Maximum Likelihood Estimation of Heterogeneous Mixtures of Gaussian and Uniform Distributions

Pietro Coretto
Department of Economics and Statistics
Università degli Studi di Salerno
pcoretto@unisa.it

Christian Hennig
Department of Statistical Sciences
University College London
chrish@stats.ucl.ac.uk

June 2009

Abstract

Existence and consistency of the Maximum Likelihood estimator of the parameters of heterogeneous mixtures of Gaussian and uniform distributions with known number of components are shown under constraints to prevent the likelihood from degeneration and to ensure identifiability. The EM-algorithm is discussed, and for the special case with a single uniform component a practical scheme to find a good local optimum is proposed. The method is compared theoretically and empirically to estimation of a Gaussian mixture with "noise component" as introduced by Banfield and Raftery (1993) to find out whether it is a worthwhile alternative particularly in situations with outliers and points not belonging to the Gaussian components.

Keywords: model-based clustering, robustness, identifiability, EM-algorithm, Hathaway constraints, noise component

1 Introduction

The original motivation behind this paper was to investigate theoretically an idea of Banfield and Raftery (1993), Fraley and Raftery (1998) for robust estimation of the parameters of a Gaussian mixture distribution and model-based clustering. In the context of maximum likelihood (ML) parameter estimation via the EM-algorithm, they suggested to add a uniform mixture component over the convex hull of the data to a mixture of Gaussians in order to catch outliers and points not belonging to any Gaussian cluster. In the present paper, for simplicity reasons, we discuss the one-dimensional setup only, so the "convex hull" is just the range of the data. They called the uniform component the "noise component". We will call their approach "R-method" (for "range") in the following.

At first sight it may seem as if the R-method yielded an ML-estimator of a mixture model with $s-1$ Gaussian and a single uniform mixture component, but this is not true.

Neither does the method define an estimator that is consistent for the parameters of such a model, as is explained in Section 3, Remark 3. Our aim is not, however, to criticise the R-method, which is rather intended to deal with robustness problems than to provide a reliable estimator for a general Gaussian-uniform mixture.

In the present paper we investigate "proper" ML-estimation in a mixture model with Gaussian and uniform components. We introduce the model and confirm its identifiability in Section 2. In Section 3 we show existence and consistency under a condition preventing the likelihood from degeneration, which is a generalisation of the one given by Hathaway (1985) for Gaussian mixtures. Note that because of the discontinuity of the likelihood function due to the uniform density, existing general results cannot be applied in a straightforward way. In these theoretical sections, we treat a general situation with q uniform and $s-q$ Gaussian mixture components.

In practice, we are interested in whether such a method provides a useful alternative to the R-method for robust Gaussian clustering, and, as Banfield and Raftery, we concentrate on a situation with a single uniform component, so $q = 1$, when discussing the EM-algorithm and a computationally feasible implementation of the method in Section 4. In Section 5.1 we give a brief overview on a comparative simulation study carried out by Coretto (2008) and apply the ML-estimator and the R-method to a small real dataset. Section 6 presents some concluding discussion.

Situations where estimation with a larger number of uniform components q is appropriate may be rare in practice and an algorithm for this is more difficult to implement, so we leave this to future research, as well as the multivariate setup and the estimation of the number of mixture components.

2 Identifiability of Gaussian/uniform mixtures

2.1 Model and basic notation

In this section we introduce the notation and the model under study. Let $0 < s < \infty$ be the number of mixture components, and let q be the number of uniform components $0 < q < s$. Let X be a real valued random variable distributed according to the following distribution function:

$$G(x;\eta) = \sum_{k=1}^{q} \pi_k U(x;\theta_k) + \sum_{l=q+1}^{s} \pi_l \Phi(x;\theta_l), \tag{1}$$

where $\eta = (\pi, \theta)$, $\pi = (\pi_1, \pi_2, \ldots, \pi_s)$, $0 < \pi_j < 1$, $\sum_{j=1}^{s} \pi_j = 1$. For $k = 1, 2, \ldots, q$. For $l = q+1, q+2, \ldots, s$, let $\theta = (\theta_1, \theta_2, \ldots, \theta_s)$, where $\theta_k = (a_k, b_k)$, a_k and b_k take

values on the real line, and $-\infty < a_k < b_k < +\infty$ (let Θ_1 be the set of such pairs). For $l = q+1, q+2, \ldots, s$, let $\theta_l = (\mu_l, \sigma_l) \in \Theta_2 = \mathbb{R} \times (0, \infty)$. Furthermore, the parameter space is denoted by $\Gamma = (0,1)^s \times \Theta_1^q \times \Theta_2^{s-q}$. $U(\bullet; \theta_k)$ is the cdf of a uniform(a_k, b_k)-distribution, $\Phi(\bullet, \theta_l)$ is the cdf of a Gaussian(μ_l, σ_l^2)-distribution. We often rewrite the model in (1) as

$$G(x, \eta) = \sum_{j=1}^{s} \pi_j F_{z_j}(x; \theta_j) \tag{2}$$

where $z_j = \{1, 2\}$ for $j = 1, 2, \ldots, s$, $z_j = 1 \Rightarrow F_{z_j} = U$, $z_j = 2 \Rightarrow F_{z_j} = \Phi$. Moreover $g(x; \eta)$ will denote the density of $G(x; \eta)$. Later we will continue to denote the components of the vector η as π_k, a_k, b_k or μ_k, σ_k, and when we use notation such as η', the components will be called correspondingly, for example, π'_k, a'_k, b'_k or μ'_k, σ'_k.

2.2 Identifiability

Identifiability is a necessary condition for the possibility to estimate the parameters of a mixture model consistently. It makes sure that no two essentially different mixture parameter vectors parameterize the same distribution.

Let $\mathscr{F} = \{F(\bullet; \theta) : \quad \theta \in \Theta\}$ be a family of distribution functions indexed by a parameter from some parameter space Θ. Let $G \in \mathscr{G}$ be a distribution on Θ. $H(x, G) = \int F(x; \theta) dG(\theta)$ defines an $\mathscr{F}$-mixture distribution. In the present paper we are dealing with finite mixtures, i.e., $\mathscr{G}$ being the set of distributions on Θ with finite support. Consequently, we write $G(\theta)$ for the mixture proportion corresponding to $F(\bullet; \theta)$. Following Teicher (1961), the mixture model generated by the family $\mathscr{F}$ with mixing distribution in $\mathscr{G}$ is said to be identifiable if for $G_1, G_2 \in \mathscr{G}$: $H(\bullet, G_1) = H(\bullet, G_2) \Leftrightarrow G_1 = G_2$.

Because we consider heterogeneous mixtures here, the parameter set Θ has to include an indication of whether F is a uniform or Gaussian distribution. We therefore define

$$\Theta = \{(z, a, b) \in \{1, 2\} \times \mathbb{R}^2 : \; z = 1 \Rightarrow a < b, \; z = 2 \Rightarrow b \in (0, \infty)\},$$
$$z = 1 \Rightarrow F(\bullet, (z, a, b)) = U(\bullet, a, b), \; z = 2 \Rightarrow F(\bullet, (z, a, b)) = \Phi(\bullet, a, b).$$

Furthermore, in order to make sure that the uniform mixture components can be identified, a further constraint is needed. Let $\mathscr{G}$ be the set of distributions G on Θ with finite support so that

$$\forall \theta_1 = (z, a_1, b_1), \theta_2 = (z, a_2, b_2) \in \Theta \text{ with } z = 1, \; G(\theta_1) > 0, \; G(\theta_2) > 0 :$$
$$\text{either } b_1 < a_2 \text{ or } b_2 < a_1. \tag{3}$$

It can easily be seen by, for example, obtaining a single uniform$(0,1)$-distribution as a mixture of any pair of uniforms on $(0,y)$ and $(y,1)$ with suitable proportions, that mixtures of uniform distributions are not identifiable without such a constraint.

Theorem 1. *The class $\mathscr{F}$ with parameter set Θ and mixing distributions in $\mathscr{G}$ as defined above is identifiable.*

Proof. (Idea.) Considering $H(\bullet, G)$, the parameters $\theta = (z, a, b)$ for which $G(\theta) > 0$ and $z = 1$, i.e., those belonging to the uniform components, have to be the points at which H is not differentiable, and because of (3), successive pairs of neighboring points define the components. The probabilities $G(\theta)$ for these components can be found by comparing the upper and lower limit of the derivative of H at these points. Define G_U to be the measure that only assigns mass $G_U(\theta) = G(\theta)$ to $\theta = (z, a, b)$ with $z = 1$, but 0 to the rest of Θ ($G_U(\theta)$ are already identified). Let $H_U(x, G_U) = \int F(x;\theta) dG_U(\theta)$, $H_G = \frac{H - H_U}{1 - G_U(\Theta)}$. H_G is a Gaussian mixture. Gaussian mixtures are identifiable by Theorem 3 in Yakowitz and Spragins (1968), and this identifies $G(\theta)$ for $\theta = (2, a, b)$. ■

The proof still holds for the more general case of mixtures of uniforms with any identifiable family of distributions with absolutely continuous cdf.

Note that identifiability of the mixing distribution G includes identification of the number of uniform and Gaussian mixture components, but it only identifies the mixture parameter vector up to permutations of the mixture components, i.e., "label switching". In this sense, the present situation is not fully identifiable.

3 Maximum likelihood estimation

In this section we will study the ML-estimation of the distribution in (1) when s and q are fixed and known. We will show that under some constraints on the parameter space the ML-estimator exists. Furthermore we will show that this estimate is strongly consistent.

3.1 Existence

Day (1969) studied finite mixtures of normal distributions. He highlighted several issues including the problem of the unboundedness of the likelihood function. Let us assume that for a given sample $\underline{X_n} = \{X_1, X_2, \ldots, X_n\}$ is an i.i.d sequence of random variables distributed according to a finite mixture of m Gaussian distributions. The log-likelihood function associated with a realization $\underline{x_n} = \{x_1, x_2, \ldots, x_n\}$ of $\underline{X_n}$ is given by

$$L_n(\xi) = \sum_{i=1}^{n} \log p(x_i; \xi)$$

where $p(x;\xi)$ is the density of a finite mixture of m Gaussian densities. Here $\xi = (\pi_1, \ldots, \pi_m, \mu_1, \sigma_1, \ldots, \mu_m, \sigma_m)$, with π_j being the proportion of the jth component, and μ_j and σ_j being the mean and standard deviation of the jth component respectively; $j = 1, 2, \ldots, m$. If we fix $\mu_j = x_j$ and take σ_j arbitrarily close to 0, then $L_n(\xi) \longrightarrow +\infty$. This means that a global maximum fails to exist. As noted by Tanaka and Takemura (2006), this problem also affects a wider class of mixtures. Particularly, it affects uniform(a, b)-mixtures as well if we fix $a = x_j$ and let $b \to a$.

The constraints $\sigma_j \geq c > 0$, for all $j = 1, 2, \ldots, m$, are frequently used (e.g., DeSarbo and Cron (1988)) to overcome this problem. However the choice of the constant c is critical. If c is chosen large enough that for some j the true $\sigma_j < c$, the ML-estimator is obviously not consistent. Tanaka and Takemura (2006) considered constraints of the type $\sigma_j \geq c_n$, $c_n = c_0 \exp(-n^d)$, $j = 1, 2, \ldots, m$. As $n \to \infty$, $c_n \to 0$. Under this type of constraints the authors showed that a sequence of ML-estimates is strongly consistent. A drawback of these restrictions is that ML-estimators are no longer scale equivariant.

Dennis (1981) proposed to constraint the parameter space imposing that $\min_{i,j} \sigma_i/\sigma_k \geq c$ for a constant $c \in (0, 1]$, $i, j = 1, 2, \ldots, s$. Hathaway (1985) showed that these constraints lead to a well posed (though somewhat difficult) optimization program and that the corresponding sequence of ML-estimates is strongly consistent and it is obviously scale equivariant.

Let now $\underline{X_n} = \{X_1, X_2, \ldots, X_n\}$ be a sequence of i.i.d random variables with distribution function $G(x;\eta)$ according to model (1). Let $\underline{x_n} = \{x_1, x_2, \ldots, x_n\}$ be a realization of $\underline{X_n}$ with associated log-likelihood function

$$L_n(\eta) = \sum_{i=1}^{n} \log g(x_i;\eta). \tag{4}$$

We denote $v_j = \sigma_j$ for $j = q+1, \ldots, s$ and $v_j = (b_j - a_j)/\sqrt{12}$ for $j = 1, 2, \ldots, q$ (the standard deviation of the jth uniform component). For $c \in (0, 1]$, we define the constrained parameter set

$$\Gamma_c = \left\{ \eta \in \Gamma : \min_{t,r} \frac{v_t}{v_r} \geq c > 0 \right\}. \tag{5}$$

Remark 1. This constraint implies that if one of the scale parameters converges to zero, all the others converge to zero at the same rate. Define $v_{min} = \min\{v_j; j = 1, \ldots, s\}$ and $v_{max} = \max\{v_j; j = 1, \ldots, s\}$. The constraint above implies $v_{min} \geq c v_{max}$. This implies that the parameters of the kth uniform components have to be such that $b_j - a_j \geq \sqrt{12} c v_{max}$. Define, for later use, $\sigma_{max} = \max\{v_j; j = q+1, \ldots, s\}$, $r_{max} = \sqrt{12} c \sigma_{max}$.

We define the constrained ML-estimator as

$$\hat{\eta}_n = \arg\max_{\eta \in \Gamma_c} L_n(\eta) \tag{6}$$

The existence of $\hat{\eta}_n$ is not immediate. Γ_c is not compact, and moreover $L_n(\eta)$ is not continuous on Γ and Γ_c.

The following lemmas and remarks are used in order to show that $L_n(\eta)$ achieves its maximum on Γ_c.

Remark 2. Let $\eta \in \Gamma_c$ be such that the parameters of the jth uniform component fixed to be $a_j = x_p < b_j = x_t$, for some $j \in \{1, \ldots, q\}$ and $p \neq t = 1, \ldots, n$. Let $N_\varepsilon^-(x_p) \equiv [x_p - \varepsilon, x_p)$ and $N_\xi^+(x_t) \equiv (x_t, x_t + \xi]$, where ε and ξ are positive real numbers fixed so that $N_\varepsilon(x_p)$ and $N_\xi(x_t)$ do not contain any data point other than x_p and x_t respectively. If $\eta' \in \Gamma_c$ coincides with η except that $a'_j \in N_\varepsilon(x_p)$ and $b'_j \in N_\xi(x_t)$, it follows that $L_n(\eta') < L_n(\eta)$. In fact, in order to maximize the log-likelihood function with respect to the parameters of the jth uniform component, we need to choose the parameters so that the length of the support of the jth uniform density is minimized for any given number of data points contained in it.

Lemma 1. *If $\underline{x_n}$ contains at least $s+1$ distinct points, then $\sup_{\eta \in \Gamma_c} L_n(\eta) = \sup_{\eta \in \bar{\Gamma}_c} L_n(\eta)$, where $\bar{\Gamma}_c$ is a compact set contained in Γ_c.*

Proof. Denote $m_n = \min\{x_i, \quad i = 1, \ldots, n\}$ and $M_n = \max\{x_i, \quad i = 1, \ldots, n\}$.
Part A. Take $\eta' \in \Gamma_c$ with $\mu'_j \leq m_n$ for some $j = q+1, \ldots, s$. Consider the vector $\eta'' \in \Gamma_c$ that is equal to η' except that $\mu''_j = m_n$. This implies that $L_n(\eta') \leq L_n(\eta'')$. By analogy, $\mu'_j > M_n$ can be ruled out.
Part B. Consider $\eta' \in \Gamma_c$ with $a'_k \leq m_n - r_{max}$ (the case $b'_k \geq M_n + r_{max}$ can be treated by analogy) for some $k = 1, \ldots, q$. Assume that $b'_k \geq m_n$ and that $a'_l > m_n$ for all other $k \neq l = 1, \ldots, q$ (otherwise the contribution of these components to the likelihood were 0 and they could be re-arranged without changing the likelihood so that $a'_l \geq m_n - q r_{max}$ $\forall l = 1, \ldots, q$). Be $\eta'' \in \Gamma_c$ equal to η' except that $a''_k = m_n - r_{max}$. By the arguments given in Remark 2 it follows that $L_n(\eta'') \geq L_n(\eta')$.
Part C. Recall Remark 1, Consider a sequence $\{\eta^t\}_{t \geq 1}$ such that $v_j^t \downarrow 0$ for all $j = 1, \ldots, s$ while all the other parameters are fixed (w.l.o.g. fix a_k^t and let $b_k^t \to a_k^t$ for $k = 1, \ldots, q$). For each $t \geq 1$, fix, w.l.o.g., $a_j^t = x_j$ for all $j = 1, \ldots, q$ and $\mu_j^t = x_j$ for all $j = q+1, \ldots, s$ (the contribution to the likelihood of any component for which this would not hold would converge to 0). By assumption the vector $\underline{x_n}$ contains at least $s+1$ points. Assume that

$\underline{x_n}$ contains just $s+1$ points (the case $n > s+1$ goes along the same lines). Then,

$$L_{s+1}(\eta^t) = \log\left(g(x_{s+1};\eta^t)\prod_{i=1}^{s} g(x_i;\eta^t)\right). \tag{7}$$

It is possible to write

$$\begin{gathered}\textstyle\prod_{i=1}^{s} g(x_i;\eta^t) = \sum_{h=1}^{s^n} \bar{\pi}(\gamma_h)\bar{g}(\underline{x_n};\gamma_h,\eta^t), \\ \text{where } \bar{g}(\underline{x_n};\gamma,\eta) = \textstyle\prod_{r=1}^{n} f_{z_r}(x_r;\theta_{j_r}), \quad \bar{\pi}(\gamma) = \prod_{r=1}^{n}\pi_{j_r},\end{gathered} \tag{8}$$

for $\gamma = (j_1, j_2, \ldots, j_n)$, $(j_1, j_2, \ldots, j_n)$ with $j_r \in \{1, 2, \ldots, s\}$ for all $r = 1, 2, \ldots, n$, and $\gamma_1, \ldots, \gamma_{s^n}$ are all possible vectors γ. Consider

$$\bar{g}(\underline{x_n};\gamma,\eta^t) = f_{z_{j_{s+1}}}(x_{s+1};\eta^t)\prod_{i=1}^{s} f_{z_{j_i}}(x_i;\eta^t).$$

If $z_{j_{s+1}} = 1$ (uniform component), $f_{z_{j_{s+1}}}(x_{s+1};\eta^t)$ and therefore $\bar{g}(\underline{x_n};\gamma,\eta^t)$ are eventually 0. If $z_{j_{s+1}} = 2$ (Gaussian component), with $\sigma = \sigma_{j_{s+1}} \downarrow 0$, there is a constant $d > 0$ so that $\prod_{i=1}^{s} f_{z_{j_i}}(x_i;\eta^t) \leq \frac{d}{\sigma^s}$, and therefore, because the Gaussian density converges faster to zero than every power of σ,

$$f_{z_{j_{s+1}}}(x_{s+1};\eta^t)\prod_{i=1}^{s} f_{z_{j_i}}(x_i;\eta^t) \leq \frac{d}{\sigma^s}\varphi(x_{s+1},\mu_{j_{s+1}},\sigma) \downarrow 0.$$

Therefore, $L_{s+1}(\eta^t) \longrightarrow -\infty$.

Part D. If one of the scale parameter gets arbitrarily large, v_j becomes arbitrarily large for all $j = 1, \ldots, s$, and both the uniform and Gaussian densities in (8) converge to zero, so $L_n(\eta^t) \longrightarrow -\infty$.

By A–D, $\sup_{\eta\in\Gamma_c} L_n(\eta) = \sup_{\eta\in\bar{\Gamma}_c} L_n(\eta)$; where $\bar{\Gamma}_c = [0,1]^s \times \bar{\Theta}_1 \times \bar{\Theta}_2$, with

$$\bar{\Theta}_1 = \left\{\theta_k \in \Theta_1 : m_n - \bar{r}_{max} \leq a_k < b_k \leq M_n + \bar{r}_{max}, k = 1, \ldots q\right\}, \tag{9}$$

and

$$\bar{\Theta}_2 = \left\{\theta_j \in \Theta_2 : m_n \leq \mu_j \leq M_n, \underline{\sigma} \leq \sigma_j \leq \overline{\sigma}, j = q+1, \ldots, s\right\}, \tag{10}$$

for some choice of the constants $\underline{\sigma}$, and $\overline{\sigma}$ such that $0 < \underline{\sigma} < \overline{\sigma} < \infty$, $\bar{r}_{max} = \sqrt{12}c\overline{\sigma}$. The sets $\bar{\Theta}_1$ and $\bar{\Theta}_2$ are now compact as well as the set $\bar{\Gamma}_c$. ■

Lemma 2. *Let $\underline{x_n}$ contain at least $s+1$ distinct points, and let $\eta^* \in \bar{\Gamma}_c$ be a local maximum for $L_n(\eta)$. Then η^* is such that for all $k = 1, 2, \ldots, q$, (a_k^*, b_k^*) either coincides*

with a pair of distinct points in $\underline{x_n}$*, or* (a_k^*, b_k^*) *is such that* $b_k^* - a_k^* = \sqrt{12}cv_{max}^*$*, where* $v_{max}^* = \max\{v_j^*, j = 1, \ldots, s\}$*, and the interval* $[a_k^*, b_k^*]$ *contains at least one data point.*

Proof. Let $\eta^*(a_k, b_k) \in \Gamma_c$ denote the parameter vector with all components equal to those of η^* except the parameters of the kth uniform component, which are set to be a_k, b_k. For a data point y, $N_\varepsilon^-(y) = [y - \varepsilon, y)$ and $N_\varepsilon^+(y) = (y, y + \varepsilon]$, where $\varepsilon > 0$ is such that $N_\varepsilon^-(y)$ and $N_\varepsilon^+(y)$ do not contain any data point. Let $\{\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(m)}\}$, $m \leq n$, be the ordered set of all distinct points of $\underline{x_n}$. Let us consider a pair of distinct data points, $\tilde{x}_{(d)}$ and $\tilde{x}_{(e)}$, with $d, e \in \{1, \ldots, n\}$, $d < e$, and $\tilde{x}_{(e)} - \tilde{x}_{(d)} \geq \sqrt{12}cv_{max}$. There are three cases: (i) the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ contains at least a pair of distinct data points; (ii) the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ does not contain any data point; (iii) the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ contains just one data point.

Case (i). Assume that the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ contains more than one distinct data points. Consider the points $\tilde{x}_{(d+1)}$ and $\tilde{x}_{(e-1)}$, with $d + 1 < e - 1$. There are two further cases: (i.a) $\tilde{x}_{(e-1)} - \tilde{x}_{(d+1)} \geq \sqrt{12}cv_{max}^*$; or (i.b) $\tilde{x}_{(e-1)} - \tilde{x}_{(d+1)} < \sqrt{12}cv_{max}^*$.

Case (i.a). First assume that $\tilde{x}_{(e-1)} - \tilde{x}_{(d+1)} \geq \sqrt{12}cv_{max}^*$. By Remark 2 conclude that for any possible $\varepsilon, \xi > 0$ and any $a_k \in N_\varepsilon^-(\tilde{x}_{(d)})$ and $b_k \in N_\xi^+(\tilde{x}_{(e)})$, $L_n(\eta^*(a_k, b_k)) < L_n(\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)}))$. Applying the same argument as above, for any possible $\varepsilon, \xi > 0$ and any $a_k \in N_\varepsilon^-(\tilde{x}_{(d+1)})$ and $b_k \in N_\xi^+(\tilde{x}_{(e-1)})$: $L_n(\eta^*(a_k, b_k)) < L_n(\eta^*(\tilde{x}_{(d+1)}, \tilde{x}_{(e-1)}))$. This means that either $\eta^*(\tilde{x}_{(d+1)}, \tilde{x}_{(e-1)})$ or $\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ are candidates for a local maximum. Case (i.b). Now assume $\tilde{x}_{(e-1)} - \tilde{x}_{(d+1)} < \sqrt{12}cv_{max}^*$. As before, for any possible $\varepsilon, \xi > 0$ and any $a_k \in N_\varepsilon^-(\tilde{x}_{(d)})$ and $b_k \in N_\xi^+(\tilde{x}_{(e)})$, it follows that $L_n(\eta^*(a_k, b_k)) < L_n(\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)}))$. Now, $(a_k, b_k) = ((\tilde{x}_{(d+1)}, \tilde{x}_{(e-1)})) \notin \Gamma_c$ since the constraint does not hold. Let us take any (a_k', b_k') such that $b_k' - a_k' = \sqrt{12}cv_{max}^*$ and $a_k' \leq \tilde{x}_{(d+1)} < \tilde{x}_{(e-1)} \leq b_k'$. The corresponding parameter η' now lies on the boundary of Γ_c. By the same argument as before, for any possible $\varepsilon, \xi > 0$ and any $a_k \in N_\varepsilon^-(a_k')$ and $b_k \in N_\xi^+(b_k')$, it follows that $L_n(\eta^*(a_k, b_k)) < L_n(\eta^*(a_k', b_k'))$. This means that either $\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ or $\eta^*(a_k', b_k')$ are candidates for a local maximum.

Case (ii). We assume that the interval $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ does not contain any data point. We can apply the same argument as before and show that $\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ is a local maximum.

Case (iii). Assume that $(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ contains just a single data point (or several identical ones). By applying the same argument as in part (i.b), we conclude that either $\eta^*(\tilde{x}_{(d)}, \tilde{x}_{(e)})$ and $\eta^*(a_k', b_k')$ are candidates for a local maximum; where a_k' and b_k' are such that $b_k' - a_k' = \sqrt{12}cv_{max}$ and $\tilde{x}_{(d)} \leq a_k'$ and $b_k' \leq \tilde{x}_{(e)}$.

Parts (i.a)–(iii) complete the proof, noting that in all cases the η^* has been updated by narrowing the supports of a uniform component, which implies that the constraint (3) remains fulfilled for all updates if it holds for η^*. ■

Theorem 2. *$L_n(\eta)$ achieves its maximum over Γ_c.*

Proof. Let η^* be a local maximum and $v^*_{max} = \max\{v^*_1, v^*_2, \ldots, v^*_s\}$. Lemma 1 implies that v^*_{max} is contained in a closed and bounded real interval. For all possible values of v^*_{max}, all possible values of the uniform parameters for which the corresponding η is a candidate for a local maximum can be obtained from Lemma 2, which leads to only finitely many possible values for $L_n(\eta)$. Let $\bar{L}_n(\bar{\eta})$ be the log-likelihood function when all uniform parameters are fixed in order to get a local or global maximum. Consider the vector of parameters $\bar{\eta} = (\pi_1, \ldots, \pi_s, \theta_{q+1}, \ldots, \theta_s)$ for any fixed uniform parameters from Lemma 2. The parameter $\bar{\eta}$ lies in $\in [0,1]^s \times \bar{\Theta}_2$ from (10). $\bar{L}_n(\bar{\eta})$ is continuous on the compact set $[0,1]^s \times \bar{\Theta}_2$, and hence has a maximum. Applying this argument for all possible values of v^*_{max} we can find all possible local maxima of $L_n(\eta)$ on Γ_c, and hence among these we get the global maximum. ■

Note that there is an ambiguity leading to non-uniqueness of the (local and therefore also global) maxima, because if $a_k - b_k = \sqrt{12}cv_{max}$ and either a_k or b_k do not coincide with a data point, any choice of a_k so that the same data points remain inside of $[a_k, b_k]$ leads to the same likelihood value.

3.2 Asymptotic analysis

The technique usually used to show consistency and asymptotic normality for ML-estimators assumes differentiability of the likelihood function besides other regularity conditions about continuity and integrability of derivatives of the likelihood function up to the third order. Here there are several problems: (i) model (1) implies a likelihood function with infinitely many discontinuity points; (ii) in order to achieve a global maximum for the log-likelihood we need to restrict the parameter space to a set Γ_c which is not compact; (iii) the distribution we want to estimate is identifiable only up to label switching.

Wald (1949) studied a general class of estimators of which ML is a particular case, and he showed strong consistency under general conditions not involving derivatives of the likelihood function. However, in Wald's approach it is assumed that the parameter space is compact and that the model is fully identifiable, which it is not in our case because of component label switching. Redner (1981) extended the results in Wald (1949). First he defined consistency for sequences of estimates of parameters of non-identifiable distributions, and then he showed the consistency of sequences of ML-estimators for such distributions. However, Redner's theory deals with compact parameter spaces. Kiefer and Wolfowitz (1956) studied the class of estimators introduced by Wald (1949) in the case when the parameter space is not compact. On the other hand the authors assume full

identifiability of the model as in Wald (1949).

Hathaway (1985) studied the strong consistency of the maximum likelihood sequence for finite mixtures of Gaussian distributions on a constrained set of the same kind as (5). The author used the theory of Kiefer and Wolfowitz (1956) with an approach similar to that employed by Redner (1981). Here we will adopt a similar approach.

Let $\eta^0 \in \Gamma_c$ denote the true parameter, i.e. $G(x, \eta^0)$ is the distribution which generated the sample $\underline{X_n}$. As in Kiefer and Wolfowitz (1956) we define a metric δ on Γ:

$$\delta(\eta, \eta_*) = \sum_{j=1}^{3s} \left| \arctan \eta^j - \arctan \eta_*^j \right|$$

for all $\eta, \eta_* \in \Gamma$ with η^j being the jth component of the vector η. We complete the set Γ_c with all limits of its Cauchy sequences. That is, $\bar{\Gamma}_c$ is the set Γ_c along with the limits of its Cauchy sequences in the sense of δ. As in Hathaway (1985) we will show that sufficient conditions given by Kiefer and Wolfowitz (1956) hold.

Let $Y = (X_1, X_2, \ldots, X_m)$ be a vector of m random variables i.i.d. according to $G(x; \eta)$. Let $g_m(y; \eta)$ the joint density of the components of Y.

Lemma 3. *We assume that $\left\{\eta^t\right\}_{t\geq 1}$ is a sequence in $\bar{\Gamma}_c$ and $\eta^* \in \bar{\Gamma}_c$. For every sequence $\eta^t \longrightarrow \eta^*$, $g_m(y; \eta^t) \longrightarrow g_m(y; \eta^*)$ holds; except perhaps on a set $E \subset \mathbb{R}^m$ which may depend on η^* and of which the Lebesgue measure is zero.*

Proof. We only have to take care of the discontinuities introduced by the uniform components. Let us take a sequence $\left\{\eta^t\right\}_{t\geq 1}$ converging to η^* in $\bar{\Gamma}_c$. If $y \in \mathbb{R}^m$, $y = (x_1, x_2, \ldots, x_m)$ is such that $x_i \neq a_k^*$ and $x_i \neq b_k^*$ for all $i = 1, 2, \ldots, m$ and $k = 1, 2, \ldots, q$, it is easy to see that the statement holds because $\mathbf{1}_{[a_k^t, b_k^t]}(x_i) \longrightarrow \mathbf{1}_{[a_k^*, b_k^*]}(x_i)$ for all k, i. This is not the case for all points $y' \in E$ where for some k and i there is some $a_k^* = x_i'$ and/or $b_k^* = x_i'$. Thus the statement above holds, in fact the set E depends on the limit point η^* and has zero Lebesgue measure. ■

The joint density of m observations $g_m(y; \eta)$ is itself a mixture of s^m components, see (8). The notation introduced there will be used in the following lemmas.

The following lemmas will be useful to show that Kiefer-Wolfowitz sufficient conditions for the consistency of the ML-estimator are satisfied for the joint density of m observations, with $m > s$. $\mathrm{E}_{\eta'} f$ denotes the expectation of the function f under the distribution G with the parameter η'.

Lemma 4. *For any $m > s$, $E_{\eta^0} \log g_m(y;\eta^0) > -\infty$.*

Proof. Let us choose h^* such that $\gamma_{h^*} = \{j^*, j^*, \ldots, j^*\}$, $j^* \in \{q+1, \ldots, s\}$. Note that $\mathrm{E}_{\eta^0} \log \varphi(x; \mu^0_{j^*}, \sigma^0_{j^*}) > -\infty$. The following chain of inequalities completes the proof:

$$\mathrm{E}_{\eta^0} \log g_m(y;\eta^0) = \mathrm{E}_{\eta^0} \log \left(\sum_{h=1}^{s^m} \bar{\pi}(\gamma_h) g(y;\gamma_h,\eta^0) \right) \geq \mathrm{E}_{\eta^0} \log(\pi(\gamma_{h^*}) \bar{g}(y;\gamma_{h^*},\eta^0)) \geq$$

$$\log \pi(\gamma_{h^*}) + \mathrm{E}_{\eta^0} \log \bar{g}(y;\gamma_{h^*},\eta^0) \geq \log \pi(\gamma_{h^*}) + \mathrm{E}_{\eta^0} \sum_{r=1}^{m} \log \varphi(x_r; \mu^0_{j^*}, \sigma^0_{j^*}) \geq$$

$$\log \pi(\gamma_{h^*}) + \sum_{r=1}^{m} \mathrm{E}_{\eta^0} \log \varphi(x_r; \mu^0_{j^*}, \sigma^0_{j^*}) > -\infty$$

■

Lemma 5. *Let X and Y be two random variables independently distributed according to G, and let E_{η^0} denote the expectation under G, then*

$$E_{\eta^0} \sup_{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+} \log \left(\frac{1}{\sigma^t} \varphi(x;\mu,\sigma)\varphi(y;\mu,\sigma) \right) < +\infty, \tag{11}$$

for any finite $t \geq 1$.

Proof. Observe

$$B(\mu,\sigma) = \log \left(\frac{1}{\sigma^t} \varphi(x;\mu,\sigma)\varphi(y;\mu,\sigma) \right) = \log \left(\frac{1}{2\pi\sigma^{t+2}} \exp\{-\frac{1}{2\sigma^2}[(x-\mu)^2 + (y-\mu)^2]\} \right)$$

for some $t \geq 1$. The maximum of $B(\mu,\sigma)$ exists on $\mathbb{R}\times\mathbb{R}_+$; this can be verified along the same line of the arguments given in proof of Lemma 1 (parts A, C). The maximum is achieved at

$$\mu^* = \frac{x+y}{2}, \quad \text{and} \quad \sigma^* = \frac{|x-y|}{\sqrt{2(t+2)}}, \text{ therefore}$$

$$B(\mu^*,\sigma^*) = \log \frac{2(t+2)^{\frac{t+2}{2}} \exp\{-(t+2)/2\}}{|x-y|^{t+2}} = \log \frac{T}{|x-y|^{t+2}}$$

for $0 < T < +\infty$, where the constant T depends on t, and

$$\mathrm{E}_{\eta^0} \sup_{(\mu,\sigma)\in\mathbb{R}\times\mathbb{R}_+} \log \frac{1}{\sigma^t} \varphi(x;\mu,\sigma)\varphi(y;\mu,\sigma) = \mathrm{E}_{\eta^0} B(\mu^*,\sigma^*), \tag{12}$$

$$\mathrm{E}_{\eta^0} B(\mu^*,\sigma^*) = \log T - (t+2) \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \log|x-y| g(x;\eta^0) g(y;\eta^0) dxdy < +\infty.$$

■

Lemma 6.

$$\textit{For any } m > s: \ E_{\eta^0} \sup_{\eta \in \bar{\Gamma}_c} \log g_m(y;\eta) < +\infty.$$

Proof. The statement holds if, for all possible indexes γ,

$$\mathrm{E}_{\eta^0} \sup_{\eta \in \bar{\Gamma}_c} \log \bar{g}(y;\gamma,\eta) < +\infty. \tag{13}$$

In order to show (13), a convenient parameterization of the uniform components in G in terms of their means and standard deviations is introduced. For all $k = 1, 2, \ldots, q$ fix $\mu_k = (a_k + b_k)/2$, $\sigma_k = (b_k - a_k)/\sqrt{12}$. With this, $u(x;\theta_k) = u(x;\mu_k,\sigma_k)$ with

$$u(x;\mu_k,\sigma_k) = \frac{1}{\sqrt{12}\sigma_k} \mathbf{1}_{[\mu_k - \sqrt{3}\sigma_k; \mu_k + \sqrt{3}\sigma_k]}.$$

Assume $m = s+1$. For each index γ all the factors of $\bar{g}(y;\gamma,\eta)$ are bounded over $\bar{\Gamma}_c$ unless σ_{z_r} approaches 0 and $\mu_{z_r} = x_r$ for some $r = 1, 2, \ldots, s+1$ and $z_r \in \{1, 2\}$. Let $\bar{g}_m(y;\gamma,\eta)$ be such that the means of $s-1$ of its components are equal to $s-1$ of the components of y. Hence for some indexes $h, t \in \{1, 2, \ldots, s+1\}$ and $z \in \{1, 2, \ldots, s+1\}$

$$\sup_{\eta \in \bar{\Gamma}_c} \log \bar{g}(y;\gamma,\eta) \leq \sup_{\eta \in \bar{\Gamma}_c} \log \left(Q \frac{1}{\sigma_z^{s-1}} f_{p_h}(x_h;\mu_z,\sigma_z) f_{p_t}(x_t;\mu_z,\sigma_z) \right), \tag{14}$$

where Q is a finite constant.

Consider the above inequality in three possible cases: (i) $z_h = z_t = 2$; (ii) $z_h = z_t = 1$; (iii) $z_h = 1$ and $z_t = 2$.

Case (i). If $z_h = z_t = 2$, then $f_{z_h} = f_{z_t} = \varphi$, applying the operator E_{η^0} on both the left and right-hand side of (14), by Lemma 5 the condition (13) holds, proving the statement.

Case (ii). If $z_h = z_t = 1$, then $f_{z_h} = f_{z_t} = u$. Introduce the function

$$\Delta_1(x_t, x_h; \mu_z, \sigma_z) = \log \frac{Q_1}{\sigma_z^{s+1}} \mathbf{1}_{[\mu_z - \sqrt{3}\sigma_z; \mu_z + \sqrt{3}\sigma_z]}(x_h) \mathbf{1}_{[\mu_z - \sqrt{3}\sigma_z; \mu_z + \sqrt{3}\sigma_z]}(x_t),$$

with Q_1 a finite constant. Note that $\Delta_1(x_t, x_h; \mu_z, \sigma_z) < T < +\infty$ for some T and any choice of μ_z and σ_z at any x_h and x_t, whence

$$\mathrm{E}_{\eta^0} \sup_{\eta \in \bar{\Gamma}_c} \Delta_1(x_t, x_h; \mu_z, \sigma_z) < T < +\infty.$$

This means that the condition (13) holds proving the statement.

Case (iii), $z_h = 1$ and $z_t = 2$, so $f_{p_h} = u$ and $f_{p_t} = \varphi$. Introduce

$$\Delta_2(x_t, x_h; \mu_z, \sigma_z) = \log \frac{Q_2}{\sigma_z^{s+1}} \mathbf{1}_{[\mu_z - \sqrt{3}\sigma_z; \mu_z + \sqrt{3}\sigma_z]}(x_h) \varphi(x_t; \mu_z, \sigma_z),$$

where Q_2 is some constant, leading to

$$\Delta_2(x_t, x_h; \mu_z, \sigma_z) = \begin{cases} \log\left(\frac{Q_2}{\sigma_z^{s+1}} \varphi(\frac{x_t - \mu_z}{\sigma_z})\right), & \text{if } x_t, x_h \in [\mu_z - \sqrt{3}\sigma_z; \mu_z + \sqrt{3}\sigma_z;]; \\ -\infty, & \text{otherwise.} \end{cases}$$

Observe for some T': $\Delta_2(x_t, x_h; \mu_z, \sigma_z) \leq T' < +\infty$ for any choice of (μ_z, σ_z) at any x_h and x_t except when $x_t = x_h$. When $x_t = x_h$, it is possible to take $\mu_z = x_t = x_h$ and $\sigma_z \downarrow 0$ making $\Delta_2(\mu_z, \sigma_z)$ approaching to $+\infty$. Note that the set of points where $x_h = x_t$ has zero Lebesgue measure in $\mathbb{R}^2$. Hence

$$\mathrm{E}_{\eta^0} \sup_{\eta \in \tilde{\Gamma}_c} \Delta_2(x_t, x_h; \mu_z, \sigma_z) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \sup_{\mu_z, \sigma_z} \Delta_2(x_t, x_h; \mu_z, \sigma_z) g(x_t; \eta^0) g(x_h; \eta^0) dx_h dx_t \leq T' < +\infty,$$

which implies (13), proving the statement. The proof is completed by noting that any $m > s + 1$ would not change the cases (i)–(iii). ■

The approach used by Redner (1981) and Hathaway (1985) to overcome the difficulty of component label switching is to work with a properly defined quotient topological space of the parameter set. Define the set

$$C(\eta') = \left\{ \eta \in \Gamma_c : \int_{-\infty}^{x} g(t; \eta) dt = \int_{-\infty}^{x} g(t; \eta') dt \quad \forall x \in \mathbb{R} \right\}.$$

Let $\tilde{\Gamma}_c$ be the quotient topological space obtained from Γ_c by identifying $C(\eta')$ to a point $\tilde{\eta}' = \eta'$. As in Redner (1981) it is possible to show strong consistency of the sequence of ML-estimates on the quotient space $\tilde{\Gamma}_c$. Let, for $\epsilon > 0$,

$$\mathscr{N}_\varepsilon(\eta') = \left\{ \eta \in \Gamma_c : \forall \eta^\star \in C(\eta') \quad \delta(\eta, \eta^\star) \leq \varepsilon \right\}.$$

Theorem 3. *For any $\varepsilon > 0$ there exists $h(\varepsilon) \in (0, 1)$ such that*

$$\Pr\left\{ \lim_{n \to \infty} \sup_{\eta \in \Gamma_c \setminus \mathscr{N}_\varepsilon(\eta^0)} \frac{\prod_{i=1}^n g(x_i; \eta)}{\prod_{i=1}^n g(x_i; \eta^0)} < h(\varepsilon)^n \right\} = 1 \tag{15}$$

Proof. Equation (15) follows from the result (2.12) in Kiefer and Wolfowitz (1956) (see also comments in Section 6 in Kiefer and Wolfowitz (1956) and the paper by Perlman (1972)), the assumptions of which can be verified using Lemmas 3, 4 and 6 and basic properties of the Gaussian and uniform distributions. ■

The result above implies convergence of the ML-estimator on the quotient space. The sequence of estimators defined in (6) is strongly consistent for $\tilde{\eta}^0$, i.e. $\hat{\tilde{\eta}}_n \xrightarrow{\text{as}} \tilde{\eta}^0$. By Theorem 1 this means that whenever n is infinitely large the sequence of estimates $\hat{\tilde{\eta}}_n$ converges almost surely to a point $\tilde{\eta}^0$ which coincides with η^0 up to permutation of the pairs (π_j^0, θ_j^0) via permutation of the indexes $j = 1, 2, \ldots, s$.

While we have shown that the sequence of ML-estimates converges with probability one to the true parameter on the quotient space, we do not provide any asymptotic normality result. Asymptotic normality cannot be expected to hold, because it does not even hold for estimation of the parameters of a homogeneous single uniform distribution, where the ML-estimator can only underestimate but never overestimate the width of the support.

Remark 3. The R-method consists, for a given dataset $\underline{x_n} = \{x_1, x_2, \ldots, x_n\}$, of setting $q = 1$ and defining an estimator $\hat{\eta}$ by fixing $\hat{a}_1 = \min \underline{x_n}$, $\hat{b}_1 = \max \underline{x_n}$. The remaining parameters (proportions for the uniform and $s - q$ Gaussian components and Gaussian parameters) are then estimated by maximizing the likelihood.

While the range of the data is the ML-estimator for a model with a single uniform distribution only, the R-method does not necessarily yield an ML-estimator for model (1) with $q = 1, s - q \geq 1$, though it may be equal to the ML-estimator for some datasets (an example where the ML-estimator is different is given in Section 6). As opposed to the homogeneous uniform model, the ML-estimator of the uniform support $[a_1, b_1]$ does not need to contain all data points, because points outside the support can still be fitted by the Gaussians.

Asymptotically it can be seen that the R-method cannot be ML, because it is not even consistent. Whatever the true value of a_1, b_1, the support of the Gaussian distributions is the whole real line, and therefore $\hat{a}_1 \to -\infty$, $\hat{b}_1 \to \infty$, which means that the noise component density vanishes asymptotically.

4 Computation via the EM algorithm

4.1 Behaviour of the EM-algorithm

We use the notation from (2) here. Be f_{z_j} the density of F_{z_j}. The EM algorithm is intended to seek a maximum for the log-likelihood function $l_n(\eta) = \sum_{i=1}^n \log g(x_i; \eta)$ over

the constrained set Γ_c. Our practical implementation and the simulations in Section 5.1 will deal with the case $q = 1$ only, but in the beginning of the present section we still allow general q.

Let the index $t = 1, 2, \ldots$ be the iteration index of the algorithm, and let us introduce the following notations:

$$w_{i,j}^{(t)} = \frac{\pi_j^{(t)} f_{z_j}(x_i; \theta_j^{(t)})}{g(x_i; \eta^{(t)})};$$

$$Q(\eta, \eta^{(t)}) = \sum_{j=1}^{s} \sum_{i=1}^{n} w_{i,j}^{(t)} \log \pi_j + \sum_{j=1}^{s} \sum_{i=1}^{n} w_{i,j}^{(t)} \log f_{z_j}(x_i; \theta_j).$$

The quantity $w_{i,j}^{(t)}$ can be interpreted as the estimated posterior probability at the iteration t that the observation x_i has been drawn from the jth mixture component. The EM-algorithm works as follows:

1. fix $\eta^{(0)} \in \Gamma_c$;
2. For all $t = 1, 2, ...,$ up to convergence do the following:
   (a) E–step: determine $Q(\eta, \eta^{(t)})$;
   (b) M–step: choose $\eta^{(t+1)} = \arg\max_{\eta \in \Gamma_c} Q(\eta, \eta^{(t)})$.

The definition (5) of Γ_c makes it difficult to find the arg max in the M-step. For simplicity, in the following, we ignore the constraint in (5) when discussing the EM-algorithm. The constraint can still be fulfilled by, in every iteration,

- computing the $\arg\max_{\eta \in \Gamma}$ in the M–step,
- checking whether $v_{min}^{(t+1)} \geq c v_{max}^{(t+1)}$ is fulfilled,
- if not, setting $v_j^{(t+1)} = c v_{max}^{(t+1)}$ for all components $j \in \{1, \ldots, s\}$ with $v_j^{(t+1)} < c v_{max}^{(t+1)}$ (this is straightforward for Gaussian components; for the uniform components we suggest to hold the interval midpoint fixed and to increase the width to r_{max}; in case of $q > 1$ some adjustment may be needed if this violates (3), which is very unlikely in practice unless q is chosen much too large for the given dataset),
- checking whether after this adjustment the likelihood of iteration $t + 1$ is still increased.
- If this is the case, the algorithm can continue, otherwise it is stopped at iteration t.

In the following, we call the combination of the EM-algorithm and the procedure to enforce the scale constraints EMC-algorithm. Note that the ECM-algorithm in case of a violation

of the constraint does not guarantee that a local optimum of the likelihood is found. However, this is difficult to achieve anyway. For pure Gaussian mixtures, the constrained algorithm proposed by Hathaway (1986) uses some modified constraints and its use of Lagrange multipliers does not easily generalize to the presence of uniform components. In a multivariate Gaussian mixture setup Ingrassia (2004) applies similar adjustments to the one proposed here in case of a violation of the constraints.

Ignoring the constraint now, the M-step at iteration t is to compute

$$\pi_j^{(t+1)} = n^{-1} \sum_{i=1}^{n} w_{i,j}^{(t)}, \qquad j = 1, 2, \ldots, s, \tag{16}$$

$$\theta_j^{(t+1)} = \arg\max_{\theta_j} \sum_{i=1}^{n} w_{i,j}^{(t)} \log f_{z_j}(x_i; \theta_j) \qquad j = 1, 2, \ldots, s. \tag{17}$$

Wu (1983) established the theory of convergence of the EM algorithm under the assumption that the function Q computed in the E–step is continuous and differentiable at any iteration in all its arguments. Theorem 4.1 in Redner and Walker (1984) offers a summary of the results in Wu (1983). Because φ is continuous with respect to θ_j the M-step is well defined for all $j = q+1, \ldots, s$. However the discontinuities introduced by the uniform components create some inconvenience. Given the sample $\underline{\mathrm{x}}^n$ we define two functions: for a constant $h \in \mathbb{R}$, $m_n(h) = \min\{x_i \in \underline{\mathrm{x}}^n : x_i \geq h\}$ and $M_n(h) = \max\{x_i \in \underline{\mathrm{x}}^n : x_i \leq h\}$, and we show that for $j \in \{1, \ldots, q\}$, the EM algorithm always makes a_j and b_j coincide with two data points (or even a single one) in the first iteration and then does not change them anymore. (Note that we here allow $a_j^{(t)} = b_j^{(t)}$ with degenerating likelihood; if this happens in practice, the EMC-algorithm enforces the scale constraints.)

Theorem 4. *For $j = 1, 2, \ldots, q$ let $\theta_j^{(0)}$ with $-\infty < a_j^{(0)} < b_j^{(0)} < +\infty$ be the initial values for the uniform parameters. Suppose that the interval $[a_j^{(0)}, b_j^{(0)}]$ contains at least one data point. Let n be fixed and finite. Then at any iteration $t = 1, 2, \ldots$ an EM solution is such that $a_j^{(t)} = m_n(a_j^{(0)}) \leq b_j^{(t)} = M_n(b_j^{(0)})$ for all $j = 1, 2, \ldots, q$.*

Proof. In iteration $t+1$ the computation of the uniform parameters is done by solving the M-step for the uniform component, which is

$$(a_j^{(t+1)}, b_j^{(t+1)}) = \arg\max_{(a,b)\in\Theta_1} \sum_{i=1}^{n} w_{i,1}^{(t)} q_i(a_j, b_j),$$

$$w_{i,j}^{(t)} = \pi_j^{(t)} \frac{\mathbf{1}_{[a_j^{(t)}, b_j^{(t)}]}(x_i)}{(b_j^{(t)} - a_j^{(t)})} \frac{1}{g(x_i; \eta^{(t)})};$$

$$q_i(a_j, b_j) = \log \frac{\mathbf{1}_{[a_j,b_j]}(x_i)}{b_j - a_j}.$$

Consider any $j \in \{1, 2, \ldots, q\}$. Consider the first iteration, i.e. $t = 1$. Any $a_j^{(t)} < a_j^{(0)}$ and $b_j^{(t)} > b_j^{(0)}$ cannot result from the M-step above. In fact, for all i such that $x_i \notin \mathbb{R}\backslash[a_j^{(0)}, b_j^{(0)}]$ we have $w_{i,j}^{(0)} = 0$, while for all i such that $x_i \in [a_j^{(0)}, b_j^{(0)}]$ it results that $(b_j^{(1)} - a_j^{(1)})^{-1} < (b_j^{(0)} - a_j^{(0)})^{-1}$. The latter implies that for every $i = 1, 2, \ldots, n$ $w_{i,j}^{(0)} q_i(a_j^{(1)}, b_j^{(1)}) < w_{i,j}^{(0)} q_i(a_j^{(0)}, b_j^{(0)})$. Therefore the solution for the M-step has to be searched in $[a_j^{(0)}, b_j^{(0)}]$. For all i such that $x_i \in [a_j^{(0)}, b_j^{(0)}]$, $w_{i,j}^{(0)} > 0$. If $x_i \notin [a_j^{(1)}, b_j^{(1)}]$ it follows that $q_i(a_j^{(1)}, b_j^{(1)}) = -\infty$. Hence, the optimal solution is thus to take the smallest interval containing all $x_i \in [a_j^{(0)}, b_j^{(0)}]$, therefore $a_j^{(1)} = m_n(a_j^{(0)})$ and $b_j^{(1)} = M_n(b_j^{(0)})$. If we assume that $a_j^{(0)}$ and $b_j^{(0)}$ are two data points, then it is easy to see that $a_j^{(1)} = a_j^{(0)}$ and $b_j^{(1)} = b_j^{(0)}$. Now since $m_n(a_j^{(0)})$ and $M_n(b_j^{(0)})$ are two data points, taking $t = 2$ and applying the same argument would lead us to conclude that $a_j^{(t)} = m_n(a_j^{(0)})$ and $b_j^{(t)} = M_n(b_j^{(0)})$ at any iteration $t = 1, 2, \ldots$. Note that because intervals are only made smaller, (3) is fulfilled finally if it is fulfilled initially. ■

Following Theorem 4.1 in Redner and Walker (1984), the EM-algorithm increases the likelihood in every single step. This holds for the EMC-algorithm as well, by definition, and by Theorem 2 the maximum of the log-likelihood function exists on Γ_c. Therefore the EMC-algorithm converges (though in the unlikely case that it is stopped prematurely because the likelihood is decreased by enforcing the scale constraints, it may not converge to a local optimum).

Assuming $q = 1$ from now on, from Theorem 4 it follows that if $[a_1^{(0)}, b_1^{(0)}]$ is chosen to be the range of the dataset as in the R-method, they are not changed throughout the EM-algorithm. This means that the R-method is a proper local likelihood maximum yielded by the EM-algorithm. However, it is often not the global maximum and not consistent, as shown in Remark 3.

Theorem 4 suggests that, while the EM-algorithm generally only produces a local maximum of the likelihood, for a heterogeneous mixture with uniform component this is a particularly severe problem, because every pair of data points (and even every single data point as long as the scale constraints are ignored) corresponds to a local maximum of the likelihood. In this sense, the EM-algorithm as well as the EMC-algorithm are not informative about the uniform component. In order to get information about it, we have to compare solutions from several runs of the EMC-algorithm started with various initializations of the uniform component.

4.2 Initialization

A way to implement the EM-algorithm to obtain a more satisfactory local maximum (or even, with luck, the global optimum) is to initialize the algorithm with the uniform component starting at every pair of data points and to eventually choose the overall likelihood maximum among the solutions. This means that the M-step only updates the proportions and Gaussian parameters, but not the ranges of the uniform.

There is also a possibility that in order to maximize the (scale constrained) likelihood, the uniform component eventually only fits a single outlier, and therefore it may be reasonable to try out further initializations with uniform supports $[x_i \pm \frac{r_{max}}{2}]$.

This, however, can be computationally infeasible. For example, with $n = 100$ we would run the EM for a Gaussian mixture 5050 times (4950 pairs of data points and 100 points). In practice we need a selection rule to reduce the number of necessary algorithm runs. Here is our proposal how to do this:

1. Define a grid of q equi-spaced points on the range of the data. For each point in the defined grid choose the nearest data point. This yields data points $x_{((1))} \leq \ldots \leq x_{((q))}$. As initializations for the uniforms take the $\frac{q(q-1)}{2}$ pairs of points from the grid and additionally $[x_{((i))} \pm \frac{r_{max}}{2}]$, $i = 1, \ldots q$. The points are not necessarily pairwise distinct, and repeated initializations may be skipped. Instead of a grid of equidistant points, it would be possible as well to use order statistics of equidistant orders, but if the uniform component is interpreted as "catching outliers and points not belonging to any Gaussian cluster", it makes sense to represent scarce regions of the dataset properly in the set of selected points.

   Note that in order to know r_{max}, the initialization of the Gaussian components has to be known.

2. The initial value of the proportion of the uniform component is fixed at 0.05. This is because we are interested in situations where the uniform distribution is interpreted as "outliers or noise" is a situation where we want to assign most of the observations to Gaussian "clusters". In some applications it may make sense to change this but of course this proportion can be increased during the algorithm anyway.

   The proportions of the other components are initialized at equal value $0.95/(s-1)$. The means and variances of the Gaussian components are initialized by trimming the 10% of observations in both the tails of the data and then applying the k-means algorithm with $s-1$ components with randomly chosen initial values. This is related to the trimmed-k-means method (Cuesta-Albertos et al. (1997)). In general, good initialization of the EM-algorithm for a Gaussian mixture alone is a complicated

issue, and alternatives to our approach exist, see for example Karlis and Xekalaki (2003).

3. Run the EMC-algorithm for every initialization of the uniform component (if time allows, it is possible to do this more than once, trying out different random initializations of the Gaussians) until it stops or until the likelihood is improved in an iteration by less than 10^{-6}, say. Eventually report the solution with the largest likelihood.

A possible choice of q is $q = 20$. The choice of c in (5) will depend on the application, but 0.1 or 0.01 could make sense to give the occurring scales some flexibility but avoid spurious solutions. Recently, Yao (2010) discussed an automatic choice of c for Gaussian mixtures.

5 Empirical experience

5.1 Simulations

Coretto (2008) carried out an extensive Monte Carlo simulation study, comparing the ML-estimator proposed here with $q = 1$ and the uniform component initialized from a grid of starting points as explained above ("G-method") with some other estimators on several different mixture models.

We summarize here the findings for mixture models with Gaussian components and a single uniform one or additional outliers concerning the comparison of the G- and the R-method, measured by misclassification rates on sample sizes $n = 50, 200, 500$. A publication of the results in more detail is in press (Coretto and Hennig (2010)).

Generally, the G-method did not do well for $n = 50$, but for larger n it was either about as good as the R-method or, with a more concentrated uniform component (either on one side of the Gaussian components or between them; such situations occur in practice, see Section 5.2)), better. In a setup with the uniform component spread further than the range that would be expected from the Gaussian distributions alone, which is the ideal situation for the R-method, corresponding to its implicit model assumption, the G-method did about equally well for $n = 500$. On the other hand, the R-method did a bit better when applied to a pure Gaussian mixture, i.e., it was attempted to fit a uniform component even though none existed in the true model.

5.2 Real dataset

As a small real data example, in Figure 1 the percentages of the Republican candidate in the 50 states of the United States in the 1968 elections are shown. A feature of this data set is that one can see points that apparently do not belong to any Gaussian shaped cluster on the left side of two such clusters. This is more convincingly fitted by a uniform component that does not span the whole range of the dataset, namely the outcome of the G-method, and the Gaussian components can be expected to be fitted more accurately without the implicit assumption of the R-method that there is some uniform "noise" in the regions of the clusters as well. Of course, the discontinuity between Gaussian clusters and uniform component in the fitted density may be seen as a disadvantage, but in terms of interpretation it is useful to have an informative interval to indicate where the observations not belonging to Gaussian clusters are.

6 Conclusion

We investigated ML-estimation in a mixture model with q uniform components and $s-q$ Gaussian components. Identifiability, existence and consistency were shown and the EM-algorithm was theoretically discussed for general q. We suggested a practical implementation for $q = 1$ and compared it to the noise component approach (R-method) by Banfield and Raftery (1993) theoretically and by a simulation study with the main focus of using the uniform component to model "points that do not belong to any cluster" in the presence of clear Gaussian clusters.

From the results, we cannot claim that the R-method should be replaced in all applications by the G-method, though the latter is at least theoretically more appealing and has benefits in some situations. On the other hand, the G-method apparently (from the simulations) overestimates the impact of the uniform component for small n and is computationally less simple (though this may be acceptable in many applications). The main disadvantage compared to the R-method is that a possible generalisation to more than one-dimensional data will be at least computationally cumbersome, because there is no easy statement anymore that makes sure that we only need pairs of points (in most cases) to find the ML-solution, while the R-package mclust (Fraley and Raftery (2006)) includes a useful implementation of the R-method for higher dimensions. Another issue not treated in the present paper is the estimation of the number of mixture components, but criteria such as the AIC and BIC are theoretically at least as appealing for the G-method as for the R-method, for which the BIC is recommended (Fraley and Raftery (2002)).

There are alternatives for robust clustering with Gaussian cluster shapes, see for ex-

ample García-Escudero et al. (2008), Coretto and Hennig (2010). Actually, as Hennig (2004) showed, the R-method is theoretically not breakdown robust, though in practice very extreme outliers are needed to spoil it. The G-method with proper EMC-algorithm can be expected to fit the uniform component around a single extreme outlier, rescuing the Gaussian clusters, but for two outliers with the distance between them converging to infinity, the same arguments as given in Hennig (2004) will again lead to breakdown. However, the G-method can be expected at its best if there are some points not belonging to any Gaussian cluster on only one side of some Gaussian clusters.

In a multivariate setup, fitting mixtures of Gaussian and uniform distributions becomes much more complicated. There is more than one method to generalize uniform distributions on intervals. For example, hyperrectangles or ellipsoids could be chosen. Hyperrectangles are not rotation invariant, but probably computationally easier, and the considerations of Section 4 could apply in some generalized form. However, the computational burden of finding good approximations to the maximum likelihood estimators for the parameters of the uniform distributions will be much worse in any case (for hyperrectangles, a multivariate grid has to be chosen). Constraints could be enforced along the lines of Ingrassia (2004), but some more decisions have to be made regarding models for the Gaussian covariance matrices (Fraley and Raftery (1998)). In terms of the theory, the more complex parameter space means that some more subtleties have to be negotiated, but we expect that this is possible in principle.

References

Banfield, J. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics 49*, 803–821.

Coretto, P. (2008). *The Noise component in Model-Based Clustering*. Ph. D. thesis, Department of Statistical Science, University College London.

Coretto, P. and C. Hennig (2010). A simulation study to compare robust clustering methods based on mixtures. *Advances in Data Analysis and Classification in press*.

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

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

Dennis, J. E. J. (Ed.) (1981). *Algorithms for nonlinear fitting*, Cambridge, England. NATO advanced Research Symposium: Cambridge University Press.

DeSarbo, W. S. and W. L. Cron (1988). A maximum likelihood methodology for cluster-wise linear regression. *J. Classification 5*, 249–282.

Fraley, C. and A. E. Raftery (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal 41*, 578–588.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*, 611–631.

Fraley, C. and A. E. Raftery (2006, September). Mclust version 3 for r: Normal mixture modeling and model-based clustering. Technical Report 504, University of Washington, Department of Statistics.

García-Escudero, L. A., A. Gordaliza, C. Matrán, and A. Mayo-Iscar (2008). A general trimming approach to robust cluster analysis. *Annals of Statistics 38*(3), 1324–1345.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics 13*, 795–800.

Hathaway, R. J. (1986). A constrained EM algorithm for univariate normal mixtures. *J. Comp. Stat. Simul. 23*, 211–230.

Hennig, C. (2004). Breakdown points for maximum likelihood estimators of locationscale mixtures. *The Annals of Statistics 32*(4), 1313–1340.

Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications 13*(2), 151–166.

Karlis, D. and E. Xekalaki (2003). Choosing initial values for the EM algorithm for finite mixtures. *Comput. Statist. Data Anal. 41*(3-4), 577–590.

Kiefer, N. M. and J. Wolfowitz (1956). Consistency of the maximum likelihood estimation in the presence of infinitely many incidental parameter. *Ann. Math. Statist. 27*(364), 887–906.

Perlman, M. D. (1972). On the strong consistency of approximate maximum likelihood estimator. In *Sixth Berkeley Symp. Math. Statist. Probab.*, Volume 1, Usa, pp. 263–282. Univ. of California Press.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics 9*, 225–228.

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

Tanaka, K. and A. Takemura (2006). Strong consistency of the mle for finite location-scale mixtures when the scale parameters are exponentially small. *Bernoulli 12*, 1003–1017.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics 32*, 244–248.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics 20*, 595–601.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics 11*(1), 95–103.

Yakowitz, S. J. and J. Spragins (1968). On the identifibility of finite mixtures. *The Annals of Mathematical Statistics 39*, 209–214.

Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference 140*(7), 2089–2098.

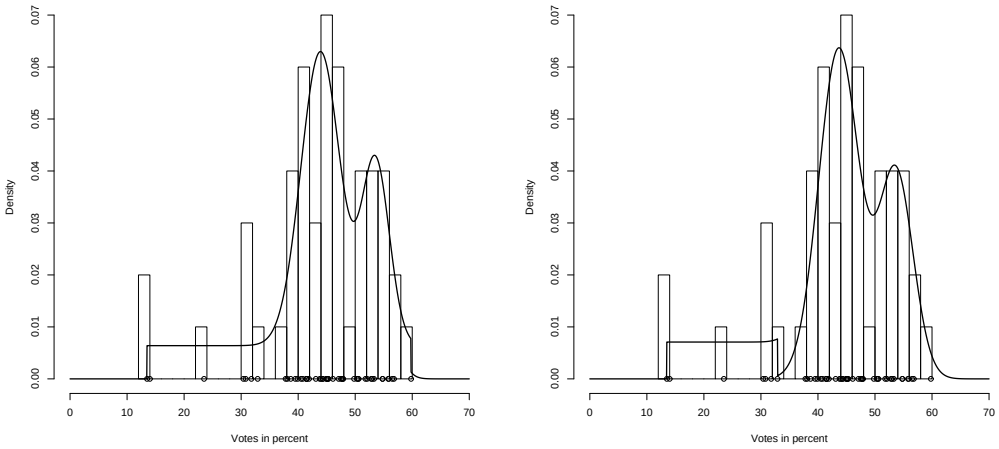

Figure 1: Votes (percentage) for the Republican candidate in the 50 states of the U.S., 1968, with estimated density with $q = 1$ uniform component and $s - q = 2$ Gaussian components, estimated by the R-method (left side) and the ML-estimator (right side).