Ann. Inst. Statist. Math.
Vol. 56, No. 4, 631–654 (2004)

TYPES OF LIKELIHOOD MAXIMA IN MIXTURE MODELS AND THEIR IMPLICATION ON THE PERFORMANCE OF TESTS*

WILFRIED SEIDEL[1] AND HANA ŠEVČÍKOVÁ[2]

[1]*Fachbereich Wirtschafts- und Organisationswissenschaften, Helmut-Schmidt-Universität Hamburg, D-22039 Hamburg, Germany,* e-mail: Wilfried.Seidel@unibw-hamburg.de
[2]*Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195, U.S.A.*

(Received May 23, 2003; revised November 25, 2003)

Abstract. In two-component mixtures of exponential distributions, different strategies for starting the likelihood maximization algorithm converge to different types of maxima. The power of an LR test of homogeneity against such a mixture strongly depends on the considered strategy, and global maximization need not result in the largest power. An explanation is given on basis of a systematic investigation of the likelihood function in a large number of simulations, using a variety of diagnostic tools. Thereby, we also gain a deeper insight into the properties of the samples that generate particular types of solutions of the likelihood equation. In particular, "spurious solutions" often occur; these are mainly responsible for the fact that global maximization may not result in a statistically meaningful estimator. Removing the smallest elements of a sample may drastically increase the power of previously inferior strategies.

Key words and phrases: Mixture models, likelihood function, likelihood ratio tests, multiple maxima, likelihood equation, spurious solutions, EM algorithm, starting values.

1. Introduction

In mixture models with restricted number of components, the likelihood function may have multiple local maxima. Examples are given in McLachlan and Peel (2000). For calculating local maxima, the EM algorithm is often applied. This is an iterative algorithm that needs a starting point. The possible existence of multiple maxima implies that the choice of its initial value(s) is an important issue. This problem has been treated by several authors; for recent discussions, see for example Biernacki *et al.* (2003) or Karlis and Xekalaki (2003).

On the other hand, one might be led to the wrong conclusion that there are several local maxima if the maximization method does not work well enough. Karlis (2001) gives an example of a sample from an exponential mixture, where at first glance the likelihood function seems to have several very different local maxima (reached from different starting values). However, if the EM is run for a large number of additional iterations, all starting values finally converge to the same estimator. This observation indicates that the flatness of the likelihood function in certain areas is an additional source of the unpleasant behaviour of likelihood maximization in mixture models.

*This research has been supported by a grant from the Deutsche Forschungsgemeinschaft.

These shortcomings have implications on the statistical performance of likelihood methods. For example, in Seidel *et al.* (2000*a*), the likelihood ratio test for homogeneity against two-component mixtures of exponential distributions is investigated by simulation. Under the alternative hypothesis, the likelihood is maximized by the EM. It is demonstrated that its starting value and also the type of stopping rule applied strongly influence the empirical properties of the test.

One might argue that these observations are not surprising, as a statistical procedure that is not based on the true likelihood maximum clearly has inferior properties. However, it is shown in Seidel *et al.* (2000*b*) and in more detail in Seidel *et al.* (2000*c*) that global maximization does not result in the best test.

As subglobal likelihood maximization may give better results, it seems to be advantageous to define a likelihood ratio test in terms of the maximization algorithm. If the latter is properly specified, the test statistic has a well defined theoretical distribution under each parameter value, therefore critical values and the power of such a test are well defined theoretical concepts. Although only the empirical behaviour of the tests has been studied in our simulation experiments, the differences between the considered variants are so evident for the observed phenomena that our conclusions seem to hold also for the theoretical behaviour. However, a satisfactory explanation was still missing. Such an explanation is given here.

It has already been observed in Seidel *et al.* (2000*b*) that the inferiority of certain tests might be caused among others by spurious solutions of the likelihood equation, resulting in artifactual components. This conjecture is strongly supported by the observations reported here. There is a typical pattern in samples from mixtures of exponentials that causes such spurious components, and removing the smallest elements of a sample may drastically increase the power of previously inferior strategies.

These and other features of likelihood functions are investigated here in a systematic way in a large number of simulations, using a variety of diagnostic tools. Thereby, we gain a deeper insight into the different kinds of solutions of the likelihood equation in two-component mixture models, into the properties of the samples that create these solutions, and into the implications on the behaviour of likelihood ratio tests. Similar observations were made in other mixture models, too. In particular, it follows from our studies that the global maximum of the likelihood function is not always a statistically meaningful solution of the likelihood equation.

2. Likelihood ratio tests for mixtures of exponentials

2.1 *The mixture model*

For $x > 0$, let

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta} \tag{2.1}$$

denote the density of the **exponential distribution** with expectation $\theta > 0$. The density of the **mixture** of m exponential distributions $f(x, \theta_1), \ldots, f(x, \theta_m)$ with mixing weights $p_1, \ldots, p_m$ $(0 \leq p_j \leq 1)$ is given by

$$f(x, P) = \sum_{j=1}^{m} p_j f(x, \theta_j),$$

the matrix

$$P = \begin{pmatrix} \theta_1 & \cdots & \theta_m \\ p_1 & \cdots & p_m \end{pmatrix}$$

denotes the parameter of the mixture. Suppose that $x_1, \ldots, x_n$ is the outcome of a random sample of size n with respect to some observation X. The log likelihood of P is

$$l(P) = \sum_{i=1}^{n} \log f(x_i, P). \tag{2.2}$$

For $m = 1$, we obtain a homogeneous population with parameter $\theta = \theta_1$; here, the log likelihood is written as $l(\theta)$.

We are especially interested in the likelihood ratio test of the null hypothesis of **homogeneity**, namely that $X \sim f(x, \theta)$ for some θ, against the alternative hypothesis that the distribution of X is a **two-component mixture** of exponential distributions, $X \sim f(x, P)$ for some P with $m = 2$. For $m = 2$, we usually set $p = p_1$. Then $p_2 = 1 - p$ and P can be written as

$$P = (\theta_1, \theta_2, p).$$

Note that the parameter P of a true two-component mixture (i.e., $\theta_1 \neq \theta_2$ and $0 < p < 1$) is identifiable in the sense that the only parameter $\tilde{P}$ that describes the same mixture is obtained from $P = (\theta_1, \theta_2, p)$ by "label switching", namely $\tilde{P} = (\theta_2, \theta_1, 1 - p)$. The situation is different for a homogeneous population with parameter θ, say: it is described by all mixtures with parameter $P = (\theta_1, \theta_2, p)$ and

$$(\theta_1 = \theta \text{ and } p = 1) \text{ or } (\theta_2 = \theta \text{ and } p = 0) \text{ or } (\theta_1 = \theta_2 = \theta). \tag{2.3}$$

Under homogeneity, the log likelihood $l(\theta)$ is maximized by $\hat{\theta} = \overline{x}$, the sample mean. Under the alternative hypothesis, a **likelihood estimator** $\hat{P}$ of P is a parameter value that is defined in a particular way in terms of $l(P)$, a special kind of a local maximizer, say. Any local maximizer $\hat{P}$ in the interior of the parameter space satisfies the **likelihood equation** $(\partial l/\partial\theta_1, \partial l/\partial\theta_2, \partial l/\partial p)(\hat{P}) = \mathbf{0}$. Let

$$\tau_i(\theta, P) = \frac{f(x_{(i)}, \theta)}{f(x_{(i)}, P)} \tag{2.4}$$

and

$$S(\theta, P) = \sum_{i=1}^{n} \tau_i(\theta, P), \tag{2.5}$$

where $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ denotes the ordered sample. Using these expressions, the likelihood equation can be rewritten as

$$\hat{\theta}_j = \frac{\sum_{i=1}^{n} x_{(i)} \tau_i(\hat{\theta}_j, \hat{P})}{S(\hat{\theta}_j, \hat{P})}, \qquad j = 1, 2 \tag{2.6}$$

and

$$\hat{p} = (\hat{p}/n)S(\hat{\theta}_1, \hat{P}). \tag{2.7}$$

This form of the likelihood equation is often used in mixture modelling as a starting point for deriving the EM algorithm, see for example Lindsay (1995), Subsection 3.2 or McLachlan and Peel (2000), Subsection 2.8.1. We additionally arrange the terms in the sums in the order of the elements $x_{(i)}$, this will be useful in the discussion of spurious solutions.

The likelihood ratio test is based on the difference d_n of the log likelihoods of $\hat{P}$ and $\hat{\theta}$. For better comparison to critical values reported in the literature, we consider as test statistic

$$2d_n = 2[l(\hat{P}) - l(\hat{\theta})]. \tag{2.8}$$

In standard statistical models, this test statistic has a χ^2-distribution. However, in mixture models this is no longer true. The nonidentifiability of the model under the null hypothesis implies a breakdown of the regularity conditions for the classical asymptotic theory, this is discussed among others by Ghosh and Sen (1985) and by Lindsay (1995). Here also exact theoretical results are developed for certain special cases. For example, in the article by Ghosh and Sen (1985), the limiting distribution for $2d_n$ in a model of normal mixtures is derived under a strong separation condition and a compactness assumption for the parameters. On the other hand, Hartigan (1985) shows that without the compactness assumption $2d_n$ is asymptotically infinite. A variety of asymptotic results has been derived under special conditions or in restricted models, see for example McLachlan and Peel (2000). It has been claimed that the separation condition of Ghosh and Sen (1985) can be removed. In fact, without assuming such a condition, Dacunha-Castelle and Gassiat (1999) show that $2d_n$ converges for a rather general class of models to a Gaussian process indexed by the closure of the convex cone of directional score functions. This generalizes also the results of Lindsay (1995). Their approach, however, does not lead to optimal assumptions. Essentially the same limiting distribution is obtained in a more general setting in Liu and Shao (2003).

2.2 *The EM algorithm*

A standard algorithm for maximizing the likelihood function in mixture models with finite number of components is the **EM algorithm**, see Böhning (2000) or McLachlan and Peel (2000). It is an iterative algorithm that starts from some externally chosen initial value $P^0 = (\theta_1^0, \theta_2^0, p^0)$ and generates a sequence $P^k = (\theta_1^k, \theta_2^k, p^k)$, $k \in \mathbb{N}$, of improved estimates. It can be derived from the likelihood equation; in our situation of a two-component mixture of exponentials, P^{k+1} can be computed from P^k by substituting P^k into the right-hand side of equations (2.6), (2.7). This results in the iteration

$$\theta_j^{k+1} = \frac{\sum_{i=1}^n x_{(i)} \tau_i(\theta_j^k, P^k)}{S(\theta_j^k, P^k)}, \qquad j = 1, 2, \tag{2.9}$$

$$p^{k+1} = (p^k/n)S(\theta_1^k, P^k). \tag{2.10}$$

The sequence $(l(P^k))_{k\in\mathbb{N}}$ is nondecreasing. For stopping the iteration, we apply a criterion based on **directional derivatives**, see Böhning *et al.* (1994) and Lindsay (1995):

$S(\theta, P) - n$ has the properties of a directional derivative of $l(P)$ in the direction of θ. Let $acc > 0$ be a given level of accuracy. In the same way as in the algorithm described in Böhning and Schlattmann (1992), we stop the iteration at $\hat{P} = P^k$ if

$$\max\{S(\theta_1^k, P^k) - n, S(\theta_2^k, P^k) - n\} < n \cdot acc \quad \text{and} \quad k \geq 3. \tag{2.11}$$

For simulating the quantiles and power of likelihood ratio tests, we chose in previous studies $acc = 10^{-5}$ and set the maximum number of EM iterations to 5000. However, for the more detailed investigation here, we choose $acc = 10^{-11}$ and set the maximum number of EM iterations to 100 000. The only exception is the *multistart* strategy (see below), which would lead to an extremely slow algorithm. A detailed discussion of the choice of accuracy can be found in Seidel and Ševčíková (2002*b*).

2.2.1 *Starting values*

We consider the following starting strategies:

- *minmax*: $P^0 = P^0_{\text{minmax}} = (x_{(1)}, x_{(n)}, 0.5)$, resulting in an estimator $\hat{P}_{\text{minmax}}$.
- *mean*: $P^0 = P^0_{\text{mean}} = (0.5\overline{x}, 1.5\overline{x}, 0.5)$, resulting in an estimator $\hat{P}_{\text{mean}}$.
- *multistart*: This strategy is based on a version proposed by Böhning (2000) (p. 69, "Gold Standard") with a modification described in Seidel and Ševčíková (2002*b*), it uses 64 starting values.
- *maximum*: the EM is started from both P^0_{minmax} and P^0_{mean}. Then $\hat{P}_{\text{maximum}}$ is the element of $\{\hat{P}_{\text{minmax}}, \hat{P}_{\text{mean}}\}$ with the larger likelihood.

Böhning and Schlattmann (1992) claim that initial values with $p = 0.5$ and well separated parameters θ_1 and θ_2 often yield the global maximum, therefore they propose to use *minmax* as a possible starting value. The *multistart* strategy is an approximation to global maximization. We found that in exponential mixture models, *maximum* results in tests with nearly the same quantiles and power as *multistart*. Therefore in this paper, we address *minmax* and *mean* as representatives of a large variety of possible starting values.

2.3 *Performance of the tests*

In Seidel *et al.* (2000*a*, 2000*b*) and, in more detail, in Seidel *et al.* (2000*c*), the performance of different versions of the likelihood ratio test for homogeneity is analyzed. For the starting strategies of Subsection 2.2.1, the theoretical distribution of the test statistic $2d_n$ (eq. (2.8)) under the null hypothesis does not depend on the parameter θ. In the papers mentioned above, simulated critical values of a level α test, i.e., the $1 - \alpha$ quantiles of $2d_n$, are shown for different sample sizes and several levels α. The quantiles based on *minmax* are always (considerably) larger than the quantiles based on *mean*, consequently *minmax* has better optimization properties under the null hypothesis. Of course, *multistart* results in the largest quantiles.

For mixing proportions $p \in \{0.1, \ldots, 0.9\}$, the power of the tests based on different starting strategies has been simulated for $P = (\theta_1, \theta_2, p)$ with $\theta_1 = 1$ as a function of θ_2, $\theta_2 > 1$. Some typical power curves are shown in Figs. 1 and 2 ($e = 0$) in Subsection 3.2. Usually, the following ranking of the strategies is observed (Fig. 1): *mean* has the largest power, *minmax* has the smallest power and is often strongly inferior. This seems to be somewhat surprising, as under the null hypothesis, *minmax* has much better optimization properties than *mean*. Perhaps the most striking result is that global maximization, represented by *multistart* or *maximum*, has always smaller power than *mean*, often it is considerably worse. We shall refer to this ranking as the "typical behaviour".

Lower contamination models, here represented by $p = 0.1$ and $\theta_2 > 1$, behave differently for small sample sizes, as Fig. 2 ($e = 0$) in Subsection 3.2 shows. For $\theta_2 \leq 10$, the typical behavior is observed. However, for $\theta_2 \geq 20$, *mean* has a smaller power than *minmax*; here, *maximum* is the best strategy.

3. Spurious components

In former studies we observed that the striking behaviour of likelihood ratio tests may be caused by the occurence of "spurious local maximizers" of the log likelihood, where an "artifactual" mixture component is fitted to a small group of sample points.

3.1 *Properties in exponential mixtures*

In samples from exponential distributions, often extremely small data points occur. In such cases it is quite common that a sharp local maximum exists at a parameter $\hat{P}$ with a very small first component $\hat{\theta}_1$ having a small mixing weight $\hat{p}$ and a second component $\hat{\theta}_2$ near $\overline{x}$. However, this is not a parameter that describes a homogeneous population in the sense of equation (2.3); typically, $l(\hat{P}) > l(\overline{x})$ holds. In samples from homogeneous populations, there is often a global maximum in such $\hat{P}$.

In the described situation, the first component of $\hat{P}$ is "spurious" in the sense that it does not correspond to a genuine group in the population from which the sample is drawn. Here, such $\hat{P}$ is called a "spurious local maximizer" or a "spurious solution of the likelihood equation". Of course, each component of an estimator $\tilde{P}$ that does not correspond to a component in the population is in some sense "spurious". However, we will reserve this term only for components which are fitted to a small number of "extreme" data points. Such phenomena have been observed in mixtures of normal distributions, see for example McLachlan and Peel (2000): a component with a very small (generalized) variance is fitted to a small group of data points located close together. This is similar to the situation considered here, as an exponential distribution with a small parameter $\hat{\theta}_1$ has small variance, too. However, we observed spurious components also in mixtures of location families of normal distributions with fixed variance, in which a separate component was fitted to one or two small or large outliers.

The character of a spurious estimator $\hat{P}$ in a mixture of exponentials can be described in more detail. Here, we give only a crude heuristic characterization, which nevertheless contains essential features of a possibly finer analysis. In samples with small $x_{(1)}$, there is often a (small) index ρ such that the first ρ order statistics are much smaller than the following elements ($x_{(\rho+1)} \approx 10x_{(\rho)}$, say). We shall refer to these elements as "small outliers". Then $\hat{\theta}_1$ is usually approximately the mean of the group of small outliers, $\hat{p}$ is proportional to the size of this group and $\hat{\theta}_2$ is approximately equal to the mean of the remainder of the sample. In this sense, $\hat{P}$ fits a spurious mixture component to the group represented by $\hat{\theta}_1$.

To give a heuristic motivation of the existence of a solution of the likelihood equation of this type, we define $w_i(\theta_1, \theta_2) = (\theta_1/\theta_2)\exp((x_{(i)}/\theta_1) - (x_{(i)}/\theta_2))$. Then $\tau_i(\theta_1, P) = 1/[p + (1-p)w_i(\theta_1, \theta_2)]$ and $\tau_i(\theta_2, P) = 1/[1 - p + p/w_i(\theta_1, \theta_2)]$. Suppose that a sample with ρ small outliers is given and consider a parameter P such that $x_{(1)} \leq \theta_1 \leq x_{(\rho)}$ and θ_2 is large compared to $x_{(\rho)}$. Then θ_1/θ_2 is very small. If the first ρ elements are located close together and if $x_{(\rho+1)}$ is much larger than $x_{(\rho)}$, we obtain in the limit (see Example 3.1 and, even more pronounced Example A.1 in Appendix A):

• For $i \leq \rho$, $\exp((x_{(i)}/\theta_1) - (x_{(i)}/\theta_2)) \approx \exp(1)$, therefore $w_i(\theta_1, \theta_2) \approx 0$, $\tau_i(\theta_1, P) \approx 1/p$ and $\tau_i(\theta_2, P) \approx 0$.

• For $i > \rho$, $x_{(i)}/\theta_1$ is large and $\exp((x_{(i)}/\theta_1) - (x_{(i)}/\theta_2))$ is large compared to θ_1/θ_2, therefore $w_i(\theta_1, \theta_2)$ is large. Consequently, $\tau_i(\theta_1, P) \approx 0$ and $\tau_i(\theta_2, P) \approx 1/(1-p)$.

With these expressions, we obtain approximate likelihood equations with solution $(\theta_1^*, \theta_2^*, p^*)$ as follows:

Equation (2.7), written as $n = S(\theta_1, P)$, has approximately the form $n = \rho/p$ with solution $p^* = \rho/n$. Equations (2.6) give approximately

$$\theta_1^* = \frac{\sum_{i=1}^{\rho} x_{(i)}/p}{\sum_{i=1}^{\rho} 1/p} = \frac{1}{\rho} \sum_{i=1}^{\rho} x_{(i)}$$

and

$$\theta_2^* = \frac{\sum_{i=\rho+1}^{n} x_{(i)}/(1-p)}{\sum_{i=\rho+1}^{n} 1/(1-p)} = \frac{1}{n-\rho} \sum_{i=\rho+1}^{n} x_{(i)}.$$

Let $\hat{P} = (\hat{\theta}_1, \hat{\theta}_2, \hat{p}) \approx (\theta_1^*, \theta_2^*, p^*)$. Then for small i,

$$\begin{aligned} f(x_{(i)}, \hat{P}) &= (p/\hat{\theta}_1) \exp(-x_{(i)}/\hat{\theta}_1) + ((1-p)/\hat{\theta}_2) \exp(-x_{(i)}/\hat{\theta}_2) \\ &\approx (p/\hat{\theta}_1) \exp(-1) + ((1-p)/\hat{\theta}_2) \exp(0) \end{aligned}$$

is of magnitude $p/\hat{\theta}_1$, and therefore the first terms in the sum defining $l(\hat{P})$ (eq. (2.2)) are large. The remaining terms are of the same magnitude as $\log f(x_{(i)}, \overline{x})$, which is the contribution of $x_{(i)}$ to the log likelihood of $\overline{x}$ under homogeneity.

Moreover, in a sample of the considered type, the EM with starting strategy *minmax* usually converges to the spurious maximizer $\hat{P}$. This will be discussed later, but it can also be made plausible by applying the above approximations to the EM iterations: For $P^0_{\text{minmax}} = (x_{(1)}, x_{(n)}, 0.5)$ we obtain after one step:

$$\theta_1^1 = \frac{\sum_{i=1}^{n} x_{(i)} \tau_i(\theta_1^0, P^0)}{\sum_{i=1}^{n} \tau_i(\theta_1^0, P^0)} \approx \frac{\sum_{i=1}^{\rho} x_{(i)}/0.5}{\sum_{i=1}^{\rho} 1/0.5} = \frac{1}{\rho} \sum_{i=1}^{\rho} x_{(i)} = \theta_1^*,$$

$$\theta_2^1 = \frac{\sum_{i=1}^{n} x_{(i)} \tau_i(\theta_2^0, P^0)}{\sum_{i=1}^{n} \tau_i(\theta_2^0, P^0)} \approx \frac{\sum_{i=\rho+1}^{n} x_{(i)}/0.5}{\sum_{i=\rho+1}^{n} 1/0.5} = \frac{1}{n-\rho} \sum_{i=\rho+1}^{n} x_{(i)} = \theta_2^*$$

and

$$p^1 = \frac{0.5}{n} \sum_{i=1}^{n} \tau_i(\theta_1^0, P^0) \approx \frac{0.5}{n} \sum_{i=1}^{\rho} \frac{1}{0.5} = \frac{\rho}{n} = p^*.$$

Example 3.1. (Three small outliers) We discuss the properties of a parameter estimate that fits a spurious component to the first three order statistics. An additional example for a highly pronounced spurious estimate is given in Appendix A. We consider a particular sample of size $n = 200$ from a homogeneous population ($\theta = 1$) with $x_{(1)} = 0.000155$ and $\overline{x} = 1.0308$. Observe that for such a sample, the expectation of $x_{(1)}$ is $1/n = 0.005$, therefore $x_{(1)}$ can in fact be considered as small.

Table 1. Contribution of sample points to likelihood equation at $\hat{P}_{\text{minmax}}$ and to the log likelihood.

		$\hat{P} = \hat{P}_{\text{minmax}}$				$\hat{P} = \hat{P}_{\text{mean}}$
i	$x_{(i)}$	$w_i(\hat{\theta}_1, \hat{\theta}_2)$	$\tau_i(\hat{\theta}_1, \hat{P})$	$\tau_i(\hat{\theta}_2, \hat{P})$	$l_i(\hat{P})$	$l_i(\hat{P})$
1	0.000155	0.000542	71.460227	0.038756	3.206438	−0.030485
2	0.000163	0.000554	71.401761	0.039554	3.186059	−0.030493
3	0.000913	0.004098	57.138012	0.234145	1.407062	−0.031220
$\Sigma_{1,2,3}$					7.799558	−0.092197
4	0.008864	6788651.8	0.00000015	1.013642	−0.065914	−0.038934
5	0.010777	1.1×10^9	0.0	1.013642	−0.067745	−0.040790
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	5.831890	∞	0.0	1.013642	−5.638952	−5.687974
Σ_{rem}					−208.621691	−205.974668
Σ_{tot}					−200.822133	−206.066866

The estimators are $\hat{P}_{\text{minmax}} = (0.000375, 1.044857, 0.013459)$ with likelihood difference $2d_n = 10.4895$ and $\hat{P}_{\text{mean}} = (1.0308, 1.0308, 0.45)$ with likelihood difference $2d_n = 0$. Table 1 shows the order statistics $x_{(i)}$ as well as the quantities $w_i(\hat{\theta}_1, \hat{\theta}_2)$, $\tau_i(\hat{\theta}_1, \hat{P})$ and $\tau_i(\hat{\theta}_2, \hat{P})$ for the first indices i with $\hat{P} = \hat{P}_{\text{minmax}}$. In addition, the contributions $l_i(\hat{P}) = \log f(x_{(i)}, \hat{P})$ of $x_{(i)}$ to the log likelihood $l(\hat{P})$ for $\hat{P} = \hat{P}_{\text{minmax}}$ and $\hat{P} = \hat{P}_{\text{mean}}$ are shown.

Clustering the sample points according to the Bayes criterion with respect to $\hat{P}_{\text{minmax}}$ (see Subsection 5.2.2) allocates the three smallest sample points to the first component and the remainder of the sample to the second one. Let us finally note that a clustering of estimators derived from all starting values of *multistart* (see Subsection 5.1.2) results in two clusters represented by $\hat{P}_{\text{minmax}}$ and $\hat{P}_{\text{mean}}$.

3.2 *Implications on the power of tests*

The likelihood function of a spurious maximizer $\hat{P}$ in a sample from a homogeneous population that follows closely the described pattern with $\rho = 2$ is displayed in Seidel *et al.* (2000*b*). There is a sharp local (even global) maximum in $\hat{P}$ with the shape of a thin needle, i.e., $l(P)$ is larger than $l(\overline{x})$ only in a very small neighbourhood of $\hat{P}$. In the usual plots of the likelihood surface (maximized over p) it is invisible due to a too coarse resolution.

To see if small outliers really affect the properties of tests, we performed the following experiments: in 10 000 samples of each of a variety of populations (a homogeneous and several mixtures), we removed the first e elements of the ordered sample and calculated the LR test statistic for *minmax* and *mean* from the remainder of the sample. Table 2 shows for $n = 200$ the simulated $1 - \alpha$ quantiles of the new test statistic under homogeneity for several values of e with $\alpha = 0.1$.

All quantiles are rapidly decreasing with increasing values of e. Moreover, the (relative) differences between *minmax* and *mean* vanish; for $n = 200$ and $e = 3$, both are almost equal. For larger values of e, the quantiles of *minmax* are even smaller than these of *mean*. For $n = 1000$, the quantiles of *minmax* are very sensitive to the choice of the accuracy level in the stopping rule of the EM (eq. (2.11)), see Seidel and Ševčíková (2003).

Table 2. $1 - \alpha$ quantiles ($\alpha = 0.1$) of $2d_n$ after removing the first e elements of the ordered sample under homogeneity.

n	e	*mean*	*minmax*
	0	2.627	3.589
	1	2.218	2.290
200	3	1.601	1.621
	5	1.172	1.166
	10	0.505	0.493

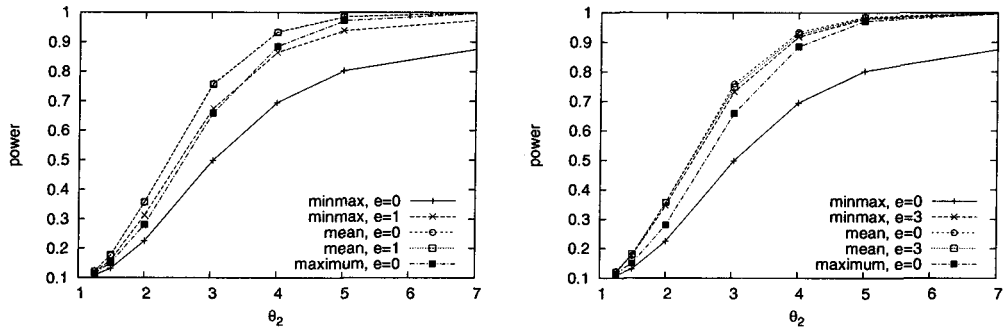


Fig. 1. Power ($\alpha = 0.1$) as function of θ_2 for $p = 0.7$ and $n = 200$ after removing the first e elements of the ordered sample. Left panel: $e = 0$ and 1; right panel: $e = 0$ and 3.

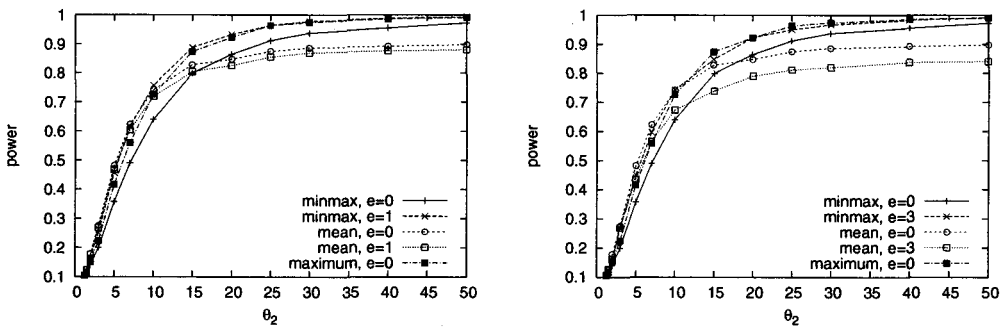


Fig. 2. Power ($\alpha = 0.1$) as function of θ_2 for $p = 0.1$ and $n = 200$ after removing the first e elements of the ordered sample. Left panel: $e = 0$ and 1; right panel: $e = 0$ and 3.

Figures 1–2 show selected simulated power curves for $n = 200$ and $e = 0$, 1 and 3. The "typical behaviour" is represented by $p = 0.7$ (Fig. 1). The power of *mean* is almost independent of e. In particular, for $e = 0$ and $e = 1$, the corresponding versions of *mean* coincide and dominate the other strategies (left panel). On the other hand, the power of *minmax* is much larger for $e = 1$ than for $e = 0$, and it is almost identical to the power of *mean* for $e = 3$.

The power for lower contamination models ($p = 0.1$) is shown in Fig. 2. For $e = 1$

and $\theta_2 \geq 10$, *minmax* has even a larger power than all other strategies for $e = 0$ and $e = 1$. For large θ_2 the power of *mean* decreases in e; for $e = 3$, it has poor power.

Similar phenomena can be observed also for $n = 1000$ (Seidel and Ševčíková (2003)). Here, it is in some cases necessary to remove a slightly larger number of elements from the sample to achieve a good power also for *minmax*. In all examples studied so far, the power of *minmax* can be drastically improved by removing low order statistics.

4. Log likelihood and EM iterations for large populations

In several simulation studies, we observed that for each particular mixture model, the likelihood functions of all samples from the same model seem to have a typical "common shape"; moreover, the sequences of EM iterations corresponding to a particular starting strategy seem to follow a typical pattern. To identify this pattern, in this section we replace the sample by a large population.

4.1 *Log likelihood*

Suppose that $x_1, \ldots, x_n$ is an i.i.d. sample from some distribution G and define

$$l_G(P) = \int_0^\infty \log f(x, P) dG(x).$$

Then (cf., eq. (2.2)) $(1/n)l(P) \to l_G(P)$ for $n \to \infty$. Consequently, the log likelihood function is for large n approximately proportional to $l_G(P)$, and therefore the function $P \mapsto l_G(P)$ represents the theoretical shape of the log likelihood function. We shall refer to it as the log likelihood of P given the population G.

To visualize the log likelihood of $P = (\theta_1, \theta_2, p)$, we maximize $l_G(P)$ with respect to the mixing weight p and display the resulting function $l_G(\theta_1, \theta_2)$. As the likelihood function is strictly concave in p, the maximization problem can be solved easily. For computational details as well as for the evaluation of the integral in the definition of $l_G(P)$, see Seidel and Ševčíková (2002*b*).

Homogeneous population. If G is an exponential distribution with parameter one, then $l_{\text{hom}}(\theta_1, \theta_2) = l_G(\theta_1, \theta_2)$ represents the "typical" log likelihood under a homogeneous population. It is shown in Fig. 3.

The function $l_{\text{hom}}(\theta_1, \theta_2)$ attains its maximum at the ridge $(]0, \infty[\times \{1\}) \cup (\{1\} \times]0, \infty[)$, which corresponds to $(]0, \infty[\times \{\overline{x}\}) \cup (\{\overline{x}\} \times]0, \infty[)$ in finite samples. There are no other local maxima.

Inhomogeneous population. Figure 4 shows $l_G(\theta_1, \theta_2)$, where G is a two-component mixture of exponential distributions with parameter $Q = (1.0, 0.33, 0.7)$.

Although this case looks similar to the previous one, this function has two local (which are also global) maxima at $(\theta_1, \theta_2) = (1.0, 0.33)$ and $(0.33, 1.0)$.

4.2 *Theoretical behaviour of the EM*

To evaluate the typical sequence of EM steps for different starting strategies, we shall replace in the spirit of Subsection 4.1 the sample by the whole population. Consider the EM iteration defined by eqs. (2.9), (2.10) and suppose again that $x_1, \ldots, x_n$ is an i.i.d. sample from some population G. Define

$$\tau(j, P^k \mid G) = \int_0^\infty \frac{f(x, \theta_j^k)}{f(x, P^k)} dG(x)$$

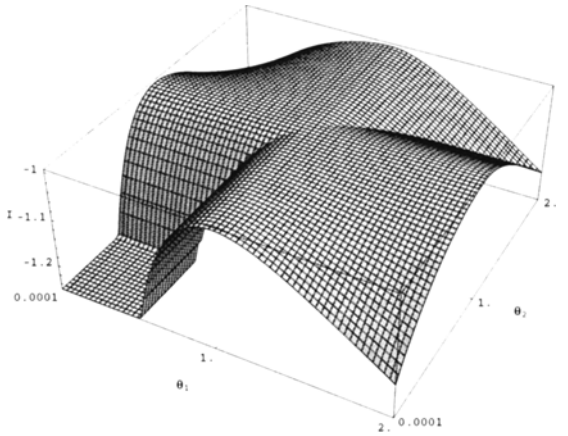


Fig. 3. A "typical" log likelihood under a homogeneous population.

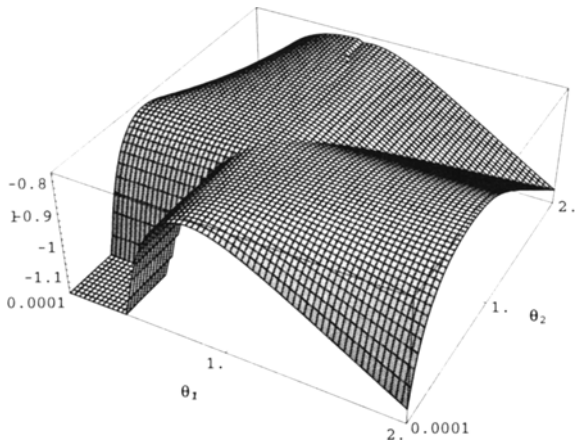


Fig. 4. Log likelihood of a two-component mixture with parameter (1.0, 0.33, 0.7).

and

$$m(j, P^k \mid G) = \int_0^\infty \frac{x f(x, \theta_j^k)}{f(x, P^k)} dG(x).$$

Then it follows similarly to Subsection 4.1 that the EM iterations can be written approximately for large n as $\theta_j^{k+1} = m(j, P^k \mid G)/\tau(j, P^k \mid G)$ and $p^{k+1} = p^k \tau(j, P^k \mid G)$. We refer to these iterations as the population version of the EM. For evaluation of the integrals, see again Seidel and Ševčíková (2002*b*).

4.2.1 *Examples*

If the EM is started from different initial values, it moves into different directions which are specific for the particular starting value. This behaviour can be studied very clearly by considering the population version of the EM.

In a particular example, we consider the strategies *minmax* and *mean* for a homogeneous population with parameter $\theta = 1$. Here, the expected value of $\overline{x}$ is $E(\overline{x}) = 1.0$,

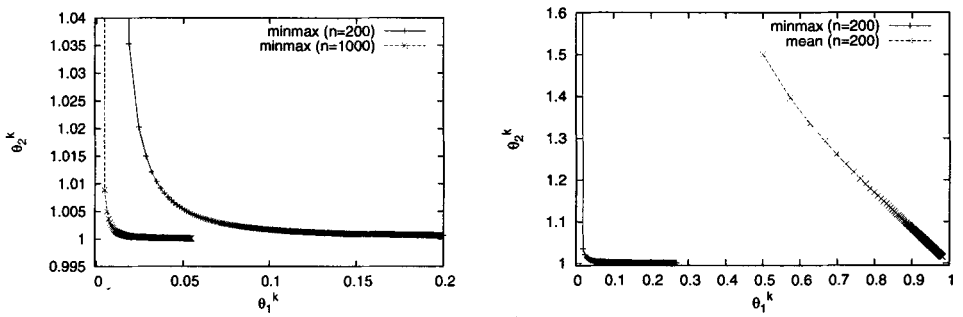


Fig. 5. Coordinates (θ_1^k, θ_2^k) of 1000 EM-iterations for different starting strategies.

so we represent the strategy *mean* by the starting value $P_{\text{mean}} = (0.5, 1.5, 0.5)$.

The expected values of $x_{(1)}$ and $x_{(n)}$ for the considered population depend on the sample size n, they are given by $E(x_{(1)}) = 1/n$ and $E(x_{(n)}) = 1+(1/2)+\cdots+(1/n)$. We consider two sample sizes: $n = 200$ and $n = 1000$. The strategy *minmax* is represented by $P_{\text{minmax}} = (E(x_{(1)}), E(x_{(n)}), 0.5)$.

Figure 5 shows for each starting strategy the coordinates (θ_1^k, θ_2^k) of the sequence P^k, $k = 1, \ldots, 1000$, of the first 1000 EM-iterations of the population version. The points of each sequence are connected by lines.

In the left panel, the curves representing the versions of the *minmax*-strategy for the two different sample sizes are shown; this is only a zoomed part of the whole plot. Obviously, *minmax* always starts in a point (θ_1^0, θ_2^0) with $\theta_1^0 \approx 0$ and θ_2^0 "large". Then θ_1^k increases very slowly, whereas θ_2^k rapidly decreases to 1 ($= \overline{x}$). Thus, (θ_1^k, θ_2^k) approaches $]0, \infty[\times\{1\}$ ($=]0, \infty[\times\{\overline{x}\}$) very rapidly, where the speed of convergence increases with increasing n.

In the right panel of Fig. 5, EM iterations for the strategy *mean* are presented in comparison to the strategy *minmax* for $n = 200$. Apparently, *mean* starts at $(0.5, 1.5)$ and approaches $(1, 1)$ ($= (\overline{x}, \overline{x})$ in finite samples) almost along the diagonal.

If there is a spurious maximum with coordinates near $(0, \overline{x})$, this is found by *minmax*, but usually not by *mean*.

We observed that in small samples, the EM typically behaves according to this scenario, especially in its first steps. However, there are cases, where the direction is suddenly changed. Then the convergence becomes very slow for a while, before the EM may move faster towards some maximum. For most of the stopping rules used in practice, in such cases the EM may terminate before a maximum is reached.

5. Diagnostic instruments

To investigate the properties of likelihood functions of two-component mixture models in a large number of replications, we use a number of automatic diagnostic instruments.

5.1 *Properties of the likelihood function*

5.1.1 *Criteria for local maxima*

To test if the log likelihood has a local maximum at some point $\hat{P}$, we use two types of criteria: an analytic criterion based on first and second order derivatives, and

a criterion based on function values only. Within certain tolerances, both are sufficient criteria for the existence of a strict local maximum near $\hat{P}$.

In particular, the analytic criterion gives us the information if $\hat{P}$ is a stationary point or not, and if there is a local maximum, minimum or a saddle point at $\hat{P}$ or if no decision about a local maximum at $\hat{P}$ can be made on basis of the second order criterion.

The other criterion evaluates $l(P)$ on a grid defined on a small cube around $\hat{P}$. If $l(\hat{P})$ is larger than $l(P)$ for all P on the surface of the cube, then according to the criterion there is a strict local maximum in the interior of the cube (not necessarily at $\hat{P}$). If there is a grid point P with $l(P) > l(\hat{P}) + 10^{-10}$, then according to the criterion, there is no local maximum at $\hat{P}$. Detailed information about the implementation of both criteria can be found in Seidel and Ševčíková (2002*b*).

As neither criteria is reliable in all situations, we state that there is a strict local maximum only if it is indicated by both criteria. Moreover, there are local maxima which cannot be strict ones. For example, if $\theta_1 = \theta_2 = \overline{x}$, then each value of p yields the same likelihood. If there is a local maximum in such a point, it cannot be strict.

5.1.2 *Clustering the results of multistart*

The *multistart* strategy defined in Subsection 2.2.1 uses 64 starting values which converge to 64 points $\hat{P}_i = (\hat{\theta}_1^i, \hat{\theta}_2^i, \hat{p}^i)$. Some of these may be approximately equal. These form a cluster, where it is taken into account, that for $\hat{\theta}_1 = \hat{\theta}_2$, the parameter $\hat{p}$ is irrelevant. Such clusters may be used as an indicator for possible local maxima of the likelihood function.

Note that $\overline{P} = (\overline{x}, \overline{x}, 0.5)$ is one of the starting points. As this is a fixed point of the EM, there is always a cluster that contains $\overline{P}$, even if it does not correspond to a local maximum.

5.2 *Properties of the sample*

5.2.1 *NPMLE*

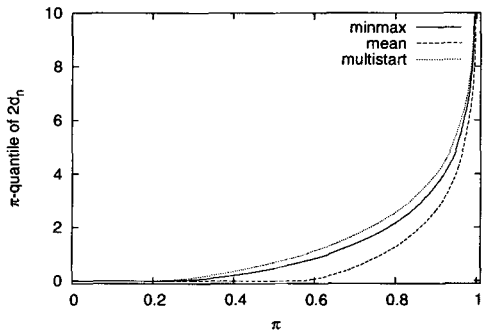
To get an idea of the nature of possible groups in the data, we calculate for each sample the nonparametric maximum likelihood estimator (NPMLE, cf., Böhning (2000) or Lindsay (1995)). The NPMLE maximizes the log likelihood in the space of all mixing distributions. There is always a version with finite support, denoted here by $\hat{P}_{\text{NPMLE}}$. To calculate $\hat{P}_{\text{NPMLE}}$ for mixtures of exponentials, we use the Intra-Simplex Direction Method (ISDM, Lesperance and Kalbfleisch (1992)) in combination with an improved method to maximize the gradient function (Seidel and Ševčíková (2002*a*)).

5.2.2 *Clustering of sample points*

According to the heuristic arguments in Section 3, spurious components should occur if for a small value of ρ, the first ρ elements in the ordered sample form a cluster that is well separated from the other points. As an indicator for such samples, we cluster the sample points according to the Bayes criterion (McLachlan and Peel (2000), Subsection 1.15.2) with respect to $\hat{P}_{\text{minmax}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{p})$. This induces an allocation of the sample points to groups g_t corresponding to $\hat{\theta}_t$, $t = 1, 2$, as follows: $x_{(i)}$ is allocated to g_1, if $\hat{p}_1 f(x_{(i)}, \hat{\theta}_1) > \hat{p}_2 f(x_{(i)}, \hat{\theta}_2)$, otherwise to g_2. If $\hat{P}_{\text{minmax}}$ corresponds to a spurious maximum, g_1 should consist of the first ρ elements for small ρ.

6. Simulation study, overview

We simulated 10 000 samples of size $n = 100$, 200 and 1000, respectively, from a homogeneous population with parameter $\theta = 1$, a mixture with parameter $P = (1, 0.33, 0.7)$



π	*minmax*	*mean*	*multistart*
0.998	11.364	9.948	11.606
0.999	13.405	11.524	13.651
1.0	19.227	15.805	19.227

Fig. 6. Quantiles of $2d_n$ as function of π under homogeneity, $n = 200$, on the set of all replications (last values are shown in the table).

and a mixture with parameter $P = (1, 30, 0.1)$. For each sample, the estimators $\hat{P}_{\text{minmax}}$, $\hat{P}_{\text{mean}}$, $\hat{P}_{\text{NPMLE}}$ and their likelihoods were calculated; for $n = 100$ and $n = 200$ also $\hat{P}_{\text{multistart}}$. For $strategy = minmax, \ldots$, we shall write $d_n(strategy)$ for the corresponding log likelihood difference.

In the following, we show the most important results for $n = 200$. For $n = 1000$, the same phenomena can be observed (exceptions will be stated explicitly), cf., Seidel and Ševčíková (2003). Here also more detailed results for $n = 200$ can be found.

6.1 *Null hypothesis, quantiles*

Figure 6 shows the simulated quantile functions of $2d_n$, i.e. the functions that assign to each $\pi \in (0, 1)$ the π-quantile, for the strategies *minmax*, *mean* and *multistart* on the set of all replications under homogeneity. Here, $2d_n(minmax)$ is in fact stochastically larger than $2d_n(mean)$; $2d_n(multistart)$ is still somewhat larger than $2d_n(minmax)$.

To explain the differences between the distributions of the test statistics, we consider certain subsets of the set Ω of all replications. Let $mean0$ ($minmax0$) denote the set of all replications where $2d_n \approx 0$ for *mean* (*minmax*). Moreover, we define

- $A = mean0 \cap minmax0$,
- $B = mean0 \setminus minmax0$ and
- $C = \Omega \setminus mean0$.

For an overview of all sets defined here and in the following, see Fig. 15 in Appendix B. For $n = 200$, set A has 2003 elements, $minmax0$ has 2231 elements and B has 3656 elements. In particular, this means that although $minmax0$ is not a subset of $mean0$, most elements of $minmax0$ also belong to $mean0$.

Figure 7 shows the quantile functions on set B (left panel) and on set C (right panel). On the set B ($2d_n(mean)$ approximately zero), $2d_n(minmax)$ and $2d_n(multistart)$ can be very large. In fact, for $\alpha = 0.1$, 0.05 and 0.01, over 30% of the replications for which $2d_n(minmax)$ exceeds its $(1 - \alpha)$-quantile come from the set B. On the other hand, on the set C the distributions of $2d_n(minmax)$ and $2d_n(mean)$ are almost equal.

We modified the strategy *minmax* by assigning the value zero to $2d_n(minmax)$ on the set $mean0$. After this modification, *minmax* had on Ω for all considered sample sizes nearly the same quantiles as *mean*. Consequently, the set B seems to be in a certain sense the source of the differences between the quantiles of *mean* and *minmax*, therefore it should be investigated more closely.

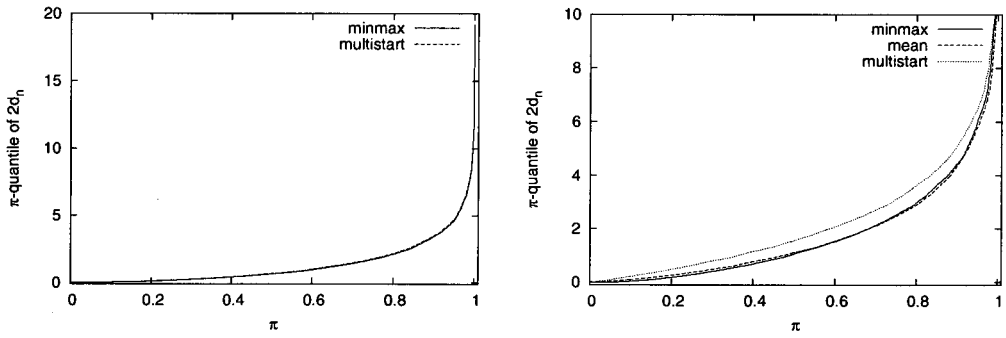


Fig. 7. Quantiles of $2d_n$ as function of π, $n = 200$, under homogeneity. Left panel: set B. Right panel: set C.

However, the set C is of interest, too. Whereas the distributions of the test statistics corresponding to *mean* and *minmax* are almost identical here, the estimators themselves and even their likelihoods may be completely different. For a closer analysis, we split the set C into subsets $C1$–$C3$ according to the following criteria:
$C1$: $\hat{P}_{\text{minmax}} \approx \hat{P}_{\text{mean}}$ and $l(\hat{P}_{\text{minmax}}) \approx l(\hat{P}_{\text{mean}})$,
$C2$: $l(\hat{P}_{\text{minmax}}) > l(\hat{P}_{\text{mean}})$,
$C3$: $l(\hat{P}_{\text{minmax}}) < l(\hat{P}_{\text{mean}})$,
see Fig. 15 in Appendix B. For defining $C2$ and $C3$, only elements of $C \setminus C1$ are considered. The set $C1$ consists of 1651 elements, whereas $|C2| = 1204$ and $|C3| = 1486$.

On each of these sets, the likelihood differences can be large. For $\alpha = 0.1$, 0.05 and 0.01, 50% of the replications for which $2d_n(mean)$ exceeds its $(1-\alpha)$-quantile come from the set $C1$, 10% from $C2$ and 40% from $C3$. For $2d_n(minmax)$, more than 30% come from $C1$, about 30% from $C2$ and also a few replications come from $C3$.

In particular, the set $C2 \cup C3$ is of interest, as the test statistics $2d_n(minmax)$ and $2d_n(mean)$ are different here. In fact, they are almost independent, see Seidel and Ševčíková (2003). The correlation coefficient between both on this set is 0.26.

A closer analysis of the samples, the likelihood functions and the estimators in the considered sets are given in Section 7.

6.2 *Alternative hypotheses*

The quantile functions of $2d_n$ for *mean*, *minmax* and *maximum* (representing global optimization) for $n = 200$ are shown in Fig. 8 for a mixture with parameter $P = (1, 0.33, 0.7)$ (left panel) and with $P = (1, 30, 0.1)$ (right panel).

For $P = (1, 0.33, 0.7)$, *mean* is in contrast to homogeneity now stochastically larger than *minmax* and *maximum* has nearly identical distribution as *mean*. On the other hand, the quantiles of *mean* under the null hypothesis are smaller than these of *minmax*. These two facts explain the larger power of *mean*. In fact, *mean* exceeds the *minmax*-quantiles even (much) more often than *minmax* exceeds the *minmax*-quantiles. On the other hand, the quantiles of *maximum* are larger than those of *mean*, and as *maximum* is stochastically almost equivalent to *mean*, it has smaller power. This explains the "typical behaviour".

For $P = (1, 30, 0.1)$ and $n = 200$, $2d_n(mean)$ is zero with probability 0.087, whereas $2d_n(minmax)$ is larger than zero almost everywhere. Therefore, the quantile function

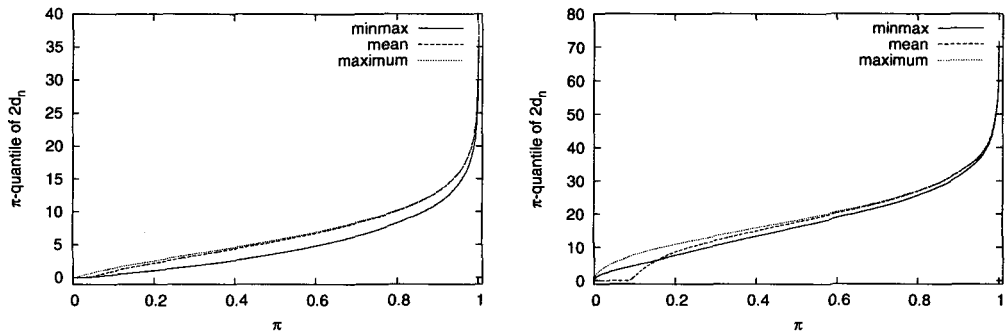


Fig. 8. Quantiles of $2d_n$ as function of π, $n = 200$, for $P = (1, 0.33, 0.7)$ (left panel) and $P = (1, 30, 0.1)$ (right panel).

of $2d_n(mean)$ is smaller than that of $2d_n(minmax)$ for small values of π. For increasing π, however, it increases faster than the latter, crosses it and then remains larger. For $\alpha = 0.1$ and 0.05, the critical values are so small that they are exceeded more often by *minmax* than by *mean*, whereas for $\alpha = 0.01$, *mean* has again the larger power.

In the following sections, we present a detailed study of the reasons for the observed behaviour of the considered strategies for $n = 200$.

7. Analysis of particular sets of samples

7.1 *Additional sets*

7.1.1 *Set B*

Consider the intervals $I(1), \ldots, I(5) = [0, 2[, [2, 4[, [4, 6[, [6, 8[, [8, \infty[$. We split the set B into subsets $B1, \ldots, B5$, where Bi is defined as the set of replications for which $2d_n(minmax) \in I(i)$, see Fig. 15 in Appendix B. The sets $B1, \ldots, B5$ consist of 2872, 516, 169, 64 and 35 replications.

7.1.2 *Alternative hypotheses*

For each of the parameters $P = (1, 0.33, 0.7)$ and $P = (1, 30, 0.1)$, we split the set of all replications into three subsets:

E: $\hat{P}_{\text{minmax}} \approx \hat{P}_{\text{mean}}$ and $l(\hat{P}_{\text{minmax}}) \approx l(\hat{P}_{\text{mean}})$.

E has 4990 elements for $P = (1, 0.33, 0.7)$ and 6747 elements for $P = (1, 30, 0.1)$.

D: The complement of E, with the subsets

$D1$: $l(\hat{P}_{\text{minmax}}) > l(\hat{P}_{\text{mean}})$
(1302 elements for $P = (1, 0.33, 0.7)$ and 1553 elements for $P = (1, 30, 0.1)$),

$D2$: $l(\hat{P}_{\text{minmax}}) < l(\hat{P}_{\text{mean}})$
(3708 elements for $P = (1, 0.33, 0.7)$ and 1700 elements for $P = (1, 30, 0.1)$),

see Fig. 15 in Appendix B. The quantile functions of $2d_n$ for *mean*, *minmax* on the different sets are shown in Fig. 9 for the mixture with parameter $P = (1, 0.33, 0.7)$ and in Fig. 10 for $P = (1, 30, 0.1)$.

7.2 *Classification of sets*

A first group is constituted by the sets of replications where $\hat{P}_{minmax} \approx \hat{P}_{mean}$ or at least $l(\hat{P}_{minmax}) \approx l(\hat{P}_{mean})$. Consequently this group consists of the sets A, $C1$ and E. The sets $B1, \ldots, B5$, $C2$, $C3$, $D1$ and $D2$ form a second group.

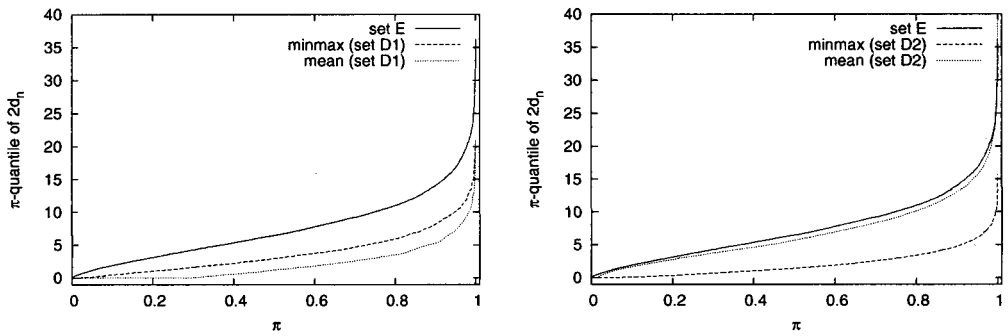


Fig. 9. Quantiles of $2d_n$ as function of π, $P = (1, 0.33, 0.7)$ and $n = 200$: sets E, $D1$ (left panel) and sets E, $D2$ (right panel).

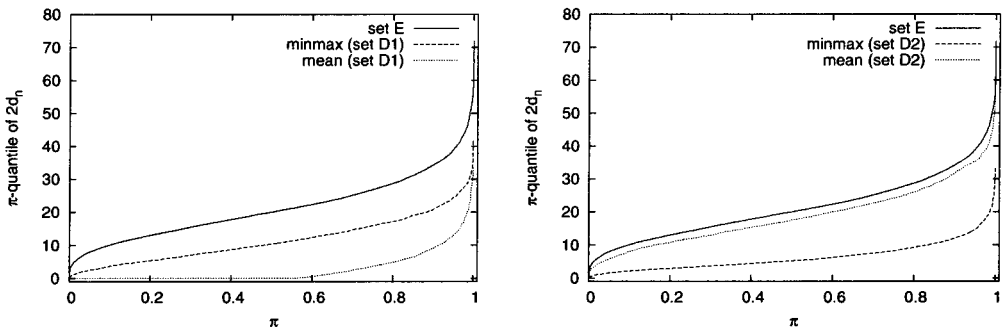


Fig. 10. Quantiles of $2d_n$ as function of π, $P = (1, 30, 0.1)$, $n = 200$: sets E, $D1$ (left panel) and sets E, $D2$ (right panel).

An important feature that characterizes the members of the groups is the presence of spurious maximizers: these are almost completely missing in the first group, whereas they play a large role in the second group. According to Section 3, spurious maximizers are caused by the occurence of very small sample points. In fact, the quantile functions of $x_{(1)}$ are in the first group much larger than in the second. The theoretical quantile function of $x_{(1)}$ in a sample of size n from a homogeneous exponentially distributed population with parameter $\theta = 1$ is given by $Q_{(1)}(\pi) = -(1/n)\log(1-\pi)$. As an example, this is compared in Fig. 11 with the empirical quantile function of $x_{(1)}$ on the set A and on the sets $B1, \ldots, B5$. On A, the empirical quantile function is much larger than $Q_{(1)}$, whereas on the sets $B1, \ldots, B5$, the quantiles are smaller, tending to extremely small values on Bi with increasing i.

Another indicator of the presence of spurious maximizers is the location of the estimators $\hat{P}_{\text{minmax}}$ and $\hat{P}_{\text{mean}}$. In a set of a samples from a homogeneous population, it is advantageous to visualize the distribution of an estimator as a scatter plot of the standardized first coordinates $(\hat{\theta}_1/\overline{x}, \hat{\theta}_2/\overline{x})$. This is done in Fig. 12 for the set $C2$. The parameter $(\hat{\theta}_1, \hat{\theta}_2)$ of $\hat{P}_{minmax}$ is located close to the axis $]0, \overline{x}] \times \{\overline{x}\}$. There are essentially two clusters, a first cluster with small values of $\hat{\theta}_1$, and a second one with large values. With the exception of a few points, the second cluster lies on the axis. In the first

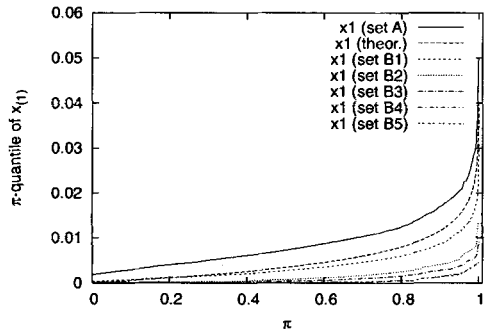


Fig. 11. Theoretical quantiles of $x_{(1)}$ in homogeneous population and observed quantiles on sets A and $B1, \ldots, B5$ as function of π.

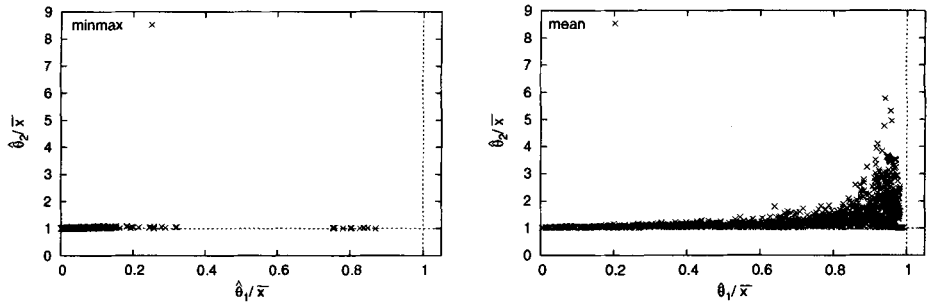


Fig. 12. Distribution of $(\hat{\theta}_1/\overline{x}, \hat{\theta}_2/\overline{x})$ for $\hat{P}_{\text{minmax}}$ (left panel) and $\hat{P}_{\text{mean}}$ (right panel) on set $C2$.

cluster, $(\hat{\theta}_1, \hat{\theta}_2)$ lies slightly above the axis. This coincides with the characterization of spurious maximizers in Subsection 3.1, and in fact, the first cluster consists of spurious maximizers, whereas the second does not.

7.2.1 *Set A*

In all replications, $(\hat{\theta}_1, \hat{\theta}_2) \approx (\overline{x}, \overline{x})$ for the estimator $\hat{P}_{\text{mean}}$. The parameter $(\hat{\theta}_1, \hat{\theta}_2)$ of $\hat{P}_{\text{minmax}}$ is located on the axis $]0, \overline{x}] \times \{\overline{x}\}$, but $\hat{P}_{\text{minmax}}$ is in no case a spurious maximizer. Rather it represents a homogeneous population with parameter $\overline{x}$ in the sense of equation (2.3).

Usually the starting points of *multistart* result according to the criterion of Subsection 5.1.2 in one single estimator, represented by $(\overline{x}, \overline{x}, 0.5)$, and the NPMLE always results in a homogeneous population with parameter $\overline{x}$. There are no small outliers, and it seems that the samples that constitute set A follow the pattern of a homogeneous population.

7.2.2 *Sets C1, E*

Let $\hat{P} = \hat{P}_{\text{minmax}} (\approx \hat{P}_{\text{mean}})$. Usually there is a strict local maximum in $\hat{P}$, and very often the likelihood function is unimodal in the sense that all starting points of *multistart* (with the exception of the saddle point $(\overline{x}, \overline{x}, 0.5)$) converge to $\hat{P}$. On set $C1$, $\hat{P}$ never contains a spurious component, and in most cases $\hat{P}_{\text{NPMLE}}$ is equal to $\hat{P}$. There it seems that the samples correspond to mixtures with two clearly distinct components

and genuinely positive mixing weights.

On set E, the NPMLE usually has between two and four components, and its likelihood can be considerably larger than the likelihood of $\hat{P}$. Often $\hat{P}_{\text{NPMLE}}$ has spurious components. For $P = (1, 0.33, 0.7)$ there are cases where $\hat{P}$ is a spurious solution with one component similar to a spurious component of $\hat{P}_{\text{NPMLE}}$, although two other components of $\hat{P}_{\text{NPMLE}}$ are reasonable estimators of the population parameters. The latter are not found by *minmax* or *mean*. Here one can also find few examples where *multistart* results in two different estimators, one spurious maximizer, which is found by both *minmax* and *mean*, and a reasonable one, which is found by some other starting point.

7.2.3 *Set B*

The estimators $\hat{P}_{\text{minmax}}$ are located near the axis $]0, \overline{x}] \times \{\overline{x}\}$, and for increasing values of i, the estimators in Bi are more and more concentrated at small values of θ_1 with $\hat{\theta}_2$ slightly larger than $\overline{x}$ at least for $i \geq 2$. In all but a very few cases, there is a strict local maximum at $\hat{P}_{\text{minmax}}$ according to both criteria in Subsection 5.1.1.

The parameter $(\hat{\theta}_1, \hat{\theta}_2)$ of $\hat{P}_{\text{mean}}$ is again located near the axis $]0, \overline{x}] \times \{\overline{x}\}$, very close to $(\overline{x}, \overline{x})$, and for $B3$–$B5$ it is practically identical to $(\overline{x}, \overline{x})$.

The diagnostic instruments in Section 5 indicate that the samples might be of the type "homogeneous population with small outliers", where the latter result in a spurious maximizer which is found by *minmax*. The starting values of *multistart* result almost always in two clusters of estimators; one corresponds to $\hat{P}_{\text{minmax}}$, the other to $\hat{P}_{\text{mean}}$. In fact, $\hat{P}_{\text{minmax}}$ usually exhibits the properties of a spurious maximizer discussed in Subsection 3.1. Whereas not all samples in the set $B1$ precisely follow this pattern, it is very pronounced for Bi with larger index i and also for samples with very small $x_{(1)}$. For example, the samples described in Examples 3.1 and A.1 belong to set $B5$.

The estimator $\hat{P}_{\text{NPMLE}}$ always has at least two components. In cases with exact two (which are the most ones) it is identical to $\hat{P}_{\text{minmax}}$.

Finally, let us describe the effect of removing the first e elements of the ordered sample (not shown here). Removing $x_{(1)}$ drastically cuts down the quantile functions of $2d_n(minmax)$ on $B1, \ldots, B5$; for $e = 5$ they are close to and for $e = 10$ practically identical to zero. This observation strongly supports the hypothesis that small data points in the samples on set B create spurious maximizers which are found by *minmax*, whereas *mean* estimates a homogeneous population.

7.2.4 *Sets C2, C3, D1, D2*

With the exception of set $D1$ for parameter $P = (1, 30, 0.1)$, which will be discussed separately, $\hat{P}_{\text{minmax}}$ usually has all properties of a spurious maximizer, whereas $\hat{P}_{\text{mean}}$ is very often a mixture with two clearly distinct components and genuinely positive mixing weights. Under the alternative hypothesis, this may be a good estimator of the population parameter. Although spurious maximizers can have a large likelihood, now the likelihood of the true two-component mixture represented by $\hat{P}_{\text{mean}}$ is often still (much) larger, especially under the alternative hypothesis.

In a large majority of replications, there is a strict local maximum in $\hat{P}_{\text{minmax}}$ and $\hat{P}_{\text{mean}}$ according to both criteria in Subsection 5.1.1. Very often, $\hat{P}_{\text{minmax}}$ and $\hat{P}_{\text{mean}}$ seem to be the only local maximizers of the log likelihood according to the clustering of the results of *multistart* (Subsection 5.1.2).

The NPMLE usually has more than two components. Frequently it has three: a very small component θ_1^* with a very small mixing weight p_1^* (having all features of a spurious component) and two additional components θ_2^* and θ_3^* with genuinely positive

mixing weights. Often then $\hat{P}_{\text{minmax}}$ is a spurious maximizer with $(\hat{\theta}_1, \hat{p})$ very similar to (θ_1^*, p_1^*) and second component slightly larger than $\overline{x}$, whereas $\hat{P}_{\text{mean}}$ is similar to the remaining two components of $\hat{P}_{\text{NPMLE}}$. In these cases, *multistart* also results in the two essentially different maxima $\hat{P}_{\text{minmax}}$ and $\hat{P}_{\text{mean}}$. Here the sample represents a two-component mixture with an additional spurious component caused by small outliers, where the spurious solution is found by *minmax* and the genuine solution by *mean*. Even in these cases, the log likelihood of the NPMLE can be considerably larger than the log likelihood of both *mean* and *minmax*.

In other cases, the NPMLE has more than three components. Especially under the alternative hypothesis, it may have two or more spurious components. In particular, spurious components with large values of θ have been observed. The components of $\hat{P}_{\text{minmax}}$ and $\hat{P}_{\text{mean}}$ need not correspond to any component of $\hat{P}_{\text{NPMLE}}$, even if the latter has only three. Often then, *multistart* results in more than two different maximizers. There may also be samples without small outliers, where none of the components of the NPMLE is a spurious component in the technical sense of Section 3. $\hat{P}_{\text{minmax}}$ is then apparently not a spurious maximizer, although it may have a component with a small parameter and a small mixing weight. In these and also in other cases, *mean* may result in an estimator with first and second component equal to $\overline{x}$.

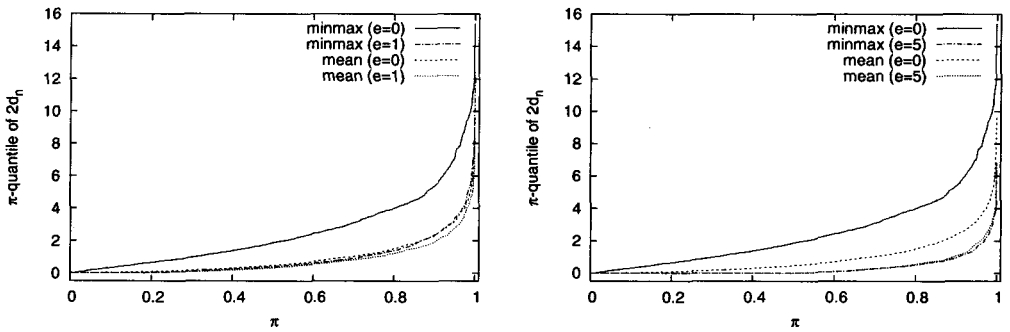


Fig. 13. Quantiles of $2d_n$ (*minmax* and *mean*) on set $C2$ as function of π after removing the first e elements of the ordered sample. Left panel: $e = 0$ and 1; right panel: $e = 0$ and 5.

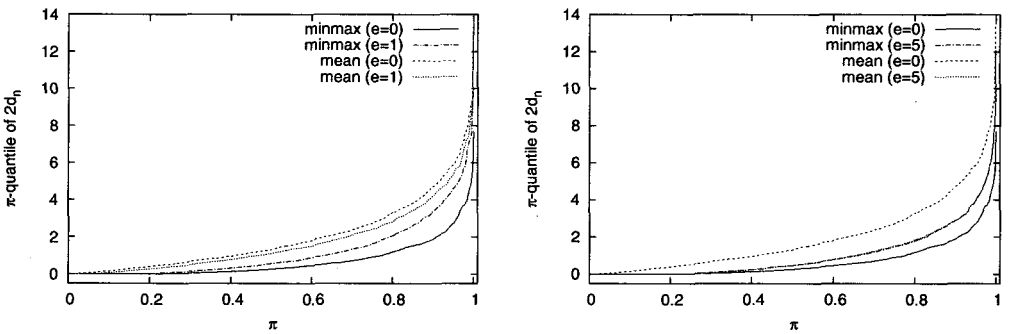


Fig. 14. Quantiles of $2d_n$ (*minmax* and *mean*) on set $C3$ as function of π after removing the first e elements of the ordered sample. Left panel: $e = 0$ and 1; right panel: $e = 0$ and 5.

Set $D1$, parameter $P = (1, 30, 0.1)$: Here again, $\hat{P}_{\text{minmax}}$ is often a spurious maximizer, and then $\hat{P}_{\text{mean}}$ may correspond to a homogeneous population with parameter $\overline{x}$ and $2d_n = 0$ or to a mixture with distinct components. However, now there are also many cases where one component of $\hat{P}_{\text{minmax}}$ has a small parameter and a small mixing weight, but both are too large for representing a spurious component. These can be considered as estimators of the lower contamination which is represented in the population by $\theta_1 = 1$ and $p = 0.1$. In these cases, $\hat{P}_{\text{mean}}$ usually corresponds again to a homogeneous population with parameter $\overline{x}$ and $2d_n = 0$.

Removing low order statistics: The influence of small outliers can be observed in Figs. 13–14, which shows the effect of removing the first e elements of the ordered sample ($e = 1$ and 5) on $2d_n(minmax)$ and $2d_n(mean)$ for the sets $C2$ and $C3$. Removing only the smallest element of the sample drastically affects $2d_n(minmax)$; on the set $C2$, it decreases, whereas it increases on the set $C3$. In the end, all curves are decreasing for increasing values of e. However, the large differences between the strategies are considerably reduced already for $e = 1$ and vanish almost completely for $e = 5$.

The same behaviour has also been observed for $D2$, both parameters, and for $D1$, parameter $P = (1, 0.33, 0.7)$. This means that here, the difference between $2d_n(minmax)$ and $2d_n(mean)$ is mainly caused by small outliers leading to spurious solutions. On the other hand, for the parameter $P = (1, 30, 0.1)$, the difference between the distributions of $2d_n(minmax)$ and $2d_n(mean)$ on the set $D1$ remains large. This is due to the fact that here, the difference between both estimators is not a consequence of spurious solutions alone.

8. Conclusion

The likelihood function in two-component exponential mixture models tends to be very flat especially if the data come from a homogeneous population, and it seems to have with a large probability not more than two essentially different local maxima. Very often it has only one.

If it has at least two, one of these is in most cases a spurious local maximum. The latter occur very often; under the null hypothesis with probability much larger than 50% (depending on the sample size), whereas under the alternative hypotheses the probability is somewhat smaller than under the null hypothesis, depending on the parameter. Spurious maximizers can have large likelihood, and they are typically found by *minmax*. On the other hand, *mean* usually results in reasonable parameter estimates. However, under the null hypothesis, these have in many cases smaller likelihood than the spurious. Therefore *minmax* results in larger quantiles.

Under the alternative hypothesis, the likelihood of the spurious can be still large, but now the regular estimates often have larger likelihoods. Since *mean* usually finds regular estimates and since the quantiles of *mean* are smaller than the quantiles of *minmax*, *mean* usually has larger power. Global optimization has in these cases smaller power than *mean*, as its quantiles are larger.

In situations where the distribution of the test statistic under the null hypothesis depends on the model parameter, quantiles have to be bootstrapped under the estimated parameter and computation time may be critical. Here it would be advantageous if one could start the EM from only one carefully chosen initial value, as for example *mean*. Our results indicate that over a wide range of parameter values, this strategy yields good estimators and powerful tests.

On the other hand, in lower contamination models, *mean* is for small sample sizes inferior to *minmax*. If one has no prior information, for which parameter set the statistical procedure should be designed, we would not recommend to use only one starting point. It seems that global optimization with elimination of spurious maximizers might be a good strategy for parameter estimation and testing.

Acknowledgements

The authors would like to thank a referee for his constructive suggestions that led to significant improvement of the presentation of the paper.

Appendix

A. *Numerical example for spurious estimates*

Example A.1. (One small outlier) As in Example 3.1, we consider a particular sample of size $n = 200$ from a homogeneous population ($\theta = 1$), now with $x_{(1)} = 0.000012975$ and $\overline{x} = 1.0784$. Compared to $E(x_{(1)}) = 1/n = 0.005$, $x_{(1)}$ is very small.

The estimators are $\hat{P}_{\text{minmax}} = (0.0000130, 1.0838, 0.0049676)$ with likelihood difference $2d_n = 8.0823$ and $\hat{P}_{\text{mean}} = (1.0784, 1.0784, 0.45)$ with likelihood difference $2d_n = 0$. Table 3 shows for the first indices i the order statistic $x_{(i)}$, the quantities $w_i(\hat{\theta}_1, \hat{\theta}_2)$, $\tau_i(\hat{\theta}_1, \hat{P})$ and $\tau_i(\hat{\theta}_2, \hat{P})$ for $\hat{P} = \hat{P}_{\text{minmax}}$ as well as the contributions $l_i(\hat{P}) = \log f(x_{(i)}, \hat{P})$ of $x_{(i)}$ to the log likelihood $l(\hat{P})$ for $\hat{P} = \hat{P}_{\text{minmax}}$ and $\hat{P} = \hat{P}_{\text{mean}}$.

Clustering the sample points according to the Bayes criterion with respect to $\hat{P}_{\text{minmax}}$ (see Subsection 5.2.2) results in a cluster $x_{(1)}$ allocated to the first component, whereas the remainder of the sample is allocated to the second component. According to the criterion in Subsection 5.1.2, the starting values of *multistart* result as in Example 3.1 in two clusters of estimators represented by $\hat{P}_{\text{minmax}}$ and $\hat{P}_{\text{mean}}$.

Table 3. Contribution of sample points to likelihood equation at $\hat{P}_{\text{minmax}}$ and to the log likelihood.

		$\hat{P} = \hat{P}_{\text{minmax}}$				$\hat{P} = \hat{P}_{\text{mean}}$
i	$x_{(i)}$	$w_i(\hat{\theta}_1, \hat{\theta}_2)$	$\tau_i(\hat{\theta}_1, \hat{P})$	$\tau_i(\hat{\theta}_2, \hat{P})$	$l_i(\hat{P})$	$l_i(\hat{P})$
1	0.000013	0.000033	200.0	0.006509	4.954145	−0.075520
Σ_1					4.954145	−0.075520
2	0.002775	9.1×10^{87}	0.0	1.004992	−0.088028	−0.078081
3	0.003657	2.9×10^{117}	0.0	1.004992	−0.088841	−0.078898
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	5.331832	∞	0.0	1.004992	−5.004969	−5.019569
Σ_{rem}					−216.014555	−215.026040
Σ_{tot}					−211.060410	−215.101560

B. *Overview of sets defined in Sections 6 and 7*

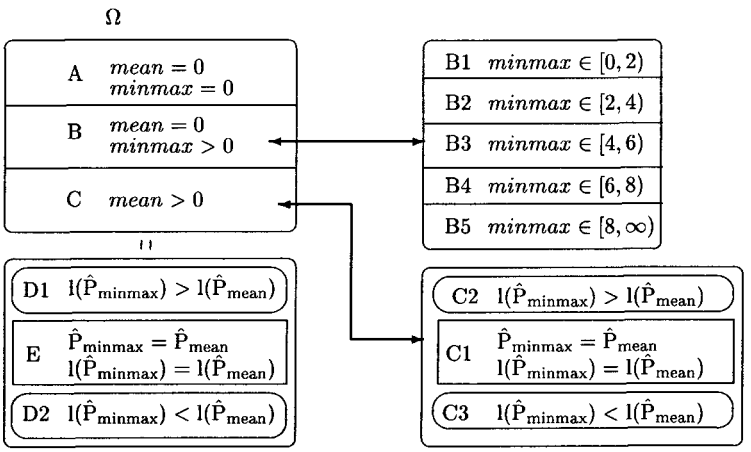


Fig. 15. Subsets overview.

REFERENCES

Biernacki, C., Celeux, G. and Govaert, G. (2003). Strategies for getting the highest likelihoods in mixture models, Special Issue on Recent Developments in Mixture Models (Guest Editors: D. Böhning and W. Seidel), *Computational Statistics & Data Analysis*, **41**, 561–575.

Böhning, D. (2000). *Computer-Assisted Analysis of Mixtures and Applications*, Chapman & Hall/CRC, Boca Raton.

Böhning, D. and Schlattmann, P. (1992). Computer-assisted analysis of mixtures (C.A.MAN): Statistical algorithms, *Biometrics*, **48**, 283–303.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. and Lindsay, B. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family, *Annals of the Institute of Statistical Mathematics*, **46**, 373–388.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes, *The Annals of Statistics*, **27**, 1178–1209.

Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (eds. L. M. Le Cam and R. A. Olshen), **2**, 789–806, Wadsworth, Belmont, California.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (eds. L. M. Le Cam and R. A. Olshen), **2**, 807–810, Wadsworth, Belmont, California.

Karlis, D. (2001). A cautionary note about the EM algorithm for finite exponential mixtures, Tech. Report, No. 150, Department of Statistics, Athens University of Economics.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the EM algorithm for finite mixtures, Special Issue on Recent Developments in Mixture Models (Guest Editors: D. Böhning and W. Seidel), *Computational Statistics & Data Analysis*, **41**, 577–590.

Lesperance, M. and Kalbfleisch, J. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution, *Journal of the American Statistical Association*, **87**, 120–126.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*, Institute of Mathematical Statistics, Hayward, California.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability, *The Annals of Statistics*, **31**, 807–832.

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

Seidel, W. and Ševčíková, H. (2002*a*). Efficient calculation of the NPMLE of a mixing distribution for mixtures of exponentials, *Discussion Papers in Statistics and Quantitative Economics*, **96**, Universität der Bundeswehr Hamburg.

Seidel, W. and Ševčíková, H. (2002*b*). Tools for analyzing and maximizing likelihood functions in mixture models, *Discussion Papers in Statistics and Quantitative Economics*, **104**, Universität der Bundeswehr Hamburg.

Seidel, W. and Ševčíková, H. (2003). A detailed investigation of likelihood maxima in two-component exponential mixture models and their implication on LR tests, *Discussion Papers in Statistics and Quantitative Economics*, **106**, Universität der Bundeswehr Hamburg.

Seidel, W., Mosler, K. and Alker, M. (2000*a*). A cautionary note on likelihood ratio tests in mixture models, *Annals of the Institute of Statistical Mathematics*, **52**, 481–487.

Seidel, W., Mosler, K. and Alker, M. (2000*b*). Likelihood ratio tests based on subglobal optimization: A power comparison in exponential mixture models, *Statistical Papers*, **41**, 85–98.

Seidel, W., Ševčíková, H. and Alker, M. (2000*c*). On the power of different versions of the likelihood ratio test for homogeneity in an exponential mixture model, *Discussion Papers in Statistics and Quantitative Economics*, **92**, Universität der Bundeswehr Hamburg.