Trimming algorithms for clustering contaminated grouped data and their robustness

María Teresa Gallegos[b] and Gunter Ritter[a,b]

a. Department of Informatics and Mathematics, University of Passau, Germany
b. VarSys, Passau, Germany

ritter@fim.uni-passau.de

Technical Report, Number MIP-0905
Department of Informatics and Mathematics
University of Passau, Germany
March 2009

Trimming algorithms for clustering contaminated grouped data and their robustness

María Teresa Gallegos[b] and Gunter Ritter[a,b]

a. Department of Informatics and Mathematics, University of Passau, Germany
b. VarSys, Passau, Germany

March 2, 2009

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

AMS subject classifications Primary 62H30, secondary 62F35

Key words and phrases statistical clustering; trimming algorithm; HDBT-ratio; outlier robustness; heteroscedasticity.

1 Introduction

1.1 Background

Statistical clustering methods start from a statistical model of the data deriving from it, in general by the ML– or MAP–paradigm, a cluster criterion to be optimized. Various problems, expected and unexpected, 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 but others containing spurious, undesirable clusters. Third, optimization of the criteria is not easy. Fourth, the 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 [13], following a proposal by Dennis [5], Evelyn Martin Lansdowne Beale, and James R. Thompson (oral communications), investigated constraints on the scale parameters V_j of a normal mixture model showing that they mitigated or even avoided some of the problems. We call them the *HDBT constraints.* In a slightly different form, Hathaway's multivariate version reads

$$V_j \succeq cV_\ell \tag{1}$$

for some constant $c > 0$ and all indices j, ℓ. The symbol $\succeq$ stands for the semi-definite ordering on the space of symmetric matrices and the constant c is necessarily bounded above by 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 $\mathbf{V} = (V_j)$ of positive-definite matrices as the maximum c for which the Constraints (1) hold. It can be computed as

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

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 partition is that of its scatter matrices. Hathaway showed in the univariate context that, besides guaranteeing the ML-estimate and its consistency, the HDBT constraints removed many undesirable local solutions.

Although there is no no formal, mathematical definition of "cluster" and "outlier," both play an important role in reality. Clustering methods deemed to be robust actually break down under the influence of a single gross outlier, García–Escudero and Gordaliza [9]. Nevertheless there are nowadays some robust trimming methods based on classification models. Cuesta–Albertos et al. [4] and García–Escudero and Gordaliza [9] 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. We [7] 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 ML estimates, too, are robust if matters are restricted to 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 cost since the very existence of an ML or MAP 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 was Rocke and Woodruff's [21] MINO. They used cardinality constraints in order to enforce the existence of ML–estimates. These constraints prevent the scatter matrices from approaching singularity if the data are in general position. In [8] we extended their method to MAP 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. [10] present a constrained heteroscedastic trimming algorithm relaxing the requirements on sphericity in [9] and of equality of shapes in [7]. They also prove convergence. However, their constraints lack affine equivariance. Here, we propose and analyse a robust, affine equivariant, heteroscedastic, full normal classification model. Specializations to normal submodels are immediate and left to the interested reader.

1.2 Outline

In Sect. 2, we start from a statistical model with "spurious" outliers deriving from it a trimmed cluster criterion. Its maximum exists if some constraint is applied. Besides cardinality constraints [21, 8], it is also possible to restrict the scale parameters of a normal model by the HDBT Constraints (1), Lemma 2.1. This approach leads to an affine equivariant *Trimmed*

Determinant Criterion, TDC. 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, ML-estimation, MAP-classification, and trimming.

Unfortunately, the optimal partitions turn out to be undesirable in many cases of real and synthetic data sets, see Figs. 1, 3, and 5. Although they provide optimal fit of estimated populations and clusters they may be unbalanced in the sense that their 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 should be estimated. 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 vs. the HDBT ratio for finding reasonable partitions together with a constant c.

The aim of a trimming algorithm is robustness. We show here that, as an additional benefit, the HDBT-constraints render the TDC *robust.* Mutatis mutandis, the properties of the homoscedastic case [7] 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 analysis of the robustness of our algorithm 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.* 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. The larger the constraint c the more robust the method turns out to be. In the final Sect. 5 we report on our experience with two data sets.

1.3 Notation

The symbol E denotes some sample space, our basic data set is $D = (x_1, \dots, x_n) \subseteq E$, and $R \subseteq D$ denotes the generic r–element subset of D, $r \leq n$. The notation $\mathbf{l} : 1..n \to 0..g$ denotes the generic assignment of the n data points to g classes $1..g$ or to 0. The assignment of data point x_i is ℓ_i. If $\ell_i = 0$ then observation i is not assigned to a class, it is *discarded.* An assignment $\mathbf{l}$ is *admissible* if $n - r$ indices are discarded. The set of all admissible assignments is denoted by Λ_r. The assignment $\mathbf{l}$ defines g clusters $C_j = C_j(\mathbf{l}) = \{i \mid \ell_i = j\}$ of cardinalities $n_j = n_j(\mathbf{l})$, $j \in 1..g$. We allow one or more clusters to be empty.

We consider a distributional model on E consisting of g components. Its parameters are denoted by $\gamma = (\gamma_1, \dots, \gamma_g)$. 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 also gather $\mathbf{m} = (m_1, \dots, m_g)$ and $\mathbf{V} = (V_1, \dots, V_g)$. We will often need the positive–semi-definite (Löwner) ordering $\preceq$ on $\mathrm{PD}(d)$.

Estimates of the parameters γ_j, m_j, and V_j w.r.t. the clusters of an assignment $\mathbf{l}$ are denoted by $\gamma_j(\mathbf{l})$, $m_j(\mathbf{l})$, and $V_j(\mathbf{l})$, respectively. We also abbreviate $\gamma(\mathbf{l}) = (\gamma_1(\mathbf{l}), \dots, \gamma_g(\mathbf{l}))$, $\mathbf{m}(\mathbf{l}) = (m_1(\mathbf{l}), \dots, m_g(\mathbf{l}))$, $\mathbf{V}(\mathbf{l}) = (V_1(\mathbf{l}), \dots, V_g(\mathbf{l}))$. A bar as in $\overline{x}$ denotes sample means and W and S denote SSP–matrices and scatter matrices, respectively, with the additional notations $W_j(\mathbf{l}) = W(C_j(\mathbf{l}))$ and $S_j(\mathbf{l}) = S(C_j(\mathbf{l}))$. Means and SSP- and scatter matrices of empty clusters are put to zero.

Finally, a $*$ indicates optimality. So, e.g., $\mathbf{l}^*$ is an optimal assignment and $m_j^* = m_j(\mathbf{l}^*)$ is the mean of its jth cluster.

2 Statistical model, criterion, and algorithm

We consider a statistical classification model for a data set D of n observations in some sample space E. At least $r \le n$ of the data are regular, i.e., they are independent draws from one of g densities $f_{\gamma_1}, \dots, f_{\gamma_g}$, each, $\gamma = (\gamma_1, \dots, \gamma_g) \in \Gamma \subseteq \Gamma_1 \times \cdots \times \Gamma_g$. The number of occurrences of each class is unknown. The remaining $n - r$ observations may, but do not have to be gross outliers.

In [7] and [8], we established parametric models with trimming for data with so-called “spurious” outliers and computed their ML- and MAP-estimators. Applying these ideas to the present situation we obtain the following trimmed MAP cluster criterion

$$-r\mathrm{H}\big((n_j(\mathbf{l})/r)_j\big) + \max_{(\gamma_j)_j} \sum_{j=1}^{g} \sum_{\ell_i = j} \ln f_{\gamma_j}(x_i) \tag{3}$$

to be maximized w.r.t. the admissible assignment $\mathbf{l}$. Here, $\sum_{j=1}^{g} \sum_{\ell_i=j} \ln f_{\gamma_j}(x_i)$ is the trimmed log–likelihood given $\mathbf{l}$. The use of the entropy H of the cluster proportions $n_j(\mathbf{l})/r$ goes back to Symons [23] and accounts for unequal cluster sizes. It distinguishes the MAP- from the ML-estimator.

It must be noted that the maximum w.r.t. $(\gamma_j)_j$ required in Eq. (3) does not exist in general. If the parameters γ_j may be chosen freely in the product space $\Gamma_1 \times \cdots \times \Gamma_g$ then the maximum, if it exists, and the sum over j commute so that the double sum reduces to

$$\sum_{j=1}^{g} \max_{\gamma} \ln f_\gamma(C_j(\mathbf{l})).$$

In normal estimation, e.g., the ML-estimate appearing here does not exist if C_j 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 scale equivariance. A second way requires that each cluster contain sufficiently many data points together with an assumption on their locations such as “general position,”[1] Rocke and Woodruff [21]. If the data are in general position and if we allow only

[1] any $d + 1$ elements are affine independent

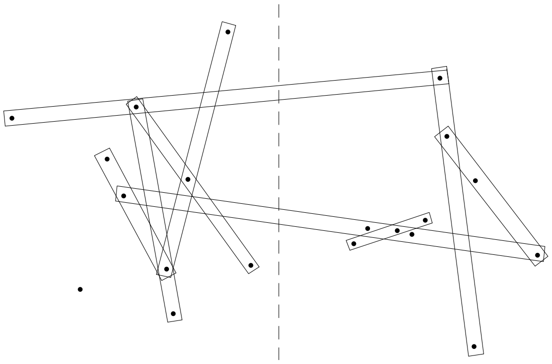

Figure 1: Two clusters of ten points, each, randomly sampled from N_{-2e_1,I_2} and N_{2e_1,I_2}, respectively (separated by the dashed line). Shown are nine almost collinear "spurious clusters." The partitions defined by them all mask the genuine partition, their negative log–posteriors falling below its value 65.96. However, the HDBT ratio (2) of the latter is 1/1.69 whereas the largest of the spurious partitions shown is 1/2757 (the cluster of five points). The optimal unconstrained solution uses the uppermost horizontal cluster and has a negative log–posterior of 60.95 but an HDBT ratio of 1/66 244.

assignments $\mathbf{l}$ with cluster sizes $\geq d+1$ then the maximum of Criterion (3) exists with free parameters and the optimal $\mathbf{l}$ minimizes the criterion

$$2r\mathrm{H}\big((n_j(\mathbf{l})/r)_j\big) + \sum_{j=1}^{g} n_j(\mathbf{l}) \ln\det S_j(\mathbf{l});$$

here $S_j(\mathbf{l})$ is the scatter matrix of cluster j w.r.t. $\mathbf{l}$. 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. Figure 1. We will, therefore, use the HDBT–constraints (1). Letting

$$\mathcal{V} = \mathcal{V}_c = \{\mathbf{V} = (V_j)_j \mid V_j \succ 0,\ V_j \succeq cV_\ell \text{ for all } j,\ell \in 1..g\},$$

we show next in the normal case that the cardinality constraints may be replaced with the HDBT constraints without giving up the maximum of Criterion (3).

2.1 Lemma

Assume that the data are in general position. If $r \geq gd+1$ then, for any assignment $\mathbf{l} \in \Lambda_r$ (some clusters may be empty), the minimum of

$$\sum_{j=1}^{g} n_j(\mathbf{l})\big(\ln\det V_j + \mathrm{tr}\, V_j^{-1} S_j(\mathbf{l})\big)$$

w.r.t. $(V_j)_j \in \mathcal{V}$ exists.

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

$$\sum_j n_j\big(\ln\det V_j + \mathrm{tr}\, V_j^{-1} S_j(\mathbf{l})\big) \geq \sum_j n_j\Big(\ln\det cV_\ell + \mathrm{tr}\, cV_\ell^{-1} S_j(\mathbf{l})\Big)$$
$$= r \ln\det cV_\ell + c\,\mathrm{tr}\, V_\ell^{-1} W(\mathbf{l}),$$

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

$$\sum_j n_j\big(\ln\det V_j + \operatorname{tr} V_j^{-1} S_j(\mathbf{l})\big) \ge r \ln\det cV_\ell + \varepsilon c \operatorname{tr} V_\ell^{-1}.$$

As $(V_j)_j$ approaches the boundary of $\mathcal{V}$, i.e., as some V_j approaches the boundary of 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 estimate tends to ∞. This proves the claim. □

Now standard normal estimation theory shows that, for any admissible assignment $\mathbf{l}$, the partial maximizer w.r.t. m_j in (3) (here, $\gamma_j = (m_j, V_j)$) is given by the sample means

$$m_j(\mathbf{l}) = \begin{cases} \overline{x}_j(\mathbf{l}), & \text{if } C_j(\mathbf{l}) \ne \emptyset, \\ \text{arbitrary, e.g. } 0, & \text{otherwise.} \end{cases}$$

Moreover, if optimization over γ is performed under the HDBT constraints, by Lemma 2.1, the whole maximum exists and equals

$$\max_{\mathbf{m},\mathbf{V}\in\mathcal{V}} \sum_{j=1}^{g} \sum_{\ell_i=j} \ln f_{m_j,V_j}(x_i) = \text{const} - \frac{1}{2}\min_{\mathbf{V}\in\mathcal{V}} \sum_{j=1}^{g} n_j(\mathbf{l})\big(\ln\det V_j + \operatorname{tr} V_j^{-1} S_j(\mathbf{l})\big).$$

This expression contains the scatter matrices $S_j(\mathbf{l})$ w.r.t. $\mathbf{l}$. Finally, the negative constrained trimmed MAP–criterion (3) becomes the *Trimmed Determinant Criterion*

$$\text{(TDC)} \qquad r\mathrm{H}\big((n_j(\mathbf{l})/r)_j\big) + \min_{\mathbf{V}\in\mathcal{V}} \frac{1}{2} \sum_{j=1}^{g} n_j(\mathbf{l})\big(\ln\det V_j + \operatorname{tr} V_j^{-1} S_j(\mathbf{l})\big).$$

We denote the minimal assignment by $\mathbf{l}^*$, $R^* = \{i \mid \ell_i \ne 0\}$ is the set of regular elements w.r.t. $\mathbf{l}^*$, and the partition of R^* associated with $\mathbf{l}^*$ is $(C_1^*, \dots, C_g^*)$. There are only few cases where the minimal parameters V_j given $\mathbf{l}$ 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 $S(\mathbf{l})$ and the TDC becomes up to an additive constant

$$r \cdot \Big\{\mathrm{H}\big((n_j(\mathbf{l})/r)_j\big) + \frac{1}{2}\ln\det S\Big\}. \tag{4}$$

Finally, a univariate case is treated in Proposition 2.3.

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

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

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

2.2 Minimum distance partitions and optimization

Several strategies for optimizing the criteria derived so far are available, among them local descent on a suitably defined graph structure on Λ_r and alternating methods of type k-means. A seeming 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 solutions and MDP's.

We propose here an alternating method of type k-means for producing MDP's. It is first useful to rewrite the posterior density, cf. (3), in a different form:

$$-r\mathrm{H}\big((n_j(\mathbf{l})/r)_j\big) + \sum_{j=1}^{g}\sum_{\ell_i=j}\ln f_{\gamma_j}(x_i) = \sum_{i:\ell_i\neq 0}\Big(\ln\frac{n_{\ell_i}}{r} + \ln f_{\gamma_{\ell_i}}(x_i)\Big) = \sum_{i:\ell_i\neq 0} u_{i,\ell_i}$$

with $u_{i,j} = \ln\frac{n_j}{r} + \ln f_{\gamma_j}(x_i)$, the posterior density of j for x_i. For given parameters γ_j, this sum is maximized w.r.t. $\mathbf{l}$ by assigning each object i according to the MAP discriminant rule and by discarding the $n-r$ observations with the overall smallest posterior probabilities. Given a labelling $\mathbf{l}$, the unconstrained maximum in Criterion (3) chooses as γ the (unconstrained) ML-estimate for the retained observations. As a consequence the following strategy improves the criterion starting from an initial admissible labelling $\mathbf{l}$. We first keep the parameters g and r fixed.

Multipoint reduction step

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

In the Classification step misfits are removed. In the Trimming step the r observations which best fit their clusters are retained. Note that both steps may leave one or more clusters empty. Iteration of the three steps will eventually become stationary since there are only finitely-many labellings and since the criterion improves. The solution attained at convergence is self–consistent (or a *(free) minimum distance partition*) in the sense that partition and parameters generate each other.

The reduction step disregards the HDBT–constraints which would need the ML-estimate w.r.t. $\mathcal{V}_c$ in the Estimation step. In fact, we do not know of a practicable analytical solution of the associated constrained optimization in Euclidean space for $d \geq 2$ and numerical methods would lead to inefficient overall algorithms. An exception are free MDP's that happen to satisfy the constraints – they are automatically *constrained* MDP's. On the other hand, the constant c is unknown and 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 $\mathbf{l}$, 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 $g = 2$. It shows that the constrained minima at the boundary of the constraints depend heavily on the constant c, another disadvantage.

2.3 Proposition

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

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

and $v_j^* = s_j,\ j \in 3..g$.

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

$$\begin{aligned} &\min_{\substack{v_1>0 \\ cv_\ell \leq v_j \leq v_\ell/c,\, \ell<j}} \sum_j f_j(v_j) = \min_{v_1>0}\Big\{f_1(v_1) + \min_{cv_1 \leq v_2 \leq v_1/c}\Big\{f_2(v_2) + \min_{\substack{cv_\ell \leq v_j \leq v_\ell/c \\ \ell<j\geq 3}} \sum_{j\geq 3} f_j(v_j)\Big\}\Big\} \\ &\geq \min_{v_1>0}\Big\{f_1(v_1) + \min_{cv_1 \leq v_2 \leq v_1/c}\Big\{f_2(v_2) + \sum_{j\geq 3} \min_{v>0} f_j(v)\Big\}\Big\} \\ &= \min_{v_1>0}\big\{f_1(v_1) + \min_{cv_1 \leq v_2 \leq v_1/c} f_2(v_2)\big\} + \sum_{j\geq 3} f_j(s_j). \end{aligned}$$

The constrained minimizer of $f_2(v_2)$ w.r.t. v_2 is

$$v_2^* = \begin{cases} s_2, & cs_2 < v_1 < s_2/c, \\ cv_1, & v_1 \geq s_2/c, \\ v_1/c, & v_1 \leq cs_2, \end{cases}$$

and we have shown

$$\min_{\substack{cv_\ell \leq v_j \leq v_\ell/c \\ \ell<j}} \sum_j f_j(v_j) \geq \min_{v_1>0}\big\{f_1(v_1) + f_2(v_2^*)\big\} + \sum_{j\geq 3} f_j(s_j). \tag{5}$$

The function $v_1 \mapsto f_2(v_2^*)$ 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 $v_1 \mapsto f_1(v_1) + f_2(v_2^*)$

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

has a minimum which is attained in the same interval where the minimum of the unimodal function $f_1(v_1)$ lies. The minimizer of the lower bound (5) turns out to be the value v_1^* given in the proposition.

We have, thus, shown that the target function cannot be less than its value at the parameters stated in the proposition. The proof will be finished if we show that these parameters satisfy the 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 $v_j^* > 0$ for all j. □

2.4 Overall algorithm and choice of the constant c

Iteration of multipoint reduction steps strives for labellings with large criteria. If the "fail" signal does not occur then the iteration stalls at some unconstrained MDP for the reasons stated before. However, it does not have to represent an interesting solution so that the process has to be replicated, possibly many times. The number of replications needed depends on the data set and on the initial assignments.

Two different outcomes of the algorithm just described are possible. It may happen that all replications output the signal "fail." This is typically the case if the data set contains very small clusters or if the number of clusters, g, has been chosen too large. An example is the attempt to partition a homogeneous data set. 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, cf. [8]

Otherwise, we obtain unconstrained MDP's and we have to decide which one to use. While the theoretical results presented in this communication are valid for all constants c, not all lead to reasonable partitions as experience shows. The desired solution cannot be determined without a further assumption. In most cases, those solutions are interesting that combine large criterion with 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 quintuplet as a region of concentration. This might suggest a partition in five plus fifteen 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 makes 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 partition: Generate a large number of (unconstrained) MDP's and display their HDBT-ratios vs. their criteria in a negative double–logarithmic plot as shown in Figs. 3 and 5. The convex hull of all MDP's will usually have a knee at its left, lower part. The extreme point at the knee is the favorite solution. 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 the favorite solution has an HDBT-ratio of a few hundred.

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

2.5 Choice of the parameters g and r

Statistical model 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 all pairs (g, r). This is a substantial reduction of the complexity of the data analytic problem. For obvious reasons, r should be chosen no larger than and close to the number of regular elements in the data set. Since there are no formal definitions of "cluster" or "outlier," their numbers are not precisely defined and there cannot be a clear answer to the question how many clusters and outliers there are. We can, however, give some guidelines for the selection of g and r.

Concerning the number of clusters, there are essentially three approaches, cf. [18, 12], *cluster validation*, the so-called *elbow criterion*, and *model selection criteria*. Cluster validation may be divided in two branches: tests and validity measures. The classical test, due to Wolfe [25], is a likelihood ratio test for the hypothesis of k clusters against $(k-1)$ clusters. Bock [2] discusses some significance tests for distinguishing between the hypothesis of a homogeneous population vs. the alternative of heterogeneity. Chen et al. [3] propose a modified likelihood ratio test for a mixture of two components vs. $g \geq 3$. Also normality tests may sometimes be beneficial in this respect, see the comprehensive review by Mecklin and Mundfrom [17]. Validity measures are functionals of partitions and usually measure the quality of cluster separation and of cluster homogeneity (or "compactness"); see, e.g., Bezdek et al. [1]. Often, the total within–cluster sum of squared distances about the centroids is used as a measure of compactness and the total between–cluster sum of squared distances for separation; cf. Milligan and Cooper [18] and the abridged presentation of their work by Gordon [12]. 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 we refer the reader to Tibshirani et al. [24].

Maximum likelihood and maximum a posteriori estimation tend towards a large number of clusters. A *model selection criterion* counteracts this tendency by subtracting a penalty term from the maximum of the log–likelihood or from the posterior log–density. Schwarz [22] 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 [16], Ch. 6. Moreover, Kéribin [15] 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 $\sum_i \ln \sum_{j=1}^{g} \pi_j f_{\gamma_j}$ 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, $q = q(g)$ is $g - 1$ (for the mixing rates) plus the sum of the dimensions of the g population models.

We propose BIC with this number q also for the classification model if separation is sufficiently good. Indeed, let $\mathbf{l}^*$ be the optimal MAP–assignment and let π^* and γ^* be the optimal mixing rates and population parameters of a mixture model under suitable constraints as in Kéribin's theorem. For any g, the optimal value of the MAP–criterion (3) is no larger than that of the mixture model: Assuming without loss $r = n$, we have

$$\begin{aligned}
&-n\mathrm{H}\big((n_j(\mathbf{l}^*)/n)_j\big) + \sum_{j=1}^{g}\sum_{\ell_i^*=j} \ln f_{\gamma_j(\mathbf{l}^*)}(x_i) = \sum_i \Big\{ \ln \frac{n_{\ell_i^*}(\mathbf{l}^*)}{n} + \ln f_{\gamma_{\ell_i^*}(\mathbf{l}^*)}(x_i)\Big\} \\
&= \ln \prod_i \frac{n_{\ell_i^*}(\mathbf{l}^*)}{n} f_{\gamma_{\ell_i^*}(\mathbf{l}^*)}(x_i) \le \ln \prod_i \sum_j \frac{n_j(\mathbf{l}^*)}{n} f_{\gamma_j(\mathbf{l}^*)}(x_i) \le \max_{\pi,\gamma} \ln \prod_i \sum_j \pi_j f_{\gamma_j}(x_i) \\
&= \ln \prod_i \sum_j \pi_j^* f_{\gamma_j^*}(x_i). \qquad (6)
\end{aligned}$$

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 $j \neq \ell_i^*$ and $\pi_j^* \approx \frac{n_j(\mathbf{l}^*)}{n}$ for all $j \in 1..g$ so that, for this g, both ends of the estimate almost meet:

$$-n\mathrm{H}\big((n_j(\mathbf{l}^*)/n)_j\big) + \sum_{j=1}^{g}\sum_{\ell_i^*=j} \ln f_{\gamma_j(\mathbf{l}^*)}(x_i) \approx \ln \prod_i \sum_j \pi_j^* f_{\gamma_j^*}(x_i). \qquad (7)$$

The combination of Kéribin's result with the estimate (6) and the approximation (7) supports BIC as a penalty term also for MAP–partitioning in the case of large data sets and good separation.

An approach to determining the number of clusters can be combined with a χ^2 goodness–of–fit test for estimating the number of outliers. In a first step, establish a table of the optimal partitions 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: Select all pairs $(g, n - r)$ for which none of the g clusters is rejected by a χ^2 goodness–of–fit test, cf. Ritter and Gallegos [20]. If no pair is accepted then the assumptions on the regular model are questionable. If g admits an acceptable pair $(g, n - r)$, keep the one with largest r 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 best g. Since the number r_g of regular observations for the selected partition with g clusters depends on g, the numbers of objects have to be normalized, e.g. to n. By consistency of parameter estimation, cf. [8], Theorems 2.1 and 2.2, the value of the MAP–criterion (3) increases approximately linearly with the number r, asymptotically, at least if there is sufficient separation. Therefore, we propose as model selection criterion with trimming the corrected BIC

$$\operatorname*{argmax}_g \Big\{ -n\,\mathrm{H}\Big(\frac{n_j(\mathbf{l}^*)}{r_g}\Big) + \frac{n}{r_g}\sum_{j=1}^{g}\sum_{\ell_i^*=j} \ln f_{\gamma_j(\mathbf{l}^*)}(x_i) - \frac{q(g)}{2}\ln n\Big\}. \qquad (8)$$

3 Robustness

Although criterion and algorithm involve trimming neither the estimates of the means nor those of the covariance matrices are robust without constraints on the HDBT ratios. In fact, no matter how r is chosen they break down under the influence of a single outlier. Just

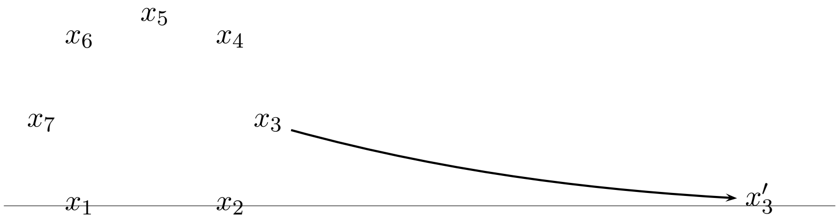

Figure 2: Non–robustness of Criterion (3) with full normal covariance matrices in the *free* heteroscedastic case. Data point x_3 is replaced with some x'_3 close to the abscissa and far away. The criterion discards x_7 generating the partition $\{(x_1, x_2, x'_3), (x_4, x_5, x_6)\}$

consider the data set consisting of seven points shown in Fig. 2. It is to be subdivided in two groups with $r = 6$, i.e., we discard one point. There are two equivalent optimal partitions, $\{(x_2, x_3, x_4), (x_5, x_6, x_7)\}$, 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 allow one point to be discarded, the criterion does not choose the "right" one. In fact, x'_3 creates together with x_1 and x_2 a cluster with a small determinant of its scatter matrix which is very attractive for the optimal partition. 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 next that the HDBT constraints do not only guarantee existence of a solution but also robustness.

3.1 Breakdown values

The finite–sample breakdown value of an estimator, Hodges [14] and Donoho and Huber [6], 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 we have a *fixed* set of regular observations at hand, but there is the disadvantage that we need two estimators, one for r data and one for n data. By contrast, in the latter we have to consider all $\binom{n}{r}$ possible replacements of $n - r$ observations but need only one estimator for n objects. We deal with replacements.

Let $\delta : \mathcal{A} \to \Theta$ an estimator on its natural domain of definition $\mathcal{A} \subseteq E^n$ of admissible data sets of length n, e.g., general position for the m.l.e. under normal assumptions. Given $m \le n$, 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 entries 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

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

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 [7] of δ w.r.t. $\mathcal{K}$ 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, the concept introduced by Donoho and Huber is 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 estimates

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

We deal here with breakdown points of the TDC assessing its robustness via the estimates of means and covariance matrices. The relatively compact subsets of the parameter space $\mathbb{R}^d$ of the means are the bounded sets. A subset of $\mathrm{PD}(d)$ is relatively compact if the eigenvalues of its members 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 positive–definite (or Löwner) ordering $\preceq$ on the vector space of symmetric matrices.

We first show that the TDC provides an asymptotically robust estimate of the covariance matrices and compute the universal breakdown point. We need a lemma. It exploits the pooled SSP-matrix $W(\mathbf{l})$.

3.2 Lemma

Let $\mathbf{V} \in \mathcal{V}$, let $\mathbf{m} \in \mathbb{R}^{gd}$, and let $\mathbf{l}$ be an admissible labelling. We have for all $\ell \in 1..g$

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

Proof.

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

□

The following theorem shows that the robustness of the covariance matrix exceeds even the number of discarded elements, $n - r$, unless r is chosen too small.

3.3 Theorem (TDC: universal breakdown point of the estimates of the covariance matrices)

Assume $r \geq gd + 1$.

(a) If $2r \geq n + g(d+1)$ then the estimates of the covariance matrices remain in a compact subset of $\mathrm{PD}(d)$ that depends only on the original data set D as at most $n - r + g - 1$ data points 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 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 remains bounded below by a strictly positive constant that depends only on the original data set D. Indeed, it is sufficient to compare the optimal solution with the partition that consists of the remaining $r-g+1$ original points in one cluster C_1 and of $g-1$ clusters $C_2 = \{y_2\}, \ldots, C_g = \{y_g\}$ of just one replacement, each. Moreover, choose $V_1 = \cdots = V_g = I_d$, $m_1 = 0$, and $m_j = y_j$, $2 \leq j \leq g$.

By assumption, we replace at most $n - r + g - 1 \leq r - (gd + 1)$ data points so that, for any assignment, at least one cluster, say ℓ, contains at least $d + 1$ original points $T \subseteq D$. This is in particular true for an optimal assignment $\mathbf{l}^*$. It follows $W(\mathbf{l}^*) \succeq W(T) \succeq \varepsilon I_d$, by general position. Lemma 3.2 and the initial remark imply

$$-r \ln \det 2\pi c V_1^* - c \operatorname{tr} W(\mathbf{l}^*)(V_1^*)^{-1} \geq 2 \ln f[R^* \mid \mathbf{l}^*, \mathbf{m}^*, \mathbf{V}^*] \geq \text{const} > -\infty.$$

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 2 that contains at least one replacement. Indeed, if no cluster contains two replacements then each cluster contains at least one and, by $r \geq gd+1$, one of them contains another element. Now, let C_ℓ be such a cluster, let $y \in C_\ell$ be a replacement, and let $x \in C_\ell$, $x \neq y$. We have

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

Now let the parameters $(\mathbf{l}^*, (m_j^*), (V_j^*))$ be optimal. Comparing them with the inferior parameters $(\mathbf{l}^*, (m_j^*), (2V_j^*))$ and noting that the entropy terms coincide, we infer from the TDC

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

The estimate $(y-x)^{\mathrm{T}}(V_\ell^*)^{-1}(y-x) \le 4dr\ln 2$ implies that 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). □

3.4 Corollary

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

As noted after Lemma 2.1, the estimates of the means are the sample means. Contrary to the covariance matrices their universal breakdown point is low. We need a lemma and denote scatter and SSP values by the letters s and w, respectively.

3.5 Lemma

Let $F \cup \{z_1, \dots, z_{g-2}\} \cup \{y_1, y_2\} \subseteq \mathbb{R}$ 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 of the parameters v_j for the partition $\mathbf{l} = \{F, \{z_1\}, \dots, \{z_{g-2}\}, \{y_1, y_2\}\}$ are

$$v_1^* = \frac{w(F) + w(\{y_1, y_2\})/c}{r} \qquad \text{and} \qquad v_j^* = c\,v_1^*,\ 2 \le j \le g.$$

Proof. Putting $s_1 = s(F)$ and $s_g = s(\{y_1, y_2\})$, the TDC requires minimizing the expression

$$n_1\Big(\ln v_1 + \frac{s_1}{v_1}\Big) + \sum_{2\le j\le g-1} \ln v_j + 2\Big(\ln v_g + \frac{s_g}{v_g}\Big) \tag{9}$$

w.r.t. $(v_1, \dots, v_g) \in \mathcal{V}$. We start with the minimum of (9) on the larger set $\mathcal{V}' = \{(v_1, \dots, v_g) \in \mathbb{R}_>^g \mid cv_1 \le v_j \le v_1/c,\ j \in 2..g\} \supseteq \mathcal{V}$. Since $\min_{cv_1\le v_j\le v_1/c} \ln v_j = \ln cv_1$, dynamic optimization shows

$$\begin{aligned}
\min_{\mathbf{v}\in\mathcal{V}'} (9) &= \min_{cv_1\le v_g\le v_1/c}\Big\{n_1\Big(\ln v_1 + \frac{s_1}{v_1}\Big) + \sum_{2\le j\le g-1}\ \min_{cv_1\le v_j\le v_1/c} \ln v_j + 2\Big(\ln v_g + \frac{s_g}{v_g}\Big)\Big\}\\
&= (g-2)\ln c + \min_{cv_1\le v_g\le v_1/c}\Big\{\Big((r-2)\ln v_1 + \frac{w_1}{v_1}\Big) + \Big(2\ln v_g + \frac{w_g}{v_g}\Big)\Big\}.
\end{aligned}$$

This is a virtual two–cluster problem. The second line of the three cases in Proposition 2.3 shows that, under the assumption $\frac{w_g}{2} \le c\frac{w_1}{r-2}$ stated in the lemma, its solution is given by the claimed values v_1^* and $v_g^* = cv_1^*$. Finally, the vector $(v_1^*, cv_1^* \dots, cv_1^*)$ lies even in $\mathcal{V}$ so that it is the minimum w.r.t. the smaller parameter set, too. □

3.6 Theorem (TDC: universal breakdown point of the sample means)
Let $g \geq 2$.

(a) If $n \geq r+1$ and $r \geq gd+2$ then the estimates of all means remain bounded by a constant that depends only on the data set D as *one* observation is arbitrarily replaced.[3]

(b) Under the standard assumption $r \geq gd+1$ there is a data set such that one sample mean of the TDC 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 $\mathbf{l}^*$ 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, this implies $\det W(\mathbf{l}^*) \to \infty$ as $\|y\| \to \infty$. We now use Lemma 3.2

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

It is well known that, given a positive-definite matrix W, the minimum of the function $V \mapsto \ln\det V + \operatorname{tr} WV^{-1}$ is $\ln\det W + d$. Hence, the right side in the inequality tends to $-\infty$ as $\|y\| \to \infty$. This contradicts the fact that the maximum posterior density and, by $r < n$, the left side is bounded below by a constant that depends only on the original data.

(b) A proof in the multivariate case requires a subtle construction of a data set. 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 [7]. 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 $r \leq n-2$ proceeding in three steps.

(α) Construction of the modified data set M:

Let x_i, $1 \leq i \leq r-g$, be strictly increasing and put $F = \{x_1, \dots, x_{r-g}\}$, let $K > 0$, and choose $z_1 < z_2 < \ldots < z_{n-r+g-2}$ such that

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

Let $0 < \varepsilon \leq \sqrt{c\frac{w(F)}{r-2}}$, let $y > z_{n-r+g-2} + \varepsilon$, define the replacements $y_{1,2} = y \pm \varepsilon$, and put $M = \{x_1, \dots, x_{r-g}, z_1, \dots, z_{n-r+g-2}, y_1, y_2\}$. Plainly, M is in general position.

Let $\mathcal{C}^*$ be the partition $\{F, \{z_1\}, \dots, \{z_{g-2}\}, \{y_1, y_2\}\}$ ($z_{g-1}, \dots, z_{n-r+g-2}$ discarded).

(β) The maximum a posteriori density of $\mathcal{C}^*$ does not depend on K and y:

Let us denote the estimates associated with $\mathcal{C}^*$ by $(R^*, (m_j^*)_{j\in 1..g}, (v_j^*)_{j\in 1..g})$. Since $w(\{y_1,y_2\}) = 2\varepsilon^2 \leq \frac{2c}{r-2}w(F)$, Lemma 3.5 shows $v_1^* = \frac{w(F)+w(\{y_1,y_2\})/c}{r}$ and $v_2^* = \ldots = v_g^* = cv_1^*$. Twice the logarithm of the corresponding posterior density equals

$$2\Big((r-g)\ln\Big(\frac{r-g}{r}\Big) + 2\ln\Big(\frac{2}{r}\Big)\Big) - r\ln v_1^* - g\ln c - r(1+\ln 2\pi).$$

[3] In the case of ties the solution is returned that has the largest discarded element.

(γ) If K is large enough then no assignment $\mathbf{l}$ of r points from the set $F \cup \{z_1, \dots, z_{n-r+g-2}\}$ 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 $\mathbf{l}$ creates a cluster $C_\ell = \mathbf{l}^{-1}(\ell)$ which contains some z_k and some other point. From (i), it follows

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

Let $((\widetilde{m}_j)_{j\in 1..g}, (\widetilde{v}_j)_{j\in 1..g})$ denote the constrained m.l.e. defined by $\mathbf{l}$. By Lemma 3.2, twice its log-likelihood is bounded above by

$$-r\ln(2\pi c\widetilde{v}_j) - c\frac{w(\mathbf{l})}{\widetilde{v}_j} \le -r\big(\ln 2\pi c^2/r + \ln w(\mathbf{l}) + 1\big) \underset{K\to\infty}{\longrightarrow} -\infty, \qquad j \in 1..g;$$

here we have used the maximum of the left side as a function of $\widetilde{v}_j$ and (10). The claim follows from (β) since there are only finitely many $\mathbf{l}$'s.

Finally, choose K as in (γ). The optimal solution retains at least one y_h causing at least one 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 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, Gordaliza [11], see also García–Escudero and Gordaliza [9], carried out experiments with trimmed g–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 sample means

Dealing with the homoscedastic case, we computed in [7] the restricted breakdown point of the sample 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 theorem of this section depends on lemmas which we first state and prove. Let $\mathcal{P} = \{P_1, ..., P_g\}$ be a partition of D and let $\emptyset \ne T \subseteq D$. The partition $T \cap \mathcal{P} = \{T \cap P_1, \dots, T \cap P_g\}$ is the *trace* of $\mathcal{P}$ in T. Let $g' \ge 1$ be a natural number and let $\mathcal{T} = (T_1, \dots, T_{g'})$ be some partition of T. The *common refinement* of $\mathcal{T}$ and $\mathcal{P}$ is denoted by $\mathcal{T} \sqcap \mathcal{P} = \{T_k \cap P_j \mid k \le g', j \le g\}$, 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

$$W(\mathcal{T}) = \sum_{j \le g'} W(T_j).$$

For all $(n_1, \dots, n_g) \in \mathbb{N}^g$ such that $n_1 + \ldots + n_g = r$, the entropy satisfies the estimate

$$-r\ln g \le -r\mathrm{H}((n_j/r)_j) \le 0. \tag{11}$$

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

4.1 Proposition

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

$$\det W(\mathcal{T}) \geq g^2 \max_{R \in \binom{D}{r}, R \supseteq T} \det\Big(\frac{1}{c^2} W(R \cap \mathcal{P})\Big). \tag{12}$$

Then the individual breakdown point of TDC 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 $(R^*, \mathbf{l}^*, (m_j^*)_{j=1}^g, (V_j^*)_{j=1}^g)$ be an optimal constrained solution for M. We will show that its 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 $\mathbf{l}^*$. Since $\#R^* = r$, $R^* = \bigcup_{j=1}^g R_j^*$ has at least $q \geq gd+1$ original observations and some R_j^* contains at least $d+1$ original observations. The proof now finishes as that of Theorem 3.3(a).

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

By (α), $\operatorname{tr} W(\mathbf{l}^*)$ remains bounded above by a constant which depends solely on the original data D. Now, let $x \in R_j^* \cap D$. We have $W(\mathbf{l}^*) \succeq (x - m_j^*)(x - m_j^*)^{\mathrm{T}}$ and, hence, $\|x - m_j^*\|^2 \leq \operatorname{tr} W(\mathbf{l}^*)$ 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^*, $j \in 1..g$, consists solely of original observations or solely of modifications. We next put $c_{d,r} = -\frac{dr}{2}(1 + \ln 2\pi)$.

(δ) $-r\mathrm{H}((n_j^*/r)_j) + \ln f[R^* \mid \mathbf{l}^*, \mathbf{m}^*, \mathbf{V}^*] < c_{d,r} - dr \ln c - \dfrac{r}{2} \ln\det \dfrac{W(\mathbf{l}^*)}{r}$,
whenever $0 < n_j^* < r$ for some j:

On account of Lemma 3.2 and the assumption, the left side is strictly bounded above by

$$-dr \ln c - \frac{dr}{2} \ln 2\pi - \frac{1}{2}\big[r \ln\det(V_1^*/c) + \operatorname{tr}\big(W(\mathbf{l}^*)(V_1^*/c)^{-1}\big)\big].$$

Part (α) and normal estimation theory now show that the function $A \mapsto r \ln\det(A/c) + \operatorname{tr}\big(W(\mathbf{l}^*)(A/c)^{-1}\big)$, $A \succeq 0$, attains its minimum at $\frac{cW(\mathbf{l}^*)}{r}$ with value $r\Big[\ln\det(\frac{W(\mathbf{l}^*)}{r}) + d\Big]$ 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 just after (γ), some cluster, say R_g^*, consists solely of replacements. Note that $r > \#(R^* \cap D) \ge q$. Let $T = R^* \cap D$ and let $\mathcal{T} = \{R_1^* \cap D, \dots, R_{g-1}^* \cap D\}$. From Steiner's formula we have the relation $W(\mathbf{l}^*) \succeq W(\mathcal{T})$ between the pooled SSP-matrices and Hypothesis (12) implies

$$\det W(\mathbf{l}^*) \ge \det W(\mathcal{T}) \ge g^2 \max_{R \in \binom{D}{r}, R \supseteq T} \det \Big(\frac{1}{c^2} W(R \cap \mathcal{P})\Big).$$

Hence,

$$2d \ln c + \ln\det \frac{W(\mathbf{l}^*)}{r} \ge 2 \ln g + \max_{R \in \binom{D}{r}, R \supseteq T} \ln\det \frac{1}{r} W(R \cap \mathcal{P}). \tag{13}$$

Now, writing $\mathbf{m}(R \cap \mathcal{P}) = (m(R \cap P_1), \dots, m(R \cap P_g))$ and $V(R \cap \mathcal{P}) = \frac{1}{r} W(R \cap \mathcal{P})$, the pooled scatter matrix, we estimate

$$\begin{aligned}
& r \ln g + \min_{R \in \binom{M \cap D}{r}} - \ln f[R \mid \mathbf{l}_{R \cap \mathcal{P}}, \mathbf{m}(R \cap \mathcal{P}), V(R \cap \mathcal{P})] \\
= \; & -c_{d,r} + r \ln g + \frac{r}{2} \min_{R \in \binom{M \cap D}{r}} \ln\det V(R \cap \mathcal{P}) \\
\le \; & -c_{d,r} + r \ln g + \frac{r}{2} \min_{T \subseteq R \in \binom{M \cap D}{r}} \ln\det V(R \cap \mathcal{P}) \\
\le \; & -c_{d,r} + r \ln g + \frac{r}{2} \max_{T \subseteq R \in \binom{D}{r}} \ln\det V(R \cap \mathcal{P}) \qquad (14) \\
\le \; & -c_{d,r} + dr \ln c + \frac{r}{2} \ln\det V(\mathbf{l}^*) \\
< \; & r \mathrm{H}((n_j^*/r)_j) - \ln f[R^* \mid \mathbf{l}^*, \mathbf{m}^*, \mathbf{V}^*],
\end{aligned}$$

where the last but one and last inequalities follow from (13) 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 partition $R \cap \mathcal{P}$ with the parameters $\mathbf{m}(R \cap \mathcal{P})$ and $V(R \cap \mathcal{P})$ for all $R \in \binom{M \cap D}{r}$. By an elementary property of the entropy, the latter is no larger than the left side of (14). This contradiction proves Claim (ϵ).

Finally, Part (β) shows that all 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 4.1 expresses actually 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

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

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 4.1.

Let $\mathcal{P}$, T, and $\mathcal{T}$ be as in Proposition 4.1. Our next, combinatorial, lemma gives conditions that secure the existence of sufficiently many elements of T in each P_j and a large intersection $T_k \cap P_j$ for some pair (k, j).

4.2 Lemma

Let $\mathcal{P} = \{P_1, \dots, P_g\}$ be a partition of D in clusters of size $\geq u$, let $T \subseteq D$ such that $q_{u,\varrho} \leq \#T < r$, and let $\mathcal{T} = \{T_1, \dots, T_{g-1}\}$ 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

$$n \geq \#T + \#P_j - \#(T \cap P_j) > \#T + u - \varrho \#T = u + (1-\varrho)\#T \geq u + (1-\varrho) q_{u,\varrho} \geq u + n - u$$

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 (b) did not hold, we would have $\#T < q_{u,\varrho}$, contradicting one of the assumptions. □

The theorem on the breakdown point 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

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

4.3 The 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 $\mathcal{T} = \{T_1, \dots, T_{g-1}\}$ of T in $g-1$ clusters, we have

$$\begin{aligned} 1 + \kappa_\varrho \min_{\substack{k,\, j \neq \ell :\\ T_k \cap P_h \neq \emptyset,\ h = j, \ell}} & (m_{T_k \cap P_j} - m_{T_k \cap P_\ell})^{\mathrm{T}} \Big(\frac{W(\mathcal{T} \sqcap \mathcal{P})}{\#T}\Big)^{-1} (m_{T_k \cap P_j} - m_{T_k \cap P_\ell}) \\ & \geq g^2 \frac{\max_{R \in \binom{D}{r}, R \supseteq T} \det \frac{1}{c^2} W(R \cap \mathcal{P})}{\det W(\mathcal{T} \sqcap \mathcal{P})}. \end{aligned} \tag{15}$$

According to Lemma 4.2(b), the minimum extends over at least one triple (k, j, ℓ), $j \neq \ell$, and by Lemma 4.2(c), the pooled scatter matrix $V(\mathcal{T} \sqcap \mathcal{P})$ is bounded below by a positive-definite matrix which depends only on D. Condition (15) is affine equivariant. We require that the minimum of the Mahalanobis distances of the submeans of P_j and P_ℓ appearing on its left-hand side should be large. Thus, Condition (15) means that the partition $\mathcal{P}$ subdivides the data set in well–separated clusters, it is the "natural" clustering 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 4.1.

4.4 Lemma

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 [7], Lemma A.3, to each $W(T_k)$, $1 \leq k < g$, with partition $\{T_k \cap P_1, \ldots, T_k \cap P_g\}$, $1 \leq j \leq g$, shows first

$$\begin{aligned} W(\mathcal{T}) &= \sum_{k=1}^{g-1} W(T_k) \\ &= \sum_{k:T_k \neq \emptyset} \Big\{ \sum_{j=1}^{g} W(T_k \cap P_j) + \sum_{1 \leq j < \ell \leq g} \frac{a_{kj} a_{k\ell}}{\#T_k} (m_{T_k \cap P_j} - m_{T_k \cap P_\ell})(m_{T_k \cap P_j} - m_{T_k \cap P_\ell})^{\mathrm{T}} \Big\}, \end{aligned}$$

where $a_{kj} = \#(T_k \cap P_j)$, $1 \leq j \leq g$, $1 \leq k < g$. Now use [7], Lemma A.1(b), and Lemma A.1 to estimate

$$\begin{aligned} &\det W(\mathcal{T}) \\ &\geq \det W(\mathcal{T} \sqcap \mathcal{P}) \cdot \\ &\qquad \Big\{ 1 + \sum_{k:T_k \neq \emptyset} \sum_{1 \leq j < \ell \leq g} \frac{a_{kj} a_{k\ell}}{\#T_k} (m_{T_k \cap P_j} - m_{T_k \cap P_\ell})^{\mathrm{T}} W(\mathcal{T} \sqcap \mathcal{P})^{-1} (m_{T_k \cap P_j} - m_{T_k \cap P_\ell}) \Big\} \\ &\geq \det W(\mathcal{T} \sqcap \mathcal{P}) \Big\{ 1 + \kappa_\varrho \min_{\substack{k, j \neq \ell : \\ T_k \cap P_h \neq \emptyset}} (m_{T_k \cap P_j} - m_{T_k \cap P_\ell})^{\mathrm{T}} \Big(\frac{W(\mathcal{T} \sqcap \mathcal{P})}{\#T} \Big)^{-1} (m_{T_k \cap P_j} - m_{T_k \cap P_\ell}) \Big\} \end{aligned}$$

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 4.4 exists. A simple reasoning shows that the bounds on u imply $\varrho < \frac{1}{g}$. Furthermore, the inequality $n-(1-\varrho)(r-1) \leq u$ implies $u \geq n-r+2$. In particular, the sizes of the natural clusters must exceed the number of discarded elements.

We finally state and prove the main result of this section. If a data set has the separation property then the TDC is much more robust w.r.t. the mean values than predicted by Theorem 3.6.

4.5 Theorem (TDC: restricted breakdown point of the sample means)

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 for the mean values w.r.t. $\mathcal{L}_{u,\varrho,c}$ satisfies

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

(b) The individual breakdown point of any data set $D \in \mathcal{L}_{u,\varrho,c}$ satisfies

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

(c) Let $2r - n \ge (g-1)gd + 1$, let $u \in \mathbb{N}$ such that $2(n-r) < u \le n/g$, and put $\varrho = \frac{u-2(n-r)}{2r-n}$ (the largest ϱ for fixed u). 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 $2(n-r) \le n/g - 1$.)

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

Proof. Part (a) is a direct consequence of Proposition 4.1 and Lemma 4.4.

(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 $C_j^* = C_j(\mathbf{l}^*)$ for some j. Lemma A.2 shows that the norm of m_j^* tends to infinity together with the narrow cluster of replacements.

(c) The hypotheses imply $\min\left\{n - r + 1, r - (g-1)gd, r + 1 - \frac{n-u}{1-\varrho}\right\} = n - r + 1$. 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 4.1. □

The following corollary of Theorem 4.5 says that the TDC of the means is asymptotically robust on well–separated, balanced data sets if the natural parameter g is used.

4.6 Corollary

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

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

5 Two studies

We illustrate the method described in Sects. 2.2, 2.4, and 2.5 with two examples and first revisit the simple data set of Fig. 1. As already seen there exist a number of minimum distance partitions with larger posterior densities than the intended partition. Fig. 3 shows the negative double-logarithmic posterior-density-HDBT-ratio plot of the MDP's found for the heteroscedastic full normal model with two clusters, no discarded elements, and unknown cluster sizes. According to the method of Sect. 2.4, the most plausible MDP is the one in the left lower corner close to (66, 0.2). It belongs indeed to the favorite partition of the data

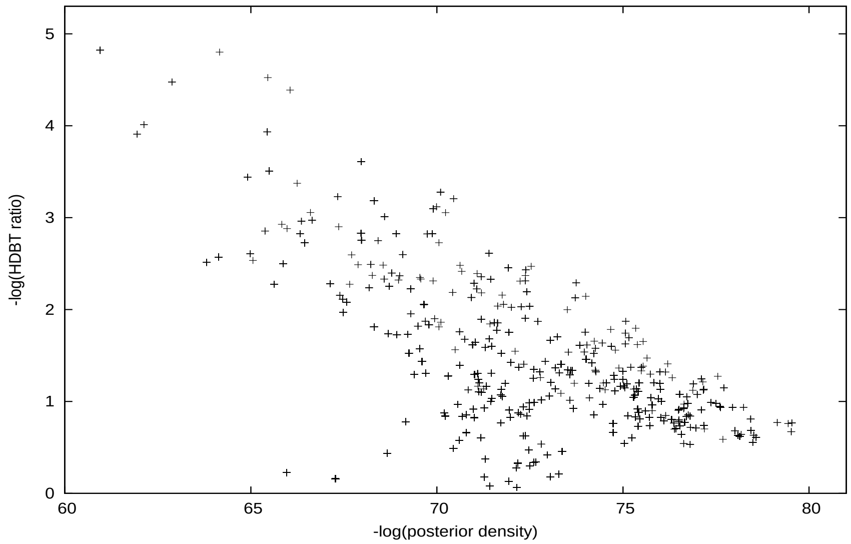

Figure 3: Synthetic data set of Fig. 1: negative double-logarithmic HDBT-ratio-posterior-density plot for a large number of minimum distance partitions with two clusters and no discarded elements.

set in 10 plus 10 elements. 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 [19] 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 w.r.t. 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 (MAP) 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 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 $d = 15$. Like almost any real data set, the present one contains outliers, see Fig. 4, and we apply the algorithm proposed in Sections 2.2, 2.4, and 2.5 with ten percent of discarded elements. The minimum cluster size was set to $d + 1 = 16$.

Fig. 5 shows the negative double-logarithmic posterior-density-HDBT-ratio plot of the MDP's of 1100 replications for six clusters. The favorite solution at the left end of the almost horizontal support line is encircled. A 2D representation of this partition is shown in Fig. 7. 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). Fig. 7 shows that the assumed number of outliers is too small.

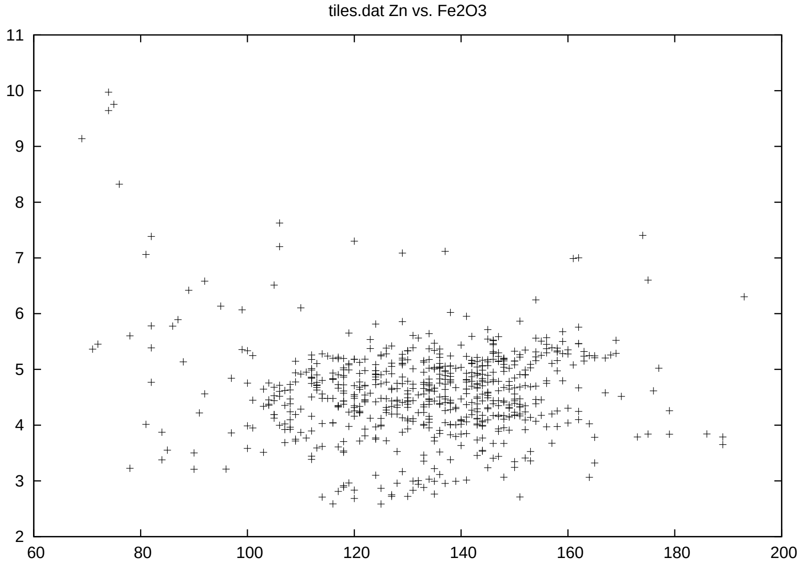

Figure 4: Tiles data: scatter plot of the features Zn and Fe_2O_3 displaying outliers

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. 6. 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.

A Appendix

A.1 Lemma

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 $j \in 1..g$, and put $a_{k\cdot} = \sum_j a_{kj}$. Then

$$\sum_{k:a_{k\cdot}>0} \frac{1}{a_{k\cdot}} \sum_{1\leq j<\ell\leq g} a_{kj}a_{k\ell} \geq \kappa_\varrho \|\mathbf{a}\|_1. \tag{16}$$

Proof. Write the left hand side of (16) as

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

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. $\beta A \geq \varrho$ elementwise, the claim follows from an elementary reasoning. □

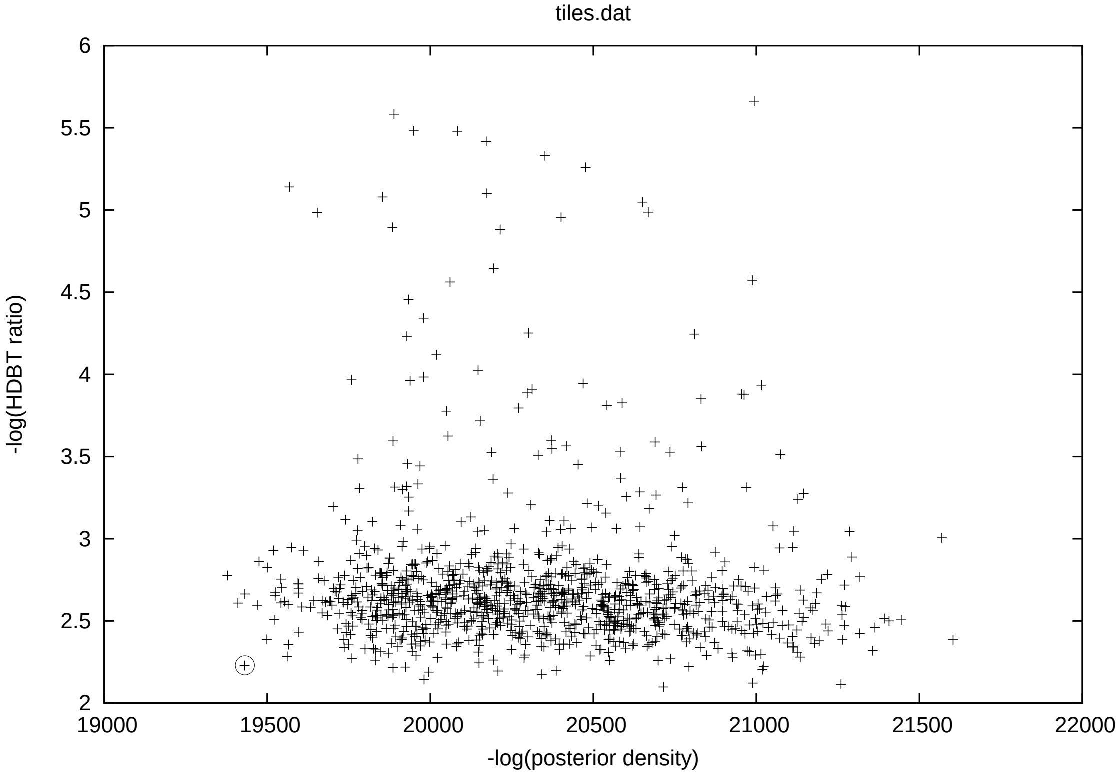

Figure 5: Tiles data: negative double-logarithmic HDBT-ratio-posterior-density plot for the minimum distance partitions of 1100 replications for the heteroscedastic, full normal model with six clusters and 66 discarded points. The encircled solution in the left lower part is most promising.

A.2 Lemma

Let $h \geq 0$ and let $k \geq 1$. 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 $y_i - y_1$, $2 \leq i \leq k$, remain bounded.

Proof. The sum of C is $\sum_{i=1}^h x_i + ky_1 + \sum_{i=2}^k (y_i - y_1)$ from which the lemma follows. $\square$

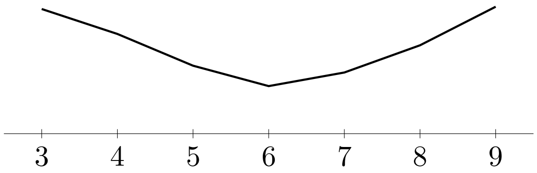

Figure 6: Tiles data: the BIC curve for the favorite solutions with three to nine clusters suggested by the posterior-density-HDBT-ratio plots.

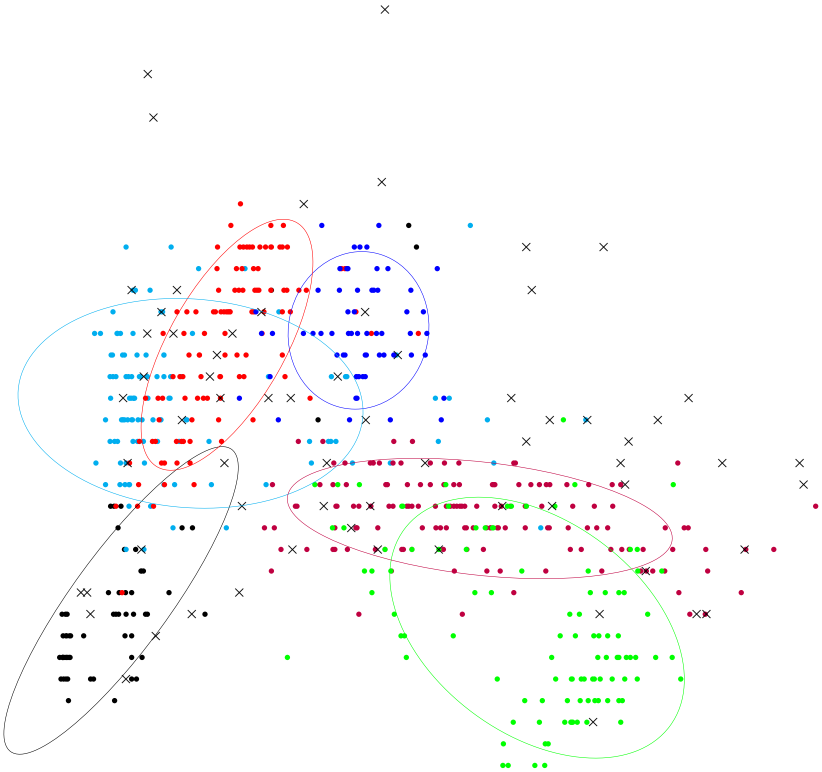

Figure 7: Tiles data: MnO-Y plot of the favorite MDP. The ellipses indicate the 0.8-quantiles of the clusters and x's stand for discarded points.

References

[1] James C. Bezdek, James Keller, Raghu Krisnapuram, and Nikhil R. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. The Handbooks of Fuzzy Sets Series. Kluwer, Boston, London, Dordrecht, 1999.

[2] Hans-Hermann Bock. On some significance tests in cluster analysis. *J. Classification*, 2:77–108, 1985.

[3] Hanfeng Chen, Jiahua Chen, and John D. Kalbfleisch. Testing for a finite mixture model with two components. *J. Royal Stat. Soc, Series B*, 66:95–115, 2004.

[4] J. A. Cuesta-Albertos, Alfonso Gordaliza, and C. Matrán. Trimmed k–means: An attempt to robustify quantizers. *The Annals of Statistics*, 25:553–576, 1997.

[5] John E. Dennis Jr. Algorithms for nonlinear fitting. In M.J.D. Powell, editor, *Nonlinear Optimization 1981*, London, New York, etc, 1982. Academic Press. (Procedings of the NATO Advanced Research Institute held at Cambridge in July 1981).

[6] David L. Donoho and Peter J. Huber. The notion of a breakdown point. In Peter J. Bickel, Kjell A. Doksum, and J.L. Hodges, Jr., editors, *A Festschrift for Erich L. Lehmann*, The Wadsworth Statistics/Probability Series, pages 157–184. Wadsworth, Belmont, CA, 1983.

[7] María Teresa Gallegos and Gunter Ritter. A robust method for cluster analysis. *Annals of Statistics*, 33:347–380, 2005.

[8] María Teresa Gallegos and Gunter Ritter. Model-based clustering with the assignment problem. Technical Report MIP-0610, Universität Passau, Fakultät für Mathematik und Informatik, 2006.

[9] Luis Angel García-Escudero and Alfonso Gordaliza. Robustness properties of k–means and trimmed k–means. *Journ. Amer. Stat. Assoc.*, 94:956–969, 1999.

[10] Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, and Augustin Mayo-Iscar. A general trimming approach to robust cluster analysis. *Ann. Stat.*, 36:1324–1345, 2008.

[11] Alfonso Gordaliza. Robustness propetrties of k–means and trimmed k–means. *Statistics & Probability Letters*, 11:387–394, 1991.

[12] A. D. Gordon. *Classification*, volume 82 of *Monographs on Statistics and Applied Probability*. CRC Press, second edition, 1999.

[13] Richard J. Hathaway. A constrained formulation of maximum–likelihood estimation for normal mixture distributions. *Ann. Stat.*, 13:795–800, 1985.

[14] J. L. Hodges Jr. Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, pages 163–186, Berkeley, 1967. Univ. California Press.

[15] Christine Kéribin. Consistent estimation of the order of mixture models. *Sankhyā*, 62, Series A:49–66, 2000.

[16] Geoffrey J. McLachlan and David Peel. *Finite Mixture Models*. Wiley, New York etc., 2000.

[17] Christopher J. Mecklin and Daniel J. Mundfrom. An appraisal and bibliography of tests for multivariate normality. *International Statistical Review*, 72(1):123–138, 2004.

[18] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.

[19] H.-J. Mucha, H.G. Bartel, and J. Dolata. Exploring Roman brick and tile by cluster analysis with validation of results. In Wolfgang Gaul and Gunter Ritter, editors, *Classification, Automation, and New Media*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 471–478, Berlin etc., 2002. Springer.

[20] Gunter Ritter and María Teresa Gallegos. Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Patt. Rec. Lett.*, 18:525–539, 1997.

[21] David M. Rocke and David L. Woodruff. A synthesis of outlier detection and cluster identification. Technical report, University of California, Davis, 1999. http://handel.cipic.ucdavis.edu/∼dmrocke/Synth5.pdf.

[22] Gideon Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6:461–464, 1978.

[23] M. J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981.

[24] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *J. Royal Stat. Soc., Series B*, 63:411–423, 2001.

[25] J.H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Res.*, 5:329–350, 1970.