=1}^{g-1} W_{T_k} \\ &= \sum_{k: T_k \neq \emptyset} \left\{ \sum_{j=1}^{g} W_{T_k \cap P_j} + \sum_{1 \leq j < \ell \leq g} \frac{a_{kj} a_{k\ell}}{|T_k|} \left(\overline{x}_{T_k \cap P_j} - \overline{x}_{T_k \cap P_\ell}\right) \left(\overline{x}_{T_k \cap P_j} - \overline{x}_{T_k \cap P_\ell}\right)^{\mathrm{T}} \right\}, \end{aligned}$$

where $a_{kj} = |T_k \cap P_j|$, $1 \leq j \leq g$, $1 \leq k < g$. Now use Gallegos and Ritter (2005), Lemma A.1(b), and Lemma A1 to obtain

$$\begin{aligned} &\det W_{\mathcal{T}} \\ &\geq \det W_{\mathcal{T} \sqcap \mathcal{P}} \cdot \left\{ 1 + \sum_{k: T_k \neq \emptyset} \sum_{1 \leq j < \ell \leq g} \frac{a_{kj} a_{k\ell}}{|T_k|} \left(\overline{x}_{T_k \cap P_j} - \overline{x}_{T_k \cap P_\ell}\right)^{\mathrm{T}} W_{\mathcal{T} \sqcap \mathcal{P}}^{-1} \left(\overline{x}_{T_k \cap P_j} - \overline{x}_{T_k \cap P_\ell}\right) \right\} \\ &\geq \det W_{\mathcal{T} \sqcap \mathcal{P}} \cdot \left\{ 1 + \kappa_\varrho \min_{\substack{k,\, j \neq \ell : \\ T_k \cap P_h \neq \emptyset}} \left(\overline{x}_{T_k \cap P_j} - \overline{x}_{T_k \cap P_\ell}\right)^{\mathrm{T}} \left(\frac{W_{\mathcal{T} \sqcap \mathcal{P}}}{|T|} \right)^{-1} \left(\overline{x}_{T_k \cap P_j} - \overline{x}_{T_k \cap P_\ell}\right) \right\} \end{aligned}$$

and the claim follows from the separation property. □

The conditions on r and u imply that the interval $[q_{u,\varrho}, r[$ contains some integer so that a set T as in Lemma 5 exists. A simple reasoning shows that the bounds on u imply $\varrho < \frac{1}{g}$.

We finally state and prove the main result of this section: the restricted breakdown point of the TDC estimates of the means. If a data set has the separation property then the TDC estimates of the means are much more robust than predicted by Theorem 2.

Theorem 3 *Let the data D be in general position, let $g \geq 2$, and let $r < n$.*

(a) *Assume $r \geq (g-1)gd+2$ and $n-(1-\varrho)(r-1) \leq u \leq n/g$. Then the restricted breakdown value of the TDC estimates of the means w.r.t. $\mathcal{L}_{u,\varrho,c}$ satisfies*

$$\beta_{\mathrm{mean}}(n, g, r, \mathcal{L}_{u,\varrho,c}) \geq \frac{1}{n} \min \left\{ n - r + 1, r - (g-1)gd, r + 1 - \frac{n-u}{1-\varrho} \right\}.$$

(b) *For any data set* $D \in \mathcal{L}_{u,\varrho,c}$, *the individual breakdown point of the TDC estimates of the means satisfies*

$$\beta_{\text{mean}}(n, g, r, D) \leq \frac{1}{n}(n - r + 1).$$

(c) *Let* $2r - n \geq (g-1)gd + 1$, *let* $u \in \mathbb{N}$ *such that* $2(n-r) < u \leq n/g$, *and put* $\varrho = \frac{u-2(n-r)}{2r-n}$. *Then*

$$\beta_{\text{mean}}(n, g, r, \mathcal{L}_{u,\varrho,c}) = \frac{1}{n}(n - r + 1).$$

(*A necessary condition for the existence of such a* u *is the inequality* $2(n-r) \leq n/g - 1$.)

(d) *Under the assumptions of* (*a*), *the TDC discards all sufficiently large replacements in a data set that satisfies the separation property* (*with any parameters*).

Proof Part (a) is a direct consequence of Proposition 2 and Lemma 5.

(b) Let M be a data set obtained from D by replacing $n - r + 1$ of its elements with a narrow and distant cluster. The modified data set contains only $r - 1$ original observations so that the optimal set R^* contains some modification. Then so does $C_j^* = C_j(\ell^*)$ for some j. Lemma A2 shows that the norm of m_j^* tends to infinity together with the narrow cluster of replacements.

(c) The hypotheses imply $\min\left\{n - r + 1, r - (g-1)gd, r + 1 - \frac{n-u}{1-\varrho}\right\} = n - r + 1$. (Note that ϱ is maximum so that the first term does not exceed the last for a given u.) Furthermore, the first condition in (a) follows from the first condition, whereas the second condition in (a) follows from the choice of ϱ and from second condition. Finally, the condition $2(n-r) < u$ implies $\varrho > 0$. The claim now follows from Parts (a) and (b).

Claim (d) follows from Part (ϵ) of the proof of Proposition 2. □

The inequality $n - (1-\varrho)(r-1) \leq u$ implies $u \geq n - r + 2$. I.e., the sizes of the natural clusters must exceed the number of discarded elements in Part (a) of Theorem 3. Moreover, the assumptions of Part (c) imply that these sizes exceed twice the number of discarded elements.

The following corollary of Theorem 3 says that the TDC estimates of the means are asymptotically robust on well-separated, balanced data sets if the parameter g is set to its natural number of clusters.

Corollary 2 *Let* $g \geq 2$, *let* $0 < \eta < \delta < 1/g$, *let* $r = \left\lceil n(1 - \frac{1}{2g} + \frac{\delta}{2}) \right\rceil$, *let* $u = \left\lceil n\left(\frac{1}{g} - \eta\right) \right\rceil$, *and let* $\varrho = \frac{\delta - \eta}{1 - \frac{1}{g} + \delta}$. *Then, asymptotically,*

$$\beta_{\text{mean}}(n, g, r, \mathcal{L}_{u,\varrho,c}) \longrightarrow \frac{1}{2}\left(\frac{1}{g} - \delta\right), \quad \text{as } n \to \infty.$$

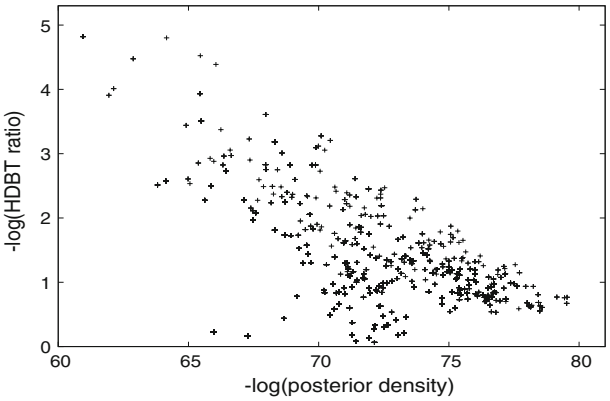

Fig. 3 Synthetic data set of Fig. 1: negative double-logarithmic HDBT-ratio-posterior-density plot for a large number of minimum distance partitions with two clusters and no discarded elements

5 Two studies

We illustrate the method described in Sects. 2.3–2.5 with two examples and first recall the simple data set of Fig. 1. As already seen there may exist minimum distance partitions the maximum posterior densities (3) of which exceed that of the desired partition. Figure 3 shows the negative double-logarithmic HDBT-ratio-posterior-density plot of the MDP's found for the heteroscedastic full normal model with two clusters, no discarded elements, and unknown cluster sizes for the synthetic data set of Fig. 1. According to the method of Sect. 2.4, the most plausible MDP is the one in the left lower region close to (66, 0.2). It belongs indeed to the desired partition of the data set in two clusters of ten elements, each. The solution close to (61, 4.8) in Fig. 3 with the largest posterior density represents the uppermost horizontal cluster in Fig. 1.

Our second example is the Tiles data set, see Mucha et al. (2002), from archeometry. It can be found under the URL www.uni-passau.de/ritter. Its objects consist presently of 660 antique roman tiles collected in the Rhine valley between Strasbourg/France and Frankfurt/Germany. Our questions are: which tiles originate from the same clay pits and how many clay pits are represented? Feature data from X-ray Fluorescence Analysis about the contents of nineteen minerals and metals are available to this end, viz., flint SiO_2, Titanium dioxide (titania) TiO_2, Aluminium oxide (aloxite) Al_2O_3, Ferric oxide (rust) Fe_2O_3, Manganese oxide MnO, Magnesium oxide (magnesia) MgO, burnt lime CaO, Sodium oxide Na_2O, Potassium oxide K_2O, vanadium V, chromium Cr, nickel Ni, zinc Zn, rubidium Rb, strontium Sr, yttrium Y, zirconium Zr, niobium Nb, and barium Ba.

Although we expect cluster sizes of a hundred or less which are not sufficient for safely estimating more than a hundred real parameters for each cluster, we used the heteroscedastic full normal model with unknown cluster sizes (maximum a posteriori) and unknown number of clusters. A look at the 2D scatter plots suggests marked correlation between some of the features: SiO_2 with MnO, CaO, Sr, and Zr, TiO_2

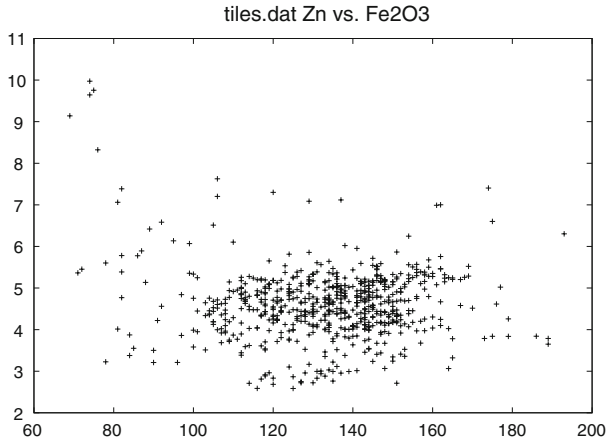

Fig. 4 Tiles data: scatter plot of the features Zn and Fe_2O_3 displaying outliers

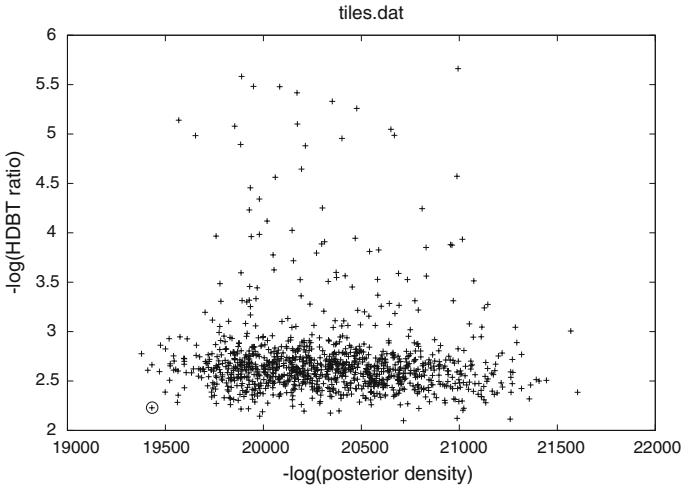

Fig. 5 Tiles data: negative double-logarithmic HDBT-ratio-posterior-density plot for the minimum distance clusterings of 1,100 replications for the heteroscedastic, full normal model with six clusters and 66 discarded points. The encircled solution in the left lower part is most promising

with Cr and Nb, CaO with Sr, and K_2O with Rb. This fact allows us to reduce the dimension of the sample space by deleting SiO_2, TiO_2, CaO, and K_2O from the feature list so that $d = 19 - 4 = 15$. Like almost any real data set, the present one contains outliers, see Fig. 4, and we apply the algorithm proposed in Sects. 2.3, 2.4, and 2.5 with ten percent of discarded elements. The minimum cluster size was set to $d + 1 = 16$.

Figure 5 shows the negative double-logarithmic HDBT-ratio-posterior-density plot of the MDP's of 1,100 replications for six clusters. The favorite solution at the left end of the almost horizontal support line is encircled. A 2D representation of this

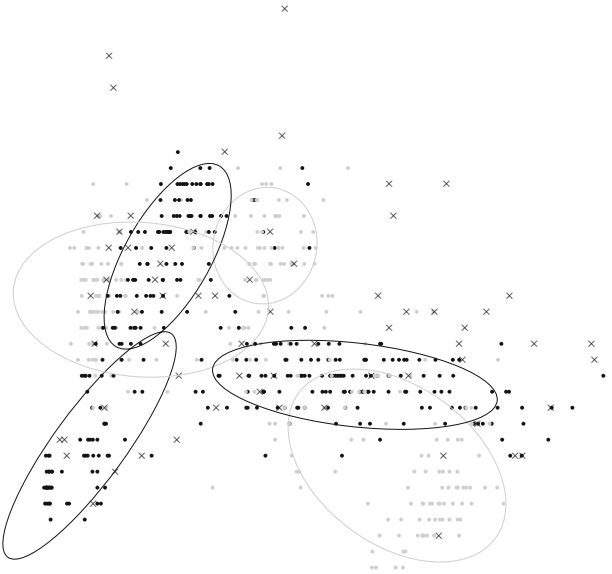

Fig. 6 Tiles data: MnO–Y plot of the favorite MDP. The ellipses indicate the 0.8-quantiles of the clusters and *crosses* stand for discarded points

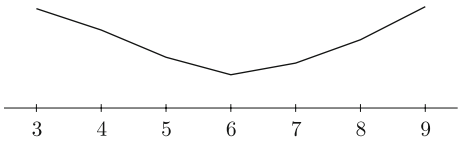

Fig. 7 Tiles data: the BIC curve for the favorite solutions with three to nine clusters suggested by the posterior-density-HDBT ratio plots

clustering is shown in Fig. 6. Its cluster sizes are 145, 111, 111, 105, 61, and 61, its HDBT ratio is 1/158. One or a few small clusters that cannot be detected by the full normal model may be hidden in the set of discarded elements (crosses). Figure 6 shows that the assumed number of outliers is too small. The oblong shape of the left lower ellipse points to two distant elements in the upper part of the figure which are assigned to this cluster but do not fit in it.

The BIC curve for the favorite solutions obtained with three to nine clusters is presented in Fig. 7. It clearly pleads for six clusters. It turns out that increasing the number of clusters by one essentially splits one group in the preceding solution.

Acknowledgments We thank H.-H. Bock for a number of hints that improved the presentation. We also thank the referees for their constructive suggestions.

Appendix

Lemma A1 *Let* $g \geq 2$, *let* $0 < \varrho \leq 1/g$, *let* $\mathbf{a} = (a_{kj})_{\substack{1 \leq k < g \\ 1 \leq j \leq g}} \in \mathbb{N}^{(g-1)\times g}$ *be such that* $\|\mathbf{a}\|_1 = \sum_{k,j} a_{kj} > 0$, *let* $\sum_k a_{kj} \geq \varrho \|\mathbf{a}\|_1$ *for all* $1 \leq j \leq g$, *and put*

$a_{k\cdot} = \sum_j a_{kj}$. *Then*

$$\sum_{k:a_{k\cdot}>0} \frac{1}{a_{k\cdot}} \sum_{1\le j<\ell\le g} a_{kj}a_{k\ell} \ge \kappa_\varrho \|\mathbf{a}\|_1. \tag{15}$$

Proof Write the left hand side of (15) as

$$\|\mathbf{a}\|_1 \sum_{k:a_{k\cdot}>0} \frac{a_{k\cdot}}{\|\mathbf{a}\|_1} \sum_{1\le j<\ell\le g} \frac{a_{kj}}{a_{k\cdot}} \frac{a_{k\ell}}{a_{k\cdot}} = \|\mathbf{a}\|_1 \sum_{k:a_{k\cdot}>0} \beta_k \sum_{1\le j<\ell\le g} A_{k,j}A_{k,\ell}.$$

Since $\beta = (a_{k\cdot}/\|\mathbf{a}\|_1)_{k:a_{k\cdot}>0}$ is a probability vector and since $A = (a_{k,j}/a_{k\cdot})_{k:a_{k\cdot}>0,\, j}$ is a stochastic matrix s.th. $\beta A \ge \varrho$ elementwise, the claim follows from an elementary reasoning. □

Lemma A2 *Let $h \ge 0$ and let $k \ge 1$. Let $C = \{x_1, \ldots, x_h, y_1, \ldots, y_k\}$ consist of h original data points and k replacements. Then the norm of the sample mean of C tends to infinity as $\|y_1\| \to \infty$ and as $y_i - y_1$, $2 \le i \le k$, remain bounded.*

Proof The sum of C is $\sum_{i=1}^h x_i + k y_1 + \sum_{i=2}^k (y_i - y_1)$ from which the lemma follows. □

References

Barnett V, Lewis T (1994) Outliers in statistical data. Wiley, Chichester

Becker C, Gather U (1999) The masking breakdown point of multivariate outlier identification rules. JASA 94:947–955

Bezdek JC, Keller J, Krisnapuram R, Pal NR (1999) Fuzzy models and algorithms for pattern recognition and image processing. The handbooks of fuzzy sets series. Kluwer, Boston

Bock H-H (1985) On some significance tests in cluster analysis. J Class 2:77–108

Chen H, Chen J, Kalbfleisch JD (2004) Testing for a finite mixture model with two components. J R Stat Soc Ser B 66:95–115

Cuesta-Albertos JA, Gordaliza A, Matrán C (1997) Trimmed k-means: an attempt to robustify quantizers. Ann Stat 25:553–576

Dennis JE Jr (1981) Algorithms for nonlinear fitting. In: Powell MJD (ed) Nonlinear optimization 1981. Procedings of the NATO Advanced Research Institute held at Cambridge in July 1981, Academic Press, London

Donoho DL, Huber PJ (1983) The notion of a breakdown point. In: Bickel PJ, Doksum KA, Hodges JL Jr (eds) A Festschrift for Erich L. Lehmann, The Wadsworth Statistics/Probability Series. Wadsworth, Belmont, pp 157–184

Gallegos MT, Ritter G (2005) A robust method for cluster analysis. Ann Stat 33:347–380

Gallegos MT, Ritter G (2009) Using combinatorial optimization in model-based clustering under spurious outliers and cardinality constraints. Comput Statist Data Anal (to appear)

García-Escudero LA, Gordaliza A (1999) Robustness properties of k-means and trimmed k-means. J Am Stat Assoc 94:956–969

García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2008) A general trimming approach to robust cluster analysis. Ann Stat 36:1324–1345

Gordon AD (1999) Classification. Monographs on statistics and applied probability, vol 82, 2nd edn. CRC Press, New York

Hathaway RJ (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Ann Stat 13:795–800

Hodges JL Jr (1967) Efficiency in normal samples and tolerance of extreme values for some estimates of location. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, pp 163–186
Kéribin C (2000) Consistent estimation of the order of mixture models. Sankhyā 62(Series A):49–66
McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
Mecklin CJ, Mundfrom DJ (2004) An appraisal and bibliography of tests for multivariate normality. Int Stat Rev 72(1):123–138
Milligan GW, Cooper MC (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika 50:159–179
Mucha H-J, Bartel HG, Dolata J (2002) Exploring Roman brick and tile by cluster analysis with validation of results. In: Gaul W, Ritter G (eds) Classification, automation, and new media. Studies in classification, data analysis, and knowledge organization. Springer, Berlin, pp 471–478
Neykov N, Filzmoser P, Dimova R, Neytchev P (2007) Robust fitting of mixtures using the trimmed likelihood estimator. Comput Stat Data Anal 52:299–308
Pollard D (1981) Strong consistency of k-means clustering. Ann Stat 9:135–140
Ritter G, Gallegos MT (1997) Outliers in statistical pattern recognition and an application to automatic chromosome classification. Patt Rec Lett 18:525–539
Rocke DM, Woodruff DL (1999) A synthesis of outlier detection and cluster identification. Technical report, University of California, Davis. http://handel.cipic.ucdavis.edu/~dmrocke/Synth5.pdf
Schroeder A (1976) Analyse d'un mélange de distributions de probabilités de même type. Revue de Statistique Appliquée 24:39–62
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Symons MJ (1981) Clustering criteria and multivariate normal mixtures. Biometrics 37:35–43
Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. J R Stat Soc Ser B 63:411–423
Wolfe JH (1970) Pattern clustering by multivariate mixture analysis. Multivar Behav Res 5:329–350