# A trimmed version of the EM–algorithm for contaminated mixtures

María Teresa Gallegos and Gunter Ritter*
Fakultät für Informatik und Mathematik
Universität Passau, Germany

September 5, 2007

**Abstract** We establish a mixture model with spurious outliers and derive its maximum–integrated–likelihood estimator. It may be computed by a trimmed version of the EM–algorithm which we call the EMT–algorithm. We analyze its properties and compute the universal breakdown values of the estimator for normal mixtures. Estimation of the covariance matrix turns out to be robust.

**AMS subject classifications** Primary 62H12, secondary 62F35

**Key words and phrases** Robust ML–estimation in mixture models, multivariate data, trimming algorithm, EM-algorithm, breakdown point

## 1 Introduction

### 1.1 History and background

Multimodal distributions arise in particular when data emanate from different causes. They occur, e.g., in pattern recognition, image processing, speech recognition, classification, and clustering. For some examples see the literature cited in the Introduction of Redner and Walker [23]. 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, properties (consistency, asymptotic normality), efficient computation, and robustness have been investigated in the past. Day [3], Sect. 7, notes that the m.l.e. in the strict sense always fails to exist in the normal, heteroscedastic case. It is sufficient to center one component at one of the data points and have its variance tend to zero to see the unboundedness of the likelihood function. On the other hand, under

*E-mail: ritter@fim.uni-passau.de.

regularity conditions and in the presence of sufficiently many data a local maximum which is even strongly consistent and asymptotically efficient always exists, see Kiefer [16] and the literature cited there. The situation is simpler in the normal *homoscedastic* case (common covariance matrix). Day [3] states that solutions to the pooled m.l. equations exist if the data set is not too small. In fact, there exists a global maximum of the likelihood function. This was proved by Hathaway [11] for one-dimensional data in a more general context and is also true in the multivariate context. Moreover, Hathaway proved strong consistency. If the mixture components are poorly separated then there exist almost always more solutions to the likelihood equations, Day [3]. The pooled m.l.e. is equivariant w.r.t. affine transformations. An important and difficult problem is that of selecting the number of components. Kéribin [15] 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.

Hasselblad [10] and Day [3] designed alternating algorithms for computing the m.l.e. in the hetero– and homoscedastic cases. Dempster et al. [4] found 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 [1] embedded EM in the general scheme of PPA–algorithms, Martinet [19] and Rockafellar [24].

Parameter estimation in mixture models can be severely affected by outliers. A cluster of remote outlying observations will, as a rule, establish a component of its own. Although the estimate of the mixture *distribution* in the sense of a suitable distance measure on the convex set of probability measures may still be close to the original, the estimate of the *parameters* of one component goes astray, at least if a fixed number of components is assumed. This fact makes robust parameter estimation in mixture models a difficult problem. It has been taken up mainly in recent years.

- McLachlan and Basford [20] propose robust estimation of the parameters of the components, e.g., by means of Huber's [14] robust M–estimators.
- McLachlan and Peel [21, 22] use mixtures of $t$–distributions (or Pearson's type VII distributions) instead of normal mixtures.
- Fraley and Raftery [7] propose an additional component uniform on the convex hull of the data in order to accommodate outliers.

However, it turns out that these methods are effective for moderate outliers only, see Hennig [12]. In fact, one gross outlier causes one mean to break down in all three methods. Actually, breakdown robustness and the general heteroscedastic mixture model with a fixed number of clusters do not go well together since a small, remote cluster of outliers looks like a regular cluster breaking one mean down. Some constraint on the covariance structure is needed. Hennig, therefore, resorts to the mixture model constrained to the set of all variances $\geq v_0$ for some $v_0 > 0$ and proposes to

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

He shows for one–dimensional data that his method adds breakdown robustness to the m.l.e., see [12], Theorem 4.11. Besides $v_0$, his estimator needs a second parameter to be carefully chosen.

We take here a different approach to robustness of parameter estimation in mixture models and restrict matters to homoscedasticity. In Section 2, we adapt the classification model [8] with "spurious outliers" and (at least) $r$ regular elements to mixtures. We derive the ML–criterion of this outlier model which is again of the trimming type. For its maximization, we design an algorithm that consists of the iterative application of an EM–step and a trimming step which we call the EMT–algorithm. In Sections 2.3–2.8, we study its convergence properties which are similar to those of the EM–algorithm.

The aim of trimming is robustness and we compute in Section 3 various (replacement) breakdown points of EMT for normal mixtures. Theorem 3.2 says that the usual breakdown value of the common covariance matrix is large. Unfortunately, the same cannot be said about the breakdown value of EMT for the means, Theorem 3.5: whereas the criterion sustains one gross outlier, there are data sets such that one mean breaks down if two observations are suitably replaced with gross outliers. One reason for this misbehavior is the stringency of the usual breakdown point. It requires proper behavior of the estimator even in the presence of data sets that are very unlikely to emanate from a mixture distribution with separated components. We will show in a forthcoming communication that the asymptotic *restricted breakdown point* [8] of EMT for the means restricted to data sets that consist of sufficiently separated clusters is positive. As a tool in some of our proofs we use the MAP–clustering associated with the mixture.

### 1.2 General notation

Given two integers $m \leq n$, the symbol $m\,..\,n$ designates the set of integers $k$, $m \leq k \leq n$. The set of all $r$–element subsets of a set $M$ is denoted by $\binom{M}{r}$. The symbol $\Delta_{g-1}$ denotes the $g-1$–dimensional unit simplex, that is, the set of all probability vectors of length $g$.

We consider data in a measurable sample space $E$, often $d$–dimensional Euclidean space $\mathbb{R}^d$. Our data set is $S = (x_1, \ldots, x_n) \in E^n$. Parameter spaces of statistical models are metric. We denote them by the upper case Greek letters $\Gamma$, $\Psi$, and $\Theta$ and parameters by the corresponding lower case letters $\gamma$, $\psi$, $\theta$, and $\vartheta$. If a random variable $X : \Omega \to E$ is distributed according to $\mu$ we write $X \sim \mu$. Its density function w.r.t. some reference measure on $E$ is denoted by $f_X$ or $f_\mu$. The conditional density of $X$ given the parameter $\gamma$ is $f_\gamma(x) = f_X[x \mid \gamma]$. We assume that it is a continuous function of $\gamma$.

The cone of positive–definite, symmetric $d \times d$–matrices is denoted by $\mathrm{PD}(d)$. Notation related to various normal cases is introduced in Sect. 2.9.

### 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* 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 log-likelihood $\ln f_{\nu_\theta}$ given the observation w.r.t. the "current" fit $\vartheta$

$$Q(\vartheta, \theta) = \mathrm{E}_{\nu_\vartheta}[\ln f_{\nu_\theta} \mid \Phi = x]. \tag{1}$$

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

$$\vartheta \leftarrow \operatorname*{argmax}_\theta Q(\vartheta, \theta). \tag{2}$$

One step in this process is called the EM–step. Dempster, Laird, and Rubin [4] 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 likelihood function.

More recently, Chrétien and Hero [1] derived the properties of the EM–algorithm from those of a *proximal point algorithm* (PPA) [19, 24] with the conditional Kullback–Leibler divergence as the penalty function. To this end they introduce the difference of the observed posterior log-likelihood and the Kullback–Leibler divergence of the complete model w.r.t. the current complete distribution $\nu_\vartheta$ conditional on the observation,

$$H(\vartheta, \theta) = \ln f_X[x \mid \theta] - D(\vartheta, \theta). \tag{3}$$

We assume that

$$D(\vartheta, \theta) = \int \nu_\vartheta[\mathrm{d}y \mid \Phi = x] \ln \frac{f_{\nu_\vartheta}[y \mid \Phi = x]}{f_{\nu_\theta}[y \mid \Phi = x]} \tag{4}$$

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 $\theta$. 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.

The parameter $\vartheta$ is called a *fixed point* of $H$ ($Q$) if $\vartheta \in \operatorname{argmax}_\theta H(\vartheta, \theta)$ ($\vartheta \in \operatorname{argmax}_\theta Q(\vartheta, \theta)$). Of course, fixed points of $H$ and $Q$ are the same.

### 1.4 EM for mixtures

Among other things, Dempster, Laird, and Rubin [4] applied the EM–algorithm to estimating the parameters $\vartheta = (\mathbf{u}, \gamma) \in \Delta_{g-1} \times \Gamma$ of a mixture distribution with density

$$f_{\mathbf{u},\gamma}(x) = \sum_j u_j f_{\gamma_j}(x). \tag{5}$$

Here, $\mathbf{u} = (u_1, \ldots, u_g) \in \Delta_{g-1}$ are its mixing parameters and $\gamma = (\gamma_1, \ldots, \gamma_g) \in \Gamma \subseteq \Gamma_0^g$ its population parameters. The complex random variable $X \sim f_{\mathbf{u},\gamma}$ is a simple function of a more easily accessible model. It is sufficient to represent each observation $X$ by randomly switching on one of $g$ random variables $Z^{(l)} \sim f_{\gamma_l}$, $l \in 1..g$, with the aid of a stochastically independent random label $L \sim \mathbf{u}$ in $1..g$, the hidden variable, so that $X = Z^{(L)}$. By the formula of total probabilities, $X$ is distributed according to the mixture distribution (5). The complete variable of the observation $X$ is thus the joint variable $Y = (L, X)$. In the case of

$n$ independent observations $X_1, \ldots, X_n$ one obtains for $S = (x_1, \ldots, x_n)$, $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$, and $\mathbf{l} = (l_1, \ldots, l_n)$

$$f_X[S \mid \mathbf{u}, \gamma] = \prod_i \sum_j u_j f_{\gamma_j}(x_i) \quad \text{and}$$
$$f_Y[\mathbf{l}, S \mid \mathbf{u}, \gamma] = \prod_i u_{l_i} f_{\gamma_{l_i}}(x_i).$$

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

$$Q((\mathbf{u}, \gamma), (\mathbf{v}, \eta)) = \sum_l \Big( \sum_i w(i,l) \Big) \ln v_l + \sum_l \sum_i w(i,l) \ln f_{\eta_l}(x_i), \tag{6}$$

the weight $w(i,l)$ being the posterior probability of the observation $x_i$ to come from component $l$ w.r.t. the parameters $\mathbf{u}$ and $\gamma$. By Bayes' formula,

$$w(i,l) = P[L_i = l \mid X_i = x_i] = \frac{u_l f_{\gamma_l}(x_i)}{\sum_j u_j f_{\gamma_j}(x_i)}. \tag{7}$$

The weights sum up to 1 over $l$, i.e., $w$ is a stochastic matrix. The entropy inequality allows to optimize Eqn. (6) w.r.t. $\mathbf{v}$ the maximum being

$$u_{\text{new},l} = \frac{1}{n} \sum_i w(x_i, l). \tag{8}$$

The EM–step starting from a stochastic matrix $(w(i,j))$ is thus split into an M–step and an E–step:

| | |
|---|---|
| E–*step*: | Compute $w(i,l)$ from the current parameters $\mathbf{u}$ and $\gamma$, cf. Eqn. (7); |
| M–*step*: | set $u_{\text{new},l} = w_l(S)/n$ and maximize $\sum_l \sum_i w(i,l) \ln f_{\eta_l}(x_i)$, cf. (6), w.r.t. $\eta$ to obtain the parameter $\gamma_{\text{new}}$. |

The EM–algorithm is iterative and alternating running as follows

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

If the m.l.e. exists then the sequence of target values converges, often to a local maximum. This always occurs in the homoscedastic normal case.

Of course, the algorithm can only be applied to models $f_{\gamma_l}$ that actually allow maximization in the M–step. The assumed continuity of the likelihood function implies that this is always the case if $\Gamma$ is compact. More can be said in general in the heteroscedastic case. Here $\Gamma$ is the $g$–fold Cartesian product of $\Gamma_0$ and one maximizes the sum $\sum_i w(i,l) \ln f_{\gamma_l}(x_i)$ separately for each $l$. If $\Gamma$ is (locally compact and) non–compact and if the likelihood function $\eta \mapsto f_\eta(x)$ vanishes as $\eta$ approaches the Alexandrov point of $\Gamma$ for all $x \in E$ then the same is true for the sum $\sum_i w(i,l) \ln f_{\eta_l}(x_i)$ and it is again plain that the maximum exists. In other cases one is interested in the largest local maximum, Kiefer [16].

# 2 The EMT–algorithm

## 2.1 A mixture model with spurious outliers and its ML criterion

We now consider a model for $r$ regular observations and $n-r$ "spurious" elements in a sample space $E$. Spuriousness [8] applies to observations that are gross, unpredictable outliers in the sense that they obey no statistical law. We feel that the best way of handling this idea in a statistical (!) framework is by assuming that each outlier comes from its own population. The following is the main assumption on the spurious outliers.

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

$$\int_{\Psi_i} f_{X_i}[x \mid \psi_i]\tau_i(\,\mathrm{d}\psi_i) = 1, \tag{9}$$

i.e., does not depend on $x$. We will later consider the parameters $\psi_i$ as 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 $\tau_i$ is Lebesgue measure on $\Psi_i$. Indeed, in this case, the conditional Lebesgue density is $f_{X_i}[x \mid \psi_i] = f_{U_i}(x - \psi_i)$ and, hence,

$$\int_{\Psi_i} f_{X_i}[x \mid \psi_i]\,\mathrm{d}\psi_i = 1.$$

(B) The parameter set $\Psi_i$ is singleton and the distribution of $X_i$ is taken as the reference measure for its density. This case includes the idea of irregular variants "uniformly distributed" on some domain.

Each of the $r$ regular observations $X_i$ comes from a mixture of $g$ populations represented by a density of the form (5). The numbers of components, $g$, and of regular objects, $r$, are considered fixed. They can be chosen by model selection criteria, Kéribin [15], and goodness–of–fit techniques that are not the subject matter of this communication. We assume that all functions $f_{\mathbf{u},\gamma}$ are strictly positive on $E$. A popular example is the homoscedastic normal model on Euclidean $d$–space with parameter space $\Gamma = \mathbb{R}^{gd} \times \{(V, \ldots, V) \mid V \in \mathrm{PD}(d)\} \approx \mathbb{R}^{gd} \times \mathrm{PD}(d)$, all normal populations on Euclidean space with a common covariance matrix.

Combining regular observations and outliers, we use the set

$$\binom{1..n}{r} \times \Delta_{g-1} \times \Gamma \times \prod_{i=1}^{n} \Psi_i.$$

as the parameter set of our model with $g$ components and $n - r$ outliers. The set $\binom{1..n}{r}$ of all $r$–element subsets of $1..n$ stands for the possible subsets of regular observations. Of course, the parametrization of the mixture model is not identifiable in the strict sense, see however the discussion in [22], Ch. 1. The density function of the $i$th observation for the parameter

values $R \in \binom{1..n}{r}$, $\mathbf{u} = (u_1, \ldots, u_g)$, $\gamma = (\gamma_1, \ldots, \gamma_g)$, and $\psi = (\psi_1, \ldots, \psi_n)$ w.r.t. a reference measure on $E$ is

$$f_{X_i}[x \mid R, \mathbf{u}, \gamma, \psi_i] = \begin{cases} f_{\mathbf{u},\gamma}(x) & \text{as in Eqn. (5)}, \quad i \in R, \\ f_{X_i}(x) & \text{as in Eqn. (9)}, \quad i \notin R. \end{cases}$$

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 likelihood for the data set $S = (x_1, \ldots, x_n)$ is

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

Considering the parameters $\psi_i$ of the outliers nuisances to be integrated out w.r.t. to the prior measures $\tau_i$ we obtain by Eqn. (9) the integrated likelihood

$$f_X[R \mid \mathbf{u}, \gamma] = \prod_{i \in R} f_{\mathbf{u},\gamma}(x_i) = \prod_{i \in R} \Big( \sum_j u_j f_{\gamma_j}(x_i) \Big), \tag{10}$$

the ML–criterion to be optimized w.r.t. the parameters $R \in \binom{1..n}{r}$, $\mathbf{u} \in \Delta_{g-1}$, and $\gamma \in \Gamma$. By the principle of dynamic optimization, the ML–estimator of the parameters is computed from

$$\operatorname*{argmax}_R \max_{\mathbf{u},\gamma} \ln f_X[R \mid \mathbf{u}, \gamma] = \operatorname*{argmax}_R \ln f_X[R \mid \mathbf{u}^*, \gamma^*], \tag{11}$$

$\mathbf{u}^*$ and $\gamma^*$ being the m.l.e.'s of $\mathbf{u}$ and $\gamma$ w.r.t. $R$.

## 2.2 The EMT–step

Our next aim is the adaptation of the EM–algorithm to the ML–criterion (10). 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 $1..n$ and the trimming step selects the $r$ elements that best fit the new parameters as the new regular elements.

*Input*: A subset $R \subseteq S$ of $r$ elements, mixing rates $(u_1, \ldots, u_g)$, and population parameters $\gamma_1, \ldots, \gamma_g$.

*Output*: A subset, mixing rates, and population parameters with improved criterion (10), cf. Proposition 2.3.

E–*step*: compute the weights $w(x, j) = \frac{u_j f_{\gamma_j}(x)}{\sum_l u_l f_{\gamma_l}(x)}$, $x \in R$, $j \in 1..g$;

M–*step*: set $u_{\text{new},j} = \frac{1}{r} \sum_{x \in R} w(x, j)$, $1 \le j \le g$, and maximize $\sum_j \sum_{x \in R} w(x, j) \ln f_{\gamma_j}(x)$ w.r.t. $\gamma \in \Gamma$ to obtain $\gamma_{\text{new}}$; (in the heteroscedastic case, each sum $\sum_{x \in R} w(x, j) \ln f_{\gamma_j}(x)$, $j \in 1..g$, may be maximized separately)

T–*step*: define $R_{\text{new}}$ to be the set of objects $x \in S$ 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 iteration of EMT–steps with a randomly or deliberately chosen initial stochastic matrix $\mathbf{w}^{(0)}$ and an $r$–element subset $R^{(0)}$ input to the M–step is the **EMT–algorithm**. It is again iterative and alternating running as follows

$$(R^{(0)}, \mathbf{w}^{(0)}) \overset{\text{M–step}}{\longrightarrow} (\mathbf{u}^{(1)}, \gamma^{(1)}) \overset{\text{T–step}}{\longrightarrow} (R^{(1)}, \mathbf{u}^{(1)}, \gamma^{(1)}) \overset{\text{E–step}}{\longrightarrow} (R^{(1)}, \mathbf{w}^{(1)}) \overset{\text{M–step}}{\longrightarrow} (\mathbf{u}^{(2)}, \gamma^{(2)}) \overset{\text{T–step}}{\longrightarrow} \cdots .$$

The rows of the initial weight matrix may be chosen uniformly from the $(g-1)$–dimensional unit simplex $\Delta_{g-1}$. An efficient procedure is OSIMP, see Fishman [6]. An alternative are randomly sampled unit vectors. If components are sufficiently separated the algorithm may also be started with parameters $(\mathbf{u}^{(0)}, \gamma^{(0)})$ from a suitable clustering algorithm. An elegant procedure for uniform generation of a subset $R^{(0)}$ appears in Knuth [17], p.136, ff. Since EMT does not necessarily find a global maximum in one sweep, the algorithm has to be replicated several or many times with different starting configurations in order to reach a high value of the integrated likelihood (10). The remarks after the statement of the EM–algorithm apply also to the EMT.

The EMT–step has two parameters, the number $g$ of components and the number of regular data points $r$. Both may be chosen by validation techniques based on goodness of fit.

The following proposition discusses monotonicity of the successive values of the target function.

### 2.3 Proposition

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

(a) An EMT–step improves the integrated likelihood $f_X[R \mid \mathbf{u}, \gamma]$ in the sense $\geq$.

(b) If $(R, \mathbf{u}, \gamma)$ is optimal then so is $(R_{\text{new}}, \mathbf{u}_{\text{new}}, \gamma_{\text{new}})$.

**Proof.** (a) Under the assumptions made, the fact that $f_X[R \mid \mathbf{u}, \gamma] \leq f_X[x \mid R, \mathbf{u}_{\text{new}}, \gamma_{\text{new}}]$ is the well–known property of monotonicity of the EM–algorithm applied to the objects in $R$, see [4], p. 8. Moreover,

$$\begin{aligned} \ln f_X[R \mid \mathbf{u}_{\text{new}}, \gamma_{\text{new}}] &= \sum_{i \in R} \ln \sum_l u_{\text{new},l} f_{\gamma_{\text{new},l}}(x_i) \leq \sum_{i \in R_{\text{new}}} \ln \sum_l u_{\text{new},l} f_{\gamma_{\text{new},l}}(x_i) \\ &= \ln f_X[R_{\text{new}} \mid \mathbf{u}_{\text{new}}, \gamma_{\text{new}}] \end{aligned}$$

by maximality of the objects in $R_{\text{new}}$.

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

We say that $(R, \mathbf{u}, \gamma)$ is a *halting point* of the EMT–step if the ML–criterion (10) remains unchanged after an EMT–step starting from it. According to Proposition 2.3, an ML–estimate is a halting point. A *critical* point of a differential function is a point where its gradient vanishes. A *face* of the simplex $\Delta_{g-1}$ is the convex combination 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 $\Delta_{g-1}$ with some subspace $L$ of $\mathbb{R}^g$ such that $\Delta_{g-1} \setminus L$ is convex or again if it is the set of points in $\Delta_{g-1}$ 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 an extreme point of the simplex.)

In the sequel we discuss the relationships between limit, halting, fixed, and critical points. We need the $H$–functional (3) w.r.t. an $r$–element subset $R \subseteq 1..n$,

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

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

### 2.4 Proposition

Assume that the sequence of successive outputs of the EMT algorithm converges with limit $(R^*,\mathbf{u}^*,\gamma^*)$. Then $(R^*,\mathbf{u}^*,\gamma^*)$ is a halting point of the EMT–step.

**Proof.** By convergence of the sequence $(R_t,\mathbf{u}_t,\gamma_t)$ to $(R^*,\mathbf{u}^*,\gamma^*)$, we have $R_t = R^*$ for eventually all $t$. It is, therefore, sufficient to consider $R_t$ fixed. Abbreviate $\theta = (\mathbf{u},\gamma)$, $\vartheta_t = (\mathbf{u}_t,\gamma_t)$, and $\vartheta^* = (\mathbf{u}^*,\gamma^*)$. From $H_{R_t}(\vartheta_t,\theta) \le H_{R_t}(\vartheta_t,\vartheta_{t+1})$ for all $\theta$ we infer

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

This is the claim. □

### 2.5 Proposition

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

**Proof.** Let us put $\vartheta^* = (\mathbf{u}^*,\gamma^*)$ and $\vartheta_{\text{new}} = (\mathbf{u}_{\text{new}},\gamma_{\text{new}})$, the output of the EMT–step starting from $(R^*,\mathbf{u}^*,\gamma^*)$. If $(R^*,\mathbf{u}^*,\gamma^*)$ is a halting point of the EMT–step then

$$f_X[R^* \mid \vartheta^*] = f_X[R^* \mid \vartheta_{\text{new}}] = f_X[R_{\text{new}} \mid \vartheta_{\text{new}}]. \tag{12}$$

The first equality implies

$$\begin{aligned} H_{R^*}(\vartheta^*,\vartheta^*) &= \ln f_X[R^* \mid \vartheta^*] = \ln f_X[R^* \mid \vartheta_{\text{new}}] = H_{R^*}(\vartheta^*,\vartheta_{\text{new}}) + D_{R^*}(\vartheta^*,\vartheta_{\text{new}}) \\ &\ge H_{R^*}(\vartheta^*,\vartheta_{\text{new}}), \end{aligned}$$

i.e., $\vartheta^*$ is a fixed point of $H_{R^*}$. □

### 2.6 Proposition

Assume that $\Gamma$ is an open subset of some Euclidean space, that $f_\gamma(x)$ is differentiable w.r.t. $\gamma$ for all $x$, and that the (conditional) Kullback–Leibler divergence $D_R(\vartheta,\theta) = D(\nu_\vartheta[\cdot \mid \Phi =$

$R], \nu_\theta[\cdot \mid \Phi = R])$ is differentiable w.r.t. $\theta$ at each point of the diagonal $\theta = \vartheta$ for all $R$. Then any halting point $(R^*, \mathbf{u}^*, \gamma^*)$ of the EMT–step has the following properties.

(i) $\gamma^*$ is a critical point of the observed likelihood function $f_X[R^* \mid \mathbf{u}^*, \gamma]$ as a function of $\gamma$.

(ii) The vector of mixing rates $\mathbf{u}^*$ is a maximum of the observed likelihood function $f_X[R^* \mid \mathbf{u}, \gamma^*]$ as a function of $\mathbf{u}$ on the face generated by $\mathbf{u}^*$.

(iii) If $\mathbf{u}^*$ is an interior point of the simplex then it is a maximum of the function $\mathbf{u} \to f_X[R^* \mid \mathbf{u}, \gamma^*]$. If, moreover, each observation $x_i$ has a strictly positive density w.r.t. $(\mathbf{u}^*, \gamma^*)$ and if the $g$ vectors

$$\big(f_{\gamma^*,1}(x_i)\big)_i, \ldots, \big(f_{\gamma^*,g}(x_i)\big)_i$$

are affine independent then it is the only maximum.

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

**Proof.** Let $\vartheta^*$ and $\vartheta_{\text{new}}$ be as defined at the beginning of the proof of Proposition 2.5. From that proposition, we know already that $\vartheta^*$ is a fixed point of $H_{R^*}$. The point $\mathbf{u}^*$ is interior to the face $F$ generated by it. Therefore, $\vartheta^*$ lies in the interior of $F \times \Gamma$. The fixed point $\vartheta^*$ 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 $\vartheta^*$, 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

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

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

Nothing has to be shown for claim (ii) if $\mathbf{u}^*$ is an extreme point of the simplex. Otherwise, note that

$$\ln f_X[R^* \mid \theta] = \sum_{i \in R} \ln \sum_j u_j f_{\gamma,j}(x_i)$$

is of the form $\sum_{i \in R} \ln(A\mathbf{u})_i$ with $A_{i,j} = f_{\gamma,j}(x_i)$. Claim (ii) now follows from concavity A.1(a) of this function restricted to the mixing parameters and from the vanishing of the gradient.

If $\mathbf{u}^*$ is an interior point then its generated face is the whole simplex and the first claim of (iii) follows from (ii). The second claim follows from Lemma A.1(b).

Claim (iv) follows directly from the second equality in Eqn. (12). □

It is often the case that the EMT–algorithm converges. The following corollary, a consequence of Propositions 2.4, 2.5, and 2.6, discusses the limit.

## 2.7 Corollary

Let the assumptions of Proposition 2.6 hold and assume that the sequence of successive outputs of the EMT algorithm converges with limit $(R^*, \mathbf{u}^*, \gamma^*)$. Then (i)–(iv) of Proposition 2.6 hold.

| 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, 0.245256, 0.254744 | –2.99999, 2.99992, 3.00005 | 1.00007 |
| –10.2809 | 1–$\alpha$–$\beta$, $\alpha$, $\beta$ | 0, 0, 0 | 10 |

Table 1: Limit points for the data set –4,–2,2,4 and the homoscedastic normal model with three components.

### 2.8 Remarks

(a) Even if the ML–estimate exists the likelihood values along the EMT–algorithm need not converge to the maximum. Here is a simple, one-dimensional, normal example. 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 limits shown in Table 1. The first limit point 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 limit point corresponds to the natural solution with two components with centers close to $-3$ and 3. One of the components is split in two very similar parts. The limit point seems to be a local maximum which is very flat in two directions, two eigenvalues of the Hessian being close to zero. The last line in the table describes a two–dimensional manifold of limit points with equal log–likelihoods. The positive semi–definite Hessian is the same at each point and has four vanishing eigenvalues. In the first two lines, the mixing rates are unique by Proposition 2.6(iii). However, 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 $\Gamma$ is not open as required in Proposition 2.6 then $(\mathbf{u}^*, \gamma^*)$ is still a fixed point as in the first part of Proposition 2.6(i) but it is only true that directional derivatives of $\gamma \mapsto f_X[R^* \mid \mathbf{u}^*, \gamma]$ at $\gamma^*$ must be $\leq 0$ in all interior directions.

### 2.9 Three normal cases

Eqn. (8) gives an analytical expression for the maximum in the M–step w.r.t. the vector $\mathbf{u}$ of mixing rates. It is well known that the population parameters $\eta_l = (m_l, V_l)$, too, can be expressed by formulae if populations are normal. The same can of course be said about the M–step in the EMT–algorithm.

The normal case needs some more notation. Here, $E = \mathbb{R}^d$ is $d$–dimensional Euclidean space so that $S \subseteq \mathbb{R}^{nd}$. The symbol $N_{m,V}$ designates the $d$–variate normal distribution with mean $m$ and covariance matrix $V$ and also its Lebesgue density. Given a subset $T \subseteq S$, the symbols $m(T)$, $W(T)$, and $V(T)$ designate its mean vector, its SSP–matrix, and its sample covariance matrix, respectively. We will mainly need *weighted* mean vectors and *weighted* pooled SSP– and covariance matrices w.r.t. a stochastic weight matrix $\mathbf{w} = (w(x,j))_{x \in S, j \in 1..g}$, $\sum_j w(x,j) =$

1 for all $x$, and a subset $T \subseteq S$. With the abbreviation

$$w_j(T) = \sum_{x \in T} w(x, j)$$

they are defined respectively as

$$m_{\mathbf{w},j}(T) = \frac{1}{w_j(T)} \sum_{x \in T} w(x, j) x,$$
$$W_{\mathbf{w},j}(T) = \sum_{x \in T} w(x, j)(x - m_{\mathbf{w},j}(T))(x - m_{\mathbf{w},j}(T))^{\mathrm{T}},$$
$$V_{\mathbf{w},j}(T) = \frac{1}{w_j(T)} \sum_{x \in T} w(x, j)(x - m_{\mathbf{w},j}(T))(x - m_{\mathbf{w},j}(T))^{\mathrm{T}}.$$

The mean value is put to zero and the covariance matrix to the identity matrix if $w_j(T)$ vanishes.

We assume throughout that the data points $S$ are in *general position*. This notion has different meanings for the three customary normal sub–populations specified by the shape of the covariance matrix: spherical, diagonal, or full. In the spherical case it means pairwise difference of all data points, in the diagonal case pairwise difference of the $k$th entries for all $k \in 1..d$, and in the full case affine independence of any $d + 1$ points in $S$. We make the standard assumption $r \geq g + 1$ in the "spherical" and "diagonal" cases and $r \geq gd + 1$ in the "full" case. General position then guarantees that all weighted variances are strictly positive and all weighted covariance matrices are positive definite in the homoscedastic case. Indeed, e.g. in the "full" case, let $\{R_1, \ldots, R_g\}$ be the MAP–*partition* of $R$ associated with the weight matrix $(w(x, j))_{x \in R, j \in 1..g}$, i.e., $x \in R_j \Leftrightarrow j = \mathrm{argmax}_\ell \, w(x, \ell)$. Then

$$w(x, j) \geq \frac{1}{g}, \quad x \in R_j, \tag{13}$$

and, hence,

$$W_{\mathbf{w},j}(T) \succeq \frac{1}{g} \sum_j \sum_{x \in R_j} (x - m_{\mathbf{w},j}(T))(x - m_{\mathbf{w},j}(T))^{\mathrm{T}} \succ 0$$

since at least one cluster $R_j$ contains at least $d+1$ elements by the pigeon hole principle. With the notation above and in the homoscedastic case, the parameter estimates in the M–step with input $\mathbf{w}$ and $R$ are

$$m_j = m_{\mathbf{w},j}(R) \qquad \text{and} \tag{14}$$

$$V = \sum_j u_j V_{\mathbf{w},j}(R) = \frac{1}{r} \sum_{j=1}^{g} \sum_{x \in R} w(x, j)(x - m_j)(x - m_j)^{\mathrm{T}}, \qquad \text{(full)} \tag{15}$$

$$v_k = V_{k,k} = \frac{1}{r} \sum_{j=1}^{g} \sum_{x \in R} w(x, j)(x_k - m_{j,k})^2, \qquad \text{(diagonal)} \tag{16}$$

$$v = \frac{1}{d} \sum_k v_k = \frac{1}{rd} \sum_{j=1}^{g} \sum_{x \in R} w(x, j) \sum_k (x_k - m_{j,k})^2. \qquad \text{(spherical)} \tag{17}$$

Furthermore, if $(\mathbf{u}, \mathbf{m}, V)$ is a fixed point of the EM–algorithm w.r.t. $R$, e.g., if $(R, \mathbf{u}, \mathbf{m}, V)$ is a halting point of EMT, see Proposition 2.5, then the criterion has the representation

$$\log f[R \mid \mathbf{u}, \mathbf{m}, V] \tag{18}$$
$$= c_{r,d} + r \sum_{j=1}^{g} u_j \log u_j - \sum_{j=1}^{g} \sum_{x \in R} w(x,j) \log w(x,j) - \begin{cases} \frac{r}{2} \log \det V, & \text{(full)} \\ \frac{r}{2} \sum_{k=1}^{d} \log v_k, & \text{(diagonal)} \\ \frac{dr}{2} \log v, & \text{(spherical)} \end{cases}$$

where $c_{r,d} = -\frac{dr}{2}(1 + \log 2\pi)$ and $w(x,j) = \frac{u_j N_{m_j,V}(x)}{\sum_l u_l N_{m_l,V}(x)}$, see Appendix A.2.

# 3 Universal breakdown points

## 3.1 Breakdown points

The finite–sample breakdown value of an estimator, Hodges [13] and Donoho and Huber [5], measures the minimum fraction of gross outliers that can *completely* spoil the estimate. Two types of breakdown points are customary, the *addition* and the *replacement* breakdown point. The former refers to the addition of $n - r$ outliers to a data set of $r$ regular observations and the latter to $n - r$ replacements in a data set of $n$ regular observations. The former is technically simpler since we have a *fixed* set of regular observations at hand, but there is the disadvantage that we need two estimators, one for $r$ data and one for $n$ data. By contrast, in the latter we have to consider all $\binom{n}{r}$ possible replacements of $n - r$ observations but need only one estimator for $n$ objects. We deal with replacements.

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

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

is not relatively compact in $\Theta$. Of course, there is no breakdown if $\Theta$ is compact. The *individual* breakdown point for the data set $S$ is the number

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

It is the minimal fraction of replacements in $S$ that may cause $\delta$ to break down. The individual breakdown point is not an interesting concept *per se* since it depends on a single data set. It tells the statistician how many gross outliers the data set $M$ under his or her study may contain without causing excessive damage if the imaginary "clean" data set that should have been observed were $S$. Now let $\mathcal{K} \subseteq \mathcal{A}$ be some subclass of admissible data sets. The *restricted* breakdown point [8] of $\delta$ w.r.t. $\mathcal{K}$ is

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

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

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

This concept depends solely on the estimator. The restricted breakdown value may be seen as a relaxed version of it. We have the estimates

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

We deal here with breakdown points of EMT for the means and the common covariance matrix. In the former case, the relatively compact subsets of the parameter space $\mathbb{R}^d$ are the bounded sets. In the latter, the parameter space is the set of all positive–definite $d \times d$–matrices and a subset is relatively compact if the eigenvalues of its members are bounded and bounded away from zero. This is equivalent to saying that the subset is bounded above and below by positive–definite matrices in the positive–definite (or Löwner) ordering $\preceq$ on the vector space of symmetric matrices.

We next show that the EMT–algorithm robustly estimates the common covariance matrix of the homoscedastic normal models described in Sect. 2.9 and compute the universal breakdown point.

### 3.2 Theorem (Universal breakdown point of EMT for the pooled covariance matrix)

Assume $2r \geq n + 2g$ in the "spherical" and "diagonal" cases and $2r \geq n + g(d+1)$ in the "full" case. (Note that these assumptions imply the standard assumptions $r \geq g + 1$ in the "spherical" and "diagonal" cases and $r \geq gd + 1$ in the "full" case.)

(a) If at most $n - r + g - 1$ points of the data set are replaced in an arbitrary but admissible way then the pooled covariance matrix output from an M–step remains bounded below by a positive–definite matrix.

(b) If at most $n - r + g - 1$ points of the data set are replaced in an arbitrary but admissible way then the optimal covariance matrix remains bounded above by a positive–definite matrix.

(c) Given any positive number $K$, $n - r + g$ points may be replaced in such a way that the largest eigenvalue of the pooled covariance matrix output from any M–step exceeds $K$.

(d) The breakdown value of the pooled scatter matrix is

$$\beta_{\text{Cov}} = \frac{n - r + g}{n}.$$

**Proof.** We give proofs in the "full" case; the other cases are similar.

(a) We first note that, by general position, there is $\varepsilon > 0$ so that the SSP matrix of any $d+1$ points of the original data $S$ dominates $\varepsilon I_d$, where $I_d$ is the $d$–dimensional unit matrix. By assumption, any $r$–element subset $z_1, \ldots, z_r$ of the modified data set contains at least $r - (n - r + g - 1) = 2r - n - g + 1 \geq gd + 1$ original points. Consider the MAP–clustering $(R_1, \ldots, R_g)$ of $z_1, \ldots, z_r$ associated with the weight matrix $\mathbf{w}$ input to the M–step. By the

pigeon hole principle, some cluster, say $R_l$, contains $\geq d+1$ original elements, say $x_1, \ldots, x_{d+1}$. Let $m_j$ and $V$ be the population parameters output from the M–step. We use (15), (13) and Steiner's formula to estimate

$$V = \frac{1}{r}\sum_{j=1}^{g}\sum_{i=1}^{r} w(i,j)(z_i - m_j)(z_i - m_j)^T \succeq \frac{1}{r}\sum_{i\in R_l} w(i,l)(z_i - m_l)(z_i - m_l)^T$$
$$\succeq \frac{1}{gr}\sum_{i\leq d+1}(x_i - m_l)(x_i - m_l)^T \succeq \frac{\varepsilon}{gr} I_d.$$

(b) We first show that, irrespective of the $n - r + g - 1$ replacements, the criterion of some admissible solution exceeds some constant value. 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 $g - 1$ 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 $V = I_d$. The integrated likelihood (10) is

$$\prod_{i=1}^{r-g+1} \frac{1}{g}\Big\{(2\pi)^{-d/2} e^{-\|x_i\|^2/2} + \sum_{j=1}^{g-1}(2\pi)^{-d/2} e^{-\|x_i - y_j\|^2/2}\Big\}$$
$$\times \prod_{i=1}^{g-1}\frac{1}{g}\Big\{(2\pi)^{-d/2} e^{-\|y_i\|^2/2} + \sum_{j=1}^{g-1}(2\pi)^{-d/2} e^{-\|y_i - y_j\|^2/2}\Big\}$$
$$\geq \prod_{i=1}^{r-g+1} \frac{1}{g}\Big\{(2\pi)^{-d/2} e^{-\|x_i\|^2/2}\Big\} \prod_{i=1}^{g-1}\frac{1}{g}(2\pi)^{-d/2}$$
$$\geq (2\pi)^{-rd/2} g^{-r} e^{-\|S\|^2/2} = c,$$

a positive constant.

Now let $(z_1, \ldots, z_n)$ be the modified data set. The likelihood of its optimal solution $(R^*, \mathbf{u}^*, \mathbf{m}^*, V^*)$, too, exceeds the constant $c$ and we have

$$c \leq \prod_{i\in R^*}\sum_{j=1}^{g} u_j^* \frac{1}{\sqrt{\det 2\pi V^*}} e^{-\frac{1}{2}(z_i - m_j^*)^T (V^*)^{-1}(z_i - m_j^*)} \leq \prod_{i\in R^*}\sum_{j=1}^{g} u_j^* \frac{1}{\sqrt{\det 2\pi V^*}} = (\det 2\pi V^*)^{-r/2}.$$

Hence, $\det V^* \leq$ const and the claim follows from Part (a).

(c) Let $(z_1, \ldots, z_n)$ be the data set modified by $n - r + g$ replacements. Let $\mathcal{R}$ be the MAP–partition of the $r$–element subset $R \subseteq 1..n$ associated with the matrix $w$ input to the M–step. By assumption, $R$ contains at least $g$ replacements. Hence, either one cluster contains two replacements or each cluster contains at least one replacement, in particular some cluster with $\geq d + 1$ elements. In any case there is an cluster of size $\geq 2$ containing a replacement $z$ and another element $y$. Let $m_j$ and $V$ be the population parameters output from the M–step. We

have by (15)

$$\begin{aligned} V =& \frac{1}{r}\sum_{j=1}^{g}\sum_{i=1}^{r} w(i,j)(z_i - m_j)(z_i - m_j)^T \succeq \frac{1}{gr}\big((z-m_l)(z-m_l)^T + (y-m_l)(y-m_l)^T\big) \\ \succeq& \frac{1}{gr}\big((z - \frac{1}{2}(z+y))(z-\frac{1}{2}(z+y))^T + (y-\frac{1}{2}(z+y))(y-\frac{1}{2}(z+y))^T\big) \\ =& \frac{1}{2gr}(z-y)(z-y)^T. \end{aligned}$$

Thus, $\operatorname{tr} V \geq \frac{1}{2gr}\|z-y\|^2$. This proves the claim since the replacements may be chosen in such a way as to be far away from all original data and from each other.

Claim (d) is immediate from (a)–(c). □

### 3.3 Corollary

(a) The maximal number of outliers that the optimal pooled scatter matrix output from the EMT–algorithm can resist is

$$\begin{aligned} &\Big\lfloor \frac{n}{2} \Big\rfloor && \text{(spherical and diagonal)}, \\ &\Big\lfloor \frac{n-g(d-1)}{2} \Big\rfloor && \text{(full)}. \end{aligned}$$

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

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

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

Part (b) of the corollary says that the EMT–algorithm attains for the pooled covariance the optimal asymptotic breakdown value of a translation equivariant estimator, Lopuhaä and Rousseeuw [18].

In view of the universal breakdown point of the location parameters we first state a lemma.

### 3.4 Lemma

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

$$\max_j \|m_j\| \longrightarrow \infty \quad \text{as } \|y_1\| \to \infty \text{ such that } \{y_i - y_1 \mid 2 \leq i \leq q\} \text{ remains bounded.}$$

**Proof.** Let $w(\cdot, j)$ be the weights from which the parameters are computed. Eqn. (14) implies

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

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

**3.5 Theorem** (Universal breakdown point of EMT for the means)

(a) If $g+1 \le r < n$ in the "spherical" and "diagonal" cases and $gd+2 \le r < n$ in the "full" case[1] then all means remain bounded by a constant that depends only on the data set as *one* observation is arbitrarily replaced.

(b) If $g \ge 2$ and if $r \ge g+2$ then there is a data set such that one mean breaks down as *two* particular observations are suitable replaced.

(c) If $g+1 \le r < n$ in the "spherical" and "diagonal" cases and $gd+2 \le r < n$ in the "full" case then $\beta_{\text{mean}}(n,r,g) = \frac{2}{n}$.

**Proof.** We restrict ourselves to proving the case of full covariance matrices, the other cases being similar.

(a) Let $S = (x_1, \ldots, x_n)$ be any data set and let $M = (x_1, \ldots, x_{n-1}, y)$ be its modification by one replacement $y$. We show that the optimal solution $(\widetilde{R}, \widetilde{\mathbf{u}}, \widetilde{\mathbf{m}}, \widetilde{V})$ for $M$ with $r < n$ under the condition that $y$ is not discarded is inferior to some solution which discards $y$ if $y$ is far away. Let $d_{\widetilde{V}}$ denote the Mahalanobis distance induced by $\widetilde{V}$, i.e., $d_{\widetilde{V}}(u,v) = \sqrt{(u-v)^{\mathrm{T}}\big(\widetilde{V}\big)^{-1}(u-v)}$, let $d_{\widetilde{V}}(u,S) = \min_{v\in S} d_{\widetilde{V}}(u,v)$ and $d_{\widetilde{V}}(S) = \max_{u,v\in S} d_{\widetilde{V}}(u,v)$ denote the distance from $u$ to $S$ and the diameter of $S$, respectively, w.r.t. $d_{\widetilde{V}}$.

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

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

We now show that the solution $(\widetilde{R}, \widetilde{\mathbf{u}}, \widetilde{\mathbf{m}}, \widetilde{V})$ is inferior to $(R, \widetilde{\mathbf{u}}, \mathbf{m}, \widetilde{V})$ if $d_{\widetilde{V}}(y,S) > 4d_{\widetilde{V}}(S)$. Comparing the integrated likelihood of the former

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

with that of the latter

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

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

Now, if $d_{\widetilde{V}}(\widetilde{m}_j,S) > d_{\widetilde{V}}(S)$ then $d_{\widetilde{V}}(x_i,x_r) \le d_{\widetilde{V}}(S) < d_{\widetilde{V}}(\widetilde{m}_j,S) \le d_{\widetilde{V}}(x_i,\widetilde{m}_j)$; if $d_{\widetilde{V}}(\widetilde{m}_j,S) \le d_{\widetilde{V}}(S)$ then

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

[1] if $d = 1$ then $g+1 \le r < n$ is of course sufficient

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 $S$. (Note that $\widetilde{V}$ and, hence, the distance $d_{\widetilde{V}}$ depend on $y$!) In other words, we have to show that the set $\{y \mid d_{\widetilde{V}}(y, S) \leq 4d_{\widetilde{V}}(S)\}$ is bounded by a constant that depends only on $S$. To this end we next show that $\widetilde{V}$ is bounded below and above by positive–definite matrices $L$ and $U$ that depend only on $S$, $g$, and $r$.

Indeed, let $(\widetilde{R}_1, \ldots, \widetilde{R}_g)$ denote the MAP–clustering of $\widetilde{R}$ w.r.t. the optimal parameters. Since $r \geq gd + 2$, there is $k$ such that $\widetilde{R}_k$ contains at least $d + 1$ original elements. We have

$$\begin{aligned} r\widetilde{V} \succeq & \sum_{j=1}^{g} \sum_{i=1}^{r-1} w(x_i, j)(x_i - \widetilde{m}_j)(x_i - \widetilde{m}_j)^{\mathrm{T}} \succeq \frac{1}{g} \sum_{x \in \widetilde{R}_k \cap S} (x - \widetilde{m}_k)(x - \widetilde{m}_k)^{\mathrm{T}} \\ \succeq & \frac{1}{g} W(\widetilde{R}_k \cap S). \end{aligned}$$

We may thus put $L = \frac{1}{gr} W(\widetilde{R}_k \cap S)$.

Furthermore, the solution $(\widetilde{R}, \widetilde{\mathbf{u}}, \widetilde{\mathbf{m}}, \widetilde{V})$ is superior to $\big(\widetilde{R}, \big(\frac{1}{g}, \ldots, \frac{1}{g}\big), (0, \ldots, 0, y), I_d\big)$, i.e.,

$$f[\widetilde{R} \mid (1/g, \ldots, 1/g), (0, \ldots, 0, y), I_d] \leq f[\widetilde{R} \mid \widetilde{\mathbf{u}}, \widetilde{\mathbf{m}}, \widetilde{V}] \leq c_{r,d} - \frac{r}{2} \log\det \widetilde{V}.$$

by (18) and Lemma A.5. On the other hand,

$$\begin{aligned} & f[\widetilde{R} \mid (1/g, \ldots, 1/g), (0, \ldots, 0, y), I_d] \\ = & g^{-r} \prod_{i<r} \Big( \sum_{j<g} N_{0,I_d}(x_i) + N_{y,I_d}(x_i) \Big) \Big( \sum_{j<g} N_{0,I_d}(y) + N_{y,I_d}(y) \Big) \\ \geq & g^{-r} \Big( \prod_{i<r} \sum_{j<g} N_{0,I_d}(x_i) \Big) N_{0,I_d}(0), \end{aligned}$$

a quantity that does not depend on $y$. Therefore, $\det \widetilde{V}$ is bounded above by a constant which depends only on $S$, $g$ and $r$. Together with lower boundedness this shows upper boundedness by a positive–definite matrix $U$.

Denoting the Mahalanobis distances w.r.t. $L$ and $U$ by $d_L$ and $d_U$, respectively, the claim finally follows from

$$d_U(y, x_1) \leq d_{\widetilde{V}}(y, x_1) \leq d_{\widetilde{V}}(y, S) + d_{\widetilde{V}}(S) \leq 5d_{\widetilde{V}}(S) \leq 5d_L(S).$$

(b) We proceed in several steps.

($\alpha$) Construction of data sets $S$ and $M$:

Let $F := \{x_1, \ldots, x_{r-g}\}$ be a set of data points in general position. We control the remainder of the data set $S$ by a constant $K_1 > 0$ and the replacements by $K_2 > 0$. Both constants are specified later. Using inductively Lemma A.4(a), we add points $z_1, \ldots, z_{n-r+g-2}$ to $F$ such that

(i) $\|z_l - z_k\| \geq K_1$ for all $l \neq k$;

(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) $\det W(H) \geq K_1$ for all $H \in \binom{F \cup \{z_1,\ldots,z_{n-r+g-2}\}}{d+1}$ that contain at least one $z_k$.

Note that (iii) implies general position of the set $F \cup \{z_1, \ldots, z_{n-r+g-2}\}$. The data set $S$ 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 Lemma A.4(b) replacing the two points $q_1$ and $q_2$ with a twin pair $y_1 \neq y_2$ such that

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

(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) $\det W(E) \geq K_2$ for all $E \in \binom{M}{d+1}$ that contain at least one $y_k$ except for $E = \{y_1, y_2\}$ if $d = 1$.

In view of Lemma 3.4, we will show that the optimal solution does not discard the outliers $y_1$ and $y_2$ if $K_1$ is chosen large enough and as $K_2 \to \infty$.

($\beta$) The maximum of the objective function 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 parameter set with likelihood bounded below by a function of $F$, $g$, and $r$. Let $R = F \cup \{y_1, y_2\} \cup \{z_1, \ldots, z_{g-2}\}$, let $u_j = 1/g$, let $m_1 = m(F)$, the usual mean of $F$, $m_2 = y_1$, $m_j = z_{j-2}$, $3 \leq j \leq g$, and let $V = I_d$. We have

$$\begin{aligned} f[R \mid \mathbf{u}, \mathbf{m}, V] &= \prod_{i=1}^{r-g} \frac{1}{g} \sum_{j=1}^{g} N_{m_j, I_d}(x_i) \prod_{i=1}^{2} \frac{1}{g} \sum_{j=1}^{g} N_{m_j, I_d}(y_i) \prod_{i=3}^{g} \frac{1}{g} \sum_{j=1}^{g} N_{m_j, I_d}(z_{i-2}) \\ &\geq g^{-r} \prod_{i=1}^{r-g} N_{m_1, I_d}(x_i) \prod_{i=1}^{2} N_{m_2, I_d}(y_i) \prod_{i=3}^{g} N_{m_j, I_d}(z_{i-2}) \\ &\geq g^{-r} (2\pi)^{-\frac{gd}{2}} e^{-\frac{1}{2}} \prod_{i=1}^{r-g} N_{m_F, I_d}(x_i) \end{aligned} \tag{19}$$

as required.

According to the combinatorial Lemma 4.2 in [8], any partition 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 classes from the $z_l's\}$ *or*
there is a class $R_j \in \mathcal{R}$ which contains some pair $\{x_i, y_k\}$ or some pair $\{z_l, u\}$, $u \neq z_l$.

($\gamma$) The MAP–clustering $\mathcal{R}$ associated with an optimal solution cannot be of the second kind if $K_1$ and $K_2$ are sufficiently large:

Choose $R_j \in \mathcal{R}$ of maximum size containing either some pair $\{x_i, y_k\}$ or $\{z_l, u\}$, $u \neq z_l$. Let us denote the two elements by $a$ and $b$. If $R_j$ is of size at least $d+1$ then we choose $d+1$ elements $E \subseteq R_j$ containing $\{a, b\}$ inferring from (iii) and (vi)

$$\det rV \geq \det \frac{1}{g} W(R_j) \geq \det \frac{1}{g} W(E) \geq \left(\frac{1}{g}\right)^d \min\{K_1, K_2\}.$$

Otherwise, by the standard assumption $r > gd$, there exists a cluster $R_l$, $l \neq j$, of size $\geq d+1$ which contains no pair $\{x_i, y_k\}$ and no $z_l$. Since $d \geq \#R_j \geq 2$ in this case, we have $R_l \neq \{y_1, y_2\}$ so that we must have $R_l \subseteq F$. Now

$$\begin{aligned} rV &\succeq \frac{1}{g}\Big(\sum_{x\in R_l}(x-m_l)(x-m_l) + (a-m_j)(a-m_j)^{\mathrm{T}} + (b-m_j)(b-m_j)^{\mathrm{T}}\Big) \\ &\succeq \frac{1}{g}\Big(W(R_l) + \frac{1}{2}(a-b)(a-b)^{\mathrm{T}}\Big), \end{aligned}$$

where the last inequality follows from Steiner's formula. From [8], Lemma A.1(b), it follows

$$\det rV \geq \frac{1}{g^d}\det W(R_l)\Big(1+\frac{1}{2}(a-b)^{\mathrm{T}}W(R_l)^{-1}(a-b)\Big) \geq \frac{1}{g^d}\det W(R_l)\Big(1+\frac{1}{2\lambda_{\max}}\|a-b\|^2\Big),$$

where $\lambda_{\max}$ is the largest eigenvalue of $W(R_l)$. Since $R_l \subseteq F$, the eigenvalues of $W(R_l)$ do not depend on $K_1$ and $K_2$. Moreover, $\|a-b\|$ tends to infinity with $K_1$ and $K_2$. Therefore, (18) and Lemma A.5 imply that the value of the objective function converges to 0 as $K_1, K_2 \to \infty$, a contradiction to (19). This proves $(\gamma)$.

Now choose $K_1$ and $K_2$ so large that the MAP–clustering 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 Remark (The case $g = 1$)

In the case of one component, the criterion (10) reduces to Rousseeuw's [26] maximum covariant 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 $\alpha$, see Rousseeuw [26], p.291. This is in harmony with our result on the scatter matrix, Theorem 3.2. For $g = 1$, reduction of the parameter $r$ has the effect that breakdown occurs at a much higher number of outliers compared with $g > 1$. The reason is that, in the case $g > 1$, the outliers may form an own cluster if they are close to each other, thus causing one mean to be large.

For $g = 1$, also the EMT–algorithm is well known. The weights are all 1 so that the E–step is trivial. The M– and T–steps reduce to Rousseeuw and Van Driessen's [27], Theorem 1, alternating C–step for computing the MCD.

## A Appendix

### A.1 Lemma

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

(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 $\Psi$ is real valued and strictly concave in the interior of $\Delta_{g-1}$.

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

(b) Under the first assumption of (b) each summand $\mathbf{u} \to \ln(A\mathbf{u})_i$ is finite in the interior of $\Delta_{g-1}$ and so is their sum $\Psi$. Under the second the mapping $\mathbf{u} \to A\mathbf{u}$ 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 $\ln A\{\frac{1}{2}\mathbf{u} + \frac{1}{2}\mathbf{v}\}_i = \ln\{\frac{1}{2}(A\mathbf{u})_i + \frac{1}{2}(A\mathbf{v})_i\} > \frac{1}{2}\{\ln(A\mathbf{u})_i + \ln(A\mathbf{v})_i\}$ and claim (b) follows. □

## A.2 Proof of formula (18)

If $(\mathbf{u}, \mathbf{m}, V)$ is a fixed point then it is output from an M–step with input $w(x,l) = \frac{u_l N_{m_l,V}(x)}{\sum_{j=1}^{g} u_j N_{m_j,V}(x)}$. We may hence compute

$$\log f[R \mid \mathbf{u}, \mathbf{m}, V] = \sum_{x \in R} \log \sum_{j=1}^{g} u_j N_{m_j,V}(x) = \sum_{x \in R} \log \frac{u_l N_{m_l,V}(x)}{w(x,l)}$$
$$= \sum_{x \in R} \{\log u_l + \log N_{m_l,V}(x) - \log w(x,l)\}.$$

This expression does not depend on $l$ and we continue

$$= \sum_{l} \sum_{x \in R} w(x,l)\{\log u_l + \log N_{m_l,V}(x) - \log w(x,l)\}$$
$$= \sum_{l} u_l \log u_l + \sum_{l} \sum_{x \in R} w(x,l) \log N_{m_l,V}(x) - \sum_{l} \sum_{x \in R} w(x,l) \log w(x,l).$$

It is now sufficient to insert $V$ from Eqns. (15)–(17) and to apply standard matrix analysis to derive (18). □

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

## A.3 Lemma

Let $(x_1 \ldots, x_m)$ be a data set in $\mathbb{R}^d$, let $\mathbf{w} = (w_1, \ldots, w_m)$ be a family of real numbers such that $w(1..m) := \sum_i^m w_i > 0$, let $b \in \mathbb{R}^d$, and let $m_{\mathbf{w}} = \frac{1}{w(1..m)} \sum_{i=1}^m w_i x_i$, the weighted mean. Then

$$\sum_{i=1}^{m} w_i (x_i - b)(x_i - b)^{\mathrm{T}} = \sum_{i=1}^{m} w_i (x_i - m_{\mathbf{w}})(x_i - m_{\mathbf{w}})^{\mathrm{T}} + \sum_{i=1}^{m} w_i \cdot (m_{\mathbf{w}} - b)(m_{\mathbf{w}} - b)^{\mathrm{T}}.$$

In particular,

$$\sum_{i=1}^{m} w_i (x_i - b)(x_i - b)^{\mathrm{T}} \succeq \sum_{i=1}^{m} w_i (x_i - m_{\mathbf{w}})(x_i - m_{\mathbf{w}})^{\mathrm{T}}.$$

### A.4 Lemma

Let $p \geq d$, let $x_1, \ldots, x_p \in \mathbb{R}^d$ be in general position, and let $K > 0$.

(a) There exists $y \in \mathbb{R}^d$ such that $\|y\| \geq K$ and $\det W(F \cup \{y\}) \geq K$ for all $d$–element subsets $F \subseteq \{x_1, \ldots, x_p\}$.

(b) There exists a pair of points $y_1, y_2 \in \mathbb{R}^d$, such that $0 < \|y_1 - y_2\| \leq 1$, $\|y_1\|, \|y_2\| \geq K$, and $\det W(F \cup \{y_1\})$, $\det W(F \cup \{y_2\}) \geq K$ for all $d$–element subsets $F \subseteq \{x_1, \ldots, x_p\}$ and $\det W(E \cup \{y_1, y_2\}) \geq K$ for all $(d-1)$–element subsets $E \subseteq \{x_1, \ldots, x_p\}$ if $d > 1$.

**Proof.** We repeatedly use the following formula. Let $A$ be a singular, positive, symmetric matrix with a one–dimensional kernel and let $w \in \mathbb{R}^d$. Then

$$\det(A + ww^{\mathrm{T}}) = (v_A^{\mathrm{T}} w)^2 \det A', \tag{20}$$

where $v_A$ is a normalized vector in the kernel of $A$ and where $A'$ is the matrix $A$ seen as an operator on its "own" space $v_A^{\perp}$. If $A$ is the SSP–matrix $W(F)$ generated by a point set $F$ then we also write $v_F = v_A$.

(a) For any point $y$, the SSP–matrix of $F \cup \{y\}$ has the representation

$$W(F \cup \{y\}) = W(F) + \frac{d}{d+1}(y - m_F)(y - m_F)^{\mathrm{T}}.$$

The SSP–matrix $W(F)$ being singular, the determinant of $W(F \cup \{y\})$ is according to (20)

$$\det W(F \cup \{y\}) = \frac{d}{d+1}\big(v_F^{\mathrm{T}}(y - m_F)\big)^2 \det W'(F). \tag{21}$$

Since there are only finitely many such subsets $F$ there exists a vector $u$ not parallel to any of the hyperplanes, i.e. $v_F^T u \neq 0$ for all such $F$. Eqn. (21) shows that $y = \alpha u$ has the required properties if $\alpha$ is sufficiently large.

(b) Let $u, w \in \mathbb{R}^d$ and let $y_1 = \beta u + w,\ y_2 = \beta u - w$. By Lemma A.3, we have

$$\begin{aligned} W(E \cup \{y_1, y_2\}) =& W(E) + W(\{y_1, y_2\}) + \frac{2(d-1)}{d+1}(m_{\{y_1,y_2\}} - m_E)(m_{\{y_1,y_2\}} - m_E)^{\mathrm{T}} \\ =& W(E) + 2ww^{\mathrm{T}} + \frac{2(d-1)}{d+1}(\beta u - m_E)(\beta u - m_E)^{\mathrm{T}}. \end{aligned}$$

Here, $W(E) + 2ww^{\mathrm{T}}$ is the SSP matrix of $E \cup \{m_E \pm w\}$. A double application of Eqn. (20), first in dimension $d-1$ then in $d$, shows that its determinant is

$$\det W(E \cup \{y_1, y_2\}) = \frac{4(d-1)}{d+1}\big(v_{E \cup \{m_E + w\}}^{\mathrm{T}}(\beta u - m_E)\big)^2 (v_E^{\mathrm{T}} w)^2 \det W''(E), \tag{22}$$

where $W''(E)$ is the matrix $W(E)$ seen as an operator on its own space.

Now, choose $w \in \mathbb{R}^d$, $\|w\| = 1/2$, not parallel to any of the $(d-2)$–dimensional affine subspaces spanned by $d-1$ points in $\{x_1, \ldots, x_p\}$ and choose $u$ not parallel to any of the hyperplanes spanned by $d$–element subsets of $\{x_1, \ldots, x_p, m_E + w\}$. Formulae (22) and (21) show that the points $y_1$ and $y_2$ defined above have the required properties if $\beta$ is large enough. □

### A.5 Lemma

If $\#R = r$ and if $u_j = \frac{1}{r}\sum_{x\in R} w(x,j)$, $j \in 1..g$, then

$$r\sum_{j=1}^{g} u_j \log u_j - \sum_{j=1}^{g}\sum_{x\in R} w(x,j)\log w(x,j) \leq 0.$$

**Proof.** The inequality follows from the entropy inequality applied to the probabilities $w(x,j)/r$ and $p(x,j) = u_j/r$ on the set $R \times 1..g$. □

## References


[1] Stéphane Chrétien and Alfred O. Hero III. Kullback proximal algorithms for maximum likelihood estimation. *IEEE Trans. Inf. Theory*, 46:1800–1810, 2000.

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

[3] N.E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474, 1969.

[4] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

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

[6] George S. Fishman. *Monte Carlo*. Springer, New York etc., 1996.

[7] Chris Fraley and Adrian E. Raftery. How many clusters? which clustering method? answers via model–based cluster analysis. *The Computer Journal*, 41:578–588, 1998.

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

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

[10] V. Hasselblad. Estimation of parameters for a mixture of normal distributions. *Technometrics*, 8:431–444, 1966.

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

[12] Christian Hennig. Breakdown points for maximum likelihood estimators of location–scale mixtures. *Ann. Stat.*, 32:1313–1340, 2004.

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

[14] Peter J. Huber. *Robust Statistics*. Wiley, New York–Chichester–Brisbane–Toronto, 1981.

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

[16] N.M. Kiefer. Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica*, 46:427–434, 1978.

[17] Donald E. Knuth. *The Art of Computer Programming*, volume 2. Addison–Wesley, Reading, Menlo Park, London, Amsterdam, Don Mills, Sydney, 2nd edition, 1981.

[18] Hendrik P. Lopuhaä and Peter J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19:229–248, 1991.

[19] B. Martinet. Régularisation d'inéquation variationnelles par approximations successives. *Rev. Française d'Inform. et de Recherche Opérationnelle*, 3:154–179, 1970.

[20] Geoffrey J. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.

[21] Geoffrey J. McLachlan and David Peel. Robust cluster analysis via mixtures of multivariate t-distributions. In *Advances in Pattern Recognition*, volume 1451 of *Lecture Notes in Computer Science*, pages 658–666. Springer, 1998.

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

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

[24] Ralph Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Contr. Optimiz.*, 14:877–898, 1976.

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

[26] Peter J. Rousseeuw. Multivariate estimation with high breakdown point. In Wilfried Grossmann, Georg Ch. Pflug, István Vincze, and Wolfgang Wertz, editors, *Mathematical Statistics and Applications*, volume 8B, pages 283–297. Reidel, Dordrecht–Boston–Lancaster–Tokyo, 1985.

[27] Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

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