Sankhyā : The Indian Journal of Statistics
2009, Volume 71-A, Part 2, pp 164-220

Trimmed ML Estimation of Contaminated Mixtures

María Teresa Gallegos
Institute for Data Analysis, Salzweg, Germany
Gunter Ritter
Universität Passau, Germany

Abstract

We establish a mixture model with "spurious" outliers and derive its maximum likelihood estimator, the maximum trimmed likelihood estimator MTLE. It may be computed with a trimmed version of the EM algorithm which we call the EMT algorithm. We analyze its properties and compute various breakdown values of the estimator for normal mixtures thereby proving robustness of the method.

AMS (2000) *subject classification.* Primary 62H12; Secondary 62F35.
Keywords and phrases. Heteroscedastic mixture models, HDBT ratio, robust maximum likelihood estimation, trimming, MTLE, breakdown point, EMT algorithm.

1 Introduction

1.1. History, background, and outline. Multimodal distributions occur in important applications of statistics, e.g., in pattern recognition, image processing, speech recognition, classification, and clustering. They arise in particular when data emanate from different causes. Some examples are offered in the literature cited in the Introduction of Redner and Walker (1984). Mixture models are useful for modeling such distributions and their decomposition in components plays a major role in the examples above. The maximum likelihood paradigm is nowadays the preferred approach to estimating their parameters.

Some issues related to the m.l.e. such as existence, efficient computation, and statistical properties such as consistency, asymptotic normality, and robustness have been investigated in the past. Day (1969), Sect. 7, notes that the heteroscedastic normal model always fails to possess an m.l.e.

in the strict sense. It is sufficient to center one component at one data point and have its variance tend to zero to see the unboundedness of the likelihood function. On the other hand, under regularity conditions and in the presence of sufficiently many data a strongly consistent and asymptotically efficient local maximum always exists, see Peters and Walker (1978) and Kiefer (1978) and the literature cited there. The situation is simpler in the normal *homoscedastic* case (common covariance matrix). Day (1969) states that solutions to the homoscedastic likelihood equations exist if the data set is not too small. If mixture components are poorly separated then the solution is almost always not unique. Hathaway (1985) circumvented the non-existence in the heteroscedastic, univariate, normal case by assuming that the variances $v_1, \ldots, v_g$ of the g components satisfy the constraints $v_j \geq cv_\ell$ for some constant $c > 0$ and all $j, \ell \in 1..g$. He refers to Dennis (1982) who, in turn, gives credit to Evelyn Martin Lansdowne Beale and James R. Thompson. On account of this origin and since they play a key rôle in our communication, we call them the Hathaway-Dennis-Beale-Thompson constraints (HDBT). Hathaway points out that they avoid also many of the spurious local maxima. Moreover, he proves strong consistency of the properly constrained m.l.e. The m.l.e. is equivariant w.r.t. affine transformations. An important problem is that of selecting the number of components. Kéribin (2000) studied conditions that ensure consistency of certain model selection criteria. The Bayesian information criterion, BIC, turns out to be a consistent maximum penalized likelihood estimator for normal mixtures. If clusters are not well separated it needs, however, many data.

Hasselblad (1966) and Day (1969) designed alternating algorithms for computing the m.l.e. in the hetero- and homoscedastic cases. Dempster et al. (1977) noticed that they were special cases of a general concept for ML estimation in complex models if the distributions can be conveniently represented by "hidden" variables. They named it the EM algorithm. Chrétien and Hero (2000) embedded EM in the general scheme of PPA algorithms, Martinet (1970) and Rockafellar (1976). Ma and Fu (2005) discuss accelerations of the EM algorithm.

Parameter estimation can be severely affected by outliers. This is in particular true for mixture models. A cluster of remote outlying observations will, as a rule, want to establish a component of its own. If the weight of the component is small, an estimate of the *mixture* itself may still be close to the original w.r.t. a suitable metric on the convex set of probability measures, cf. Toma (2007). But the estimate of the *parameters* of one component goes

astray. This fact makes robust parameter estimation in mixture models a difficult problem. It has been taken up mainly in recent years.

- McLachlan and Basford (1988) propose robust estimation of the parameters of the components, e.g., by means of Huber's (1981) robust M-estimators.
- McLachlan and Peel (1998, 2000) use mixtures of t-distributions (or Pearson's type VII distributions) instead of normal mixtures.
- Fraley and Raftery (2002) propose an additional component uniform on the convex hull of the data in order to accommodate outliers.

Hennig (2004), p. 1326, analyzes the performance of these methods with known and unknown numbers of components on data sets with outliers and inliers. He notices that all are ineffective in the presence of gross outliers, at least if the number of components is known and fixed. In fact, one such outlier causes one mean to break down in all three methods. As a remedy, he proposes to

- modify Fraley and Raftery's approach by an additional component with a certain improper uniform distribution.

Hennig (2004), Theorem 4.11, shows for one-dimensional data that his method adds breakdown robustness to the m.l.e.. Besides a lower bound on the variances which has to be carefully chosen, Hennig's estimator needs the coefficient of the improper uniform distribution, b, as a second parameter. He does not transform his ideas to an implementable algorithm.

We take here a different approach to robust parameter estimation in mixture models following in Section 2.1 our statistical model of "spurious" outliers, see Gallegos and Ritter (2005, 2010). It leaves the primary rôle to the regular populations and allows to derive trimming algorithms. We use it here again in order to establish a mixture model with outliers and r retained observations. The likelihood function of this model is the *trimmed likelihood*, Neykov and Neytchev (1990). It uses only the retained observations. For its maximization we design an algorithm that consists of the alternating application of EM and trimming steps, the EMT algorithm. In Sections 2.3 – 2.9, we study its properties.

Denoting the positive-semidefinite (Löwner) ordering on the space of symmetric matrices by $\succeq$, we reformulate the HDBT constraints for the multivariate, normal model with covariance matrices $V_1, \ldots, V_g$ as

$$V_j \succeq cV_\ell \quad \text{for all} \quad j, \ell \in 1..g. \tag{1.1}$$

The constant c is necessarily ≤ 1. While restricting the relationships between sizes and shapes of the g scale parameters V_j, they impose no restriction on their absolute values. In other words, V_1 may be any positive-definite matrix. Two choices of the constant c are of particular interest: $c = 1$ specifies homoscedasticity and the constrained heteroscedastic case with $c \ll 1$ allows much freedom between scale parameters. We show in Proposition 2.13 that the trimmed ML criterion for the HDBT-constrained normal model possesses a maximum.

In general, the constant c is not a priori given and, in Section 2.14, we address its estimation. The *HDBT ratio* (Gallegos and Ritter, 2009) of the g-tuple $\mathbf{V} = (V_1, \ldots, V_g)$ is the largest constant c such that all HDBT constraints (1.1) are satisfied. It is a measure of *balance* of the g components and can be computed as

$$r_{\text{HDBT}}(\mathbf{V}) = \max\{c \mid V_j \succeq c\, V_\ell \text{ for all } j, \ell\} = \min_{j,\ell,k} \lambda_k\big(V_\ell^{-1/2} V_j V_\ell^{-1/2}\big), \tag{1.2}$$

where $\lambda_1(A), \ldots, \lambda_d(A)$ denote the d eigenvalues of a symmetric d by d matrix A. The larger the HDBT ratio of a mixture is the more similar component scales are. In many applications, the solutions with the largest likelihoods, i.e., with the highest fit, are spurious and undesirable suffering from low balance. The desired solution often does not enjoy the best fit but is much more balanced. We are thus facing a biobjective optimization problem: seek a fitting and balanced solution. In Section 2.14, we propose an exploratory method for simultaneously estimating the HDBT ratio and determining the trimmed m.l.e. for HDBT-constrained, heteroscedastic normal mixtures, the *(constrained) normal MTLE*.

The aim of trimming is robustness. We compute in Sections 3 and 4 various (replacement) breakdown points of the normal MTLE. It turns out that the constraints do not only guarantee existence of a solution but also its robustness. Theorem 3.2 says that the normal MTLE of the covariance matrices tolerates even more outliers than there are trimmed elements, as long as at least half of the data set is regular. This implies a strictly positive asymptotic breakdown point. Unfortunately, the (usual universal) breakdown value of the estimates of the means is almost zero, Theorem 3.5: whereas the criterion resists one gross outlier in any data set, there are data sets such that one mean breaks down if two observations are replaced with special outliers. One reason for this misbehavior is the stringency of the

universal breakdown point. It requires proper performance of the estimator even when it is applied with an inappropriate number of components. For instance, one cannot expect breakdown robustness in a clear structure of two components if three components are assumed. We, therefore, follow here again a concept already introduced and investigated by Gallegos and Ritter (2005, 2009), in the framework of clustering, the *restricted* breakdown point. We show in Theorem 4.4 and its corollary that the asymptotic breakdown value of the normal MTLE of the means applied to data sets with well-separated clusters is strictly positive if their natural number of clusters is used. Not surprisingly, the better the components are separated, the more uniform the mixing rates are, and the larger the HDBT constant c is, the more robust the estimator turns out to be; see the comment before Theorem 4.4.

1.2 General notation. Given two integers $m \leq n$, the symbol $m\,..\,n$ designates the set of integers k s.th. $m \leq k \leq n$. The set of all r-element subsets of a set M is denoted by $\binom{M}{r}$. The symbol Δ_{g-1} denotes the $(g-1)$-dimensional unit simplex, that is, the set of all probability vectors of length g. The convex hull of a subset $T \subseteq \mathbb{R}^d$ is denoted by $\operatorname{conv} T$.

We are given a data set $D = \{x_1, \ldots, x_n\}$ of n points x_i in a measurable sample space E, often d-dimensional Euclidean space $\mathbb{R}^d$. Given a subset $R \subseteq D$, we denote the elements of R written as a tuple by $\mathbf{x}_R$ and $\mathbf{x} = \mathbf{x}_D = (x_1, \ldots, x_n)$. Parameter spaces of statistical models are metric. We denote them by the upper case Greek letters Γ, Ψ, and Θ and parameters by the corresponding lower case letters γ, ψ, θ, and ϑ. If a random variable $X : \Omega \to E$ is distributed according to a probability measure μ we write $X \sim \mu$. Its density function w.r.t. some (fixed) reference measure on E is denoted by f_X or f_μ. The conditional density of X given the parameter γ is $f_\gamma(x) = f_X[x \mid \gamma]$. We assume that it is a continuous function of γ.

1.3 The EM algorithm. The EM algorithm computes the ML (or MAP) estimate of the parameter $\vartheta \in \Theta$ of a complex statistical model $X \sim \mu_\vartheta$ by representing it as a measurable function $X = \Phi(Y)$ of a so-called *complete-data* model $Y \sim \nu_\vartheta$ that is easier to handle. At its heart is the so-called Q-functional, the conditional expectation of the complete-data log-likelihood $\ln f_{\nu_\theta}$ given the observation w.r.t. the "current" fit ϑ

$$Q(\vartheta, \theta) = \mathrm{E}_{\nu_\vartheta}[\ln f_{\nu_\theta} \mid \Phi = x].$$

Intuitively, one wishes to maximize the function $\theta \mapsto \ln f_Y[y \mid \theta]$. Since it is not observed, the EM algorithm recursively maximizes $Q(\vartheta, \cdot)$ if possible,

i.e.,

$$\vartheta \leftarrow \operatorname*{argmax}_{\theta} Q(\vartheta, \theta).$$

One step in this process is called an EM-step. Dempster, Laird, and Rubin (1977) showed that each iteration of an EM-step increases the observed likelihood (in the sense $\geq$) and if the iterations stall then, under regularity conditions, they do so at a critical point of the observed likelihood function.

More recently, Chrétien and Hero (2000) derived the properties of the EM algorithm from those of a *proximal point algorithm* (PPA), see Martinet (1970) and Rockafellar (1976), with the conditional Kullback-Leibler divergence as the penalty function. To this end they introduce the difference of the observed log-likelihood and the Kullback-Leibler divergence of the complete model w.r.t. the current complete distribution ν_ϑ conditional on the observation,

$$H(\vartheta, \theta) = \ln f_X[x \mid \theta] - D(\vartheta, \theta). \tag{1.3}$$

Assume that the divergence

$$D(\vartheta, \theta) = \int \nu_\vartheta[\,\mathrm{d}y \mid \Phi = x] \ln \frac{f_{\nu_\vartheta}[y \mid \Phi = x]}{f_{\nu_\theta}[y \mid \Phi = x]}$$

is finite and jointly continuous. A simple algebraic transformation shows that the difference between the functionals Q and H does not depend on the variable θ. Therefore, the EM algorithm may also be represented by the PPA recursion

$$\vartheta \leftarrow \operatorname*{argmax}_{\theta} H(\vartheta, \theta)$$

from which its properties flow.

For a parameter ϑ, the statements $\vartheta \in \operatorname{argmax}_\theta Q(\vartheta, \theta)$ and $\vartheta \in \operatorname{argmax}_\theta H(\vartheta, \theta)$ are equivalent. Such a parameter ϑ is called a *fixed point* (of H or of Q).

1.4 EM for mixtures. Among other things, Dempster, Laird, and Rubin (1977) applied the EM algorithm to estimating the parameters $\vartheta = (\mathbf{u}, \gamma) \in \Delta_{g-1} \times \Gamma$ of a mixture with density

$$f_{\mathbf{u},\gamma}(x) = \sum_j u_j f_{\gamma_j}(x), \tag{1.4}$$

mixing rates $\mathbf{u} = (u_1, \ldots, u_g) \in \Delta_{g-1}$, and population parameters $\gamma = (\gamma_1, \ldots, \gamma_g) \in \Gamma \subseteq \Gamma_1 \times \cdots \times \Gamma_g$, Γ_j being the parameter space of some family of distributions on E. The random variable $X \sim f_{\mathbf{u},\gamma}$ is a simple function

of a model that is more easily accessible. It is sufficient to take g random variables $Z^{(j)} \sim f_{\gamma_j}$, $j \in 1..g$, and a stochastically independent random label $L \sim \mathbf{u}$ with values in $1..g$, the hidden variable. By the formula of total probability, $Z^{(L)}$ is distributed according to the mixture (1.4), i.e., $Z^{(L)} \sim X$. The complete variable associated with the observation X is the joint variable $Y = (L, X)$. In the case of n observations $D = \{x_1, \ldots, x_n\}$ from independent random variables $X_1, \ldots, X_n$ one obtains with $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$, and $\boldsymbol{\ell} = (\ell_x)_{x \in D}$

$$f_X[\mathbf{x} \mid \mathbf{u}, \gamma] = \prod_{x \in D} \sum_j u_j f_{\gamma_j}(x) \quad \text{and}$$
$$f_Y[\boldsymbol{\ell}, \mathbf{x} \mid \mathbf{u}, \gamma] = \prod_{x \in D} u_{l_x} f_{\gamma_{\ell_x}}(x) = \prod_j \prod_{l_x = j} u_j f_{\gamma_j}(x).$$

A further simple computation shows that, with another pair of parameters $\theta = (\mathbf{v}, \eta) \in \Delta_{g-1} \times \Gamma$, the functional Q becomes

$$Q((\mathbf{u}, \gamma), (\mathbf{v}, \eta)) = \sum_j \Big(\sum_x w_j(x) \Big) \ln v_j + \sum_j \sum_x w_j(x) \ln f_{\eta_j}(x), \tag{1.5}$$

the weight $w_j(x)$ being the posterior probability of the observation x to come from component j w.r.t. the parameters $\mathbf{u}$ and γ. By Bayes' formula,

$$w_j(x) = P[L_i = j \mid X_i = x] = \frac{u_j f_{\gamma_j}(x)}{\sum_\ell u_\ell f_{\gamma_\ell}(x)}. \tag{1.6}$$

The weights sum up to 1 w.r.t. j, i.e., $\mathbf{w} = (w_j(x))_{x,j}$ is a stochastic matrix. The entropy inequality allows to optimize Eq. (1.5) w.r.t. $\mathbf{v}$. The maximum is

$$u_{\text{new},j} = \frac{1}{n} \sum_{x \in D} w_j(x) = w_j(D)/n, \tag{1.7}$$

where we have used the abbreviation $w_j(T) = \sum_{x \in T} w_j(x)$, $T \subseteq D$. The EM-step is thus split into an E-step and an M-step:

E-*step*: Compute $w_j(x)$ from the current parameters $\mathbf{u}$ and γ, cf. Eq. (1.6);

M-*step*: set $u_{\text{new},j} = w_j(D)/n$ and maximize $\sum_j \sum_x w_j(x) \ln f_{\eta_j}(x)$, cf. (1.5), w.r.t. $\eta = (\eta_1, \ldots, \eta_g)$ to obtain the parameter γ_{new}.

The EM algorithm, a suite of EM-steps, is iterative and alternating. If started from an M-step with a stochastic weight matrix $\mathbf{w}^{(0)}$, it proceeds as follows:

$$\mathbf{w}^{(0)} \to (\mathbf{u}^{(1)}, \gamma^{(1)}) \to \mathbf{w}^{(1)} \to (\mathbf{u}^{(2)}, \gamma^{(2)}) \to \mathbf{w}^{(2)} \to (\mathbf{u}^{(3)}, \gamma^{(3)}) \to \cdots .$$

The sequence of target values converges often to a local maximum even if the m.l.e. exists. The latter is always the case if components are homoscedastic normal.

Of course, the algorithm can only be applied to models f_{γ_j} that actually allow maximization in the M-step. The assumed continuity of the likelihood function implies that this is always the case if Γ is compact. Optimization is easier if $\Gamma = \prod_j \Gamma_j$ is a g-fold Cartesian product. One then maximizes the sum $\sum_x w_j(x) \ln f_{\eta_j}(x)$ separately for each j. If Γ is (locally compact and) non-compact and if the likelihood function $\eta \mapsto f_\eta(x)$ vanishes for all $x \in D$ as η approaches the Alexandrov point of Γ then the same is true for the sum $\sum_{x\in D} w_j(x) \ln f_{\eta_j}(x)$ and it is again plain that the maximum exists. In other cases one is interested in special local maxima, Dennis (1982). See also the discussion of the normal case below.

2 The EMT Algorithm

2.1 A mixture model with spurious outliers and its trimmed likelihood function. Given a natural number $r \le n$, we now establish an ML estimator for mixtures which yields reasonable results if the data set contains at least r regular observations. The remaining $\le n - r$ observations may (but do not have to) be gross, unpredictable outliers which obey no statistical law and which we call "spurious" (Gallegos and Ritter, 2005, 2009, 2010). We feel that the best way of handling this idea in a statistical (!) framework is by assuming that each outlier i comes from its own Bayesian population with random parameter ψ_i. The following is the main assumption on the spurious outliers.

($\mathrm{SV_o}$) A *spurious outlier* $X_i : \Omega \to E$, $i \in 1..n$, obeys a parametric model f_{ψ_i} with parameter $\psi_i \in \Psi_i$ such that the likelihood integrated w.r.t. some prior measure τ_i on Ψ_i satisfies

$$\int_{\Psi_i} f_{\psi_i}(x)\tau_i(\,\mathrm{d}\psi_i) = 1, \quad x \in E, \tag{2.1}$$

i.e., does not depend on x. In some sense, this equality says that the density of X_i is flat. Since each spurious outlier is obseved only once we cannot,

and do not wish to, estimate the parameters ψ_i and will later consider them nuisances. There are two important and sufficiently general situations where ($\mathrm{SV_o}$) holds.

(A) The sample space is Euclidean, $E = \mathbb{R}^d$, $\Psi_i = E$, the outliers obey a *location model*

$$X_i = U_i + \psi_i$$

with some (unknown) random noise $U_i : (\Omega, P) \to E$, and τ_i is Lebesgue measure on Ψ_i. Indeed, in this case, the conditional Lebesgue density is $f_{\psi_i}(x) = f_{U_i}(x - \psi_i)$ and, hence, $\int_{\Psi_i} f_{\psi_i}(x)\,\mathrm{d}\psi_i = 1$.

(B) The parameter sets Ψ_i consist of one point, each, and the distribution of X_i is the reference measure on E so that $f_{X_i} = 1$. This case includes the idea of irregular objects "uniformly distributed" on some domain.

Each *regular* observation X_i comes from a mixture of g populations represented by a density function of the form (1.4). All densities f_{γ_j} are strictly positive on E. Popular examples are normal models on Euclidean d-space with parameter space $\Gamma = \mathbb{R}^{gd} \times \mathcal{V}$ and $\mathcal{V} \subseteq \mathrm{PD}(d)^g$, where $\mathrm{PD}(d)$ stands for the cone of symmetric, positive-definite d by d matrices. The HDBT constraints (1.1) are characterized by $\mathcal{V} = \mathcal{V}_c := \{\mathbf{V} \in \mathrm{PD}(d)^g \mid V_j \succeq cV_\ell \text{ for all } j, \ell \in 1..g\}$.

The parameter set of our model is

$$\tbinom{D}{r} \times \Delta_{g-1} \times \Gamma \times \textstyle\prod_{i=1}^n \Psi_i,$$

the set $\binom{D}{r}$ of all r-element subsets of D standing for the possible $\binom{n}{r}$ subsets of regular objects. Of course, the parametrization of the mixture model is not identifiable in the strict sense, see however the discussion in McLachlan and Peel (2000), Ch. 1. The density function of the ith observation for the parameters $R \in \binom{D}{r}$, $\mathbf{u} = (u_1, \ldots, u_g)$, $\gamma = (\gamma_1, \ldots, \gamma_g)$, and $\psi = (\psi_1, \ldots, \psi_n)$ w.r.t. some fixed reference measure on E is

$$f_{X_i}[x \mid R, \mathbf{u}, \gamma, \psi] = \begin{cases} f_{\mathbf{u},\gamma}(x), & \text{see Eq. (1.4)}, \quad i \in R, \\ f_{\psi_i}(x), & \text{see Eq. (2.1)}, \quad i \notin R. \end{cases}$$

We assume that the sequence of observations $(X_i)_{i=1}^n$ is statistically independent but not necessarily i.i.d. unless there are no outliers, $n = r$. By the product formula, the joint likelihood for the data set $D = \{x_1, \ldots, x_n\}$ is

$$f_X[\mathbf{x} \mid R, \mathbf{u}, \gamma, \psi] = \prod_{i \in R} f_{\mathbf{u},\gamma}(x_i) \prod_{i \notin R} f_{\psi_i}(x_i).$$

Considering the parameters ψ_i of the outliers nuisances to be integrated out w.r.t. to the prior measures τ_i we obtain with Eq. (2.1) the *trimmed likelihood*

$$f[\mathbf{x}_R \mid \mathbf{u}, \gamma] = \prod_{i \in R} f_{\mathbf{u},\gamma}(x_i) = \prod_{i \in R} \Big(\sum_j u_j f_{\gamma_j}(x_i) \Big), \tag{2.2}$$

the ML criterion to be optimized w.r.t. the parameters $R \in \binom{D}{r}$, $\mathbf{u} \in \Delta_{g-1}$, and $\gamma \in \Gamma$. We have, thus, statistically justified the idea of trimming the likelihood (instead of the data) which goes back to Neykov and Neytchev (1990). It was placed into a broader context by Hadi and Luceño (1997), and applied to mixtures by Neykov et al. (2007). If, for each $R \in \binom{D}{r}$, the m.l.e.'s $\mathbf{u}^*$ and γ^* of $\mathbf{u}$ and γ w.r.t. R exist then, by the principle of dynamic optimization, the ML estimator of the parameters is the *Maximum Trimmed Likelihood Estimator*

$$\operatorname*{argmax}_R \max_{\mathbf{u},\gamma} \ln f[\mathbf{x}_R \mid \mathbf{u}, \gamma] = \operatorname*{argmax}_R \ln f[\mathbf{x}_R \mid \mathbf{u}^*, \gamma^*]. \tag{MTLE}$$

The numbers g and r are parameters of the model and of the algorithm below. We comment in Section 2.14 how to exploit them in order to estimate the numbers of components and of outliers in the data set. Both are unknown in most applications.

2.2 The EMT-step. Our next aim is generating (local) maxima of the trimmed likelihood function (2.2) w.r.t. $\mathbf{u}$ and γ. We extend the EM algorithm to contaminated mixtures proposing the following EMT-step, a suite of an E-, an M-, and a T-step. The E- and M-steps are carried out w.r.t. an r-element subset of D and lead to new parameters while the trimming step retains the r elements that best conform to the new parameters.

Input	:	An initial subset $R \subseteq D$ of r elements, mixing rates $(u_1, \dots, u_g)$, and initial population parameters $\gamma_1, \dots, \gamma_g$.
Output	:	A subset, mixing rates, and population parameters with improved criterion (2.2), cf. Proposition 2.3.
E-*step*	:	compute the weights $w_j(x) = \frac{u_j f_{\gamma_j}(x)}{\sum_\ell u_\ell f_{\gamma_\ell}(x)}$, $x \in R$, $j \in 1 \dots g$;

M-*step* : set $u_{\text{new},j} = w_j(R)/r$, $1 \le j \le g$, and maximize $\sum_j \sum_{x\in R} w_j(x) \ln f_{\gamma'_j}(x)$ w.r.t. $\gamma' \in \Gamma$ to obtain γ_{new}; (possibly without constraints, see the end of Section 2.11; in the case of a product $\Gamma = \prod_j \Gamma_j$, each sum $\sum_{x\in R} w_j(x) \ln f_{\gamma'_j}(x)$, $j \in 1..g$, is maximized separately)

T-*step* : define R_{new} to be the set of data points $x \in D$ with the r largest values of $f_{\mathbf{u}_{\text{new}},\gamma_{\text{new}}}(x) = \sum_j u_{\text{new},j} f_{\gamma_{\text{new},j}}(x)$.

The **EMT algorithm** is the iteration of EMT-steps. Like the EM algorithm it is iterative and alternating proceeding as follows

$$(R^{(0)}, \mathbf{w}^{(0)}) \overset{\text{M-step}}{\longrightarrow} (\mathbf{u}^{(1)}, \gamma^{(1)}) \overset{\text{T-step}}{\longrightarrow} (R^{(1)}, \mathbf{u}^{(1)}, \gamma^{(1)}) \overset{\text{E-step}}{\longrightarrow} (R^{(1)}, \mathbf{w}^{(1)}) \overset{\text{M-step}}{\longrightarrow} \cdots .$$

We may start it from the M-step with a randomly or expediently chosen r-element subset $R^{(0)}$ and a stochastic matrix $\mathbf{w}^{(0)}$ as initial quantities. An elegant procedure for uniform generation of a subset $R^{(0)}$ appears in Knuth (1981), p.136, ff. The rows of the initial weight matrix $w^{(0)}$ may be chosen uniformly from the $(g-1)$-dimensional unit simplex Δ_{g-1}. An efficient procedure is OSIMP, see Fishman (1996). An alternative is a set of randomly sampled unit vectors or the output obtained from some clustering algorithm. If components are sufficiently separated, the algorithm may also be started from the E-step with initial parameters $(\mathbf{u}^{(0)}, \gamma^{(0)})$. We assume that all initial mixing rates are strictly positive. They preserve this property during iteration. The iteration is successfully stopped as soon as the trimmed likelihood (2.2) is close to convergence or with a failure as there is indication that convergence will not take place. If the m.l.e. exists, see Proposition 2.13, then convergence always takes place.

The remarks after the statement of the EM algorithm for mixtures apply also to the EMT algorithm. As the likelihood function, its trimmed version has many local maxima. We next analyze the behavior of the algorithm in this regard. We say that $(R, \mathbf{u}, \gamma)$ is a *halting point* of the EMT-step if the ML criterion (2.2) remains unchanged after an EMT-step starting from it. The EMT algorithm starting from a halting point has this point as a possible output and the algorithm is stopped. A *limit point* $(R, \mathbf{u}, \gamma)$ is a point of convergence of the EMT algorithm starting from some initial parameters. A *critical* point of a differential function is a point where its gradient vanishes. There are relationships between fixed, halting, limit, critical, and optimal points. Moreover, the successive values of the target function are monotone as the following proposition shows.

2.3 PROPOSITION. *Let the statistical model be as described at the beginning of this section.*

(a) The EMT-step either improves the trimmed likelihood $f[\mathbf{x}_R \mid \mathbf{u}, \gamma]$ *or does not change it.*

(b) If $(R, \mathbf{u}, \gamma)$ *is optimal then so is* $(R_{new}, \mathbf{u}_{new}, \gamma_{new})$ *and* $(R, \mathbf{u}, \gamma)$ *is a halting point.*

PROOF. (a) The inequality $f[\mathbf{x}_R \mid \mathbf{u}, \gamma] \leq f[\mathbf{x}_R \mid \mathbf{u}_{\text{new}}, \gamma_{\text{new}}]$ is the well-known fact that the EM algorithm is monotone, here applied to the data set R, see Dempster et al. (1977), p. 8. Moreover,

$$\begin{aligned}\ln f[\mathbf{x}_R \mid \mathbf{u}_{\text{new}}, \gamma_{\text{new}}] &= \sum_{x \in R} \ln \sum_{\ell} u_{\text{new},\ell} f_{\gamma_{\text{new},\ell}}(x) \\ &\leq \sum_{x \in R_{\text{new}}} \ln \sum_{\ell} u_{\text{new},\ell} f_{\gamma_{\text{new},\ell}}(x) \\ &= \ln f[\mathbf{x}_{R_{\text{new}}} \mid \mathbf{u}_{\text{new}}, \gamma_{\text{new}}]\end{aligned}$$

by maximality of the data points in R_{new}.

(b) follows from the increasing property (a).

We need the H-functional (1.3) w.r.t. an r-element subset $R \subseteq D$,

$$H_R((\mathbf{u}, \gamma), (\mathbf{v}, \eta)) = \ln f[\mathbf{x}_R \mid \mathbf{v}, \eta] - D_R((\mathbf{u}, \gamma), (\mathbf{v}, \eta)),$$

where $D_R((\mathbf{u}, \gamma), (\mathbf{v}, \eta))$ is the Kullback-Leibler divergence of the complete model w.r.t. R conditional on $[\Phi = \mathbf{x}_R]$.

2.4 PROPOSITION

(a) If $(R^*, \mathbf{u}^*, \gamma^*)$ *is a halting point of the EMT-step then* $(\mathbf{u}^*, \gamma^*)$ *is a fixed point w.r.t.* R^* *(see Section 1.3.)*

(b) If $(\mathbf{u}^*, \gamma^*)$ *is the unique fixed point w.r.t.* R^* *then* $(R^*, \mathbf{u}^*, \gamma^*)$ *is a halting point*

PROOF. Let us put $\vartheta^* = (\mathbf{u}^*, \gamma^*)$ and $\vartheta_{\text{new}} = (\mathbf{u}_{\text{new}}, \gamma_{\text{new}})$, the output of the EM-step starting from (R^*, ϑ^*).

(a) If (R^*, ϑ^*) is a halting point of the EMT-step then

$$\begin{aligned}H_{R^*}(\vartheta^*, \vartheta_{\text{new}}) = \ln f[\mathbf{x}_{R^*} \mid \vartheta_{\text{new}}] - D_{R^*}(\vartheta^*, \vartheta_{\text{new}}) \leq \ln f[\mathbf{x}_{R^*} \mid \vartheta_{\text{new}}] \\ \leq \ln f[\mathbf{x}_{R_{\text{new}}} \mid \vartheta_{\text{new}}] = \ln f[\mathbf{x}_{R^*} \mid \vartheta^*] = H_{R^*}(\vartheta^*, \vartheta^*),\end{aligned}$$

i.e., ϑ^* is a fixed point w.r.t. R^*.

(b) By assumption, we find $\vartheta_{\text{new}} = \vartheta^*$ after the EM-step starting from (R^*, ϑ^*) and the claim follows from the definition of the T-step.

2.5 Proposition. *Limit and halting points are the same.*

Proof. Let the sequence $(R_t, \mathbf{u}_t, \gamma_t) = (R_t, \vartheta_t)$ converge to $(R^*, \mathbf{u}^*, \gamma^*) = (R^*, \vartheta^*)$. Since we have $R_t = R^*$ for eventually all t, it is sufficient to consider R_t fixed. Abbreviate $\theta = (\mathbf{v}, \eta)$ and $\vartheta_t = (\mathbf{u}_t, \gamma_t)$. From $H_{R_t}(\vartheta_t, \theta) \leq H_{R_t}(\vartheta_t, \vartheta_{t+1})$ for all θ we infer

$$H_{R_t}(\vartheta^*, \theta) = \lim_{t\to\infty} H_{R_t}(\vartheta_t, \theta) \leq \lim_{t\to\infty} H_{R_t}(\vartheta_t, \vartheta_{t+1}) = H_{R_t}(\vartheta^*, \vartheta^*).$$

This shows that limit points are halting points.

Conversely, if (R^*, ϑ^*) is a halting point then ϑ^* is a fixed point w.r.t. R^*. We may, therefore, choose $\vartheta_{\text{new}} = \vartheta^*$ in the EM-step w.r.t. R^* and $R_{\text{new}} = R^*$ in the subsequent T-step. This proves that (R^*, ϑ^*) is a limit point.

The following proposition investigates optimality of the mixing rates. It applies in particular to fixed points. A *face* of the simplex Δ_{g-1} is the convex hull of a non-empty set of unit vectors in $\mathbb{R}^g$. A subset $F \subseteq \Delta_{g-1}$ is a face if it is the non-empty intersection of Δ_{g-1} with some hyperplane H of $\mathbb{R}^g$ such that $\Delta_{g-1} \setminus H$ is convex or again if it is the set of points in Δ_{g-1} where some linear form on $\mathbb{R}^g$ assumes its minimum. To each non-empty subset $M \subseteq \Delta_{g-1}$ there is a smallest face that contains it, the face *generated* by M. The face generated by a subset that contains an interior point of the simplex is the whole simplex. The face generated by one point contains this point in its interior. (This is also true if the point is extremal.)

2.6 Proposition. *Let $R \subseteq D$, $|R| = r$, let $\gamma \in \Gamma$, and let $\widetilde{\mathbf{u}} \in \Delta_{g-1}$.*

(a) The following statements are equivalent.

(i) The EM-step with input $(R, \widetilde{\mathbf{u}}, \gamma)$ retrieves $\widetilde{\mathbf{u}}$;
(ii) the vector $\widetilde{\mathbf{u}}$ is an extreme point of the simplex or a critical point of the function $\mathbf{u} \mapsto f[\mathbf{x}_R \mid \mathbf{u}, \gamma]$ restricted to the face generated by it;
(iii) the function $\mathbf{u} \mapsto f[\mathbf{x}_R \mid \mathbf{u}, \gamma]$ restricted to the face generated by $\widetilde{\mathbf{u}}$ is maximal at $\widetilde{\mathbf{u}}$.

(b) Let the equivalent conditions (i)–(iii) be satisfied. If $\widetilde{\mathbf{u}}$ is an interior point of the simplex then it is a maximum of the function $\mathbf{u} \mapsto f[\mathbf{x}_R \mid \mathbf{u}, \gamma]$.

If, moreover, the g vectors

$$\big(f_{\gamma_1}(x)\big)_{x\in R},\dots,\big(f_{\gamma_g}(x)\big)_{x\in R}$$

are affine independent then it is the only maximum.

(c) Any fixed point $(\widetilde{u},\gamma)$ w.r.t. R satisfies the equivalent conditions in (a).

PROOF. Part (a) is immediate if $\widetilde{\mathbf{u}}$ is extremal. Otherwise, assume without loss of generality $\widetilde{u}_g \neq 0$, let F be the face generated by $\widetilde{\mathbf{u}}$, and let $\widetilde{\mathbf{w}}$ be returned from the E-step with input $(R,\widetilde{\mathbf{u}},\gamma)$. Since $f_{\gamma_j}(x) > 0$ by general assumption, the partial derivative of the function

$$F \to \mathbb{R}, \quad \mathbf{u} \mapsto \ln f[\mathbf{x}_R \mid \mathbf{u},\gamma] = \sum_{x\in R} \ln\Big[\sum_{j\neq g} u_j f_{\gamma_j}(x) + \big(1-\sum_{j\neq g} u_j\big) f_{\gamma_g}(x)\Big],$$

w.r.t. $j \neq g$ such that $\widetilde{u}_j \neq 0$ shows that $\widetilde{\mathbf{u}}$ is critical if and only if $\widetilde{w}_j(R) = \text{const}\cdot\widetilde{u}_j$ for all j. The equivalence of (i) and (ii) now follows from $u_{\text{new},j} = \widetilde{w}_j(R)/r$.

In view of the implication (ii)⇒(iii) note that

$$\ln f[\mathbf{x}_R \mid \theta] = \sum_{x\in R} \ln \sum_j u_j f_{\gamma_j}(x)$$

is of the form $\sum_{x\in R}\ln(A\mathbf{u})_x$ with $A_{x,j} = f_{\gamma_j}(x)$. Assertion (iii) now follows from concavity A.1 (a) of this function restricted to the mixing parameters and from (ii). The sense (iii)⇒(ii) is plain.

(b) If $\widetilde{\mathbf{u}}$ is an interior point then the face generated by it is the whole simplex and the first claim follows from (a). The second claim follows from Lemma A.1(b).

(c) Let $\widetilde{\mathbf{w}}$ be defined by $(\widetilde{u},\gamma)$ and let $(u_{\text{new}},\gamma_{\text{new}})$ be the parameters after an EM-step starting from $(\widetilde{u},\gamma)$. Since both pairs maximize $Q_R((\widetilde{u},\gamma),\cdot)$ and since $\widetilde{w}_j(R) = r u_{\text{new},j}$, Eq. (1.5) shows

$$r\sum_j u_{\text{new},j}\ln u_{\text{new},j} \;+\; \sum_j\sum_{x\in R}\widetilde{w}_j(x)\ln f_{\gamma_{\text{new},j}}(x) \;=\; r\sum_j u_{\text{new},j}\ln\widetilde{u}_j \;+ \\ +\sum_j\sum_{x\in R}\widetilde{w}_j(x)\ln f_{\gamma_j}(x).$$

The likelihood grows with the EM-step, hence

$$\sum_j u_{\text{new},j} \ln u_{\text{new},j} \le \sum_j u_{\text{new},j} \ln \widetilde{u}_j$$

and the entropy inequality shows $u_{\text{new}} = \widetilde{u}$, i.e., (a)(i).

Our next proposition discusses the population parameters and the set of retained observations of a halting point.

2.7 Proposition. *Let $(R^*, \mathbf{u}^*, \gamma^*)$ be a halting point of the EMT-step.*

(a) Assume that Γ is an open subset of some Euclidean space, that $f_\gamma(x)$ is differentiable w.r.t. γ for all x, and that the (conditional) Kullback-Leibler divergence $D_R(\vartheta, \theta) = D(\nu_\vartheta[\cdot \mid \Phi = \mathbf{x}_R], \nu_\theta[\cdot \mid \Phi = \mathbf{x}_R])$ is differentiable w.r.t. θ at each point of the diagonal $\theta = \vartheta$ for all R. Then γ^* is a critical point of the function $\gamma \mapsto f[\mathbf{x}_{R^*} \mid \mathbf{u}^*, \gamma]$.

(b) R^* is consistent with the output $(\mathbf{u}_{new}, \gamma_{new})$ of the EM-step starting from $(R^*, \mathbf{u}^*, \gamma^*)$.

Proof. Let ϑ^* and ϑ_{new} be as defined at the beginning of the proof of Proposition 2.4. From that proposition, we know already that ϑ^* is a fixed point w.r.t. R^*. The point $\mathbf{u}^*$ is interior to the face F generated by it. Therefore, ϑ^* lies in the interior of $F \times \Gamma$. The fixed point ϑ^* maximizes the H-functional $\theta \mapsto H_{R^*}(\vartheta^*, \theta)$ and minimizes the (conditional) Kullback-Leibler divergence $\theta \mapsto D_{R^*}(\vartheta^*, \theta)$ since it vanishes there. By interiority of ϑ^*, the gradients of both functions restricted to $F \times \Gamma$ vanish at this point. Thus, the gradient of the restriction to $F \times \Gamma$ of the observed log-likelihood

$$\theta \mapsto \ln f[\mathbf{x}_{R^*} \mid \theta] = H_{R^*}(\vartheta^*, \theta) + D_{R^*}(\vartheta^*, \theta),$$

too, vanishes at ϑ^*, this representation being valid at least near $\theta = \vartheta^*$. This completes Claim (a).

Claim (b) follows directly from the estimate $f[\mathbf{x}_{R_{\text{new}}} \mid \vartheta_{\text{new}}] = f[\mathbf{x}_{R^*} \mid \vartheta^*] \le f[\mathbf{x}_{R^*} \mid \vartheta_{\text{new}}]$.

The EMT algorithm often converges. The following corollary, a consequence of Propositions 2.5 – 2.7, summarizes properties of the limit.

2.8 Corollary. *Let the assumptions of Proposition 2.7 (a) hold and assume that the sequence of successive outputs of the EMT algorithm converges with limit $(R^*, \mathbf{u}^*, \gamma^*)$. Then Propositions 2.6(b), (c) and 2.7(a), (b) apply to $(R^*, \mathbf{u}^*, \gamma^*)$.*

log-likelihood	**u**	**m**	v
–8.39821	0.5, 0.25, 0.25	–3.00000, 2.07741, 3.92258	0.57442
–8.44833	0.5, $0.5-\alpha$, α	–2.99999, 2.99999, 2.99999	1.00007
–10.2809	$1-\alpha-\beta$, α, β	0, 0, 0	10

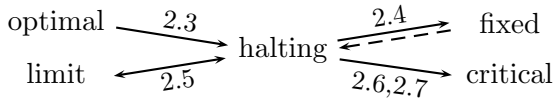
Table 1: Halting points for the data set $-4,-2,2,4$ and the homoscedastic normal model with three components.

2.9 REMARKS. (a) In general, the trimmed likelihood function has several or many local maxima. Here is a simple, one-dimensional, normal example with an infinite number of halting points. The data set consists of the four points $-4,-2,2,4$. It has two obvious clusters. Running the EM algorithm (no outliers) with the homoscedastic model and $g=3$ we find the halting points shown in Table 1. The first is the global maximum. It essentially uses the two negative observations for one component and each of the two positive ones for the remaining components. The second line represents a continuum of halting points corresponding to the natural solution with two components and means close to -3 and 3. One of the components is split in two very similar parts. These halting points lie in a region that is very flat in two directions, two eigenvalues of the Hessian being close to zero. The last line describes a two-dimensional manifold of halting points with equal log-likelihoods. The positive semi-definite Hessian is the same at each point and has four vanishing eigenvalues. In the first line, the mixing rates are unique by Proposition 2.6(b). Each of the first two lines induces a number of symmetrical, equivalent solutions.

(b) Modifications to the M-step are possible. It is not necessary to go to the maximum in the M-step. Each improvement in the M-step or in the T-step improves the observed likelihood.

(c) If Γ is not open as required in Proposition 2.7(a) and if γ^* is at the boundary of Γ then it is only true that *directional* derivatives of $\gamma \mapsto f[\mathbf{x}_{R^*} \mid \mathbf{u}^*, \gamma]$ at γ^* must be ≤ 0 in all interior directions.

(d) There are the following relationships between the various interesting parameters.



2.10 Normal components. In the normal case, $E = \mathbb{R}^d$ is d-dimensional Euclidean space so that $D \subseteq \mathbb{R}^{nd}$. The symbol $\mathrm{N}_{m,V}$ designates the d-variate normal distribution with mean m and covariance matrix V and also its Lebesgue density. The normal model is characterized by $\gamma = (\mathbf{m}, \mathbf{V})$, $\mathbf{m} = (m_1, \ldots, m_g)$, $\mathbf{V} = (V_1, \ldots, V_g)$, and the trimmed likelihood (2.2) becomes the *trimmed normal-mixture likelihood*

$$f[\mathbf{x}_R \mid \mathbf{u}, \mathbf{m}, \mathbf{V}] = \prod_{x \in R} \sum_{j=1}^{g} u_j \mathrm{N}_{m_j, V_j}(x). \tag{2.3}$$

In the homoscedastic case, $V_1 = \cdots = V_g = V$, and the HDBT-constrained heteroscedastic case is characterized by $\mathbf{V} \in \mathcal{V}_c$ with $c < 1$.

Besides the *full* model, two submodels specified by shape and orientation of the populations are customary in each of the two cases, *diagonal* covariance matrices $V_j = \mathrm{diag}(v_{j,1}, \ldots, v_{j,d})$ and *spherical* covariance matrices $V_j = v_j I_d$.

We make the *standard assumption* that the data points D are in *general position*. This concept has different meanings for the three normal sub-populations mentioned above. In the spherical case it means pairwise difference of all data points, in the diagonal case pairwise difference of all d entries of any two data points, and in the full case affine independence of any $k \leq d + 1$ points in D. We also assume throughout $r \geq gd + 1$ in the "full" case and $r \geq g + 1$ in the "diagonal" and "spherical" cases.

We have to introduce some notation. Given a subset $T \subseteq D$, the symbols $\overline{x}_T$, W_T, and S_T designate the *sample mean vector*, the *SSP matrix*, and the *scatter matrix* of T, respectively. We also need *weighted* analogs of these statistics w.r.t. a weight vector $w = (w(x))_{x \in T}$ of real numbers $w(x) \geq 0$ (in most cases $w = w_j$, the j'th column in a stochastic weight matrix $\mathbf{w} = (w_j(x))_{x \in T, j \in 1..g}$, $\sum_j w_j(x) = 1$ for all $x \in T$). Writing $w(T) = \sum_{x \in T} w(x)$, cf. the definition after Eq. (1.7), we define them, respectively, as

$$\begin{aligned}
\overline{x}_T(w) &= \frac{1}{w(T)} \sum_{x \in T} w(x)x \quad (= 0, \text{ if } w(T) = 0), \\
W_T(w) &= \sum_{x \in T} w(x)(x - \overline{x}_T(w))(x - \overline{x}_T(w))^{\mathrm{T}}, \text{ and} \\
S_T(w) &= \frac{1}{w(T)} W_T(w) \quad (= I_d, \text{ if } w(T) = 0).
\end{aligned}$$

The *pooled weighted SSP matrix* and the *pooled weighted scatter matrix* w.r.t. a stochastic weight matrix $\mathbf{w}$ are, respectively,

$$W_T(\mathbf{w}) = \sum_j W_T(w_j) \text{ and } S_T(\mathbf{w}) = \frac{1}{|T|} W_T(\mathbf{w}).$$

If weights are binary then $\mathbf{w}$ defines in a natural way a partition of T and $W_T(\mathbf{w})$ and $S_T(\mathbf{w})$ reduce to the ordinary pooled quantities.

We will repeatedly use the MAP partition $\{T_1, \dots, T_g\}$ of some subset $T \subseteq D$ w.r.t. some stochastic weight matrix $\mathbf{w} = (w_j(x))_{x \in T, j \in 1..g}$: $x \in T_\ell \Leftrightarrow \ell = \operatorname{argmax}_j w_j(x)$. The obvious estimate

$$w_j(x) \geq 1/g \quad \text{for all } x \in T_j \tag{2.4}$$

and Steiner's formula A.3 imply for all j

$$W_T(w_j) \succeq \sum_{x \in T_j} w_j(x)(x - \overline{x}_T(w_j))(x - \overline{x}_T(w_j))^{\mathrm{T}} \succeq \frac{1}{g} W_{T_j}. \tag{2.5}$$

2.11 The M-step in the constrained normal cases. As with all distributional models, the optimal mixing rates u_j in the M-step can be directly computed by the analytical expression (1.7) applied with R instead of D. A routine argument using Steiner's identity A.3 shows that, with the notation above, the estimates of the location parameters in the M-step with input $\mathbf{w}$ and R are in all cases

$$m_j = \overline{x}_R(w_j). \tag{2.6}$$

In the *homoscedastic*, normal case, $c = 1$, the estimate of the common scale parameter V, too, can be represented in closed form in the M-step. Normal estimation theory shows that the minimizer of the scale parameter given R and $\mathbf{w}$ is here

$$V = S_R(\mathbf{w}) \text{ (full)}, \quad v_k = V(k,k) \text{ (diagonal)}, \quad v = \frac{1}{d} \sum_k v_k \text{ (spherical)}, \tag{2.7}$$

$k \in 1 \dots d$. Furthermore, the trimmed likelihood (2.3) of a fixed point $(\mathbf{u}^*, \mathbf{m}^*, V^*)$ w.r.t. R of the homoscedastic EM algorithm assumes the form

$$\begin{aligned} \ln f[\mathbf{x}_R \mid \mathbf{u}^*, \mathbf{m}^*, V^*] \;=\; & c_{d,r} + r \sum_{j=1}^{g} u_j^* \ln u_j^* - \sum_{j=1}^{g} \sum_{x \in R} w_j(x) \ln w_j(x) \\ & - \frac{r}{2} \ln \det V^*, \end{aligned} \tag{2.8}$$

with $c_{d,r} = -\frac{dr}{2}(1 + \ln 2\pi)$ and $w_\ell(x) = \frac{u_\ell^* \mathrm{N}_{m_\ell^*, V^*}(x)}{\sum_{j=1}^g u_j^* \mathrm{N}_{m_j^*, V^*}(x)}$. Just insert $\mathbf{m}^*$ and V^* from Eqs. (2.6) and (2.7), respectively, into Lemma A.2, note that $w_\ell(R) = ru_j^*$, see Proposition 2.6(c), and apply standard matrix analysis. The equality applies in particular to any halting point $(R, \mathbf{u}^*, \mathbf{m}^*, V^*)$ of the EMT algorithm, see Proposition 2.4(a).

The HDBT-*constrained heteroscedastic* case is less obvious. The routine argument above shows that the parameter $\mathbf{V} \in \mathcal{V}_c$ returned from the M-step with input $\mathbf{w}$ is the solution to the minimization problem

$$\operatorname*{argmin}_{\mathbf{V} \in \mathcal{V}_c} \sum_j w_j(R) \big(\ln \det V_j + \operatorname{tr} S_R(w_j) V_j^{-1} \big). \tag{2.9}$$

However, for $c < 1$, we do not know a representation of the V_j's as in the homoscedastic formulae (2.7). In general, it depends on the unknown constant c which must be estimated. We will deal with this question in Section 2.14 using free, i.e., *unconstrained, local minima*. The estimates of the scale parameters in the *free heteroscedastic* case are known to be

$$V_j = S_R(w_j) \ \text{(full)}, \quad v_{j,k} = V_j(k,k) \ \text{(diagonal)}, \quad v_j = \frac{1}{d} \sum_k v_{j,k} \ \text{(spherical)}, \tag{2.10}$$

$j \in 1 \ldots g$, $k \in 1 \ldots d$. The equivalent to Eq. (2.8) for a *free* fixed point $(\mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*)$ w.r.t. R is

$$\begin{aligned} \ln f[\mathbf{x}_R \mid \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*] \ = \ & c_{d,r} + r \sum_{j=1}^g u_j^* \ln u_j^* - \sum_{j=1}^g \sum_{x \in R} w_j(x) \ln w_j(x) \\ & - \frac{r}{2} \sum_j u_j^* \ln \det V_j^* \end{aligned}$$

where, $\mathbf{w}$ is the matrix of posterior probabilities defined by Eq. (1.6) for $\mathbf{u}^*$ and $\gamma = (\mathbf{m}^*, \mathbf{V}^*)$.

It depends on the situation, whether constraints between mixture components such as the HDBT constraints (1.1) should be used in the M-step. If constraints are *unknown* (e.g., an unknown constant c in (1.1)) then we recommend free parameter estimation in the M-step resorting to Section 2.14 in order to estimate the HDBT ratio together with the mixture. If constraints are *known* and if there is a computationally efficient way of computing the constrained maximum in the M-step then use them. An example is the homoscedastic normal model where the estimate of the covariance matrix is the pooled weighted SSP matrix given by Eq. (2.7). If they are known

but there is no computationally efficient method for constrained parameter estimation then we recommend to disregard the known constraints in the M-step and to use from all replications the largest local maximum that satisfies the constraints. This avoids universal optimization paradigms such as gradient descent, the Metropolis algorithm, etc. which would lead to very slow overall algorithms.

The following lemma is crucial for our theoretical analyses. It makes up for the missing representation of the minimizer of (2.9).

2.12 LEMMA. *Let R be some data set in $\mathbb{R}^d$ of cardinality r and let $\mathbf{V} \in \mathcal{V}_c$.*

(a) With $w_\ell(x) = \frac{u_\ell N_{m_\ell,V_\ell}(x)}{\sum_{j=1}^g u_j N_{m_j,V_j}(x)}$, $x \in R$, $1 \le \ell \le g$, we have for all j

$$2 \ln f[\mathbf{x}_R \mid \mathbf{u}, \mathbf{m}, \mathbf{V}] \le -r \ln \det 2\pi c V_j - c \operatorname{tr} W_R(\mathbf{w}) V_j^{-1}.$$

(b) If $|R \cap D| \ge gd + 1$ and if D is in general position, there is a constant $K > 0$ that depends only on D such that, for all j,

$$2 \ln f[\mathbf{x}_R \mid \mathbf{u}, \mathbf{m}, \mathbf{V}] \le -r \ln \det 2\pi c V_j - cK \operatorname{tr} V_j^{-1}. \tag{2.11}$$

PROOF. (a) By Lemma A.2, the HDBT constraints, and Steiner's formula A.3 we have

$$\begin{aligned}
2 \ln f[\mathbf{x}_R \mid \mathbf{u}, \mathbf{m}, \mathbf{V}] &\le 2 \sum_\ell \sum_{x \in R} w_\ell(x) \ln \mathrm{N}_{m_\ell,V_\ell}(x) \\
&= -\sum_\ell \sum_{x \in R} w_\ell(x) \big(\ln \det 2\pi V_\ell + (x - m_\ell)^{\mathrm{T}} V_\ell^{-1} (x - m_\ell) \big) \\
&\le -\sum_\ell \sum_{x \in R} w_\ell(x) \big(\ln \det 2\pi c V_j + c(x - m_\ell)^{\mathrm{T}} V_j^{-1} (x - m_\ell) \big) \\
&= -r \ln \det 2\pi c V_j - c \operatorname{tr} \sum_\ell \sum_{x \in R} w_\ell(x) (x - m_\ell)(x - m_\ell)^{\mathrm{T}} V_j^{-1} \\
&\le -r \ln \det 2\pi c V_j - c \operatorname{tr} W_R(\mathbf{w}) V_j^{-1}.
\end{aligned}$$

(b) Let $\{R_1, \dots, R_g\}$ be the MAP partition of R w.r.t. $\mathbf{w}$ in (a). By assumption on $|R \cap D|$ there is a subset R_ℓ that contains at least $d + 1$ elements of D. By general position, W_{R_ℓ} is regular and, by Eq. (2.5), $W_R(\mathbf{w}) \succeq W_{R,\ell}(\mathbf{w}) \succeq \frac{1}{g} W_{R_\ell} \succeq K I_d$ with some constant $K > 0$ that depends only on D. The claim therefore follows from (a).

We next show that an m.l.e. for the trimmed, constrained normal model exists.

2.13 Proposition. *If D and r satisfy the standard assumptions made in Section 2.10 then the HDBT-constrained trimmed normal-mixture likelihood* (2.3) *possesses a maximum w.r.t. R, $\mathbf{u}$, $\mathbf{m}$, and $\mathbf{V} \in \mathcal{V}_c$.*

Proof. We prove the existence of a maximum for every subset $R \subseteq D$. Let us first show that it is sufficient to consider mean values m_j in the convex hull $\operatorname{conv} R$. Let $V_j \in \mathrm{PD}(d)$, $m \in \mathbb{R}^d \setminus \operatorname{conv} R$, and let m' be the Euclidean projection of $V_j^{-1/2} m$ on the compact, convex set $V_j^{-1/2} \operatorname{conv} R$. Then $V_j^{1/2} m' \in \operatorname{conv} R$ and we have for $x \in R$

$$\begin{aligned}(x - V_j^{1/2} m')^{\mathrm{T}} V_j^{-1} (x - V_j^{1/2} m') &= \|V_j^{-1/2} x - m'\|^2 \\ &< \|V_j^{-1/2}(x - m)\|^2 \\ &= (x - m)^{\mathrm{T}} V_j^{-1} (x - m).\end{aligned}$$

In view of the trimmed normal likelihood (2.3), this proves the first claim.

Now fix R and apply Lemma 2.12(b). The well-known behavior of the upper bound (2.11) as a function of V_j and the HDBT constraints imply that the mixture likelihood vanishes at the boundary of $\mathrm{PD}(d)^g$ uniformly for all $\mathbf{u}$ and $\mathbf{m}$. The proposition follows since there are only finitely-many sets R.

We call the maximizing parameters R^*, $\mathbf{u}^*$, $\mathbf{m}^*$, $\mathbf{V}^* \in \mathcal{V}_c$ for any constant c the *Maximum Trimmed Normal Likelihood Estimates*, (normal MTLE's). Note that the expression "normal MTLE" includes in the present paper the HDBT constraints.

2.14 The favorite solution: estimation of the HDBT ratio and of the numbers of outliers and components. We next use the parameters r ang g appearing in the trimmed likelihood (2.2) and in the EMT-step 2.2 together with the HDBT ratios (1.2) of limit points of the EMT algorithm in order to estimate the unknown numbers of outliers and components, an HDBT constraint, and the mixture. Let us first determine a constant c together with a solution keeping r and g fixed. Hathaway (1985) states that the *HDBT-constrained* m.l.e. consistently estimates the parameters of the mixture if their HDBT ratio exceeds the constant c. Unfortunately, application of his theorem requires this constant. What is more, constrained optima are not easily computed although any free halting point (2.10) that satisfies the HDBT constraints is a constrained halting point. We, therefore, rather

recur to the theorems of Peters and Walker (1978) and Kiefer (1978). Their results say that a strongly consistent *free* local maximum of the trimmed likelihood (2.2) exists and this is what we seek. Often, it is not the largest one, see Section 2.14. Selecting it from all local maxima needs a further assumption.

As already stated in the introduction, we expect the sizes and shapes of the scale parameters not to deviate excessively from each other. In most normal cases it is not only good *fit*, i.e., a large trimmed likelihood, but also sufficient *balance*, i.e., a large value of the HDBT ratio (1.2) that characterizes a reasonable solution. We wish to find the best-fitting among the well-balanced solutions. In order to solve this *biobjective optimization* problem, we propose the following graphical procedure which was introduced by Gallegos and Ritter (2009) in the related context of statistical clustering. We replicate the algorithm starting from many different randomly or expediently generated parameters in order to find a good number of (unconstrained) local maxima. One of them is the favorite solution. In order to select it, we create a negative double-logarithmic plot of the HDBT ratios of their covariance matrices vs. their likelihoods. The left and lower parts of the convex hull of all points often form a knee. The extreme points close to its bend are candidates for the favorite solution. Sometimes this solution is even unique. The plot provides also some guidance about the number of replications needed: run the EMT algorithm until the convex hull has stabilized.

The favorite solution can be detected with the aid of a criterion. Since it belongs to an extreme point in the left lower part of the convex hull of the plot, it is found by touching the point set in the plot from below with a supporting line of (non-unique) negative slope. A suitable slope depends on the data set. Let $(-\sigma, 1)$, $\sigma > 0$, be a vector in the direction of the support line pointing northwest. The graphical method is equivalent to maximizing the criterion

$$\log f[\mathbf{x}_{R^*} \mid \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*] + \sigma \log r_{\mathrm{HDBT}}(\mathbf{V}^*) \tag{2.12}$$

w.r.t. all limit points $(R^*, \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*)$ of the EMT algorithm. In the homoscedastic case, the theoretical HDBT ratio is 1 and the second term is omitted in criterion (2.12).

By way of illustration, Figure 1 shows a data set of 20 points sampled from the normal mixture $\frac{1}{2}\mathrm{N}_{-2e_1,I_2} + \frac{1}{2}\mathrm{N}_{2e_1,I_2}$, $e_1 = (1, 0)$. The framed three-point constellation defines one component of the largest local maximum of the (unconstrained) likelihood function for $r = n$, $g = 2$; see also the caption.

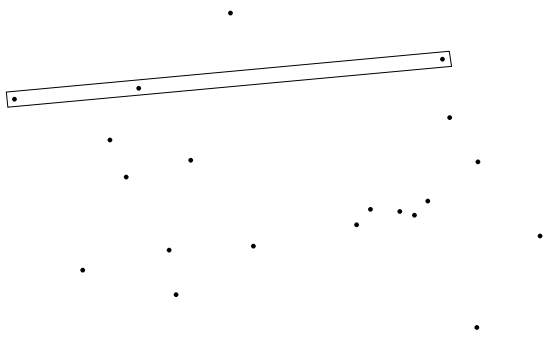


Figure 1: A synthetic data set of 20 points randomly sampled from the normal mixture $\frac{1}{2}\mathrm{N}_{-2e_1,I_2} + \frac{1}{2}\mathrm{N}_{2e_1,I_2}$. There are no outliers. The framed point triple generates one of the two components of the solution with the smallest local minimum of the negative log-likelihood. It is 60.93 whereas that of the favorite local minimum close to the sampling distribution is 65.93. The HDBT ratio (1.2) of the former is as small as 1/66,562 whereas that of the latter is 1/1.71.

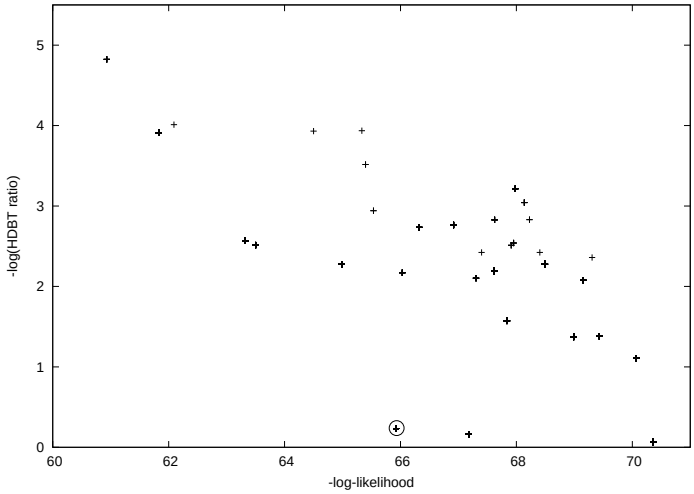


Figure 2: Synthetic data set of Fig. 1: Negative double-logarithmic HDBT-ratio-likelihood plot for a number of local minima with two components and no discarded elements. The favorite solution close to the sampling distribution is encircled.

This solution, however, is undesirable. Figure 2 shows the negative double-logarithmic HDBT-ratio-likelihood plot. The favorite solution close to the sampling population is the best-fitting among the well-balanced ones. Its estimated parameters are $u_1 = 0.501,\ u_2 = 0.499, m_1 = (2.039, -0.790),\ m_2 =$

$(-1.743, -0.493)$, and $V_1 = \begin{pmatrix} 0.476 & \\ -0.034 & 0.817 \end{pmatrix}$, $V_2 = \begin{pmatrix} 0.808 & \\ -0.082 & 1.310 \end{pmatrix}$, its negative log-likelihood is 65.93 and its HDBT ratio 1/1.71. By contrast, the estimated parameters of the best-fitting but spurious solution are $u_1 = 0.851$, $u_2 = 0.149$, $m_1 = (0.362, -0.880)$, $m_2 = (-1.096, 0.721)$, and $V_1 = \begin{pmatrix} 3.649 & \\ -0.143 & 0.885 \end{pmatrix}$, $V_2 = \begin{pmatrix} 5.666 & \\ 0.532 & 0.050 \end{pmatrix}$, its negative log-likelihood being 60.93 and its HDBT ratio as low as 1/66,562. The smaller component is generated by the framed almost collinear point triple in Figure 1. As Figure 2 shows, there are nine other solutions that fit better than the favorite one, but with much inferior balance. They are generated by other almost collinear constellations in the data set.

Up to now, the parameters r and g were fixed. They can be used for model selection, i.e., for estimating the numbers of outliers and of components in the data set. A first option is the application of validation techniques, e.g. based on goodness of fit of the mixture estimated for the parameter pair; see the literature cited in Redner and Walker (1984), p.202, and McLachlan and Peel (2000), Sect. 3.5. The parameter g allows us to estimate the number of components, e.g., by means of Gallegos and Ritter's (2010) *corrected BIC*. It is based on the model selection criterion BIC, see Kéribin (2000). A way of estimating both numbers simultaneously was recently proposed by Neykov et al. (2007), the *trimmed BIC*. The method uses a table of BIC values for solutions indexed by g and r returning the parameter values where the minima w.r.t. g stabilize. In Table 2, it is applied to optimal solutions w.r.t. criterion (2.12). No matter what criterion is used, the method requires solutions close to the optimum since otherwise the stabilizing point may not be clear. In real-world applications this generally poses a serious problem.

3 Universal Breakdown Points

3.1 Breakdown points. 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 $n - r$ arbitrary elements 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 needs a sequence of estimators

since the addition increases the number of objects. The latter needs only the estimator for the given number of objects. We deal with replacements.

Let $\delta : \mathcal{A} \to \Theta$ be a statistic on its natural domain of definition $\mathcal{A} \subseteq E^n$. In the context of a normal m.l.e., $\mathcal{A}$ is defined by the standard assumptions. Given a natural number $m \leq n$, we say that $M \in \mathcal{A}$ is an m-modification of $D \in \mathcal{A}$ if it arises from D by modifying m entries in an (admissible but otherwise) arbitrary way. An estimator δ "breaks down with the (n-element) data set 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.[1] The *individual* breakdown point for D is the number

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

If there is such an m then it is the minimal fraction of replacements in D that may cause δ to break down, otherwise ∞. 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.

Donoho and Huber's breakdown point is the *universal* breakdown point

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

This concept depends solely on the estimator. We cannot always expect the universal breakdown point to be large. An example is provided in Sect. 4. We therefore introduced the *restricted* breakdown point (Gallegos and Ritter, 2005) of δ w.r.t. some subclass $\mathcal{K}$ of admissible data sets. It is

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

The restricted breakdown point depends on δ and on 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 $\mathcal{K}$. The restricted breakdown value may be seen as a relaxed version of the universal since we have the estimates

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

[1] Of course, no breakdown is possible if Θ is compact.

Plainly, $\beta(\delta) = \beta(\delta, \mathcal{A})$. The asymptotic breakdown point of δ is $\liminf_{n\to\infty} \beta(\delta)$.

In order to compute breakdown points in special situations one has to know the system of compact subsets of the target space of the statistic. We deal here with breakdown points of the means and of the covariance matrices. The relatively compact subsets of the parameter space $\mathbb{R}^d$ of the means are the bounded ones. A subset of $\mathrm{PD}(d)$ is relatively compact if it is bounded above and below by positive-definite matrices in the positive-definite (or Löwner) ordering $\preceq$ on the vector space of symmetric matrices. This is equivalent to saying that the eigenvalues of all its members are uniformly bounded and bounded away from zero.

We first show that the MTLE's of the covariance matrices for the constrained normal models described in Sect. 2.10 are robust and compute their individual breakdown point.

3.2 THEOREM. *(Individual breakdown point of the normal MTLE's of the covariance matrices.)*

(a) Assume $2r \geq n + g(d+1)$ *in the "full" case and* $2r \geq n + 2g$ *in the "diagonal" and "spherical" cases. (Note that these assumptions imply the standard assumptions on* r *made in Sect. 2.10.) The normal MTLE's of the covariance matrices remain in a compact subset of* $PD(d)$ *that depends only on* D *as at most* $n - r + g - 1$ *data points are replaced in an arbitrary but admissible way.*

(b) The covariance matrices returned from any M-step break down as $n - r + g$ data points are suitably replaced.

(c) Under the assumption of (a) the individual breakdown value of the normal MTLE's of the covariance matrices is for all data sets D

$$\beta_{\mathrm{Cov}}(n, g, r, D) = \frac{1}{n}(n - r + g).$$

PROOF. We give proofs in the "full" case; the others are similar.

(a) We first show that, no matter what the modified data set M is, the maximum trimmed likelihood remains bounded below by a strictly positive constant. The constant is determined by a simple solution that is sufficient for our purpose. We choose as R the remaining $n - (n - r + g - 1) = r - g + 1$ original observations and $g - 1$ of the replacements. Without loss of generality, let the original data be $x_1, \ldots, x_{r-g+1}$ and the replacements

$y_1, \ldots, y_{g-1}$. Let $u_1 = \cdots = u_g = \frac{1}{g}$, $m_1 = 0$, $m_j = y_{j-1}$, $j \in 2..g$, and $V_j = I_d$. The trimmed likelihood is

$$\begin{aligned}
&\prod_{i=1}^{r-g+1} \frac{1}{g}\Big\{(2\pi)^{-d/2} e^{-\|x_i\|^2/2} + \sum_{j=1}^{g-1} (2\pi)^{-d/2} e^{-\|x_i - y_j\|^2/2}\Big\} \\
&\qquad\qquad \times \prod_{i=1}^{g-1} \frac{1}{g}\Big\{(2\pi)^{-d/2} e^{-\|y_i\|^2/2} + \sum_{j=1}^{g-1} (2\pi)^{-d/2} e^{-\|y_i - y_j\|^2/2}\Big\} \\
&\geq \prod_{i=1}^{r-g+1} \frac{1}{g}\Big\{(2\pi)^{-d/2} e^{-\|x_i\|^2/2}\Big\} \prod_{i=1}^{g-1} \frac{1}{g}(2\pi)^{-d/2} \geq (2\pi)^{-dr/2} g^{-r} e^{-\|D\|^2/2} = C_D,
\end{aligned}$$

a strictly positive constant.

Now, by assumption, any r-element subset R of the modified data set contains at least $r - (n - r + g - 1) = 2r - n - g + 1 \geq gd + 1$ original points. Therefore, Lemma 2.12(b) may be applied and we have thus shown

$$2 \ln C_D \leq 2 \ln f[\mathbf{x}_R \mid \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*] \leq -r \ln \det 2\pi c V_j^* - K_D \operatorname{tr}\big(V_j^*\big)^{-1}$$

for $1 \leq j \leq g$, with some constant K_D that depends only on D. It is well known that the set of matrices $V_j \in \mathrm{PD}(d)$ for which the right side is bounded below is compact.

(b) Let M be the data set D modified by $n - r + g$ replacements and let $(R_1, \ldots, R_g)$ be the MAP partition of the r-element subset $R \subseteq M$ associated with the stochastic matrix $\mathbf{w}$ input to the M-step. By assumption, R contains at least g replacements. Hence, *either* one cluster contains at least two replacements *or* each cluster contains at least one replacement, in particular some cluster with $\geq d + 1$ elements. In any case the partition has a cluster R_ℓ containing a replacement y and some other element x. Eq. (2.5) implies

$$\begin{aligned}
&W_R(w_\ell) \succeq \frac{1}{g} W_{R_\ell} \\
\succeq &\frac{1}{g}\big((y - \frac{1}{2}(y+x))(y - \frac{1}{2}(y+x))^{\mathrm{T}} + (x - \frac{1}{2}(y+x))(x - \frac{1}{2}(y+x))^{\mathrm{T}}\big) \\
= &\frac{1}{2g}(y - x)(y - x)^{\mathrm{T}}.
\end{aligned}$$

Now, let $(\mathbf{u}, \mathbf{m}, \mathbf{V})$, $\mathbf{V} \in \mathcal{V}_c$, be the parameter computed from $\mathbf{w}$ in the M-step. Since also $2\mathbf{V} \in \mathcal{V}_c$, minimality (2.9) of $(\mathbf{u}, \mathbf{m}, \mathbf{V})$ implies

$$\begin{aligned}0 \le & \sum_j w_j(R)\big\{ \ln\det 2V_j + \operatorname{tr} S_R(w_j)(2V_j)^{-1} - \ln\det V_j - \operatorname{tr} S_R(w_j)V_j^{-1}\big\} \\ = & rd\ln 2 - \frac{1}{2}\sum_j \operatorname{tr} W_R(w_j)V_j^{-1} \le rd\ln 2 - \frac{1}{2}\operatorname{tr} W_R(w_\ell)V_\ell^{-1} \\ \le & rd\ln 2 - \frac{1}{4g}(y-x)^{\mathrm{T}}V_\ell^{-1}(y-x).\end{aligned}$$

The estimate $(y-x)^{\mathrm{T}}V_\ell^{-1}(y-x) \le 4grd\ln 2$ obtained proves that the smallest eigenvalue of V_ℓ^{-1} approaches zero arbitrarily closely if the replacements are chosen in such a way as to be far away from all original data and from each other.

Part (c) follows from (a) and (b).

It is interesting to remark that the normal MTLE's of the covariance matrices withstand $g-1$ more outliers than there are discarded elements, $n-r$. The constraints effect that outliers that are spread out may each create a component of their own and outliers located close together may create a common component. In each case the covariance matrices of the optimal mixture do not completely break down.

3.3 COROLLARY *(a) The maximal number of outliers that the normal MTLE's of the covariance matrices can resist is*

$$\begin{cases} \left\lfloor \dfrac{n-g(d-1)}{2} \right\rfloor - 1 & \textit{(full)}, \\ \left\lfloor \dfrac{n}{2} \right\rfloor - 1 & \textit{(diagonal and spherical)}. \end{cases}$$

The parameter r has to be set to $\left\lceil \frac{n+g(d+1)}{2} \right\rceil$ and $\left\lceil \frac{n}{2} \right\rceil + g$, respectively.

(b) The asymptotic breakdown point in each case is $1/2$.

PROOF. (a) We are asking for the largest integer $n-r+g-1$ under the constraint $2r \ge n + g(d+1)$ ("full" case) and $2r \ge n + 2g$ ("diagonal" and "spherical" cases). This proves Part (a) and (b) is immediate.

According to Davies (1987), the maximal asymptotic breakdown value of any affine equivariant covariance estimator is 1/2. Part (b) of the corollary says that the MTLE's of the covariance matrices share this value.

We next prove that, despite constraints and trimming, the *universal* breakdown point of the sample mean is small. We first state a lemma.

3.4 Lemma. *Let $1 \le q \le r$, and let $R = (x_1, \dots, x_{r-q}, y_1, \dots, y_q)$ consist of $r-q$ original data points x_i and q replacements y_i, and let $(\mathbf{u}, \mathbf{m}, \mathbf{V})$ be parameters computed in the M-step for R, e.g. a fixed point. Then*

$$\max_j \|m_j\| \longrightarrow \infty \quad \text{as } \|y_1\| \to \infty \text{ such that } y_i - y_1,\ 2 \le i \le q,$$

remain bounded.

Proof. Let w_j be the weights that induce the parameters. Eq. (2.6) implies

$$\sum_j m_j \Big(\sum_{i=1}^{r-q} w_j(x_i) + \sum_{i=1}^{q} w_j(y_i) \Big) = \sum_{i=1}^{r-q} x_i + \sum_{i=1}^{q} y_i = \sum_{i=1}^{r-q} x_i + q y_1 + \sum_{i=1}^{q} (y_i - y_1)$$

and the claim follows since the quantities in parentheses on the left side remain bounded.

3.5 Theorem. *(Universal breakdown point of the normal MTLE's of the means.)* Let $g \ge 2$.

(a) Assume $r < n$ and $r \ge gd + 2$ in the "full" and $r \ge g + 1$ in the "diagonal" and "spherical" cases. The normal MTLE's of all means remain bounded by a constant that depends only on the data set D as one *observation is arbitrarily replaced.*

(b) Assume $r \ge g + 2$ and $r \ge gd + 1$ in the "full" case. There is a data set in general position such that the normal MTLE of one mean breaks down as two *particular observations are suitably replaced.*

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

Proof. We restrict ourselves again to considering full covariance matrices, the other cases being similar.

(a) Let $M = (x_1, \dots, x_{n-1}, y)$ be a modification of D by one admissible replacement y. We show that the optimal solution $(\widetilde{R}, \widetilde{\mathbf{u}}, \widetilde{\mathbf{m}}, \widetilde{\mathbf{V}})$ for M with $r < n$ under the condition that y is not discarded is inferior to some solution which discards y if y is sufficiently distant. Let $\widetilde{d}_j$ denote the Mahalanobis distance induced by $\widetilde{V}_j$, i.e., $\widetilde{d}_j(u, v) = \sqrt{(u-v)^{\mathrm{T}} \widetilde{V}_j^{-1} (u-v)}$, let $\widetilde{d}_j(u, D) =$

$\min_{v\in D}\widetilde{d}_j(u,v)$ and $\widetilde{d}_j(D)=\max_{u,v\in D}\widetilde{d}_j(u,v)$ denote the distance between u and D and the diameter of D w.r.t. $\widetilde{d}_j$, respectively.

Without loss of generality, $\widetilde{R}=(x_1,\dots,x_{r-1},y)$. Let $R=(x_1,\dots,x_r)$ and let

$$m_j=\begin{cases}x_r, & \text{if } \widetilde{d}_j(\widetilde{m}_j,D)>\widetilde{d}_j(D),\\ \widetilde{m}_j, & \text{otherwise.}\end{cases}$$

We now show that the solution $(\widetilde{R},\widetilde{\mathbf{u}},\widetilde{\mathbf{m}},\widetilde{\mathbf{V}})$ is inferior to $(R,\widetilde{\mathbf{u}},\mathbf{m},\widetilde{\mathbf{V}})$ if y is such that $\widetilde{d}_j(y,D)>3\widetilde{d}_j(D)$ for all j. Comparing the trimmed likelihood of the former

$$\Big(\prod_{i=1}^{r-1}\sum_{j=1}^{g}\widetilde{u}_j\mathrm{N}_{\widetilde{m}_j,\widetilde{V}_j}(x_i)\Big)\sum_{j=1}^{g}\widetilde{u}_j\mathrm{N}_{\widetilde{m}_j,\widetilde{V}_j}(y)$$

termwise with that of the latter

$$\Big(\prod_{i=1}^{r-1}\sum_{j=1}^{g}\widetilde{u}_j\mathrm{N}_{m_j,\widetilde{V}_j}(x_i)\Big)\sum_{j=1}^{g}\widetilde{u}_j\mathrm{N}_{m_j,\widetilde{V}_j}(x_r),$$

we see that it is sufficient to show $\widetilde{d}_j(x_i,x_r)<\widetilde{d}_j(x_i,\widetilde{m}_j)$, $i<r$, if j is such that $\widetilde{d}_j(\widetilde{m}_j,D)>\widetilde{d}_j(D)$ and $\widetilde{d}_j(x_r,\widetilde{m}_j)<\widetilde{d}_j(y,\widetilde{m}_j)$ in the opposite case.

Now, if $\widetilde{d}_j(\widetilde{m}_j,D)>\widetilde{d}_j(D)$ then $\widetilde{d}_j(x_i,x_r)\le\widetilde{d}_j(D)<\widetilde{d}(\widetilde{m}_j,D)\le\widetilde{d}_j(x_i,\widetilde{m}_j)$; if $\widetilde{d}_j(\widetilde{m}_j,D)\le\widetilde{d}_j(D)$ then

$$\begin{aligned}\widetilde{d}_j(y,\widetilde{m}_j)&\ge\widetilde{d}_j(y,D)-\widetilde{d}_j(\widetilde{m}_j,D)\\&>3\widetilde{d}_j(D)-\widetilde{d}_j(\widetilde{m}_j,D)\\&\ge\widetilde{d}_j(D)+\widetilde{d}_j(\widetilde{m}_j,D)\\&\ge\widetilde{d}_j(x_r,\widetilde{m}_j).\end{aligned}$$

In order to prove that the means remain bounded, we still have to prove that the locations of the replacement y where it is not necessarily discarded are bounded by a constant that depends only on D. (Note that $\widetilde{V}_j$ and, hence, the distance $\widetilde{d}_j$ depends on y!) In other words, we have to show that the sets $\{y\mid\widetilde{d}_j(y,D)\le3\widetilde{d}_j(D)\}$ are bounded by constants that depend only on D. To this end we next show that $\widetilde{V}_j$ is bounded below and above by positive-definite matrices L_j and U_j that depend only on D.

Indeed, the optimal parameters $(\widetilde{R}, \widetilde{\mathbf{u}}, \widetilde{\mathbf{m}}, \widetilde{\mathbf{V}})$ are superior to the parameters $\big(\widetilde{R}, \big(\frac{1}{g}, \dots, \frac{1}{g}\big), (0, \dots, 0, y), I_d\big)$, i.e.,

$$\begin{aligned} &f[\mathbf{x}_{\widetilde{R}} \mid \widetilde{\mathbf{u}}, \widetilde{\mathbf{m}}, \widetilde{\mathbf{V}}] \geq f[\mathbf{x}_{\widetilde{R}} \mid (1/g, \dots, 1/g), (0, \dots, 0, y), I_d] \\ =&g^{-r} \prod_{i<r} \Big(\sum_{j<g} \mathrm{N}_{0,I_d}(x_i) + \mathrm{N}_{y,I_d}(x_i) \Big) \Big(\sum_{j<g} \mathrm{N}_{0,I_d}(y) + \mathrm{N}_{y,I_d}(y) \Big) \\ \geq&g^{-r} \Big(\prod_{i<r} \sum_{j<g} \mathrm{N}_{0,I_d}(x_i) \Big) \mathrm{N}_{0,I_d}(0) =: c_D. \end{aligned}$$

The constant c_D does not depend on y. Since $r \geq gd+2$, $\widetilde{R}$ contains at least $gd+1$ original elements and Lemma 2.12(b) shows

$$2 \ln c_D \leq 2 \ln f[\mathbf{x}_{\widetilde{R}} \mid \widetilde{\mathbf{u}}, \widetilde{\mathbf{m}}, \widetilde{\mathbf{V}}] \leq -r \ln \det 2\pi c \widetilde{V}_j - K_D \operatorname{tr} \widetilde{V}_j^{-1},$$

i.e., the right side is bounded below by a constant that depends only on D. Its behavior as a function of $\widetilde{V}_j$ provides two matrices L_j and U_j as required. Denoting the Mahalanobis distances w.r.t. L_j and U_j by d_{L_j} and d_{U_j}, respectively, the claim finally follows from

$$d_{U_j}(y, x_1) \leq \widetilde{d}_j(y, x_1) \leq \widetilde{d}_j(y, D) + \widetilde{d}_j(D) \leq 4\widetilde{d}_j(D) \leq 4 d_{L_j}(D).$$

(b) We proceed in several steps.

(α) Construction of data sets D and M:

Let $F := \{x_1, \dots, x_{r-g}\}$ be a set of data points in general position. We complete F to a data set D by points which we control by a constant $K_1 > 0$ and we control the two replacements by another constant $K_2 > 0$. Both constants are specified later. Using Lemma A.6, it is possible to inductively add points $z_1, \dots, z_{n-r+g-2}$ to F such that

(i) $\|z_\ell - z_k\| \geq K_1$ for all $\ell \neq k$;

(ii) $\|x_i - z_k\| \geq K_1$ for all $i \in 1..(r-g)$ and all $k \in 1..(n-r+g-2)$;

(iii) $W_H \succeq c_F I_d$ for all $H \in \binom{F \cup \{z_1, \dots, z_{n-r+g-2}\}}{d+1}$,

with some constant c_F that depends only on F. The set of x's and z's is of size $n-2 \geq d+1$. (For $d=1$ this estimate follows from $r \geq g+2$ and for $d \geq 2$ it follows from $r \geq gd+1$.) Thus, (iii) implies general position of

the points constructed so far. The data set D is completed by two arbitrary points q_1, q_2 in general position. In order to obtain our modified data set

$$M := F \cup \{z_1, \ldots, z_{n-r+g-2}\} \cup \{y_1, y_2\}$$

we use again Lemma A.6 replacing the two points q_1 and q_2 with a twin pair $y_1 \neq y_2$ such that

(iv) $\|y_1 - y_2\| = 1$;

(v) $\|u - y_k\| \geq K_2$ for all $u \in F \cup \{z_1, \ldots, z_{n-r+g-2}\}$ and for $k = 1, 2$;

(vi) $W_T \succeq c_F I_d$ for all $T \in \binom{M}{d+1}$ that contain at least one y_k.

Conditions (iii) and (vi) taken together imply $W_T \succeq c_F I_d$ for all $d+1$-element subsets $T \subseteq M$. The idea was to place the points z_ℓ and the replacements y_k in directions that guarantee general position as homotheties are applied. In view of Lemma 3.4, we will show that the optimal solution does not discard the outliers y_1 and y_2 if K_1 and K_2 are chosen large enough.

(β) The maximum of the trimmed likelihood for the modified data set M is bounded below by a constant that depends only on F, g, and r:

It is sufficient to construct a subset $R \subseteq M$ and a parameter set with likelihood bounded below by a function of F, g, and r. We let $R = F \cup \{z_1, \ldots, z_{g-2}\} \cup \{y_1, y_2\}$, $u_j = 1/g$, $m_j = z_j$, $1 \leq j \leq g-2$, $m_{g-1} = 0$, $m_g = y_1$, and $V = I_d$. Using $g \geq 2$ and (iv) we have

$$\begin{aligned} &f[\mathbf{x}_R \mid \mathbf{u}, \mathbf{m}, V] \\ &\quad = \prod_{i=1}^{r-g} \frac{1}{g} \sum_{j=1}^{g} \mathrm{N}_{m_j, I_d}(x_i) \prod_{i=1}^{g-2} \frac{1}{g} \sum_{j=1}^{g} \mathrm{N}_{m_j, I_d}(z_i) \prod_{i=1}^{2} \frac{1}{g} \sum_{j=1}^{g} \mathrm{N}_{m_j, I_d}(y_i) \\ &\quad \geq g^{-r} \prod_{i=1}^{r-g} \mathrm{N}_{0, I_d}(x_i) \prod_{i=1}^{g-2} \mathrm{N}_{z_i, I_d}(z_i) \prod_{i=1}^{2} \mathrm{N}_{y_1, I_d}(y_i) \\ &\quad \geq g^{-r} (2\pi)^{-\frac{gd}{2}} e^{-\frac{1}{2}} \prod_{i=1}^{r-g} \mathrm{N}_{0, I_d}(x_i) \end{aligned}$$

as required.

The assumptions of the combinatorial Lemma 4.2 by Gallegos and Ritter (2005) are satisfied, so any partition $\mathcal{R}$ of any subset of M of size r in g clusters has either the form

$\mathcal{R} = \{\{x_1, \ldots, x_{r-g}\}, \{y_1, y_2\}, g-2$ one-point clusters from the $z_k's\}$ *or* there is a cluster $R_\ell \in \mathcal{R}$, $|R_\ell| \geq 2$, which contains some pair $\{x_i, y_h\}$ or some z_k.

(γ) The MAP partition $\mathcal{R}$ associated with the weight matrix $\mathbf{w}$ of an optimal solution $(R, \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*)$ is of the first kind if K_1 and K_2 are sufficiently large:

Assume on the contrary that $\mathcal{R}$ is of the second kind. Choose R_ℓ containing a pair x_i, y_h or z_k, u with some $u \neq z_k$. By (i), (ii), and (v), R_ℓ contains two distant elements so that Eq. (2.5) implies

$$g \operatorname{tr} W_R(\mathbf{w}) \geq g \operatorname{tr} W_R(w_\ell) \geq \operatorname{tr} W_{R_\ell} \underset{K_1, K_2 \to \infty}{\longrightarrow} \infty. \tag{3.1}$$

Moreover, by $r \geq gd + 1$ there exists j such that $|R_j| \geq d + 1$ and we infer from Eq. (2.5), (iii), and (vi)

$$gW_R(\mathbf{w}) \succeq gW_R(w_j) \succeq W_{R_j} \succeq c_F I_d. \tag{3.2}$$

Now, Lemma 2.12 (a), (β), and Eq. (3.2), show that the quantities

$$\begin{aligned} 2 \ln f[\mathbf{x}_R \mid \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*] &\leq -r \ln \det 2\pi c V_j^* - c \operatorname{tr} W_{R^*}(\mathbf{w})(V_j^*)^{-1} \\ &\leq -r \ln \det 2\pi c V_j^* - \text{const} \operatorname{tr} (V_j^*)^{-1} \end{aligned}$$

all remain bounded below by a constant that depends only on F. The second estimate shows that V_j^* lies in a compact subset of $\mathrm{PD}(d)$ that is independent of the choice of the points z_k and of the replacements. Therefore, the first estimate shows that $\operatorname{tr} W_R(\mathbf{w})$ is bounded above by a constant that depends only on F. This contradiction to Eq. (3.1) proves (γ).

Finally choose K_1 and K_2 so large that the MAP partition of any optimal solution is of the first kind. In particular, the solution does not discard the replacements. According to Lemma 3.4, at least one mean breaks down as $K_2 \to \infty$.

(c) follows from (a) and (b).

3.6 The case $g = 1$ In the case of one component, the criterion (2.2) reduces to Rousseeuw's (1985) maximum covariance determinant, MCD, for robust estimation of location and scatter. If $\alpha < 0.5$ then its asymptotic breakdown point with parameter $r = \lceil (1-\alpha)n \rceil$ is known to be α, see Rousseeuw (1985), p.291. This is in harmony with our result on the scatter matrices, Theorem 3.2. For $g = 1$, reduction of the parameter r has the effect

that breakdown of the mean occurs at a much higher number of outliers compared with the case $g > 1$ stated in Theorem 3.5. The reason is that, in the case $g > 1$, the outliers may establish a component of their own if they are close to each other, thus causing one mean to diverge.

For $g = 1$, not only the criterion but also the algorithm is known. The weights are all 1 so that the E-step is trivial. The M- and T-steps of EMT reduce to Rousseeuw and Van Driessen's (1999), Theorem 1, alternating C-step for computing the MCD.

4 Restricted Breakdown Point of the Means

Theorem 3.5 and the preceding remark state that the asymptotic breakdown value of the normal MTLE's of the means is zero, an at first sight disappointing result for a trimming algorithm. It is, however, not the estimator that has to be blamed but the stringent universal breakdown point. Besides allowing any kind of contamination it makes a statement about any data set, even it is unlikely to emanate from a g-component model. García-Escudero and Gordaliza (1999) conjectured from simulations in the classification framework that it is hard to break down "clear" cluster structures if the natural number of clusters is chosen as the parameter g. Robustness of the means depends also on the structure of data set. This phenomenon was mathematically analyzed by us (Gallegos and Ritter, 2005, 2009) in the cases of homo- and heteroscedastic normal clustering. We transfer this result to the heteroscedastic normal mixture model under the HDBT constraints by considering the restricted breakdown value w.r.t. a subclass of data sets with an inherent cluster structure, see Sect. 4.3.

We need more notation. Let $\mathcal{P} = \{P_1, \ldots, P_g\}$ be a partition of D and let $\emptyset \neq T \subseteq D$. The partition $\mathcal{P} \cap T = \{P_1 \cap T, \ldots, P_g \cap T\}$ 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 a partition of T. The *common refinement* of $\mathcal{P}$ and $\mathcal{T}$ is denoted by $\mathcal{P} \sqcap \mathcal{T} = \{P_j \cap T_k \mid j \leq g, k \leq g'\}$, a partition of T; some clusters may be empty. The *pooled SSP matrix* $\sum_k W_{T_k}$ of $\mathcal{T}$ is denoted by $W_{\mathcal{T}}$ and the *pooled scatter matrix* of $\mathcal{T}$ is $S_{\mathcal{T}} = \frac{1}{|T|} W_{\mathcal{T}}$. We denote the set of all stochastic matrices over the index set $T \times (1..g')$ by $\mathcal{M}(T, g')$. For a data set T and $\boldsymbol{\alpha} \in \mathcal{M}(T, g')$, $W_T(\boldsymbol{\alpha}) = \sum_{k \leq g'} W_T(\alpha_k)$ is the *pooled weighted SSP matrix*, cf. Section 2.10. Given a partition $\mathcal{Q} = \{Q_1, \ldots, Q_g\}$ of some subset $Q \subseteq D$, we also define

$$W_{\mathcal{Q}}(\boldsymbol{\alpha}) = \sum_{j \leq g} W_{Q_j}(\boldsymbol{\alpha}) = \sum_{j \leq g} \sum_{k \leq g'} W_{Q_j}(\alpha_k).$$

If $\boldsymbol{\alpha}$ is binary then $W_{\mathcal{Q}}(\boldsymbol{\alpha})$ is the pooled SSP matrix of the common refinement of $\mathcal{Q}$ and the partition defined by $\boldsymbol{\alpha}$. The proof of the theorem of this section depends on lemmas which we next state and prove. The following one states a basic condition which implies robustness of the means. In Theorem 4.4, we will see that it actually means a separation property of the data set.

4.1 Lemma. *Let $g \geq 2$ and $r < n$, assume $r > gd + 1$ in the full and $r > g + 1$ in the diagonal and spherical normal cases, and let $q \in \max\{2r - n, gd+1\} \,..\, (r-1)$ (full case) and $q \in \max\{2r - n, g+1\} \,..\, (r-1)$ (diagonal and spherical cases). Assume that the data set D possesses a partition $\mathcal{P}$ in g clusters such that for all $T \subseteq D$, $q \leq |T| < r$, and all $\boldsymbol{\alpha} \in \mathcal{M}(T, g-1)$*

$$\det W_T(\boldsymbol{\alpha}) \geq g^2 \max_{R \in \binom{D}{r}, R \supseteq T} \det \frac{1}{c^2} W_{\mathcal{P} \cap R} \qquad \textit{(full and diagonal)} \quad (4.1)$$

$$\operatorname{tr} W_T(\boldsymbol{\alpha}) \geq g^{2/d} \max_{R \in \binom{D}{r}, R \supseteq T} \operatorname{tr} \frac{1}{c^2} W_{\mathcal{P} \cap R}. \qquad \textit{(spherical)}$$

Then the individual breakdown point of the normal MTLE's of the means satisfies

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

Proof. in the "full" case. Let M be any data set obtained from D by modifying at most $r - q$ elements. Let $(R^*, (u_j^*)_{j=1}^g, (m_j^*)_{j=1}^g, (V_j^*)_{j=1}^g)$ be an optimal solution for M and let $\mathbf{w}^* = (w_j^*(x))_{x,j}$ be the related posterior probabilities. We will show that the means m_j^* are bounded by a number that depends solely on the original data D. Our proof proceeds in several steps. We write $\overline{\mathbf{x}}_{\mathcal{P} \cap R} = (\overline{x}_{P_1 \cap R}, \ldots, \overline{x}_{P_g \cap R})$.

$$(\alpha) \quad f[\mathbf{x}_{R^*} \mid \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*] \geq \max_{R \in \binom{M \cap D}{r}} f[\mathbf{x}_R \mid g^{-1}, \overline{\mathbf{x}}_{\mathcal{P} \cap R}, S_{\mathcal{P} \cap R}].$$

Since $q \geq 2r - n$ there exists a subset $R \subseteq M \cap D$ of size r and the claim follows from optimality of $(R^*, \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*)$.

Now let $\mathcal{R}^* = (R_1^*, \ldots, R_g^*)$ be an MAP partition of R^* w.r.t. the optimal solution and let

$$\lambda_{\min} := \min\{\lambda \mid \lambda \text{ eigenvalue of } W_C,\ C \subseteq D,\ |C| = d + 1\},$$

a constant > 0 that depends only on the data set D.

(β) The matrices V_j^*, $j \in 1 \dots g$, are bounded above and below by positive-definite matrices that depend only on D, not on the replacements:

Since $|R^*| = r$, $R^* = \bigcup_{j=1}^g R_j^*$ has at least $q \geq gd + 1$ original observations. By the pigeon hole principle, there exists $j \in 1..g$ such that $|R_j^* \cap D| \geq d+1$. By Eq. (2.5), this implies the lower estimate

$$W_{R^*}(\mathbf{w}^*) \succeq W_{R^*}(\mathbf{w}_j^*) \succeq \frac{1}{g} W_{R_j^*} \succeq \frac{\lambda_{\min}}{g} I_d.$$

Applying 2.12(a) to the optimal solution, we infer from (α) that the expression $-\frac{r}{2} \ln \det 2\pi c V_j^* - \frac{c}{2} \lambda_{\min} \operatorname{tr} (V_j^*)^{-1}$ remains bounded below by a constant which depends solely on D, n, g, and r. The well-known behavior of this function of V_j^* implies that the matrices V_j^*, $j \in 1 \dots g$, remain bounded above and below in the Löwner ordering.

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

Let $x \in R_j^* \cap D$. We have $W_{R^*}(\mathbf{w}^*) \succeq \frac{1}{g}(x - m_j^*)(x - m_j^*)^{\mathrm{T}}$ and, hence, $\|x - m_j^*\|^2 \leq g \operatorname{tr} W_{R^*}(\mathbf{w}^*)$. Part ($\gamma$) will be proved if we show that $\operatorname{tr} W_{R^*}(\mathbf{w}^*)$ has an upper bound that does not depend on the modifications. By Lemma 2.12,

$$c \operatorname{tr} W_{R^*}(\mathbf{w}^*)(V_j^*)^{-1} \leq -r \ln \det 2\pi c V_j^* - 2 \ln f[\mathbf{x}_{R^*} \mid \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*],$$

and the claim follows from (α) and (β).

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

This is proved like (γ) with x replaced by the replacement.

It follows from (γ) and (δ) that, in the long run, each set R_j^* consists solely of original observations or solely of modifications, i.e., either $R_j^* \cap D = \emptyset$ or $R_j^* \cap (M \setminus D) = \emptyset$.

(ϵ) If R_j^* contains some replacement then $w_j^*(x) \to 0$ for all $x \in R^* \cap D$ as the replacement tends to ∞:

Let $x \in R_\ell^* \cap D$. By Eq. (2.4), we have $u_\ell^* = \frac{1}{r}\sum_{z\in R^*} w_\ell^*(z) \geq 1/gr$ and, hence,

$$\begin{aligned} w_j^*(x) &= \frac{u_j^*(\det V_j^*)^{-1/2} e^{-\frac{1}{2}(x-m_j^*)^{\mathrm{T}}(V_j^*)^{-1}(x-m_j^*)}}{\sum_k u_k^*(\det V_k^*)^{-1/2} e^{-\frac{1}{2}(x-m_k^*)^{\mathrm{T}}(V_k^*)^{-1}(x-m_k^*)}} \\ &\leq gr\sqrt{\frac{\det V_\ell^*}{\det V_j^*}}\,\frac{e^{-\frac{1}{2}(x-m_j^*)^{\mathrm{T}}(V_j^*)^{-1}(x-m_j^*)}}{e^{-\frac{1}{2}(x-m_\ell^*)^{\mathrm{T}}(V_\ell^*)^{-1}(x-m_\ell^*)}}. \end{aligned}$$

The claim follows from (β), (γ), and (δ).

(ζ) $\ln f[\mathbf{x}_{R^*} \mid \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*] \leq c_{d,r} - dr\ln c - \frac{r}{2}\ln\det S_{R^*}(\mathbf{w}^*)$:

By 2.12(a),

$$\begin{aligned} &2\ln f[\mathbf{x}_{R^*} \mid \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*] \\ &\qquad \leq -dr\ln 2\pi c^2 - \big[r\ln\det(V_1^*/c) + \operatorname{tr}(V_1^*/c)^{-1}W_{R^*}(\mathbf{w}^*)\big] \\ &\qquad \leq -dr\ln 2\pi c^2 - \min_{A\succeq 0}\big[r\ln\det A + \operatorname{tr}A^{-1}W_{R^*}(\mathbf{w}^*)\big]. \end{aligned}$$

Now, standard normal estimation theory shows that the function

$$A \mapsto r\ln\det A + \operatorname{tr}A^{-1}W_{R^*}(\mathbf{w}^*),\ A \succ 0,$$

attains its minimum at $\frac{1}{r}W_{R^*}(\mathbf{w}^*)$ with value $r\big[\ln\det\frac{W_{R^*}(\mathbf{w}^*)}{r} + d\big]$. This is the claim.

(η) There is $K > 0$ such that R^* contains no modification y, $\|y\| > K$:

Assume on the contrary that, for all $K > 0$, there is $y \in R^* \setminus D$, $\|y\| > K$, $y \in R_g^*$, say. Let $\varepsilon > 0$ and let K be so large that $w_g^*(x) \leq \varepsilon$ for all $x \in R^* \cap D$, cf. (ϵ). Putting $\alpha_j^*(x) = \frac{w_j^*(x)}{1-w_g^*(x)}$, $x \in R^* \cap D$, $j < g$, we have $\boldsymbol{\alpha}^* \in \mathcal{M}(R^* \cap D, g-1)$ and

$$\begin{aligned} W_{R^*}(\mathbf{w}^*) &\succeq \sum_{j=1}^{g-1}\sum_{x\in R^*\cap D} w_j^*(x)(x-m_j^*)(x-m_j^*)^{\mathrm{T}} \\ &\succeq \min_{x\in R^*\cap D}(1-w_g^*(x))\sum_{j=1}^{g-1}\sum_{x\in R^*\cap D}\alpha_j^*(x)(x-m_j^*)(x-m_j^*)^{\mathrm{T}} \\ &\succeq (1-\varepsilon)\sum_{j=1}^{g-1} W_{R^*\cap D}(\alpha_j^*) = (1-\varepsilon)W_{R^*\cap D}(\boldsymbol{\alpha}^*) \end{aligned}$$

by Steiner's formula A.3. We modified at most $r-q$ elements and R^* contains at least one modification. Hence, $q \le |R^* \cap D| < r$ and we may apply Hypothesis (4.1) to continue

$$\begin{aligned}&\det W_{R^*}(\mathbf{w}^*)\\ &\ge (1-\varepsilon)^d \det W_{R^*\cap D}(\boldsymbol{\alpha}^*) \ge (1-\varepsilon)^d g^2 \max_{R\in\binom{D}{r}, R\supseteq R^*\cap D} \det \frac{1}{c^2} W_{\mathcal{P}\cap R}.\end{aligned}$$

Since ε is arbitrary we conclude

$$\det W_{R^*}(\mathbf{w}^*) \ge g^2 \max_{R\in\binom{D}{r}, R\supseteq R^*\cap D} \det \frac{1}{c^2} W_{\mathcal{P}\cap R}$$

and

$$2d\ln c + \ln\det S_{R^*}(\mathbf{w}^*) \ge 2\ln g + \max_{R\in\binom{D}{r}, R\supseteq R^*\cap D} \ln\det S_{\mathcal{P}\cap R}.$$

Now, for all $R \subseteq D$,

$$\begin{aligned}&\ln f[\mathbf{x}_R \mid g^{-1}, \overline{\mathbf{x}}_{\mathcal{P}\cap R}, S_{\mathcal{P}\cap R}]\\ &= \sum_{\ell=1}^{g} \sum_{x\in P_\ell\cap R} \ln \frac{1}{g} \sum_{j=1}^{g} (\det 2\pi S_{\mathcal{P}\cap R})^{-1/2} e^{-1/2(x-\overline{\mathbf{x}}_{P_j\cap R})^{\mathrm{T}} S_{\mathcal{P}\cap R}^{-1}(x-\overline{\mathbf{x}}_{P_j\cap R})}\\ &> -r\ln g - \frac{r}{2}\ln\det 2\pi S_{\mathcal{P}\cap R} - \frac{1}{2}\sum_{\ell=1}^{g}\sum_{x\in P_\ell\cap R} (x-\overline{\mathbf{x}}_{P_\ell\cap R})^{\mathrm{T}} S_{\mathcal{P}\cap R}^{-1}(x-\overline{\mathbf{x}}_{P_\ell\cap R})\\ &= c_{d,r} - r\ln g - \frac{r}{2}\ln\det S_{\mathcal{P}\cap R}.\end{aligned}$$

The last two estimates and Part (ζ) combine to show

$$\begin{aligned}\max_{R\in\binom{M\cap D}{r}} \ln f[\mathbf{x}_R \mid g^{-1}, \overline{\mathbf{x}}_{\mathcal{P}\cap R}, S_{\mathcal{P}\cap R}] &> c_{d,r} - r\ln g - \frac{r}{2} \min_{R\in\binom{M\cap D}{r}} \ln\det S_{\mathcal{P}\cap R}\\ &\ge c_{d,r} - r\ln g - \frac{r}{2} \max_{R\in\binom{D}{r}, R\supseteq R^*\cap D} \ln\det S_{\mathcal{P}\cap R}\\ &\ge c_{d,r} - dr\ln c - \frac{r}{2}\ln\det S_{R^*}(\mathbf{w}^*)\\ &\ge \ln f[\mathbf{x}_{R^*} \mid \mathbf{u}^*, \mathbf{m}^*, \mathbf{V}^*],\end{aligned}$$

a contradiction to (α). This proves Claim (η).

Finally, (η) shows that the means m_j^* are convex combinations of elements from D and replacements in the centered ball of radius K. This proves the lemma.

In order to remove the dependence on $\boldsymbol{\alpha}$ in Lemma 4.1 we need a definition. Assume $g \geq 2$, let $\varrho > 0$, and let $u \leq n/g$ be an integer. We define the real number

$$q_{u,\varrho} = \begin{cases} \max\left\{2r - n, (g-1)gd + 1, \frac{n-u}{1-\varrho}\right\} & \text{(full)} \\ \max\left\{2r - n, (g-1)g + 1, \frac{n-u}{1-\varrho}\right\} & \text{(diagonal and spherical).} \end{cases}$$

Plainly $\frac{q_{u,\varrho}}{(g-1)g} > d$ in the case of full covariance matrices and > 1 in the diagonal and spherical cases. Before stating the separation property we prove a combinatorial lemma. One of its assumptions is $q_{u,\varrho} \leq r - 1$. This is equivalent to $n > r > (g-1)gd+1$ and $n - (1-\varrho)(r-1) \leq u$. Combined with $u \leq n/g$ the last estimate implies also $\varrho \leq 1/g$.

4.2 Lemma. *Assume $q_{u,\varrho} \leq r-1$. Let $\mathcal{P} = \{P_1, \ldots, P_g\}$ be a partition of D in clusters of size $\geq u$, let $T \subseteq D$ such that $q_{u,\varrho} \leq |T| < r$ (the assumption on $q_{u,\varrho}$ implies the existence of such a subset T), 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 $|P_j \cap T| \geq \varrho|T|$.
(b) There are clusters T_k and P_j such that $|T_k \cap P_j| \geq \frac{q_{u,\varrho}}{(g-1)g}$.

Proof. (a) Assume on the contrary that $|P_\ell \cap T| < \varrho|T|$. From $D \supseteq T \cup P_\ell$ we infer

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

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

(b) The observations in T are spread over the $(g-1)g$ disjoint subsets of the form $T_k \cap P_j$. If (b) did not hold, we would have $|T| = \sum_{k,j} |T_k \cap P_j| < q_{u,\varrho}$, a contradiction.

Define

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

Given two subsets $S, T \subseteq \mathbb{R}^d$, $d(S,T)$ denotes their Euclidean distance and $d(S)$ the Euclidean diameter of S.

4.3. The separation property. Assume $q_{u,\varrho} \leq r-1$ and let c be the HDBT constant. We denote by $\mathcal{L}_{u,\varrho,c}$ the system of all d-dimensional admissible data sets D of size n with the following *separation property*:

D possesses a partition $\mathcal{P}$ in g subsets of size at least u such that, for all subsets $T \subseteq D$, $q_{u,\varrho} \le |T| < r$,

$$1+\kappa_\varrho \min_{\substack{s_h \in \operatorname{conv} P_h \cap T \\ \ell \neq j}} (s_\ell - s_j)^{\mathrm{T}} S_{\mathcal{P}\cap T}^{-1}(s_\ell - s_j) \qquad \text{(full and diagonal}^2\text{ cases)}$$

$$\ge g^2 \frac{\max_{R\in\binom{D}{r}, R\supset T} \det \frac{1}{c^2} W_{\mathcal{P}\cap R}}{\det W_{\mathcal{P}\cap T}} \Big\{ 1 + \frac{g-2}{2d(g-1)} \max_{\substack{s,s' \in P_j \\ 1\le j\le g \\ \mathcal{T}}} (s-s')^{\mathrm{T}} S_{\mathcal{P}\sqcap\mathcal{T}}^{-1}(s-s') \Big\}^d, \tag{4.2}$$

$$1+\kappa_\varrho \frac{\min_{\ell\neq j} d^2(\operatorname{conv}(P_j\cap T), \operatorname{conv}(P_\ell \cap T))}{\operatorname{tr} S_{\mathcal{P}\cap T}} \qquad \text{(spherical case)}$$

$$\ge g^{2/d} \frac{\max_{R\in\binom{D}{r}, R\supset T} \operatorname{tr} \frac{1}{c^2} W_{\mathcal{P}\cap R}}{\operatorname{tr} W_{\mathcal{P}\cap T}} \Big\{ 1 + \frac{g-2}{2(g-1)} \frac{\max_{1\le j\le g} d^2(P_j\cap T)}{\min_{\mathcal{T}} \operatorname{tr} S_{\mathcal{P}\sqcap\mathcal{T}}} \Big\}.$$

where $\mathcal{T}$ runs over all partitions of T in $g-1$ clusters. Lemma 4.2(b) shows that the inverse matrices appearing in Eq. (4.2) exist. The factor on the right hand side of (4.2) may be replaced with the expression

$$\exp\Big\{ \frac{g-2}{2(g-1)} \max_{\substack{s,s' \in P_j \\ 1\le j\le g \\ \mathcal{T}}} (s-s')^{\mathrm{T}} S_{\mathcal{P}\sqcap\mathcal{T}}^{-1}(s-s') \Big\},$$

a number independent of dimension d. The partition $\mathcal{P}$ appearing in the separation property plays the role of a partition of the data set in well-separated clusters. Note that condition (4.2) is affine equivariant. The set $\mathcal{L}_{u,\varrho,c}$ increases with decreasing u and increasing $\varrho \le 1/2$ and c.

Roughly speaking, a data set D has the separation property if it is composed of well separated clusters. Moreover, it helps if population shapes are balanced, i.e., the HDBT ratio c is close to 1, and if cluster sizes are balanced, i.e., u is large so that κ_ϱ and ϱ may be chosen large. Note, however, that κ_ϱ is bounded since $\varrho \le 1/g$.

If a data set has the separation property, then the normal MTLE's of the means are much more robust than predicted by the universal breakdown value in Theorem 3.5.

[2]In the diagonal case a similar estimate can also be given by working with the components of $\mathbb{R}^d$ separately.

4.4. THEOREM. *(Restricted breakdown point of the normal MTLE's of the means) Let $g \geq 2$, $r < n$.*

(a) Assume $r \geq (g-1)gd+2$ in the full and $r \geq (g-1)g+2$ in the diagonal and spherical cases, respectively, and let $u \in \mathbb{N}$ s.th. $n-(1-\varrho)(r-1) \leq u \leq n/g$ (hence $\varrho < 1/g$ by $n > r$). Then the restricted breakdown value of the normal MTLE's of the means w.r.t. $\mathcal{L}_{u,\varrho,c}$ satisfies

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

(b) The individual breakdown point of any data set D satisfies $\beta_{mean}(n, g, r, D) \leq \frac{1}{n}(n - r + 1)$.

(c) Let $2r - n \geq (g-1)gd + 1$ in the full and $2r - n \geq (g-1)g + 1$ in the diagonal and spherical cases, assume $2(n-r) \leq n/g - 1$. Let $u \in \mathbb{N}$ be s.th. $2(n-r) < u \leq n/g$ and put $\varrho = \frac{u-2(n-r)}{2r-n}$. Then

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

(d) If the hypotheses of (a) are satisfied and if the data set is of the class $\mathcal{L}_{u,\varrho,c}$ then the normal MTLE of R discards all replacements that are large enough.

PROOF. (a) The assumptions imply $q_{u,\varrho} \leq r - 1$. Let $T \subseteq D$ such that $q_{u,\varrho} \leq |T| < r$ and let $\boldsymbol{\alpha} \in \mathcal{M}(T, g-1)$. In order to shorten notation we use the abbreviation

$$d(h, j, k, \ell) := \overline{x}_{P_j \cap T}(\alpha_h) - \overline{x}_{P_\ell \cap T}(\alpha_k).$$

For $j \leq g$ such that $|P_j \cap T| > 0$ let

$$A_{T,j}(\boldsymbol{\alpha}) = \frac{1}{|P_j \cap T|} \sum_{1 \leq h < k < g} \alpha_h(P_j \cap T)\alpha_k(P_j \cap T)d(h, j, k, j)d(h, j, k, j)^{\mathrm{T}}.$$

Applying Lemma A.5 with $g' = g - 1$, we obtain $W_{P_j \cap T} = W_{P_j \cap T}(\boldsymbol{\alpha}) + A_{T,j}(\boldsymbol{\alpha})$ and

$$W_{\mathcal{P} \cap T} = W_{\mathcal{P} \cap T}(\boldsymbol{\alpha}) + \sum_j A_{T,j}(\boldsymbol{\alpha}) = W_{\mathcal{P} \cap T}(\boldsymbol{\alpha}) + A_T(\boldsymbol{\alpha}). \tag{4.3}$$

For $k < g$ such that $\alpha_k(T) > 0$ let

$$B_T(\alpha_k) = \frac{1}{\alpha_k(T)} \sum_{1\le j<\ell\le g} \alpha_k(P_j \cap T)\alpha_k(P_\ell \cap T) d(k,j,k,\ell) d(k,j,k,\ell)^{\mathrm{T}}.$$

Applying Lemma A.4 with $w(x) = \alpha_k(x)$ and $T_j = P_j \cap T$, we have for $\alpha_k(T) > 0$ the identity $W_T(\alpha_k) = W_{\mathcal{P}\cap T}(\alpha_k) + B_T(\alpha_k)$ and hence

$$W_T(\boldsymbol{\alpha}) = W_{\mathcal{P}\cap T}(\boldsymbol{\alpha}) + \sum_{k:\alpha_k(T)>0} B_T(\alpha_k). \tag{4.4}$$

According to Gallegos and Ritter (2005), Lemma A.1(b), $\det(A+\sum_h y_h y_h^{\mathrm{T}}) \ge (1+\sum_h y_h^{\mathrm{T}} A^{-1} y_h) \det A$ for all $A \in \mathrm{PD}(d)$. Applying this estimate, and using Eqs. (4.4) and (4.3), we infer

$$\begin{aligned}
&\det W_T(\boldsymbol{\alpha}) \\
&\ge \det W_{\mathcal{P}\cap T}(\boldsymbol{\alpha}) \Big(1 + \sum_{k:\alpha_k(T)>0} \frac{1}{\alpha_k(T)} \sum_{1\le j<\ell\le g} \alpha_k(P_j \cap T)\alpha_k(P_\ell \cap T) \times \\
&\qquad\qquad d(k,j,k,\ell) W_{\mathcal{P}\cap T}^{-1} d(k,j,k,\ell)^{\mathrm{T}} \Big) \\
&= \det W_{\mathcal{P}\cap T}(\boldsymbol{\alpha})(1 + r_T(\boldsymbol{\alpha})).
\end{aligned} \tag{4.5}$$

We next estimate the two factors in (4.5) rendering them devoid of $\boldsymbol{\alpha}$. Since $\sum_k \frac{\alpha_k(T)}{|T|} \frac{\alpha_k(P_j\cap T)}{\alpha_k(T)} = \sum_k \frac{\alpha_k(P_j\cap T)}{|T|} = \frac{|P_j\cap T|}{|T|} \ge \varrho$ according to Lemma 4.2 (a), Lemma A.8(b) may be applied to the expression

$$\begin{aligned}
&\sum_{k:\alpha_k(T)>0} \frac{1}{\alpha_k(T)} \sum_{1\le j<\ell\le g} \alpha_k(P_j \cap T)\alpha_k(P_\ell \cap T) \\
&\qquad = |T| \sum_{k:\alpha_k(T)>0} \frac{\alpha_k(T)}{|T|} \sum_{1\le j<\ell\le g} \frac{\alpha_k(P_j \cap T)}{\alpha_k(T)} \frac{\alpha_k(P_\ell \cap T)}{\alpha_k(T)}
\end{aligned}$$

implying

$$r_T(\boldsymbol{\alpha}) \ge \kappa_\varrho |T| \min_{\substack{s_h \in \operatorname{conv} P_h \cap T \\ \ell \ne j}} (s_\ell - s_j)^{\mathrm{T}} W_{\mathcal{P}\cap T}^{-1} (s_\ell - s_j). \tag{4.6}$$

Next use Lemmas A.8(a) and A.9 to estimate

$$
\begin{aligned}
&\operatorname{tr} W_{\mathcal{P}\cap T}^{-1}(\boldsymbol{\alpha}) A_T(\boldsymbol{\alpha}) \\
=&\sum_{j\le g} \frac{1}{|P_j \cap T|} \sum_{1\le h<k<g} \alpha_h(P_j\cap T)\alpha_k(P_j\cap T) \operatorname{tr} W_{\mathcal{P}\cap T}^{-1}(\boldsymbol{\alpha}) d(h,j,k,j) d(h,j,k,j)^{\mathrm{T}} \\
=&\sum_{j\le g} \frac{1}{|P_j \cap T|} \sum_{1\le h<k<g} \alpha_h(P_j\cap T)\alpha_k(P_j\cap T) d(h,j,k,j)^{\mathrm{T}} W_{\mathcal{P}\cap T}^{-1}(\boldsymbol{\alpha}) d(h,j,k,j) \\
\le&\frac{|T|}{2}\frac{g-2}{g-1} \max_{\substack{s,s'\in P_j \\ 1\le j\le g}} (s-s')^{\mathrm{T}} W_{\mathcal{P}\cap T}^{-1}(\boldsymbol{\alpha})(s-s') \\
\le&\frac{|T|}{2}\frac{g-2}{g-1} \max_{\mathcal{T}} \max_{\substack{s,s'\in P_j \\ 1\le j\le g}} (s-s')^{\mathrm{T}} W_{\mathcal{P}\cap \mathcal{T}}^{-1}(s-s'). \qquad (4.7)
\end{aligned}
$$

The identity $\det(A+B) = \det A \cdot \det(I_d + A^{-1/2} B A^{-1/2})$ and the arithmetic-geometric inequality $\det C \le \left(\frac{1}{d}\operatorname{tr} C\right)^d$, valid for $A \succ 0$, $B \in \mathbb{R}^{d\times d}$, and $C \succeq 0$, yield the inequality $\det(A+B) \le \det A\left(1 + \frac{1}{d}\operatorname{tr} A^{-1}B\right)^d$. Its application to Eq. (4.3) shows together with Eq. (4.7)

$$
\begin{aligned}
&\det W_{\mathcal{P}\cap T} \le \det W_{\mathcal{P}\cap T}(\boldsymbol{\alpha}) \Big\{ 1 + \frac{1}{d} \operatorname{tr} W_{\mathcal{P}\cap T}^{-1}(\boldsymbol{\alpha}) A_T(\boldsymbol{\alpha}) \Big\}^d \qquad (4.8) \\
\le&\det W_{\mathcal{P}\cap T}(\boldsymbol{\alpha}) \Big\{ 1 + \frac{1}{2d}\frac{g-2}{g-1} \max_{\mathcal{T}} \max_{\substack{s,s'\in P_j \\ 1\le j\le g}} (s-s')^{\mathrm{T}} S_{\mathcal{P}\sqcap\mathcal{T}}^{-1}(s-s') \Big\}^d .
\end{aligned}
$$

Eqs. (4.5) and (4.6), the separation property, and Eq. (4.8) finally combine to show Eq. (4.1) and Claim (a) follows from Lemma 4.1,

(b) Let M be a set obtained from D by replacing $n-r+1$ of its elements with a narrow and distant cluster. M contains only $r-1$ original observations so that each r-element subset of M contains one modification, in particular, each optimal set R^*. Let $\{R_1^*, \ldots, R_g^*\}$ be an associated MAP partition. Then some R_j^* contains at least one modification and Lemma 3.4 shows that the norm of m_j^* tends to infinity together with the compact cluster of replacements.

(c) The general assumption $r < n$ yields the estimate $2r - n \le r - 1$ which in turn shows that the first condition in (a) is fulfilled and that the given ϱ satisfies $1 - \varrho = (n-u)/(2r-n) \ge (n-u)/(r-1)$. It follows that the second condition in (a), too, is satisfied and that ϱ is the largest number s.th. $q_{u,\varrho} = 2r - n$. The claim on β_{mean} now follows from (a).

Claim (d) follows from Part (η) of the proof of Lemma 4.1.

The following corollary is a consequence of Theorem 4.4. We formulate it in the case of full covariance matrices. Combined with Cor. 3.3 it says that EMT is asymptotically robust on data sets with a well-separated, balanced cluster structure if the natural parameter g is used.

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

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

4.6 REMARKS. (a) The inequality $n - (1 - \varrho)(r - 1) \leq u$ required in Theorem 4.4(a) implies $u \geq n - r + 2$. That is, the sizes of the natural clusters must exceed the number of discarded elements in Theorem 4.4(a). Moreover, the assumptions of Part (c) imply that these sizes exceed twice the number of discarded elements.

(b) Although we have formulated the trimmed likelihood function and the EMT algorithm for general statistical models, we have stated and proved our robustness results only for various normal models. Extensions to other models are possible but not straightforward, in general. The proofs in the present paper depend on the crucial Lemma 2.12. We sketch how this lemma and much of the robustness theory can be extended to a whole family of elliptical distributions including the normal case leaving the details to the interested reader.

Denote the d-variate elliptical density function with radial function $\varphi\colon \mathbb{R}_+ \to \mathbb{R}_>$, mean vector $m \in \mathbb{R}^d$, and scale parameter $V \in \mathrm{PD}(d)$ by

$$f_{\varphi,m,V}(x) = \frac{1}{\sqrt{\det V}}\varphi\big((x - m)^{\mathrm{T}} V^{-1}(x - m)\big), \quad x \in \mathbb{R}^d,$$

In the normal case, $\varphi(s) = (2\pi)^{-d/2} e^{-s/2}$. Lemma 2.12 remains valid if the normal densities N_{m_j,V_j} are replaced with elliptical densities f_{φ,m_j,V_j} provided that

(i) φ is strictly decreasing and logarithmically concave.

This condition implies $\lim_{s\to\infty} s^t \varphi(s) = 0$ for all $t \in \mathbb{R}$. We infer that Proposition 2.13 holds true, i.e., a maximum of the HDBT-constrained trimmed likelihood function exists. Under (i), Theorem 3.2(a) remains also true for the estimates of the scale parameters. The estimated means returned

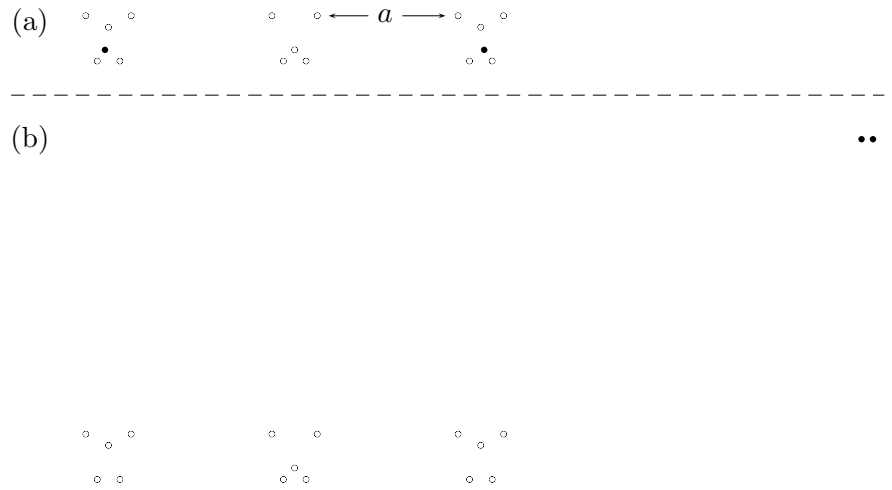


Figure 3: (a) The data set D and (b) its modification shown at the critical distance $a = 12.42$ where transition between robustness and breakdown occurs.

from any M-step lie in the convex hull of R. Theorem 3.5, too, remains valid under the condition (i).

Any concave function is differentiable from the right and, from (i), we infer $\lim_{s\to\infty} s(\ln\varphi)'(s) = -\infty$. This is needed to show also the robustness results stated in Theorem 3.2(b) and (c) for the estimates of the scale parameters. Finally, Lemma 4.1 (full case) remains true if the inequality (4.1) is replaced with

$$\min_{A\in \mathrm{PD}(d)} \big(\ln\det A - 2\ln\varphi\big(\mathrm{tr}\left\{W_T(\boldsymbol{\alpha})A^{-1}\right\}\big)\big)$$
$$\geq \max_{R\in\binom{D}{r}, R\supseteq T} \Big[\ln g^2 \det\frac{1}{c^2}W_{\mathcal{P}\cap R} - \frac{2}{r}\sum_{j=1}^{g}\sum_{x\in P_j\cap R}\ln\varphi\big((x-\overline{x}_{\mathcal{P}\cap R,j})^{\mathrm{T}} S_{\mathcal{P}\cap R}^{-1}(x-\overline{x}_{\mathcal{P}\cap R,j})\big)\Big].$$

In the normal case, this inequality reduces to inequality (4.1). Examples where the condition (i) is satisfied are light-tailed elliptical distributions with $\varphi(s) = c_1\cdot e^{-c_2\cdot s^p}$, $p\geq 1$, $c_2 > 0$.

4.7 Illustrations. Theorem 4.4 requires well-separated components in order to guarantee robustness. We give first an example which shows that such an assumption is necessary. Consider the two-dimensional data set D consisting of the $n = 17$ regular points

$$D = \left\{ \begin{pmatrix} -a-6 \\ 4 \end{pmatrix}, \begin{pmatrix} -a-4 \\ 3 \end{pmatrix}, \begin{pmatrix} -a-2 \\ 4 \end{pmatrix}, \begin{pmatrix} -a-5 \\ 0 \end{pmatrix}, \begin{pmatrix} -a-4.3 \\ 1 \end{pmatrix}, \begin{pmatrix} -a-3 \\ 0 \end{pmatrix} \right\}$$
$$\cup \left\{ \begin{pmatrix} -2 \\ 4 \end{pmatrix}, \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}$$
$$\cup \left\{ \begin{pmatrix} a+2 \\ 4 \end{pmatrix}, \begin{pmatrix} a+4 \\ 3 \end{pmatrix}, \begin{pmatrix} a+6 \\ 4 \end{pmatrix}, \begin{pmatrix} a+3 \\ 0 \end{pmatrix}, \begin{pmatrix} a+4.3 \\ 1 \end{pmatrix}, \begin{pmatrix} a+5 \\ 0 \end{pmatrix} \right\}.$$

displayed in Fig. 3(a). As expected, the classical ML estimate with the full *homo*scedastic model and $g = 3$ determines the left, the middle, and the right components. Now assume that the two stars * in Fig. 3(a) have been grossly mismeasured as shown in Fig. 3(b). Their mutual distance is now 1. Then for $a \geq 12.43$, the homoscedastic EMT algorithm with $g = 3$ and $r = 15$ produces a reasonable estimate discarding both outliers. However for $a \leq 12.42$, it discards the two (original) observations interior to the convex hulls of the right and middle clusters thus producing as MTLE two slim horizontal components complemented by a component determined by the two outliers. Its negative log-likelihood is 73.3324 whereas that of the natural solution is 73.3355. Using upper and lower bounds on the Mahalanobis distance square appearing on the left side of Eq. (4.2) we estimated that the lower bound of values a for which the separation property is satisfied lies between 1763 and 3841. For the spherical model this interval is $[60, 61]$.

This special and contrived example represents the worst case. First, the two replaced elements are purposefully chosen. Second, they are replaced with two close outliers aligned horizontally. Third, a further removal of two points makes the remaining original points two narrow horizontal clusters which determine the solution if there is not enough separation between the three original clusters.

The MTLE may be robust also w.r.t. the means even when applied to substantially overlapping components. To underpin this contention, we randomly generate in a second experiment 54 four-dimensional data sets of 150+100+50 regular data points, each, randomly drawn from N_{0,I_4}, N_{3e_1,V_2}, and N_{6e_2,V_3}, respectively, where

$$V_2 = \begin{pmatrix} 1 & & & \\ -1.5 & 4 & & \\ -1 & -2 & 9 & \\ 0 & -1 & 3 & 16 \end{pmatrix}, \; V_3 = \mathrm{diag}(1, 16, 9, 4).$$

$n-r$ / g	0	1	2	3	4	5	6	7	8
1	**222.1**	**219.4**	216.1	212.6	208.5	205.4	202.4	199.4	195.8
2	227.8	228.6	**210.9**	**207.0**	**203.9**	**200.9**	**197.7**	**194.8**	**191.4**
3	229.4	225.7	217.3	213.8	211.8	208.1	206.2	201.9	199.4
4	238.9	235.5	227.7	222.5	219.8	216.1	212.4	209.0	207.0

Table 2: Table of trimmed BIC values for various numbers of components and of discarded elements. The boldface numbers mark the columnwise minima. They stabilize at the values $g = 2$ and $n - r = 2$.

To each data set we add 30 "outliers" randomly drawn from $\mathrm{N}_{0,10^4 I_4}$. It turns out that the normal MTLE (best among 500 replications of the EMT algorithm) with full covariance matrices subject to the HDBT constraints (1.1) with $c = 200$ and with the parameters $g = 3$ and $r = 300, 295, 290, 285, 280$ correctly identifies all outliers in all 270 runs. These findings plead again for its robustness. The assumption of too many outliers is not harmful. Of course, we must not assume less than 30 outliers.

It is hardly possible to predict the number of replications necessary to achieve the optimum. In most cases, the optimum among the 500 replications appears during the first 50 trials, often earlier, sometimes substantially later. One replication consists on the average of about 50 EMT-steps for which our C++ implementation of the algorithm needs 0.04 sec on a 2 GHz processor.

Mixture modelling is also used for the purpose of data clustering. As our last illustration, we randomly draw 60 samples from the normal mixture $0.8\,\mathrm{N}_{2e_1,I_d} + 0.2\,\mathrm{N}_{-2e_1,I_d}$ and add two outliers at the points $(7, 0.1)$ and $(7, -0.1)$ in order to discuss the effect of different choices of the number of discarded elements. Moreover, we compare the clusterings created by EMT and by Fraley and Raftery's (2002) MCLUST.

We first determine the favorite solution according to Section 2.14. Table 2 shows the BIC values for the optimal solutions w.r.t. criterion (2.12). The minima in each column stabilize at the value $g = 2$ and this takes place at $n-r = 2$. We conclude that the data set was sampled from a two-component mixture and contains two outliers. Figure 4 presents the MAP clusterings for the optimal mixtures obtained with $g = 2$ components and zero, one, two, and eight discarded elements (asterisks). The first two images reveal a marked instability as long as the number of outliers exceeds the number of discarded elements. Everything can happen. Whereas the clustering that retains all elements looks reasonable, although the outliers could not be recognized, the clustering that discards one element is in disorder. The

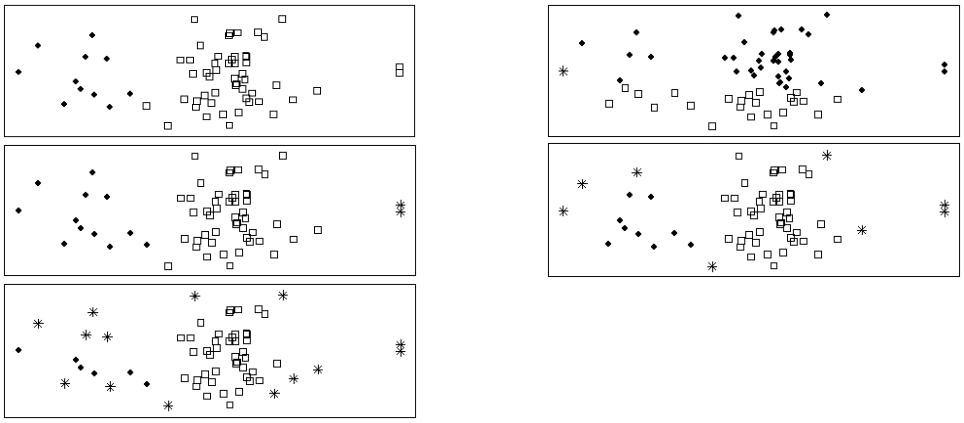


Figure 4: Sixty data points randomly sampled from $0.8\,\mathrm{N}_{2e_1,I_2} + 0.2\,\mathrm{N}_{-2e_1,I_2}$ with two additional outliers. The images show five clusterings determined by MAP discriminant analysis on the basis of estimated heteroscedastic, full normal mixtures. First two rows: mixtures estimated with the present EMT, two components and zero, one, two, and eight discarded elements shown as asterisks. The left image in the second row belongs to the favorite solution selected with trimmed BIC, see Table 2, according to Section 2.14. Bottom: a typical mixture estimated with mclustBIC (VVV) for two components plus noise (asterisks). There is a spurious component and the solution clearly overestimates the number of noise elements.

retained objects are separated by a horizontal line instead of a vertical one. The situation changes suddenly as the number of discarded elements reaches the number of outliers, second row, left. This solution is correct. Increasing the number of discarded elements further does not dramatically change the clustering, see the second row, right. The method of Section 2.14 returns a reasonable result although we discard eight elements while there are actually only two outliers. It just discards also a few extreme elements in each cluster.

The procedure mclustBIC of the MCLUST package (Version 3.3.1) implements EM algorithms for various normal models and has three main characteristics: (i) it initializes the EM algorithm with a partition obtained from normal ML-clustering optimized by hierarchical agglomeration; (ii) it provides an additional uniform component for handling noise; (iii) it uses the BIC to determine the number of components. We replicated mclustBIC 100 times with two full normal components (VVV) and with random initializations of the outliers at a Bernoulli rate of 3.2% (the true value). All MAP clusterings contain spurious clusters similar to that in the image on the bottom of Figure 4. Obviously, the algorithm gets trapped in unbalanced

solutions containing small, elongated components, a result of the overfitting m.l.e.. Gross outliers usually do not conform to a statistical model, not even to the uniform one. This is another reason why mclustBIC does not recognize the correct two clusters taking instead a large number of regular observations for noise (asterisks).

4.8 Summary and Discussion. The design of automatic methods for data analysis is still considered a true challenge, in particular in view of real applications. Dougherty and Brun (2004) write: "Although used for many years, data clustering has remained highly problematic – and at a very deep level." We believe that methodological progress in data analysis will arise from statistical models and assumptions by means of statistical paradigms. Algorithms designed for homoscedastic models are stable but should be applied only when it is a priori known that the underlying data-generating mechanism enjoys this property. Most of the data analytic world is heteroscedastic and here the phenomenon of spurious solutions appears. Although it has been known for a long time, algorithms have essentially ignored it.

The present communication focuses on two main topics: spurious solutions in connection with heteroscedastic models and gross outliers. We solve the first problem by proposing a compromise between model fit measured by the likelihood and model balance measured by the HDBT ratio. This leads to a method that does not primarily seek a large likelihood. The examples and illustrations clearly show its advantage. Our theorems in Sections 3 and 4 show that balance offers also a solution to the second problem, outlier robustness. A main characteristic of many outliers is that they do not conform to any statistical population, see also the discussion of outlier types by Ritter and Gallegos (1997). Otherwise, they could be modeled with an additional component of the mixture. Trimming the likelihood offers a simpler and more effective way of dealing with gross contaminations than mixture models with outliers do. Our main Theorems 3.2 and 4.4 state that the HDBT-constrained MTLE provides substantial protection against gross outliers while being at the same time affine equivariant.

The weakness of our method is the high time consumption of EMT which it shares with the EM algorithm. Real applications of only medium size require hundreds of thousands of replications of EMT runs. Although the EM algorithm, mainly due to its elegance, has superseded all other optimization paradigms since its appearance in mixture estimation the question still remains whether there is no computationally more efficient tool. Finally, our

method could be sharpened by other population models than the normal, such as heavy-tailed or asymmetric ones. But this would be at the cost of an even higher time consumption.

A Appendix

A.1 LEMMA. *Let $A \in \mathbb{R}^{r\times g}$ be a matrix with all entries ≥ 0.*

(a) The function $\Psi : \Delta_{g-1} \to \mathbb{R}\cup\{-\infty\}$, $\Psi(\mathbf{u}) = \sum_i \ln(A\mathbf{u})_i$, is concave.

(b) If no row of A vanishes and if the g columns of A are affine independent then Ψ is real valued and strictly concave in the interior of Δ_{g-1}.

PROOF. (a) Each of the summands $\mathbf{u} \mapsto \ln(A\mathbf{u})_i$ is concave as a function with values in $\mathbb{R} \cup \{-\infty\}$ and so is their sum.

(b) Under the first assumption of (b) all summands $\mathbf{u} \mapsto \ln(A\mathbf{u})_i$ are finite in the interior of Δ_{g-1} and so is their sum Ψ. Under the second the mapping $\mathbf{u} \mapsto A\mathbf{u}$ is one to one. Hence, if $\mathbf{u} \neq \mathbf{v}$ then there is an index i such that $(A\mathbf{u})_i \neq (A\mathbf{v})_i$ and, by strict concavity of the logarithm, we have $\ln A\big\{\frac{1}{2}\mathbf{u}+\frac{1}{2}\mathbf{v}\big\}_i = \ln\big\{\frac{1}{2}(A\mathbf{u})_i+\frac{1}{2}(A\mathbf{v})_i\big\} > \frac{1}{2}\big\{\ln(A\mathbf{u})_i+\ln(A\mathbf{v})_i\big\}$ and Claim (b) follows.

The following formula is found in Hathaway (1986).

A.2 LEMMA. *Let $(R, \mathbf{u}, \gamma)$ be any parameter triple. With the weight matrix $\mathbf{w}$ defined by Eq. (1.6), we have the representation*

$$\begin{aligned}&\ln f[\mathbf{x}_R \mid \mathbf{u}, \gamma]\\ &= \sum_\ell w_\ell(R)\ln u_\ell - \sum_\ell \sum_{x\in R} w_\ell(x)\ln w_\ell(x) + \sum_\ell \sum_{x\in R} w_\ell(x)\ln f_{\gamma_\ell}(x).\end{aligned}$$

Moreover, $\ln f[\mathbf{x}_R \mid \mathbf{u}, \gamma] \leq \sum_\ell \sum_{x\in R} w_\ell(x)\ln f_{\gamma_\ell}(x)$.

PROOF. Since $\sum_\ell w_\ell(x) = 1$, Eq. (2.2) implies

$$\begin{aligned}\ln f[\mathbf{x}_R \mid \mathbf{u}, \gamma] &= \sum_{x\in R} \ln \sum_{j=1}^{g} u_j f_{\gamma_j}(x) = \sum_{x\in R}\sum_\ell w_\ell(x)\ln \sum_{j=1}^{g} u_j f_{\gamma_j}(x)\\ &= \sum_{x\in R}\sum_\ell w_\ell(x)\ln\frac{u_\ell f_{\gamma_\ell}(x)}{w_\ell(x)}.\end{aligned}$$

This is the first claim and the second follows from it by the entropy inequality applied to the probabilities $(w_\ell(x))_\ell$ and $(u_\ell)_\ell$.

The following lemma is of Steiner's type. We omit its elementary proof.

A.3 Lemma. *Let T be a (finite) data set in $\mathbb{R}^d$, let $w = (w(x))_{x\in T}$ be a family of real numbers such that $w(T) := \sum_x w(x) > 0$, let $b \in \mathbb{R}^d$, and let $\overline{x}(w) = \frac{1}{w(T)} \sum_{x\in T} w(x)x$, the weighted mean. Then*

$$\begin{aligned}&\sum_x w(x)(x-b)(x-b)^{\mathrm{T}}\\ =&\sum_x w(x)(x-\overline{x}(w))(x-\overline{x}(w))^{\mathrm{T}} + w(T)\cdot(\overline{x}(w)-b)(\overline{x}(w)-b)^{\mathrm{T}}.\end{aligned}$$

In particular, $\sum_x w(x)(x-b)(x-b)^{\mathrm{T}} \succeq \sum_x w(x)(x-\overline{x}(w))(x-\overline{x}(w))^{\mathrm{T}}$.

Our next lemma reduces a global weighted "within-samples" SSP matrix to a sum of local weighted "within-samples" SSP matrices and a weighted "between-samples" SSP matrix. In its proof we use the identity

$$\sum_{j,\ell,t} a_j a_\ell a_t (m_\ell - m_j)(m_t - m_j)^{\mathrm{T}} = A \sum_{j<\ell} a_\ell a_j (m_\ell - m_j)(m_\ell - m_j)^{\mathrm{T}} \quad \text{(A.1)}$$

valid for any real numbers a_t, $A = \sum a_t$, and any vectors m_t. Indeed, exploiting symmetries, we have

$$\begin{aligned}&\sum_{j,\ell,t} a_j a_\ell a_t (m_\ell - m_j)(m_t - m_j)^{\mathrm{T}}\\ &\quad= \sum_{j,\ell,t} a_j a_\ell a_t (m_\ell m_t^{\mathrm{T}} - m_j m_t^{\mathrm{T}} - m_\ell m_j^{\mathrm{T}} + m_j m_j^{\mathrm{T}})\\ &\quad= \sum_{j,\ell,t} a_j a_\ell a_t (m_j m_j^{\mathrm{T}} - m_\ell m_j^{\mathrm{T}}) = A \sum_{j,\ell} a_j a_\ell (m_j - m_\ell) m_j^{\mathrm{T}}\\ &\quad= A \sum_{j<\ell} a_j a_\ell (m_j - m_\ell) m_j^{\mathrm{T}} + A \sum_{j>\ell} a_j a_\ell (m_j - m_\ell) m_j^{\mathrm{T}}\\ &\quad= A \sum_{j<\ell} a_\ell a_j (m_\ell - m_j)(m_\ell - m_j)^{\mathrm{T}}.\end{aligned}$$

A.4 Lemma. *Let $g \geq 1$, let T be a Euclidean data set, let $\{T_1, \ldots, T_g\}$ be a partition of T, let $w = (w(x))_{x\in T}$ be a family of real numbers s.th. $w(T) > 0$, and put $w_j = w_{|T_j}$. With the notation agreed upon in Section 2.10 we have*

$$W_T(w) = \sum_{j=1}^{g} W_{T_j}(w_j) + \frac{1}{w(T)} \sum_{1\leq j<\ell\leq g} w(T_j)w(T_\ell)\big(\overline{x}_{T_j}(w_j) - \overline{x}_{T_\ell}(w_\ell)\big)\big(\overline{x}_{T_j}(w_j) - \overline{x}_{T_\ell}(w_\ell)\big)^{\mathrm{T}}.$$

PROOF. By Lemma A.3 applied to all subsets T_j,

$$W_T(w) = \sum_j \sum_{x\in T_j} w(x)(x-\overline{x}_T(w))(x-\overline{x}_T(w))^{\mathrm{T}}$$

$$= \sum_j W_{T_j}(w_j) + \sum_j w(T_j)(\overline{x}_{T_j}(w_j)-\overline{x}_T(w))(\overline{x}_{T_j}(w_j)-\overline{x}_T(w))^{\mathrm{T}}. \quad \text{(A.2)}$$

From $\sum_\ell w(T_\ell) = w(T)$ we infer

$$\overline{x}_T(w) - \overline{x}_{T_j}(w_j) = \frac{1}{w(T)}\sum_\ell w(T_\ell)\overline{x}_{T_\ell}(w_\ell) - \overline{x}_{T_j}(w_j)$$

$$= \frac{1}{w(T)}\sum_\ell w(T_\ell)\big(\overline{x}_{T_\ell}(w_\ell) - \overline{x}_{T_j}(w_j)\big)$$

and, hence, the second sum on the right of (A.2) equals

$$\frac{1}{w(T)^2}\sum_j w(T_j)\Big(\sum_\ell w(T_\ell)(\overline{x}_{T_\ell}(w_\ell)-\overline{x}_{T_j}(w_j))\Big)\Big(\sum_\ell w(T_\ell)(\overline{x}_{T_\ell}(w_\ell)-\overline{x}_{T_j}(w_j))\Big)^{\mathrm{T}}$$

$$= \frac{1}{w(T)^2}\sum_{j,\ell,t} w(T_j)w(T_\ell)w(T_t)(\overline{x}_{T_\ell}(w_\ell)-\overline{x}_{T_j}(w_j))(\overline{x}_{T_t}(w_t)-\overline{x}_{T_j}(w_j))^{\mathrm{T}}.$$

The claim now follows from Eq. (A.2) and the identity (A.1).

A.5 LEMMA. *For any nonempty subset $U \subseteq D$ and $\boldsymbol{\alpha} \in \mathcal{M}(U, g')$, we have*

$$W_U = W_U(\boldsymbol{\alpha}) + \frac{1}{|U|}\sum_{1\le h<k\le g'} \alpha_h(U)\alpha_k(U)\big(\overline{x}_U(\alpha_h)-\overline{x}_U(\alpha_k)\big)\big(\overline{x}_U(\alpha_h)-\overline{x}_U(\alpha_k)\big)^{\mathrm{T}}.$$

PROOF. This follows from Lemma A.4 if we put $g = g'$, $T_j = U$ for all j, $T =$ the disjoint union of all T_j's, and $w(x) = \alpha_j(x)$, $x \in T_j$. (The family w is thus the linearization of the weight matrix $\boldsymbol{\alpha}$.) It follows $w(T_j) = \alpha_j(U)$, $\overline{x}_{T_j}(w) = \overline{x}_U(\alpha_j)$, $w(T) = \sum_{x\in T} w(x) = \sum_j \sum_{x\in T_j} \alpha_j(x) = |U|$, and $\overline{x}_T(w) = \overline{x}_U$, and the claim is a term-by-term translation of Lemma A.4.

A.6 LEMMA. *Let $E = \{x_0, \dots, x_d\}$ be a set of $d+1$ points in $\mathbb{R}^d$. If its convex hull contains a ball of radius r around its mean vector then $W_E \succeq 2\,r^2 I_d$.*

PROOF. Without restriction let the mean of E be the origin and let W_E be diagonal. By assumption, the centered r-ball $B_r(0)$ satisfies

$$B_r(0) \subseteq \Big\{ \sum_{i=0}^d \lambda_i x_i \mid \lambda_i \ge 0, \sum_{i=0}^d \lambda_i = 1 \Big\}.$$

Orthogonal projection to the coordinate axes shows $[-r,r] \subseteq \{ \sum_{i=0}^d \lambda_i x_{i,k} \mid \lambda_i \geq 0, \sum_{i=0}^d \lambda_i = 1\}$, $1 \leq k \leq d$, i.e., the interval $[-r,r]$ is contained in the convex hull of the set $\{x_{0,k}, \ldots, x_{d,k}\}$, i.e., $\min_i x_{i,k} \leq -r$, $\max_i x_{i,k} \geq r$ for all k. This implies $W_E(k,k) = \sum_{i=0}^d x_{i,k}^2 \geq 2\,r^2$.

A.7 Lemma. *Let $g \geq 2$. For any stochastic matrix $(a_{h,j})_{h,j} \in \mathbb{R}^{(g-1)\times g}$ we have*

$$\sum_j \min_h \sum_{\ell \neq j} a_{h,\ell} \geq 1.$$

Proof. For each column j, consider the line h with minimal sum $\sum_{\ell\neq j} a_{h,\ell}$. Since there are g columns but only $g-1$ lines, the pigeon hole principle implies that some line appears for two j's and the two related sums cover all elements in this line.

A.8 Lemma. *Let $g \geq 2$.*

(a) The maximum of the function $\mathbf{a} \mapsto \sum_{1\leq h<k\leq g} a_h a_k$ w.r.t. all vectors $\mathbf{a} = (a_1, \ldots, a_g) \in [0,1]^g$ is attained at the point $(1/g, \ldots, 1/g)\sum_k a_k$ and has the value $\frac{g-1}{2g}\big(\sum_k a_k\big)^2$.

(b) Let $\varrho \leq 1/g$ be a nonnegative real number. The minimum of the expression

$$\sum_{1\leq k<g} \beta_k \sum_{1\leq j<\ell\leq g} a_{k,j} a_{k,\ell}$$

w.r.t. all probability vectors $\beta \in \mathbb{R}^{g-1}$ and all stochastic matrices $A = (a_{k,j})_{k,j} \in \mathbb{R}^{(g-1)\times g}$ such that $\beta^T A \geq \varrho$ (pointwise) is assumed at $\beta^* = (1)$ and $A^* = (1-\varrho, \varrho)$ if $g = 2$ and at $\beta^* = \big(\frac{1-2\varrho}{g-2}, \ldots, \frac{1-2\varrho}{g-2}, 2\varrho\big)^T$ and

$$A^* = \begin{pmatrix} 1 & & & & 0 & 0 \\ & 1 & & 0 & 0 & 0 \\ & & \ddots & & 0 & 0 \\ & 0 & & 1 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1/2 & 1/2 \end{pmatrix},$$

if $g \geq 3$. The minimum is the number κ_ϱ defined before Sect. 4.3.

Proof. (a) It is sufficient to determine the maximum of the sum $\sum_{1\leq h<k\leq g} a_h a_k$ subject to the constraint $\sum_t a_t = 1$. Now, $2\sum_{1\leq h<k\leq g} a_h a_k = \big(\sum_k a_k\big)^2 - \|\mathbf{a}\|^2 = 1 - \|\mathbf{a}\|^2$ and the point on the plane $\sum_t a_t = 1$ with minimal distance to the origin is $(1/g, \ldots, 1/g)$. The claim follows.

(b) The case $g = 2$ being simple we let $g \geq 3$. Using Lemma A.7 we estimate

$$2\sum_{1\leq k<g}\beta_k\sum_{1\leq j<\ell\leq g}a_{k,j}a_{k,\ell} = \sum_{1\leq k<g}\beta_k\sum_{j\neq\ell}a_{k,j}a_{k,\ell} = \sum_j\sum_k\beta_k a_{k,j}\sum_{\ell\neq j}a_{k,\ell}$$
$$\geq \sum_j\sum_k\beta_k a_{k,j}\min_h\sum_{\ell\neq j}a_{h,\ell} \geq \varrho\sum_j\min_h\sum_{\ell\neq j}a_{h,\ell} \geq \varrho = 2\kappa_\varrho.$$

That is, the least upper bound is $\geq \kappa_\varrho$. The remaining claims are plain.

The following lemma frees the estimates from the stochastic matrix $\boldsymbol{\alpha}$ replacing it with SSP matrices. Part (b) could be proved by the extremal property of the partitions contained in $\mathcal{M}(T, g')$. The present proof uses a disintegration technique that provides more insight.

A.9 LEMMA. *Let $\mathcal{P}$ be a partition of D, let T be a nonempty subset of D, let $g' \geq 2$, and let $\boldsymbol{\alpha} \in \mathcal{M}(T, g')$.*

(a) There exists a finite sequence $(r_m)_{m=1}^t$, $t \leq g'|T|$, of strictly positive numbers r_m such that $\sum r_m = 1$ and a finite sequence $\big(\mathcal{T}^{(m)}\big)_{m=1}^t$ of partitions $\mathcal{T}^{(m)} = \{T_1^{(m)}, \ldots, T_{g'}^{(m)}\}$ of T in g' subsets (some may be empty) such that $W_{\mathcal{P}\cap T}(\boldsymbol{\alpha}) \succeq \sum_{m=1}^t r_m W_{\mathcal{P}\sqcap\mathcal{T}^{(m)}}$.

(b) Assume that, for each m, there exist indices k and j such that $|P_j \cap T_k^{(m)}| > d$. Then

(i) The matrix $W_{\mathcal{P}\cap T}(\boldsymbol{\alpha})$ is regular and $W_{\mathcal{P}\cap T}(\boldsymbol{\alpha})^{-1} \preceq \sum_{m=1}^t r_m W_{\mathcal{P}\sqcap\mathcal{T}^{(m)}}^{-1}$.

(ii) Given $y \in \mathbb{R}^d$ there exists a partition $\mathcal{T}$ of T in g' clusters such that, for all $\boldsymbol{\alpha} \in \mathcal{M}(T, g')$, $y^T W_{\mathcal{P}\cap T}(\boldsymbol{\alpha})^{-1} y \leq y^T W_{\mathcal{P}\sqcap\mathcal{T}}^{-1} y$.

PROOF. (a) Let r_1 be the smallest non-zero entry in the matrix $(\alpha_k(x))$. In each line x, subtract the number r_1 from any of its smallest non-zero entries to obtain a new matrix $\boldsymbol{\alpha}^{(1)}$ with entries ≥ 0. All its row sums are equal and it contains at least one additional zero. Let $T_k^{(1)} = \big\{x \in T \mid \alpha_k^{(1)}(x) \neq \alpha_k(x)\big\}$, $k \leq g'$. Now continue this procedure with $\boldsymbol{\alpha}^{(1)}$ instead of $\boldsymbol{\alpha}$ and so on. It stops after at most $g'|T|$ steps with the zero matrix and we have constructed a representation

$$\alpha_k(x) = \sum_{m=1}^t r_m \mathbf{1}_{T_k^{(m)}}(x).$$

of $\boldsymbol{\alpha}$. Moreover, $\sum r_m = 1$. Define $\mathcal{T}^{(m)} = \big(T_1^{(m)}, \ldots, T_{g'}^{(m)}\big)$. Summing up over $x \in T$ we have

$$\begin{aligned}
&W_{P_j\cap T}(\alpha_k)\\
&= \sum_{x\in P_j\cap T} \alpha_k(x)(x - \overline{x}_{P_j\cap T}(\alpha_k))(x - \overline{x}_{P_j\cap T}(\alpha_k))^{\mathrm{T}}\\
&= \sum_{m=1}^{t} r_m \sum_{x\in P_j\cap T} \mathbf{1}_{T_k^{(m)}}(x)(x - \overline{x}_{P_j\cap T}(\alpha_k))(x - \overline{x}_{P_j\cap T}(\alpha_k))^{\mathrm{T}}\\
&= \sum_{m=1}^{t} r_m \sum_{x\in P_j\cap T_k^{(m)}} (x - \overline{x}_{P_j\cap T}(\alpha_k))(x - \overline{x}_{P_j\cap T}(\alpha_k))^{\mathrm{T}}\\
&\succeq \sum_{m=1}^{t} r_m \sum_{x\in P_j\cap T_k^{(m)}} \big(x - \overline{x}_{P_j\cap T_k^{(m)}}\big)\big(x - \overline{x}_{P_j\cap T_k^{(m)}}\big)^{\mathrm{T}}
\end{aligned}$$

by Steiner's classical formula. Now sum up over j and k to obtain (a).

(b) By assumption, each SSP matrix $W_{\mathcal{P}\sqcap\mathcal{T}^{(m)}}$ is regular. Estimate (i) therefore follows from (a) and monotone decrease and convexity of the matrix inversion on PD(d), cf. Marshall and Olkin (1979), E.7.c.

(ii) Among the finitely many partitions of T there is some $\mathcal{T}$ such that $y^{\mathrm{T}}W_{\mathcal{P}\sqcap\mathcal{T}}^{-1}y$ is maximal. Hence, by (i),

$$y^{\mathrm{T}}W_{\mathcal{P}\cap T}^{-1}(\boldsymbol{\alpha})y \le \sum_{m=1}^{t} r_m y^{\mathrm{T}}W_{\mathcal{P}\sqcap\mathcal{T}^{(m)}}^{-1}y \le y^{\mathrm{T}}W_{\mathcal{P}\sqcap\mathcal{T}}^{-1}y.$$

References

Chrétien, S. and Hero, A.O., III (2000). Kullback proximal algorithms for maximum likelihood estimation. *IEEE Trans. Inf. Theory*, **46**, 1800–1810.

Davies, P.L. (1987). Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.*, **15**, 1269–1292.

Day, N.E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., Ser. B*, **39**, 1–38.

Dennis, J.E., Jr. (1982). Algorithms for nonlinear fitting. In *Nonlinear Optimization 1981 (Procedings of the NATO Advanced Research Institute held at Cambridge in July 1981)*, (M.J.D. Powell, ed.). Academic Press.

DONOHO, D.L. and HUBER, P.J. (1983). The notion of a breakdown point. In *A Festschrift for Erich L. Lehmann*, (P.J. Bickel, K.A. Doksum and J.L. Hodges Jr., eds.), The Wadsworth Statistics / Probability Series, 157–184. Wadsworth, Belmont, CA.

DOUGHERTY, E.R. and BRUN, M. (2004). A probabilistic theory of clustering. *Patt. Rec.*, **37**, 917–925.

FISHMAN, G.S. (1996). *Monte Carlo.* Springer, New York.

FRALEY, C. and RAFTERY, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, **97**, 611–631.

GALLEGOS, M.T. and RITTER, G. (2005). A robust method for cluster analysis. *Ann. Statist.*, **33**, 347–380.

GALLEGOS, M.T. and RITTER, G. (2009). Trimming algorithms for clustering contaminated grouped data and their robustness. *Adv. Data Anal. Classif.*, **3**, 135–167.

GALLEGOS, M.T. and RITTER, G. (2010). Using combinatorial optimization in model-based trimmed clustering with cardinality constraints. *Comput. Statist. Data Anal.*, **54**, 637–654. DOI 10.1016/j.csda.2009.08.023.

GARCIA-ESCUDERO, L.A. and GORDALIZA, A.. (1999). Robustness properties of k-means and trimmed k-means. *J. Amer. Statist. Assoc.*, **94**, 956–969.

HADI, A.S. and LUCEÑO, A. (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Comput. Statist. Data Anal.*, **25**, 251–272.

HASSELBLAD, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics*, **8**, 431–444.

HATHAWAY, R.J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.*, **13**, 795–800.

HATHAWAY, R.J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statist. Probab. Lett.*, **4**, 53–56.

HENNIG, C. (2004). Breakdown points for maximum likelihood estimators of location-scale mixtures. *Ann. Statist.*, **32**, 1313–1340.

HODGES, J.L., JR. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 163–186. Univ. California Press.

HUBER, P.J. (1981). *Robust Statistics.* Wiley, New York.

KÉRIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā, Series A*, **62**, 49–66.

KIEFER, N.M. (1978). Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica*, **46**, 427–434.

KNUTH, D.E. (1981). *The Art of Computer Programming*, volume 2. Addison-Wesley.

MA, J. and FU, S. (2005). On the correct convergence of the EM algorithm for Gaussian mixtures. *Patt. Rec.*, **38**, 2602–2611.

MARSHALL, A.W. and OLKIN, I. (1979). *Inequalities: Theory of Majorization and its Applications, v. 143 of Mathematics in Science and Engineering.* Academic Press, New York.

MARTINET, B. (1970). Régularisation d'inéquations variationnelles par approximations successives. *Rev. Francaise d'Inform. et de Recherche Opérationnelle,* **3**, 154–179.

MCLACHLAN, G.J. and BASFORD, K.E. (1988). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

McLachlan, G.J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In *Advances in Pattern Recognition*, *Lecture Notes in Computer Science*, **1451**, 658–666. Springer.

McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Neykov, N., Filzmoser, P., Dimova, R. and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Comput. Statist. Data Anal.*, **52**, 299–308.

Neykov, N.M. and Neytchev, P.N. (1990). A robust alternative of the maximum likelihood estimator. In *COMPSTAT 1990 – Short Communications*, 99-100.

Peters, B.C. Jr. and Walker, H.F. (1978). An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.*, **35**, 362–378.

Redner, R.A. and Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195–239.

Ritter, G. and Gallegos, M.T. (1997). Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Patt. Rec. Lett.*, **18**, 525–539.

Rockafellar, R.T. (1976). Monotone operators and the proximal point algorithm. *SIAM J. Contr. Optim.*, **14**, 877–898.

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, (W. Grossmann, G.Ch. Pflug, I. Vincze and W. Wertz, eds.), **8B**, 283–297. Reidel.

Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–223.

Toma, A. (2007). Minimum Hellinger distance estimators for some multivariate models: influence functions and breakdown point results. *Comptes Rendus Mathematique*, **345**, 353–358.

María Teresa Gallegos
Institute for Data Analysis
Salzweg, Germany

Gunter Ritter
Fakultät für Informatik und Mathematik
Universität Passau, Germany
E-mail: ritter@fim.uni-passau.de

Paper received: May 2008; revised: November 2009.