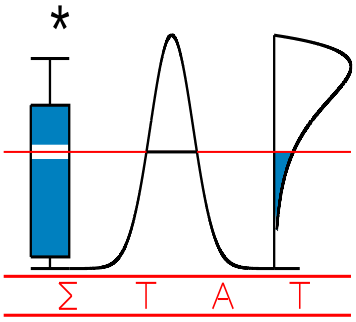
T E C H N I C A L

R E P O R T

0355

LIKELIHOOD ESTIMATION OF FINITE MIXTURES

ANDRIES, E., CROES, K., VERBEKE, G., DE SCHEPPER, L. and G. MOLENBERGHS



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Likelihood estimation of finite mixtures

E. Andries†

Limburgs Universitair Centrum, Institute for Materials Research (IMO), Diepenbeek, Belgium.

K. Croes

Xact, Genk, Belgium.

G. Verbeke

Catholic Univ Louvain, Ctr Biostat, Louvain, Belgium.

L. De Schepper

Limburgs Universitair Centrum, IMO, Diepenbeek, Belgium.

G. Molenberghs

Limburgs Universitair Centrum, Center for Statistics, Diepenbeek, Belgium.

Summary. Maximum likelihood (ML) estimation of general finite mixtures is, due to the singularities in the likelihood, a common but very difficult problem. In particular, a classical ML estimator does not exist. In this paper, standard methods dealing with the unbounded likelihood are reviewed. Important drawbacks of these methods are discussed. An alternative method, rooted in the literature, and known as likelihood estimation, is proposed. This likelihood estimation procedure is similar in many respects to maximum likelihood and has good statistical properties for the problem at hand. Often, this estimate can be found as the largest local maximum of the likelihood. It is shown how this approach can be of attraction in the finite mixture problem. Further, the important problem of a so-called spurious maximum as likelihood estimate is tackled. By means of key examples, spurious maxima are characterized. Our perception of the underlying problem is given and guidelines are proposed to deal with these spurious maxima in practice.

Keywords: Likelihood estimation; General finite mixture; Spurious maximum; Local maximum; Conditions of Cramér

1. Introduction

In today's applied research, mixtures of distributions have become extremely popular. They exist in a wide range of forms, are used in a wide class of applications, and although the related literature is huge already, it is still growing. One of the main methods to estimate these distributions is maximum likelihood (ML) estimation. Of interest here, are (univariate) general finite mixtures where the ML method tends to break down. The density of a general M-component mixture is denoted by:

$$f_M(x|\boldsymbol{\theta}) = \sum_{m=1}^{M} \pi_m f(x|\mu_m, \sigma_m), \tag{1}$$

†*Address for correspondence:* Limburgs Universitair Centrum, Institute for Materials Research (IMO), Materials Physics Division, Wetenschapspark 1,B-3590 Diepenbeek,Belgium.
E-mail: Ellen.Andries@luc.ac.be

with $\sum_{m=1}^{M} \pi_m = 1$, $f(x|\mu_m, \sigma_m)$ the density of a 2-parameter distribution, μ_m a scale and σ_m a shape parameter. In this paper, we will mostly consider the component densities to be normal with mean μ_m and standard deviation σ_m, but other distributions such as the Weibull or gamma can equally well be used. Note, however, that literature on finite mixtures with other component distributions exists but is relatively limited.

It is well known that ML estimation of normal mixtures with unequal means and variances is problematic. This is due to the fact that the likelihood function corresponding to such a mixture is unbounded at some points, also referred to as singularities, on the edge of the parameter space. For example, take $(y_1, \cdots, y_n)$, a sample from a 2-component normal mixture with likelihood $L(\boldsymbol{\theta}, \mathbf{y})$ given by:

$$L(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^{n} \left[\pi_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\left(\frac{y_i - \mu_1}{\sigma_1}\right)^2} + (1 - \pi_1) \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{1}{2}\left(\frac{y_i - \mu_2}{\sigma_2}\right)^2} \right]. \tag{2}$$

It is then easily seen that the likelihood goes to infinity whenever $\mu_1 = y_i$ and σ_1 approaches zero, with the other parameters having arbitrary values. Clearly, these "maxima" are pathological and do not correspond to useful mixtures. Moreover they are inconsistent estimates and cannot be regarded as maximum likelihood estimates (MLEs), since due to its unboundedness, the likelihood does not have a global maximum (Lehmann, 1980). Nevertheless, some authors prefer to call these maxima inconsistent MLEs (Duda and Hart, 1973). We believe that calling it this name should be avoided.

Very often the variance parameter is taken to be equal for all components because the singularities of the likelihood $L(\boldsymbol{\theta}, \mathbf{y})$ then disappear (if the sample has a size larger than 1 and not all values are equal). This restrictive assumption can be a satisfactory solution in some applications but in general it obviously is not (Fisher *et al.*, 2000; Joyce *et al.*, 1976). Although the common way to go was and still is not to use mixtures with unequal means and variances, some methods are proposed in the literature to satisfactorily tackle the problem of an unbounded likelihood. In some sense, these techniques try to regularize the problem by removing the unboundedness. We believe they all suffer, however, from some important drawbacks. The next section reviews and discusses some standard methods to deal with a restricted ML estimation of such mixtures.

However, there exists a local maximum estimation method with good statistical properties. This theory is rooted in the literature, acknowledged by some authors, but still rarely applied. By going back to Cramér, we will review in Section 3 that well-behaved estimates as a solution of the likelihood equations (LEQs) do exist for mixtures with density (1), despite the non-existence of the MLE. Moreover, they have similar behavior as MLEs in the case of mixtures with an equal variance parameter. This does not solve the entire problem, however, since the likelihood function has multiple roots and it is not specified which local maximum of the likelihood is the proper one. It will be discussed that for finite normal mixtures, the largest local maximum of the likelihood function corresponds to these well-behaved estimates.

It ought to be mentioned that not everyone agrees on this: McLachlan and Peel (2000), amongst others, argue that one should first skip some spurious or illogical maxima before selecting the largest local one. Although, such maxima are indeed an issue in likelihood estimation, the way they are handled so far, seems to be flawed. The result aimed at in Section 4 is to clarify the situation. By means of some key examples, the problem is situated and spurious maxima are characterized. It is discussed that an overall view of the surface of the likelihood function is required in order to obtain an idea about the credibility of the

likelihood estimate, found as the largest local maximum of the likelihood function. Based on the surface of the likelihood function, samples are classified as being highly unstable, unstable or stable. The latter is used to deal with spurious maxima in practice.

2. Removing the unboundedness

As explained in the introduction, ML estimation for a general M-component mixture turns out to be problematic in a classical sense. More precisely, an MLE, defined as the global maximum of the likelihood, does in many cases not exist. In the past, several authors have tried to regularize the problem by removing the singularities of the likelihood in order to still obtain a, perhaps modified, MLE. The most common approaches are based either on adapting the likelihood (§2.1) or on restricting the parameter space (§2.2).

2.1. Adaptation of the likelihood function

Cox and Hinkley (1974) pointed out that the anomaly of an infinite likelihood function would disappear if one would take into account the inherent grouped nature of the data. In practice, all observations are discrete and therefore a continuous model is only a theoretical concept. Similarly, Aitkin (2001) states that the unboundedness of the likelihood arises from its approximation to the actual grouped data likelihood.

From their point of view, the infinity problem results from a misspecification of the likelihood. As such, the problem could then be solved through a more principled construction of the likelihood function. It should be built as:

$$L(\boldsymbol{\theta}, \mathbf{y}) = \prod_{i=1}^{n} \left[F_M(y_i + \delta/2) - F_M(y_i - \delta/2) \right], \tag{3}$$

with $F_M(x) = \sum_{m=1}^{M} \pi_m F(x|\mu_m, \sigma_m)$, the cumulative distribution function (cdf) of the mixture, $F(x)$ the cdf of the mixture component and δ the grouping interval or the measurement instrument's precision with which y_i is measured. As a consequence, this likelihood is bounded between 0 and 1. Moreover, if a global maximum exists, it corresponds to a consistent MLE.

Although this approach seems reasonable, it suffers from some important drawbacks. Numerically, it can be demanding to manipulate a likelihood composed of differences of cumulative distributions instead of densities. Secondly, some authors state that the argument of discrete data does not necessarily get to the real issue. Whether or not it is possible in practice, it is still legitimate to suppose that the observations are intrinsically continuously distributed and that discreteness is the approximation (Cheng and Iles, 1987). Further, the original likelihood (in the continuous case) was composed of density contributions $f(x;\theta)$, derived from probability elements $\mathrm{P}(x \in dx) = f(x,\theta)dx$ (Cramér, 1946), obviating the need to be bounded above by 1. Also, despite the fact that the infinite spikes of the likelihood do not yield useful estimates, the infinity is not counter-intuitive. Indeed, if one of the variances goes to zero, the corresponding component of the mixture becomes discrete with a contribution to the joint distribution that will be "infinitely" greater then a continuous one (e.g., you cannot better fit a point than by assigning the entire mass to it). Thirdly, even if one considers (3) being the correct specification of the likelihood, how should one then choose the precision δ? In rare cases, this value is known as being the precision of the measurement system. But for most cases, the value of δ is unknown and one would be

unable to choose it without an unacceptably high amount of arbitrariness. In addition, the global maximum of the likelihood function does not always exist. This leads to a problem which is similar to the unbounded likelihood problem (Section 4).

Instead of concentrating on the MLE, Cheng and Amin (1983) and Ranneby (1984), propose a new kind of estimator based on the likelihood that does not suffer from the infinity problem. They introduce the maximum product of spacings (MPS) estimator, obtained by maximizing the following product of spacings instead of densities:

$$\mathrm{H}(\boldsymbol{\theta}; \mathbf{y}) = \prod_{i=1}^{n+1} \left[F_M(y_{(i)}) - F_M(y_{(i-1)}) \right], \tag{4}$$

with $y_{(i)}$, the ordered observations and $y_{(0)} = -\infty$, $y_{(n+1)} = +\infty$. Titterington (1985), however, discussed that this method is in essence a maximum likelihood method based on grouped data, hence exhibiting the exact same drawbacks discussed earlier.

2.2. Restriction of the parameter space

Another way of bounding the likelihood is by restricting the parameter space. The singularities of the likelihood are situated on the edge of the parameter space. Hence, by constraining the latter such that problematic points are excluded, a bounded likelihood over the restricted space can be obtained. However, one still has to prove the existence of a consistent MLE for this kind of restricted likelihood problems.

By interrelating the standard deviations of the different mixture components, one can prevent them to become zero. In this way, a constrained parameter space without singularities is obtained. Moreover, Quandt and Ramsey (1978), amongst others, noted the existence of a consistent MLE in case a relationship between the standard deviations of the true mixture component densities was known and incorporated as constraints.

One of the most popular forms of constraints, although it is often not recognized as such, is the imposition of equal standard deviations among the different components of the mixture. For this kind of mixtures the existence of a consistent MLE is well known. But, as noted earlier, although this assumption may be justified in some cases, for most cases it is certainly not a satisfactory solution.

Constraints of the form $\sigma_i = k_{ij}\sigma_j$, with k_{ij} known constants, are another possibility. Exact knowledge of the constants, however, is rare. Alternatively, inequality constraints can be imposed as is done by Hathaway (1985). He introduced the following inequalities:

$$\sigma_i \geq c\sigma_{i+1}, i = 1, \ldots, (M-1); \sigma_M \geq c\sigma_1, \text{with } c \in]0, 1]. \tag{5}$$

For this restricted likelihood problem, Hathaway proved the existence of a global maximum of the likelihood regardless of the value of c. Further, he showed the consistency of such a global maximum if the constrained parameter space contained the true parameter. In other words, a consistent MLE exists if the true parameter is in the restricted parameter space. Hathaway (1986) also adapted the EM-algorithm to incorporate restrictions.

Another way to restrict the parameter space and exclude singularities is to work directly with compact subsets. For these likelihood problems, Redner and Walker (1984) proved the existence of a consistent MLE for the normal mixture problem over any compact subspace containing the true parameter.

Although this kind of approach seems reasonable as well, it suffers from some important drawbacks. First, numerically it is more demanding to optimize a function over a constrained parameter space. Second, restrictions are sometimes too limiting. For example, imposing equality of the standard deviations is an easy way to proceed, but in a lot of cases it is implausible. An important question is how to choose a constrained parameter space such that it contains the real parameter, although the latter is unknown. Also the choice of the value c and the choice of the compact subset, without knowledge of the true parameter, is problematic. Finally, like the other methods aimed at removing the unboundedness of the likelihood, the real problem is essentially circumvented only.

3. An alternative: likelihood estimation

One of the main reasons for the popularity of the maximum likelihood method are the good statistical properties of the corresponding estimators in the sense that they are consistent, asymptotically efficient and normally distributed under suitable regularity conditions. MLEs are defined as the global maxima of the likelihood function. The latter, however, do not exist for the likelihood corresponding with the general M-component mixture. Nevertheless, the likelihood has local maxima. This provides a way to avoid the search for a global maximum. First, empirical evidence was found, for example by Quandt (1972) and Duda and Hart (1973), that a local maximum, more specifically the largest one, corresponds to reasonable parameter estimates. Later, Sundberg (1974), for incomplete data from an exponential family and Kiefer (1978), for a switching regression model and Lehmann (1983), for general situations, provided a solid basis for such an approach. They all proved the existence of a consistent sequence of roots of the likelihood equations for their particular problem. These likelihood equations (LEQs) are obtained by equating to 0 the partial derivatives of the logarithm of the likelihood function with respect to its parameters.

In what follows a review of this theory is given (§3.1), together with our perception of how it can be used to obtain parameter estimates with good statistical properties in case of normal mixtures without any restriction on the relationship between the parameters (§3.2).

3.1. Review

Cramér (1946) discussed the method of maximum likelihood for one-parameter distributions. Although he first defines the MLE as the value which renders the likelihood as large as possible, his final definition of an MLE is different from the classical one. Moreover, he states: "**Any** *solution of the likelihood equation will be called a maximum likelihood estimate of the unknown parameter*". With this definition in mind, he proves that under certain general conditions the likelihood equation has **a** (but not any) solution that converges in probability to the true parameter value as the sample size goes to infinity, hence is consistent. Further, this solution is also asymptotically efficient and normally distributed. In other words, Cramér proved the existence of a solution of the likelihood equation (LEQ) with good statistical properties and called it an MLE.

In 1948, Huzurbazar showed, under the same conditions as Cramér, that with probability going to one as the sample size goes to infinity, such a consistent root is unique and corresponds to a local maximum of the likelihood. Thus, if a density satisfies the conditions of Cramér, a local maximum of the likelihood possesses the required statistical properties. This result provides a useful alternative to the condition of global maxima only.

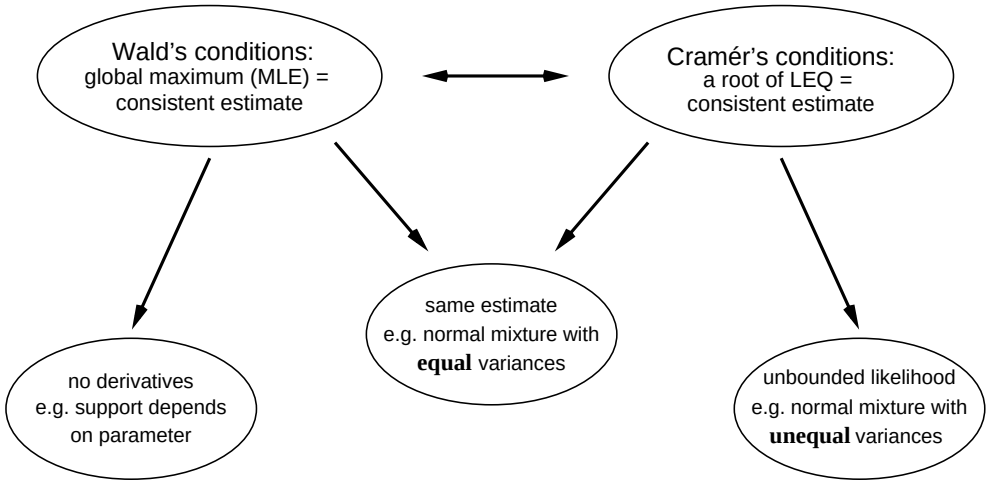


Fig. 1. Likelihood estimation

Wald (1949) gave a proof of consistency of the classical MLE, i.e., with the usual meaning attached to the global maximum of the likelihood. His proof was based on totally different and more demanding assumptions compared to Cramér's conditions. In essence, Wald does not use differentiability assumptions; even the LEQs do not have to exist. In addition Wald notes that Cramér is only proving the consistency of a local maximum, in contrast to his proof of the consistency of a global maximum.

Figure 1 gives an overview of these different approaches of likelihood estimation. On the one hand, there are the conditions of Wald ensuring the consistency of a global maximum (MLE). On the other hand, there are the conditions of Cramér guaranteeing the consistency of a local maximum. Depending on the assumptions a parametric family fulfills, we have the following three possibilities:

Both conditions hold. They lead to the same estimate. This is the case for a lot of 2-parameter distributions such as the normal, Weibull, gamma,..., but also for a mixture with equal variances for the components.

Only Wald's conditions hold. This is typical for distributions that do not have a derivative at some points in the parameter space. An example is when the support depends on some parameter, like the uniform distribution.

Only Cramér's conditions hold. Even if the global maximum exists, it does not have good statistical properties, but at least one local maximum does. This is often the case for distributions with singularities situated on the edge of the parameter space, such as for a general M-component mixture.

Importantly, this figure shows that whether we either work with a mixture with equal variances or with unequal variances, in essence the same kind of estimate is obtained from the likelihood equations, in spite of the convention of terminology to only call the first an MLE. The latter will be referred to as a likelihood estimate (LE).

Note that Cramér's results were for the one-parameter case (in contrast to Wald). Aitchison and Silvey (1958) generalize his results to the multi-parameter case, whereas Chanda

(1954) (proven by Tarone and Gruenhage, 1975) extends the uniqueness theorem of Huzurbazar. The conditions are straightforward extensions of the one-parameter case.

3.2. Multiple roots

In spite of these results, problems are not entirely solved for the general mixture model. As is known from regular ML estimation, the LEQs for a number of models suffer from a multiple root problem. The same is true for a mixture model. These multiple roots lead to an additional problem since Cramér's theory only states the existence of a consistent root of the LEQ. No results are available on which root to specify.

Basically, for mixtures, there are two types of roots. In the first place, there are multiple roots caused by the non-identifiability of the parameters in the model. Indeed, although the family of normal (finite) mixtures (equal or unequal variance case) is identifiable (Teicher, 1963), the parameters are not due to the arbitrariness of the numbering of components of the mixture. Moreover each permutation of the component labels provides another root, resulting in at least $M!$ roots for the likelihood equations. This problem, however, is not of great concern and can be avoided, for example, by ordering the sizes of the different means or by introducing an equivalence relation in the sample space making the true parameter identifiable relative to its equivalence class. On the other hand, a second class of roots is of more concern. Day (1969) stated that any pair, triplet,... of distinct observations sufficiently close together, would generate a local maximum of the likelihood, resulting in several roots for the likelihood equations. But his comment that therefore ML estimation breaks down is not warranted as observed previously. These roots are fundamentally different from each other.

According to the theory, the LEQs contain a "unique consistent" root. Note that unique here refers to a unique equivalence class. Also, there is a certain ambiguity in this uniqueness statement (Perlman, 1983). In case the conditions of Wald are satisfied, the global maximum or MLE is known to be consistent. So, there is a criterion based on the value of the likelihood to discriminate between several roots. But, if only Cramér's conditions hold, which is the case for normal mixtures with unequal variances, this criterion cannot be applied. Nevertheless, for these mixtures, the root corresponding to the largest (finite) local maximum of the likelihood is consistent. This can be shown in several ways using results described in §2.2. Indeed, both the propositions of Hathaway (1985) and Redner and Walker (1984) on the consistency of the global maximum in a constrained parameter space, imply the consistency of the largest local maximum. Consequently, also for normal mixtures with unequal variances, there exists a criterion based on the likelihood value to choose a consistent root. It is the same kind of criterion as for ML estimation. In the following, the likelihood estimate (LE) will refer to this root. Note that this result can be generalized to other types of mixtures satisfying both Cramér's conditions over the entire parameter space and Wald's conditions over any compact subspace.

4. The problem of spurious maxima

So far we have discussed that under the conditions of Cramér the LEQs contain a consistent root, which is unique, has exactly the same properties as an MLE, and called an LE, perhaps at the expense of a slight efficiency loss. Further, for many general finite mixtures, including finite normal mixtures, this solution corresponds to the largest local maximum of the likelihood. Nevertheless, McLachlan and Peel (2000) argue, amongst others, that

Table 1. Some local maxima of the likelihood function of the simulated sample from McLachlan and Peel (2000), with maxima in bold obtained by them. The first 3 maxima are the largest local maxima. The last row gives the MLE obtained from estimating a normal distribution.

maximum	μ_1	σ_1	μ_2	σ_2	π_1	Log Likelihood
1 (LE)	-0.830	0.000400	1.062	1.329	0.0198	-163.886
2	2.517	0.000650	0.995	1.338	0.0196	-165.528
3	**-2.161**	**0.00850**	**1.088**	**1.277**	**0.0196**	**-165.937**
4	0.908	1.405	1.712	0.467	0.855	-170.248
5	0.961	1.393	1.622	0.281	0.904	-170.254
6	**-0.701**	**0.948**	**1.383**	**1.114**	**0.172**	**-170.558**
MLE	1.025	1.342				-171.294

this largest local likelihood criterion cannot be followed since a so-called spurious maximum can be chosen as LE. In particular, it was noted, when estimating a general finite normal mixture, that for some samples the largest local maximum of the likelihood could correspond to a maximum with implausible values for the parameters. Note that the presence of "implausible" maxima in the likelihood function for these mixtures was already observed by Day (1969).

To make more clear what is meant with a spurious maximum, we look at an example given by McLachlan and Peel (2000). They generated a sample of size 100 from a 2-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$, $\mu_2 = 2$ and $\pi_1 = 0.5$. A normal QQ-plot of this sample is shown in Figure 2a. Two maxima of the likelihood function were located. In Table 1, which gives the parameter values of several maxima of this likelihood, they are referred to as maximum 3 and 6, respectively. Clearly, of these two, maximum 3 has the largest likelihood value. But, it also has a value for π_1 which is about $2/100$ and a very small value for the shape parameter σ_1. Moreover, the first component of the mixture (corresponding to maximum 3) is related to a subgroup of only 2 successive data points of the ordered sample. As such, it is highly unlikely that one would consider that maximum 3 reflects the "truth". This maximum is related to a pure random cluster of data points in the sample and therefore it is called spurious. Further, the other maximum found (i.e., maximum 6 in Table 1) was considered to be the LE, due to the fact that its parameter values are much more plausible. Also, when the sample was binned into 7 intervals of equal width, apparently the parameter values of the MLE then obtained are close to the parameter values of maximum 6, confirming their conclusion that maximum 6 was the LE. Hereby, binning the sample was regarded as a procedure to remove spurious maxima since the occurrence of these maxima in the likelihood function was attributed to the continuous nature of the data.

We also scanned the whole parameter space for solutions of the LEQs. A lot more maxima than indicated by McLachlan and Peel (2000), are found. Note that it is out of the scope of this paper to show how these maxima can be obtained. We developed some methods based on the specific nature of a finite mixture model to scan the parameter space in a well-reasoned way for solutions of the LEQs. These are explained in Andries *et al.* (2004), introducing a starting value method for the finite (log)normal mixture model. Some of the maxima found are given in Table 1. The first 3 maxima are the largest local maxima of the likelihood function, the last 3 are the only maxima which have "plausible" parameter values. Between maximum 3 and 4 more than 20 other maxima are situated.

Clearly, the largest local maximum was not obtained by McLachlan and Peel (2000) and

Table 2. Local maxima of the likelihood function for several binned samples from the simulated sample of McLachlan and Peel (2000). The symbol $\to 0$ indicates that the maximum would be attained in $\sigma = 0$. The second column refers to the label of the maxima in Table 1.

# classes	max	μ_1	σ_1	μ_2	σ_2	π_1	Log Likelihood
80	3	-2.199	0.0609	1.116	1.245	0.0274	-418.529
	1	-0.515	0.00273	1.086	1.333	0.0384	-419.108
	5	0.996	1.362	1.736	0.00770	0.959	-420.683
	4	0.959	1.398	1.614	0.272	0.898	-421.199
	6	-0.790	0.933	1.339	1.138	0.147	-421.726
50	3	-2.198	$\to 0$	1.117	1.248	0.0276	-371.523
	5	0.992	1.368	1.699	$\to 0$	0.953	-374.233
	4	0.985	1.375	1.672	0.0974	0.941	-374.244
	6	-0.827	0.894	1.345	1.136	0.146	-374.834
20	3	-2.161	$\to 0$	1.110	1.254	0.0270	-281.169
	4-5	0.946	1.408	1.638	0.205	0.893	-282.770
	6	-0.814	0.906	1.369	1.118	0.159	-283.671
7	3	-2.216	$\to 0$	1.145	1.224	0.0293	-177.786
	4-5	0.875	1.369	2.0748120	0.0793	0.836	-177.787
	6	-0.507	0.927	1.541	1.011	0.239	-178.105

the likelihood function contains many more maxima than two. However, maximum 1 and 3 are similar in nature: both are truly spurious. As such, there is still the problem that the LE is not believable or reflecting the truth. But, as noted from Table 1, there is also no a priori reason to take maximum 6 as the LE. Why not choosing maximum 4 or 5? Indeed, both have a larger likelihood value and their parameter values also seem plausible. The only motivation for choosing maximum 6 as LE, is that it is the maximum *closest* to the true values, with closest defined by some distance measure. However in real examples, one does not know the true values, which underscores that there are no good grounds to choose maximum 6 as the LE.

The argument that spurious maxima are due to the continuous nature of data is not warranted either. Indeed, binning the sample into a number m of intervals with equal width or introducing a measurement error δ for the data (Section 2.1), will not solve the problem of spurious and multiple maxima of the likelihood function. The appearance of multiple maxima of the likelihood function is due to the specific nature of a general mixture model. In particular, it models clusters within a sample, whether these clusters are real or random. Of course, the number of maxima of the likelihood, and so also of spurious maxima, found will decrease when m becomes smaller (or δ becomes larger), since clusters of the sample with a small within variation will be "smoothed out". But, how far can we decrease m (or increase δ) without smoothing out the "real" subdivision of the sample?

As an example, we binned the sample shown in Figure 2a into 80, 50, 20 and 7 classes of equal width. In Table 2, some maxima of the likelihood functions for each binned sample are given. From this table it is not only clear that spurious maxima do not necessarily disappear when binning, but also that for this kind of samples apparently there are some problems with the MLE too. Namely, for the samples binned into 50, 20 or 7 classes, the supremum of the likelihood function is never attained within the parameter space. The largest value of the likelihood would be obtained for $\sigma_1 = 0$ or $\sigma_2 = 0$. This value for σ, however, does not belong to the parameter space. Consequently, a global maximum of the likelihood function does not exist, and so also the MLE. Note that this is similar to the case

of the unbounded likelihood for the unbinned sample (which also corresponded to a value of 0 for one of the σ parameters). Also here another local maximum has the same statistical properties as the MLE. Further, binning the sample does not solve the problem of which maximum to choose as the LE. As noted from Table 2, more than 1 plausible maximum is present in the likelihood of the binned sample. Again, there is no reason to choose the maximum related to maximum 6 as MLE. Clearly, although binning can be useful in a sensitivity analysis (with respect to the results of the unbinned sample), it cannot be used to pick out the real cluster in a sample.

To summarize, estimation procedures that pick out a maximum of the likelihood function with "plausible" parameter values or look for a maximum with parameter values that are close to the parameter values of the MLE obtained from a binned sample, are subjective, will not lead to a consistent sequence of estimators and make the inference results unreliable. As such, we do not recommend them. Nevertheless we cannot neglect that there is sometimes a problem with the LE, in the sense that it does not reflect the true parameter values. In the following we will try to clarify the situation. By means of some examples, we will first draw a picture of the global problem. This is followed by a discussion and an attempt to charactize the problem. To end, some guidelines are given of how to deal with spurious maxima in practice.

4.1. Examples

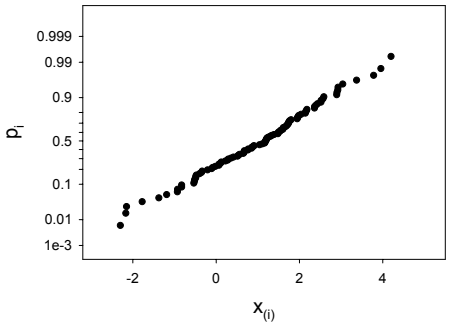
4.1.1. Highly unstable samples

Consider again the sample shown in Figure 2a and discussed in detail previously. As mentioned before, the problem for this sample was that the parameter values of the LE were implausible or not reflecting the truth. The reason for this spurious LE is the lack of information available within the sample in order to fit a 2-component mixture. In other words, although this particular finite mixture model (i.e., the mixture with parameter values $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = \sigma_2$ and $\pi_1 = 0.5$) is theoretically identifiable (Teicher, 1963), numerically for this sample it is not. This can be observed in several ways:

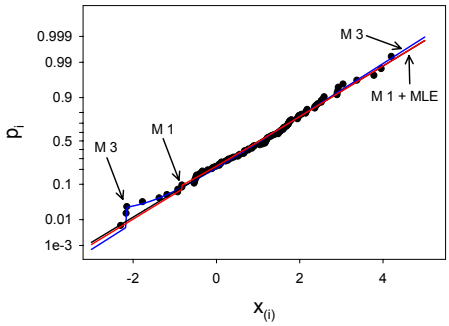
- A single normal distribution can be used to model the sample satisfactorily. A test of normality does not reject the null hypothesis for $\alpha = 0.005$. The correlation between the sample quantiles $x_{(i)}$ and $\Phi^{-1}[(i-0.375)/(n+0.25)]$ is 0.9957. Moreover, without prior knowledge that this sample is simulated from a 2-component mixture, no one would fit a 2-component mixture to this sample.

- In Figure 2b the fit of the MLE obtained from a single normal distribution and the fits of the two spurious maxima 1 and 3 are shown. Apart from a small deviation in a small number of data points, the fits can hardly be distinguished. While one of the two components of the mixtures (corresponding to these spurious maxima) fits exactly 2 or 3 data points, the other component, for which the parameter values of the scale and shape parameter resembles the parameter values of the MLE, has to fit the rest of the sample. Figure 2c shows also the fit of the MLE of a single normal distribution, but now with the fits of the three plausible maxima given in Table 1. Again, the distinction between all 4 fits is minimal, certainly within the range of the data.

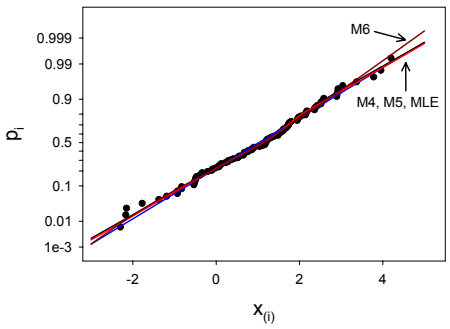
- Figure 2d depicts (on normal distribution probability scales) the cumulative distribution function (cdf) of the true 2-component normal mixture with the cdf of a normal



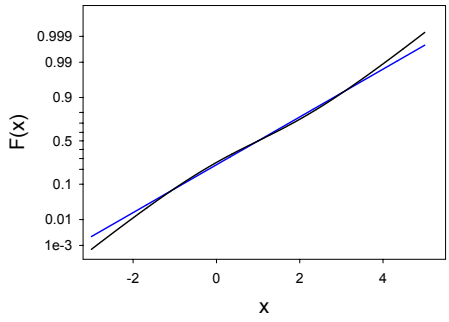
(a) Normal QQ-plot of the sample.



(b) M1-M3: fit of the 1^{st} and 3^{th} maximum of Table 1, MLE: fit of a normal distribution.

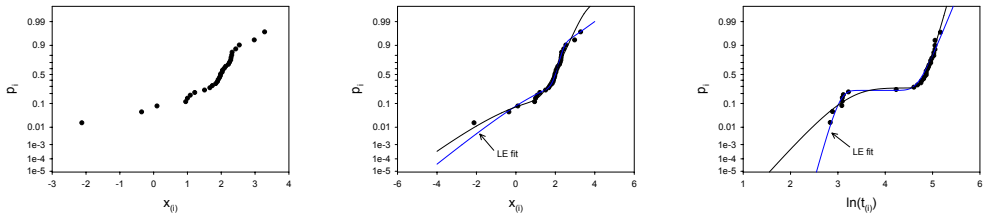


(c) MLE: fit of a normal distribution, M4-M5-M6: fit of the $4^{th}, 5^{th}$ and 6^{th} maximum of Table 1.



(d) Cdf of the true mixture and the cdf of a single normal distribution with the same mean and standard deviation of the mixture.

Fig. 2. Simulated sample of size 100 from McLachlan and Peel (2000). The true parameter values are $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$, $\mu_2 = 2$ and $\pi_1 = 0.5$.



(a) Normal QQ-plot of a simulated sample of size 30 from a 2-component normal mixture. True parameter values are $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$ and $\pi_1 = 0.2$.

(b) LE fit of the sample of Figure 3a and fit of the second largest maximum of the likelihood function (maximum 2 in Table 3).

(c) Lognormal QQ-plot of a real failure time sample of size 29 with the LE-fit and fit of the 2^{nd} largest maximum of the likelihood function.

Fig. 3. A simulated and a real unstable sample.

distribution with the same mean and standard deviation of the mixture. As can be observed, except for the extreme tail ends, these two distributions can hardly be distinguished. As a result, to identify a sample as coming from this mixture distribution, the sample size has to be huge (i.e., over the 1000).

Thus, unless the sample size is unduly large, a single normal distribution can equally well be used to fit a sample generated from this particular 2-component normal mixture. As a result, solutions of the LEQs for which one of the two components of the mixture, fits exactly 2 or 3 data points, will correspond to maxima at the top of the likelihood function, i.e., maxima with a large likelihood value.

This sample is a typical example of what we define as a *highly unstable* sample with respect to the 2-component normal mixture. This means that the largest local maximum of the likelihood function can be altered through some minor perturbations in the sample, that there are several maxima of the likelihood function with about the same large likelihood value (Table 1) and that there is no maximum which dominates the likelihood function.

4.1.2. Unstable samples

Figure 3a shows the normal QQ-plot of a sample of size 30, simulated from a 2-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$ and $\pi_1 = 0.2$. The 5 largest maxima of the likelihood function are given in Table 3. At first sight, there seems to be no problem. The parameter values of the LE are credible and its fit is acceptable (Figure 3b). Nevertheless, the parameter values of the LE are not at all in the neighborhood of the true values, while those of the 2^{nd} largest maximum are *closest* to the true values. This means that the LE is also spurious, i.e., its parameter values do not reflect the truth, in spite of the fact that it is not possible to derive it from the parameter values itself.

Table 3. The 5 largest local maxima of the likelihood of the simulated sample of size 30 shown in Figure 3a. The last row shows the MLE obtained when estimating a normal distribution.

maximum	μ_1	σ_1	μ_2	σ_2	π_1	Log Likelihood
1 (LE)	1.219	1.394	2.088	0.224	0.428	-32.941
2	0.171	1.457	2.028	0.498	0.168	-34.574
3	1.615	1.078	2.284	0.0323	0.849	-38.791
4	1.976	0.00367	1.699	1.053	0.0615	-40.060
5	2.316	0.00484	1.677	1.043	0.0607	-40.308
MLE	1.716	1.022				-43.228

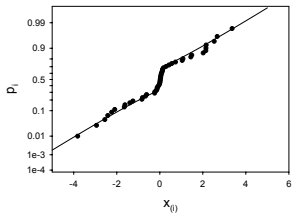
As a result, the problem is the same as in Section 4.1.1, only less pronounced. The reason why the LE is spurious, is also the same: the sample size n is too small to distinguish the true distribution. Moreover, the true 2-component normal mixture can numerically not be identified. But, while in the previous section, the mixture could not be distinguished from a single normal distribution, here it is clear from the QQ-plot in Figure 3a that a normal distribution would not be appropriate, i.e., a straight line will not fit the sample satisfactorily. However, there are two different 2-component mixtures which can be hardly distinguished within the range of data. Outside this range, differences become marked. Consequently, conclusions drawn will depend highly on which of the mixtures (i.e., which of the two maxima at the top of the likelihood) is chosen. For example, the null hypothesis of equal shape parameters versus the alternative of unequal shape parameters would be rejected with the likelihood ratio test (lrt) if the first maximum was taken (LRT-value = 6.461), but accepted if the second maximum was considered (LRT-value = 3.195) on a 95% level. Also the difference in estimation of the low quantiles could influence the decision whether a physical component is accepted as reliable or not.

This sample is an example of an *unstable* sample, i.e., a few maxima, mostly with plausible parameter values, are at the top of the likelihood function (Table 3). Small perturbations in the sample can alter the largest maximum of the likelihood into one of the other maxima at the top of the likelihood.

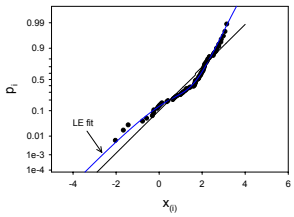
To end this section, Figure 3c shows that the situation discussed here, does occur in real terms. A lognormal QQ-plot of a failure time sample of size 29, obtained from an experiment carried out at the Institute for Materials Research (IMO), is given. The LE-fit of a 2-component lognormal mixture is shown, together with the fit of the 2^{nd} largest maximum. As noted, the difference, between the two mixtures, with respect to the estimation of low quantiles, will be large.

4.1.3. Stable samples

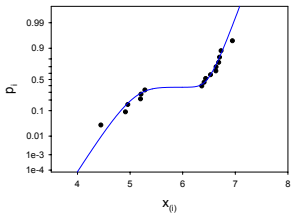
Figure 4a depicts the normal QQ-plot of a simulated sample of size 50 from a 2-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 0.1$, $\mu_2 = 0$, $\sigma_2 = 2$ and $\pi_1 = 0.5$. Table 4 gives the 4 largest local maxima of the likelihood function. The LE has plausible parameter values and does reflects the true values. It is also the maximum closest to these true values. Most other maxima found have implausible values for the parameters. Apparently, the sample is large enough to distinguish one specific 2-component mixture. Moreover, this sample is an example of a *stable* sample, i.e., the likelihood function is dominated by one maximum, other maxima are pushed into the background. Small perturbations in the sample will not alter the largest local maximum of the likelihood function. Further, the



(a) Normal QQ-plot of a simulated sample of size 50 with LE fit. The true parameter values are $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$ and $\pi_1 = 0.2$.



(b) Normal QQ-plot of a simulated sample of size 80 with LE fit and fit of the 2^{nd} largest maximum. True parameter values are $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$ and $\pi_1 = 0.4$.



(c) Lognormal QQ-plot of a real failure time sample of size 16 with the LE-fit.

Fig. 4. Simulated and real stable samples.

Table 4. The 4 largest local maxima of the likelihood of the simulated sample of size 50 shown in Figure 4a.

maximum	μ_1	σ_1	μ_2	σ_2	π_1	Log Likelihood
1 (LE)	0.0100	0.0982	-0.0353	1.793	0.338	-73.709
2	0.0379	0.000316	-0.0224	1.490	0.0397	-82.350
3	2.147	0.00576	-0.154	1.398	0.0582	-82.426
4	0.0189	0.00170	-0.0215	1.4887517	0.0382	-85.627

difference between the first and the second maximum is large.

Another example of a stable sample is shown in Figure 4b. It is a sample of size 80, simulated from a 2-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, $\sigma_2 = 0.5$ and $\pi_1 = 0.4$. The 4 largest local maxima of the likelihood function are tabulated in Table 5. Here, the largest local maximum is not so dominant as for the previous sample, but it is resistant to small perturbations. Further, the difference between the 1^{st} and 3^{th} maximum is considerable and the 2^{nd} maximum has implausible parameter values with a fit that differs a lot from the LE fit (Figure 4b). No other maximum with reasonable parameter values is found at the top of the likelihood function. Also here, the LE is the maximum closest to the true values.

Obviously, for these two examples, the LE can be trusted. Although many more maxima with mostly implausible parameter values are present in the likelihood function, none of them bother. The samples contain enough information, i.e., the sample size is large enough, to numerically distinguish the underlying distribution. To conclude, Figure 4c gives a lognormal QQ-plot of a real failure time sample together with the LE fit of a 2-component lognormal distribution. It is an example of stable sample encountered in practice.

Table 5. The 4 largest local maxima of the likelihood of the simulated sample of size 50 shown in Figure 4b.

maximum	μ_1	σ_1	μ_2	σ_2	π_1	Log Likelihood
1 (LE)	0.360	1.075	2.068	0.511	0.367	-111.619
2	1.681	3.328e-005	1.436	1.139	0.0250	-112.388
3	1.681	0.00267	1.431	1.151	0.0458	-116.285
4	1.680	0.00139	1.433	1.145	0.0356	-116.855

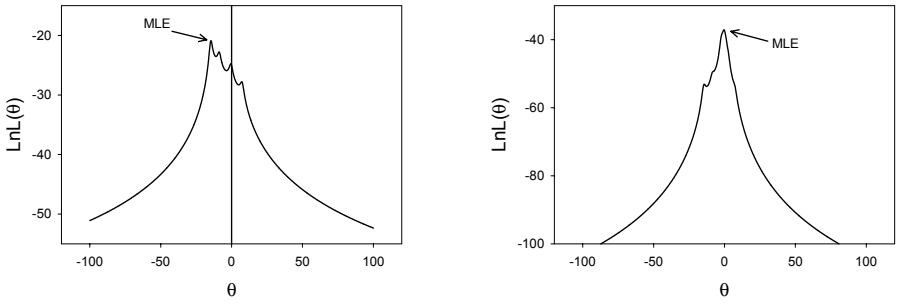
4.2. Discussion

Previous examples made clear the problem, touched upon already by some, but never treated in detail. Namely, when estimating a general 2-component normal mixture to a sample, for certain samples the LE does not reflect the true parameter values. In other words, the LE is unreliable. Sometimes, this is clear from the parameter values of the LE itself, but equally well it may not be. Mostly, for these samples, a 2-component mixture was numerically not identifiable. Moreover, we characterized them as being *unstable* with respect to a 2-component mixture. As mentioned yet, the reason for this non-identifiability or unreliable LE, is a too small sample size (if the true model is a 2-component mixture). The cause is twofold. First, there is the fact that the "consistency" property of the LE is an asymptotic concept. This means that only for a sufficiently large sample size the estimators looked at will approach the true parameter values. For small sample sizes, on the contrary, nothing is known about the performance of a consistent estimator. Note that for small sample sizes, Hosmer (1973) already indicated that the MLE could be unreliable in case of normal mixtures. Second, there would be no problem if the LEQs had only one root. However, due to the nature of the mixture model itself, the LEQs contain many roots.

Having said this, the problem of an unreliable LE is not related to the case of likelihood estimation only or to the mixture model only. It is inherent to all consistent estimators obtained as a solution of the LEQs derived for a certain distribution. As such, it can just as well happen in case of classical ML estimation. A simple example to demonstrate this is the case of the one-parameter Cauchy location distribution (Barnett, 1966; Reeds, 1985). Its density function is given by:

$$f(x) = \frac{1}{\pi[1 + (x - \theta)^2]} \quad (-\infty < x, \theta < \infty), \tag{6}$$

with θ a location parameter. This parametric family fulfills both the conditions of Cramér and Wald (Perlman, 1983). Therefore, the MLE exists, is consistent and can be found as a root of the LEQ. This equation, however, has usually more than 1 root or the likelihood function has more than 1 maximum. Here, the presence of multiple maxima is related to the absence of finite moments for the Cauchy location distribution. In particular, Reeds (1985) showed that anomalous local maxima are related to outlying values of the sample which arrive frequently due to the heavy tails of the Cauchy distribution. Similar to the case of the mixture model, it is not possible to distinguish an anomalous root (i.e., a spurious root) from a proper one in case the sample size is too small. As an example, Figure 5a depicts the logarithm of the likelihood function of a sample of size 5, generated from the Cauchy location distribution with location parameter $\theta = 0$. As noted, the likelihood function has 4 maxima. The MLE corresponds to an anomalous root, since its parameter value is quite far from 0 and it is the maximum farthest from the true value. According to the definition introduced in Section 4.1.1, this is an unstable sample (with respect to the Cauchy location distribution). Indeed, leaving out only one data point, will easily switch the global



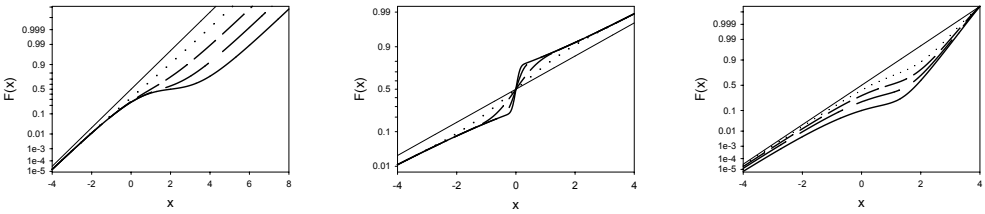
(a) Simulated sample of size 5.

(b) Simulated sample of size 9.

Fig. 5. Logarithm of the likelihood function of simulated samples from the Cauchy location distribution with true parameter value $\theta = 0$.

maximum of the likelihood function into one of the other 3 maxima. If the sample size of this sample is increased to 9, however, the sample becomes stable as shown in Figure 5b. One maximum, i.e., the one closest to the true value, dominates the likelihood function. Clearly, the same behavior is observed as for the examples discussed in Section 4.1. Namely, for small sample sizes, the MLE cannot be trusted, while for large samples the MLE is reliable. Importantly, the value of "small" and "large" depends highly on the distribution used. For the one-parameter Cauchy distribution, a small sample size means a value not larger than about 6, while a large sample size is from about 10 onwards. As such, in practice for this distribution there will be no problems with the MLE, since usually the sample size will be larger than 10. For the finite general mixture model, on the contrary, for some mixtures a size of 50 will be large enough, while for others 1000 or even 10000 will not be sufficient (Section 4.3). Consequently, the credibility of the LE is an important issue there.

Further it is interesting to observe a similarity between the surfaces of the likelihood for small and large sample sizes in case the LEQs have multiple roots and in case they only have one root. On the one hand, for a small sample size, the likelihood function will have a (very) flat curvature in case the LEQs have a unique root. The flatness of the surface of the likelihood in case the LEQs have multiple roots, is expressed through several maxima which are at the top of the likelihood. One could think of a bumpy surface (Figure 5a). On the other hand, the likelihood function will have a sharp curvature for a large sample size in the one root case, while for the multiple root case the sharpness of the likelihood is expressed through one dominating maximum (Figure 5b). This means that, while in the one root case, a small sample size can be noticed through the large value of the standard errors (which are in relation to the curvature of the likelihood function), this is not entirely true for the multiple root case. There, it is important, as will be discussed in the following section, to not only focus on the largest maximum, but to obtain an overall view of the surface of the likelihood function. This is the only way to obtain information about the credibility of the LE.



(a) Group 1: The dotted line corresponds to $\mu_2 = 1$, the short dashed to $\mu_2 = 2$, the long dashed to $\mu_2 = 3$, the solid to $\mu_2 = 4$ and the straight line to a normal distribution with $\mu = 0, \sigma = 1$.

(b) Group 2: The dotted line corresponds to $\sigma_1 = 1$, the short dashed to $\sigma_1 = 0.5$, the long dashed to $\sigma_1 = 0.2$, the solid to $\sigma_1 = 0.1$ and the straight line to an normal distribution with $\mu = 0, \sigma = 2$.

(c) Group 3: The dotted line corresponds to $\pi_1 = 0.8$, the short dashed to $\pi_1 = 0.6$, the long dashed to $\pi_1 = 0.4$, the solid to $\pi_1 = 0.2$ and the straight line to a normal distribution with $\mu = 0, \sigma = 1$.

Fig. 6. Cumulative distribution functions on a normal probability scale for the three groups of parameter values.

4.3. Guidelines

Apparently, for small sample sizes, the property of consistency for a likelihood estimator when multiple roots are present in the LEQs, is not enough to guarantee that the estimator is meaningful. As suggested several times in previous sections, a spurious maximum can, on a purely theoretical basis, be defined as any maximum not *closest* to the true values, with *closest* defined by some distance measure. As such, for each sample, there is only one proper maximum. Importantly, for some sample size n on, this maximum will be equal to the LE or MLE due to their consistency property. However, the sample size required such that this holds, depends highly on the mixture used. To demonstrate the dependency between the specific mixture and the sample size needed and to obtain an idea about how large this sample size has to be, we carried out a small simulation study.

Samples are generated from a 2-component normal mixture model with 12 different sets of parameter values divided into 3 groups of 4. In each group, only one parameter is varied in order to study one aspect of the identifiability of the mixture, i.e., how well its two component distributions can be identified from the mixture or how much the plot of its cdf, when placed on a normal probability scale, deviates from a straight line. This is related to mainly three aspects of the mixture: the difference in scale parameter of the two component distributions, the size of the ratio of the two shape parameters and to a lesser degree the size of the proportion parameter. In the first group, the scale parameter of the second component, μ_2, is varied. It takes the values $1, 2, 3$, and 4. The values of the other parameters are $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$ and $\pi_1 = 0.5$. As noted from Figure 6a, the larger the value of μ_2 is, the better the component distributions can be identified from the mixture distribution (or the more the plot of the cdf of the mixture deviates from a straight line). For the second group, all parameters, except the shape parameter of the first component,

Table 6. The number of times out of 1000 (k) that the largest local maximum of the likelihood is a spurious maximum for the first group of parameter values.

Sample size n	k (first group)			
	$\mu_2 = 1$	$\mu_2 = 2$	$\mu_2 = 3$	$\mu_2 = 4$
20	951	933	842	589
50	988	971	794	279
100	996	972	587	35
200	997	970	213	0
300	998	964	51	0
400	997	942	12	0
500	999	888	2	0
1000	1000	642	0	0

Table 7. The number of times out of 1000 (k) that the largest local maximum of the likelihood is a spurious maximum for the second group of parameter values.

Sample size n	k (second group)			
	$\sigma_1 = 1$	$\sigma_1 = 0.5$	$\sigma_1 = 0.2$	$\sigma_1 = 0.1$
20	943	845	435	208
50	964	622	60	10
100	970	227	1	0
200	910	14	0	0
300	812	0	0	0
400	744	0	0	0
500	616	0	0	0
1000	151	0	0	0

are kept fixed. The values for the parameters here are $\mu_1 = \mu_2 = 0$, $\sigma_2 = 2$, $\pi_1 = 0.5$ and $\sigma_1 = 0.1,\ 0.2,\ 0.5,\ 1$. In spite of the equality of the component means, the components of the mixture can still be clearly identified if the ratio of the two shape parameters deviates sufficiently from 1. The more it deviates from 1, the better the mixture can be identified (Figure 6b). In the last group, the proportion parameter is altered from a small value (0.2) over two average values (0.4 and 0.6) to a large value (0.8). The values for the other parameters are $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$ and $\sigma_2 = 0.5$. It is clear form Figure 6c that also the value of π_1 has an influence on the identifiability of the mixture. The latter improves for smaller values of π_1. The reverse would be true if $\sigma_1 < \sigma_2$.

For each set of parameter values, sample sizes of $20, 50, 100, 200, 300, 400, 500$ and 1000 were used, with 1000 simulations in each case. Results are summarized in Tables 6, 7 and 8. The possible spurious nature of the LE was assessed through a comparison with the maximum closest to the true values. Here, *closest* is defined by the Euclidean distance, but with the shape and proportion parameters rescaled such that their domain is the same as for the scale parameters. The tabulated value k is then the number of times out of 1000 that the LE is spurious. The value of k should go to 0 when n increases.

Clearly, for all sets of parameter values, except for one set shown in the first column of Table 6, the value of k shows finally a decreasing trend. The dependency between the identifiability of the mixture and the sample size required such that the LE is the maximum closest to the true values, is evident from the tables. Moreover, the value of k goes relatively

Table 8. The number of times out of 1000 (k) that the largest local maximum of the likelihood is a spurious maximum for the third group of parameter values.

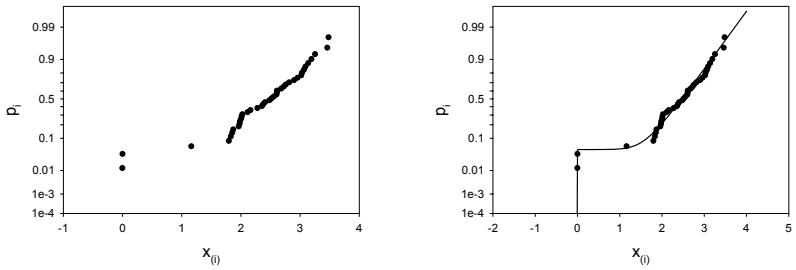
Sample size n	k (third group)			
	$\pi_1 = 0.8$	$\pi_1 = 0.6$	$\pi_1 = 0.4$	$\pi_1 = 0.2$
20	906	803	691	643
50	927	712	408	287
100	901	475	101	64
200	787	100	25	28
300	637	23	13	15
400	428	10	9	17
500	271	3	5	10
1000	9	0	3	4

fast to 0 for mixtures that can be clearly identified (i.e., the last one or two columns in each table). For some sets of parameter values a sample size of 100 or lower would be sufficient, while for others a sample size of 200 is required. But, for some poorly identifiable mixtures, although k shows at last a decreasing trend, 0 is not reached for even a sample size of 1000. The worst case is the mixture with parameter values $\mu_1 = 0$, $\mu_2 = 1$, $\sigma_1 = \sigma_2 = 1$ and $\pi_1 = 0.5$ (1^{st} column in Table 6), where k does not show at all a decreasing trend before a sample size of 1000. For the given sample sizes, it even gets worse as n increases. For example, for $n = 1000$, in none of the generated samples, the LE was equal to the maximum closest to the true values. The reason is clear: this specific mixture can be hardly distinguished from a single normal distribution. As seen in Figure 6a, the cdf of this mixture is practically a straight line. It is doubtful that any sample of this particular mixture distribution will be ever identified as coming from a mixture.

In summary, from some sample size onwards, the LE will be a good estimator. But the sample size required depends highly on how well the component distributions of the true mixture can be identified. For some mixtures, a very small sample size will be sufficient, but for others even a huge sample size will not do.

Although in theory the definition of a spurious maximum sounds nice, in practice there is one big problem: the "truth" is not known. It is not possible to search for the maximum closest to the true values. It is possible, however, to search for the LE. As shown, if the sample size is large enough, the LE will be the maximum closest to the true values. In other words, it will not be spurious. But, the sample size required is not known either. Fortunately, the stability of a sample gives an excellent idea whether the sample size is large enough, i.e., whether the LE can be trusted. We derived some easy to use but important guidelines. They are based on the fact that not only the LE has to be looked at, but also other maxima of the likelihood function.

- The sample is *highly unstable*, i.e., many maxima from which a lot have implausible parameter values are at the top of the likelihood function (Section 4.1.1). If the true distribution is a 2-component mixture distribution the sample size is far too small to detect this mixture. One can select from several options, apart from proceeding with the LE: look for prior information (for example, physical background), increase sample size or use a simpler model. For example, for the sample used in Section 4.1.1, a normal distribution would equally well fit this sample. Moreover, an increase of sample size would not help in this case, unless it would be huge.



(a) Normal QQ-plot.

(b) LE fit of a 2-component mixture.

Fig. 7. Simulated sample of size 40 from a 2-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 0.005$, $\mu_2 = 2.5$, $\sigma_2 = 0.5$ and $\pi_1 = 0.06$

- The sample is *unstable*, i.e., a few maxima which have mostly credible parameter values, are dominating the likelihood function (Section 4.1.2). Generally, if the true distribution is a 2-component mixture distribution, the sample size is somewhat too small to distinguish between several 2-component mixtures. Often, a worst-case scenario can be used: based on the few maxima dominating the likelihood function, several analyses are carried out. The one with worst results (with respect to what is asked) is taken. Again prior information or an increase of the sample size could help. For example, for the real sample shown in Section 4.1.2, information of other experiments led to the 2^{nd} maximum (and not the LE) as the proper one.
- The sample is *stable*, i.e., one maximum dominates the likelihood function or the largest maximum is followed by maxima which have only implausible parameter values (Section 4.1.3). There is nothing in such a case, suggesting that the LE cannot be trusted.

The stability of a sample (with respect to any model) tells a lot about the credibility of the LE or MLE. Obviously, the likelihood function will have to be scanned for local maxima in a well-reasoned way. One way to deal with that in case of 2-component (log)normal, Weibull or smallest extreme value mixtures, is explained in Andries *et al.* (2004), a paper constructing starting values for these mixtures. Further, an extension of these guidelines to mixtures with more than 2 components or other component distributions is evident.

To end this discussion, note that not all maxima with a very small value for the proportion parameter are spurious. This occurs, for example, in case the sample has a small group of outliers. In this situation, the guidelines represented above, do hold equally well. As an example, consider the sample shown on a normal QQ-plot in Figure 7a. This sample of size 40 is simulated from a 2-component normal mixture with parameter values $\mu_1 = 0$, $\sigma_1 = 0.005$, $\mu_2 = 2.5$, $\sigma_2 = 0.5$ and $\pi_1 = 0.06$. As observed, the sample has a subgroup of two outlying data points. Table 9 gives the 3 largest local maxima of the likelihood function. Based on the difference in likelihood value between the first and the second maximum, this

Table 9. The 3 largest local maxima of the likelihood of the simulated sample of size 40 shown in Figure 7.

maximum	μ_1	σ_1	μ_2	σ_2	π_1	Log Likelihood
LE	-0.00257	0.00143	2.502	0.530	0.0450	-27.452
2	2.608	0.000151	2.365	0.769	0.0497	-37.104
3	2.606	0.00355	2.361	0.776	0.0677	-40.979

sample is stable. So, in spite of the small value of the proportion parameter, the LE can be trusted, and indeed it reflects the true parameter values.

Acknowledgments

This work was supported by the Flemish Science Foundation (IWT).

References

Andries E., Croes K., De Schepper L. and Molenberghs G. (2004) Likelihood estimation of a mixture of two (log)normal or Weibull distributions: an automatic procedure to generate starting values. *In preparation.*

Aitchison J. and Silvey S.D. (1958) Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, **29**, 813-828.

Aitkin M. (2001) Likelihood and bayesian analysis of mixtures. *Statistical Modelling*, **1**, 287-304.

Barnett V. D. (1966) Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. *Biometrika*, **53**, 151-165.

Chanda K. C. (1954) A note on the consistency and maximum of the roots of likelihood equations. *Biometrika*, **41**, 56-61.

Cheng R. C. H. and Amin N. A. K. (1983) Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society: Series B*, **45**, 394-403.

Cheng R. C. H. and Iles T. C. (1987) Corrected maximum likelihood in non-regular problems. *Journal of the Royal Statistical Society: Series B*, **49**, 95-101.

Cox D. R. and Hinkley D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Cramér H. (1946) *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

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

Duda R. O. and Hart P. E. (1973) *Pattern Classification and Scene Analysis*. New York: Wiley.

Fisher A. H., Abel A., Lepper M., Zitzelsberger A. E. and von Glasgow A. (2000) Experimental Data and Statistical models for bimodal EM failures. *IEEE International Reliability Physics Symposium Proceedings.*

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. *Journal of Statistical Computation and Simulation*, **23**, 211-230.

Hosmer D. W. Jr. (1973) On MLE of the parameters of a mixture of two normal distributions when the sample size is small. *Communications in Statistics*, **1**, 217-227.

Huzurbazar V. S. (1948) The likelihood equations, consistency, and the maximum of the likelihood function. *Annals of Eugenetics*, **14**, 185.

Joyce W. B., Dixon R. W. and Hartman R. L. (1976) Statistical characterization of the lifetimes of continuously operated (Al,Ga)As double-heterostructure lasers. *Applied Physics Letters*, **28**, 684-686.

Kiefer N. M. (1978) Discrete parameter variation: efficient estimation of a switching regression model. *Econometrica*, **46**, 427-433.

Lehmann E. L. (1980) Efficient likelihood estimators. *The American Statistician*, **34**, 233-235.

Lehmann E. L. (1983) *Theory of Point Estimation.* New York: Wiley.

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

Perlman M. D. (1983) The limiting behavior of multiple roots of the likelihood equations. In *Recent Advances in Statistics: Papers in honor of Herman Chernoff on his Sixtieth Birthday*, pp. 339-370. New York: Academic Press.

Quandt R. E. (1972) A new approach to estimating switching regressions. *Journal of the American Statistical Association*, **76**, 306-310.

Quandt R. E. and Ramsey J. B. (1978) Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, **73**, 730-751.

Ranneby B. (1984) The maximum spacings method: an estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics*, **11**, 93-112.

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

Reeds J. A. (1985) Asymptotic number of roots of cauchy Location Likelihood equations. *The Annals of Statistics*, **13**, 775-784.

Sundberg R. (1974) Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics: Theory and Applications*, **1**, 49-58.

Tarone R. D. and Gruenhage G. (1975) A note on the uniqueness of roots of the likelihood equations for vector-valued parameters. *Journal of the American Statistical Association*, **70**, 903-904.

Teicher H. (1963) Identifiability of finite mixtures. *Annals of Mathematical Statistics*, **34**, 1265-69.

Titterington D. M. (1985) Comment on "Estimating parameters in continuous univariate distributions". *Journal of the Royal Statistical Society: Series B*, **47**, 115-116.

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