ELSEVIER Computational Statistics & Data Analysis 41 (2003) 577–590

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Choosing initial values for the EM algorithm for finite mixtures

Dimitris Karlis*, Evdokia Xekalaki

Department of Statistics, Athens University of Economics and Business, 76 Patision St., 10434, Athens, Greece

Received 1 February 2002; received in revised form 1 March 2002

Abstract

The EM algorithm is the standard tool for maximum likelihood estimation in finite mixture models. The main drawbacks of the EM algorithm are its slow convergence and the dependence of the solution on both the stopping criterion and the initial values used. The problems referring to slow convergence and the choice of a stopping criterion have been dealt with in literature and the present paper deals with the initial value problem for the EM algorithm. The aim of this paper is to compare several methods for choosing initial values for the EM algorithm in the case of finite mixtures as well as to propose some new methods based on modifications of existing ones. The cases of finite normal mixtures with common variance and finite Poisson mixtures are examined through a simulation study.

Keywords: Moment estimates; Bootstrap root search; Finite mixtures; Aitken acceleration

1. Introduction

Assume that the model is a finite mixture model of the form $f(x)=\sum_{j=1}^{k} p_j f(x|\theta_j)$, where $0 \leqslant p_j \leqslant 1$, $j=1,\ldots,k$, are the mixing proportions with $\sum_{j=1}^{k} p_j = 1$ and θ_j can be either a scalar or a vector of parameters. Among the various estimation methods considered in the literature for finite mixtures, the maximum likelihood (ML) method via the EM algorithm (Dempster et al., 1977) has dominated the field, not only because

* Corresponding author. Tel.: +3010-8203-503; fax: +3010-8203-162.
E-mail addresses: karlis@hermes.aueb.gr (D. Karlis), exek@hermes.aueb.gr (E. Xekalaki).

0167-9473/03/$ - see front matter
PII: S0167-9473(02)00177-9

of its simplicity relative to other methods, but also because of its interesting features: It exhibits monotonic convergence, it leads to estimates within the admissible range if the initial values are within the admissible range, and it can be given a simple and natural statistical interpretation (see Böhning, 1999; McLachlan and Peel, 2000 for details). On the other hand, there are drawbacks connected with the general EM algorithm. These include slow convergence, the need for a suitable stopping rule that can detect whether the algorithm has reached the maximum, and the choice of initial values in order to reach the global maximum in fewer iterations.

The issue of slow convergence has been dealt with by various authors whose proposals to alleviate the problem are mainly based on Aitken's approach (McLachlan and Peel, 2000) or on creating different data augmentations (Pilla and Lindsay, 2001). The issue of choosing a suitable stopping rule has also been given a lot of attention in the literature. Several criteria have been proposed and the effect of stopping early has been examined (Seidel et al., 2000a). The criteria used as stopping rules for the algorithm can be based on the relative change of the parameters and/or of the loglikelihood, indicating lack of progress rather than actual convergence, (Lindstrom and Bates, 1988), on Aitken's acceleration scheme (Böhning et al., 1994; Pilla et al., 2001) or on the gradient function (Lindsay, 1983, 1995; Pilla and Lindsay, 2001). The common characteristic of all the aforementioned criteria is that the algorithm stops iterating when the value of the chosen criterion becomes smaller than a specified constant. The smaller this constant, the more severe the criterion. Of course when the loglikelihood is trapped in a flat area, any criterion is likely to stop the algorithm early, while if the algorithm were to keep running, there would be possibly a substantial improvement in the likelihood.

When the algorithm has been trapped in a flat area, a rather naive strategy would be to keep iterating, hoping that the algorithm will locate the global maximum after a large number of iterations. Clearly, it would be preferable to start from different initial values and stop after a specified, relatively small, number of iterations are reached. Then one may keep iterating only from the 'solution' that has led to the largest value of the loglikelihood.

As far as the problem of choosing initial values is concerned, the literature abounds in methods proposed for making the choice that will lead to a global maximum. (A brief review of such methods is given in Section 2.) However, no comparative study of initial value strategies appears to have been made. The present paper focuses on this issue. In particular, a simulation study is presented that compares several different approaches for choosing initial values for the EM algorithm in the case of finite normal and finite Poisson mixtures. Some new approaches are proposed based mainly on modifications of existing ones. Different stopping rules are used in order to examine the effect of the initial value strategy in attaining the global maximum.

The material of the paper is structured as follows. In Section 2 existing methods for selecting initial values in fitting finite mixtures are reviewed. Section 3 focuses on the case of finite normal mixtures, containing a simulation comparison among different methods. Section 4 discusses the case of finite Poisson mixtures. A brief discussion and concluding remarks are given in Section 5.

2. Choosing initial values—a review

The choice of initial values is of great importance in the algorithm-based literature as it can heavily influence the speed of convergence of the algorithm and its ability to locate the global maximum. Laird (1978) proposed a grid search for setting the initial values. Leroux (1992) suggested the use of supplementary information in order to form clusters whose means were used as initial values. McLachlan (1988) proposed the use of principal component analysis for selecting initial values for the case of multivariate mixtures. Another clustering idea is described by Woodward et al. (1984).

Finch et al. (1989) proposed that, for a two-component normal mixture, only the mixture proportion needs to be given an initial value, as the rest of the parameters can be estimated automatically based on this value. Their idea was that, given the mixing proportion p, the sample is separated into two parts, one containing the first $[np]$ observations assumed to belong to the first component of the mixture and one containing the remaining observations, assumed to belong to the second component ($[a]$ stands for the integer part of a). The mean of the observations in the first part of the sample is used as an initial value for the mean of the first component of the mixture, while that of the second part is used as an initial value for the mean of the second component. Atwood et al. (1992) examined 5 different possible choices of p based on different partitions of the data in groups. Böhning (1999) proposed an initial partition of the data by maximizing the within sum of squares criterion.

Böhning et al. (1994) proposed to start with well-separated values as, in their experience, the algorithm could then converge faster. This was verified by our simulation study, but, for the finite normal mixture case, it requires a relatively small initial variance in addition to the far apart initial means. For the finite Poisson mixtures, as well as the finite exponential mixtures, where the value of the parameter corresponding to the mean determines the variance too, it is not easy to find such 'well separated' initial components.

Another natural choice is to begin with estimates obtained by other estimation methods, like the moment method. Furman and Lindsay (1994a, b) and Lindsay and Basak (1993) considered such starting values for normal mixtures. Fowlkes (1979) proposed some graphical and ad hoc methods for choosing initial values in the case of normal mixtures. Seidel et al. (2000a, b, c) examined some other choices for initial values for finite exponential mixtures. Practice has shown that it is preferable to start from several different initial values in order to ensure that the global maximum is obtained. Böhning (1999, p. 69) proposed a grid search over a large parameter space as a strategy to find several different initial values. In Sections 3 and 4, a comparison of some of the previously mentioned methods is provided.

3. A simulation comparison—finite normal mixtures

Consider the case of k-component normal mixtures. To avoid the problem of unbounded likelihoods, all the components are assumed to have a common variance. In our simulation study, several different sets of initial values for the means, denoted

as $\mu_j^{(0)}$, $j = 1,\ldots,k$, were compared. These were: (a) The 'true' values (applicable only in cases where the data were simulated from the assumed k-component normal mixture); (b) Random starting points (means are generated from uniform distributions over the data range, the mixing proportions are generated from a Dirichlet distribution and the variance is generated from a uniform distribution ranging from 0 to the value of the sample variance); (c) The 'best' of ten different random starting points (for each set of initial values, the loglikelihood was calculated and the set with the largest likelihood was considered as the 'best' and was used as the starting point); (d) Values obtained by Finch et al.'s method (described in the previous section) with equal mixing proportions and an initial value for σ^2 given by $\sigma_0^2 = s^2 - \hat{\sigma}^2(\mu)$ where $\hat{\sigma}^2(\mu) = \sum_{j=1}^{k}(\mu_j^{(0)} - \bar{x})^2/k$; (e) Values obtained by a moment matching method that uses equal mixing proportions and component means given by $\bar{x} \pm (j/2)s$, $j = 1,\ldots,k/2$, for k even and $j = 0,\ 1,\ldots,(k-1)/2$, for k odd. The variance is determined as in (d) above. The method has been used by Seidel et al. (2000a); (f) Moment estimates. (a detailed description of the procedure is given by Furman and Lindsay, 1994a); (g) Values obtained by a method based on the range which starts with equal mixing proportions, component means equal to $\mu_j^{(0)} = \min X_i + dj/(k+1)$, where d is the sample range, and variance specified as above, leading to well separated initial values, a useful strategy as suggested by Böhning et al. (1994) (note that in all the above cases if the variance estimate σ_0^2 came out to be negative, $s^2/2$ was used as an initial value of the variance); (h) Values obtained by a new method, which assigns the initial values to the latent vector $\mathbf{w_i} = (w_{i1}, w_{i2}, \ldots, w_{ik})$ by setting $w_{ij} = 1$ if the ith observation belongs to the jth component and 0 otherwise. Thus, the method distributes observations to specific components. Initially, the percentiles that split the whole data set in equal parts are considered as initial values for the means. Then, each observation is assigned to the component with the initial mean that is closer to the observation, in absolute distance. Thus, each initial vector $\mathbf{w_i}$ has one element equal to 1, and $(k-1)$ elements equal to 0. The initial values of the parameters are set using the M-step of the EM algorithm. Clearly, more sophisticated versions of this approach (which is similar in nature to the method proposed by Finch et al.) could be used.

In addition, four additional sets of random points were used so as to increase the probability that at least one of the methods would lead to the global maximum. The results show that the initial choice of the variance is quite important. Methods that led to a large initial variance often failed to converge to the global maximum, or they required an extremely large number of iterations.

It should be noted that the 'true' values were considered among the competitors since, often samples are required to be generated from a mixture with known parameter values, as for example, in the case of bootstrap likelihood ratio tests (McLachlan, 1987; Karlis and Xekalaki, 1999).

In order to examine the performance of the methods for selecting initial values defined above, the following criteria were used: The number of iterations until convergence was attained, which gives an indication about the computing time needed. The cost for building up the initial guess is usually negligible. The mean number of iterations needed until convergence is reported. Note that the standard EM algorithm was used without any acceleration scheme. To reduce the effect of stopping the

algorithm too early, a rather strict stopping criterion for the EM iterations was used. Iterations were stopped when the relative change in the loglikelihood was smaller than 10^{-12}. Another criterion for a good initial guess is its ability to lead to the global maximum. The likelihood surface for mixture models is known to have many local maxima, which are not global. To check if the global maximum has been obtained, the following procedure was followed:

For each of the different sets of initial values applied to the given sample (12 different sets of initial values), several values of the maximized loglikelihoods were obtained. The maximum value over all these values, say $L_{\max}$, was then considered as the 'global' maximum. The values of the parameters corresponding to it are denoted by $\theta_{\max}$. The jth set of initial values is regarded as having succeeded in locating the 'global' maximum if the estimates given by this set of initial values, say θ_j, and the maximized loglikelihood of this set, say L_j, satisfy the conditions (a) $\max|\theta_{\max} - \theta_j| < 10^{-5}$ and (b) $|(L_{\max} - L_j)/L_{\max}| < 10^{-5}$. It is implicitly assumed that the global maximum has been obtained by at least one of the methods.

In the case of two-component normal mixtures, four different values for $\Delta = |\mu_1 - \mu_2|/\sigma$ were considered with $p = 0.1, 0.5, 0.7$ and $\sigma = 1$, $\mu_2 = 0$. In addition, a three-component normal mixture was used in order to examine the behavior of the initial values in the case of an incorrect model. For the case of three-component normal mixtures, the configurations are given in Tables 1 and 2, for $\mu_2 = 0$, $\mu_3 = -\mu_1$ and $\sigma = 1$.

Table 1 provides the proportions of times (out of 100 replications) the various methods succeeded in locating the 'global maximum', while Table 2 gives the ratios of the numbers of iterations required, till convergence, by the various methods relative to another method (reference method) that is considered as performing better (the moment method for the two-component case and the 'true' values method for the three-component case). The mean number of iterations and its standard error, of the reference method, are reported in order to provide an idea of the magnitude of the number of iterations required. Since failures to locate the 'global maximum' usually led to a large number of iterations, only values corresponding to the cases the method succeeded in reaching the 'global maximum' are reported. (This explains the presence of empty cells). For the three-component mixtures, the moment estimates were not considered due to the computational difficulty in obtaining them.

From Tables 1 and 2, one can see that if the algorithm starts from random points, the performance is poorer. The moment estimates method shows the best behavior with respect to locating the global maximum, thus verifying the findings of Furman and Lindsay (1994b). However, it requires more iterations till convergence, compared to other methods with a large probability of locating the global maximum, as can be seen from Table 2. For the case of well-separated components, the majority of methods considered performed well. On the other hand, if the model is incorrect, all the methods have difficulties in locating the global maximum and it is quite interesting that they all perform in quite a similar manner. The 'true' values method behaves well if the components are not too close. However, its high rate of success makes it an appealing choice in cases where repetitive application of the EM is required as, for example, in the case of bootstrap likelihood ratio tests. The new method proposed provides very good results and, because of its simplicity, can be extended to the case of models

Table 1
Normal mixture case—proportions of times (based on 100 replications) the various methods succeeded in locating the "global maximum"

2-*Component mixture*										
Distribution		Sample size	TRUE	Random	Best of ten	Finch	Moment matching	Moment	Well separated	New
Δ	P									
		50	—	0.59	0.58	0.91	0.18	0.91	0.43	0.90
0	—	100	—	0.64	0.69	0.97	0.65	0.97	0.4	0.96
		500	—	0.61	0.65	0.98	0.75	0.99	0.51	0.99
		50	0.54	0.64	0.65	0.99	0.99	0.99	0.48	0.5
1	0.1	100	0.55	0.58	0.72	0.94	0.91	0.93	0.42	0.55
		500	0.74	0.61	0.62	0.88	0.85	0.88	0.42	0.71
		50	0.98	0.61	0.64	0.98	0.96	0.98	0.58	0.97
1	0.5	100	0.97	0.65	0.69	0.98	0.94	0.98	0.49	0.97
		500	0.96	0.7	0.71	0.95	0.9	0.96	0.07	0.99
		50	0.95	0.64	0.74	0.98	0.98	0.97	0.62	0.98
3	0.1	100	1	0.64	0.7	0.99	0.99	0.99	0.32	0.99
		500	1	0.61	0.69	1	1	1	0.08	1
		50	1	0.73	0.82	1	1	1	0.84	1
3	0.5	100	1	0.83	0.81	1	1	1	0.65	1
		500	1	0.81	0.87	1	1	1	0.29	1
		50	1	0.74	0.83	1	1	0.96	1	1
5	0.5	100	1	0.82	0.84	1	1	1	1	1
		500	1	0.76	0.87	1	1	1	0.99	1
3-*component mixture*		50	—	0.79	0.75	0.74	0.74	0.78	0.76	0.78
$\mu_1 = -1, \mu_2 = 0, \mu_3 = 1,$		100	—	0.71	0.73	0.73	0.71	0.72	0.71	0.72
$p_1 = p_2 = p_3 = \frac{1}{3}, \sigma = 1$		500	—	0.72	0.76	0.74	0.74	0.7	0.77	0.75

3-component mixture

Distribution		Sample size	TRUE	Random	Best of ten	Finch	Moment matching	Moment	Well separated	New
Δ	(p_1, p_2, p_3)									
		50	0.90	0.03	0.66	0.95	0.01	—	0.01	0.89
1	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	100	0.90	0.06	0.62	1	0.04	—	0.04	0.90
		500	0.91	0.02	0.63	1	0	—	0.02	0.90
		50	0.63	0	0.61	0.93	0	—	0	0.57
1	0.7,0.2,0.1	100	0.64	0.03	0.55	0.9	0.03	—	0.01	0.59
		500	0.78	0	0.43	0.78	0.01	—	0.01	0.59
		50	1	0	0.51	1	0	—	0	1.00
3	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	100	1	0	0.55	1	0	—	0	1.00
		500	1	0	0.47	1	0	—	0	1.00
		50	0.98	0	0.27	0.34	0		0	0.31
3	0.7,0.2,0.1	100	1	0	0.13	0.15	0	—	0	0.12
		500	1	0	0.20	0	0	—	0	0.00

with more than two components. The latter is often not feasible, with the method that uses moment estimates as the initial values, since these are not as easily obtainable in this case. Finally, another point is that the standard errors associated with the number of iterations are quite high, making many of the differences statistically insignificant.

Table 2
Normal mixture case—ratios of the mean numbers of iterations until convergence for each method relative to those for a reference method. The mean number of iterations and its standard error for the reference method are reported in the last two columns

Distribution									Reference method	
Δ	P	Sample size	TRUE	Random	Best of ten	Finch	Moment matching	Well separated	New	Moment mean	Moment st. error
		50	—	1.05	1.27	1.05	7.67	1.75	1.17	3055	688
0	—	100	—	1.90	1.71	1.15	30.04	1.79	0.93	578	141
		500	—	1.35	0.61	0.82	11.51	2.47	0.81	777	308
		50	4.38	2.65	2.27	1.09	2.06	2.01	3.23	352	105
1	0.1	100	0.59	2.56	1.00	1.47	0.94	3.12	0.28	1385	791
		500	0.40	1.03	1.19	1.06	1.24	0.89	0.28	4058	1562
		50	1.30	1.86	2.20	2.36	3.57	2.39	1.19	323	74.8
1	0.5	100	1.94	1.65	1.11	1.21	2.34	2.74	1.36	923	199
		500	1.26	1.88	1.42	0.79	1.63	1.07	1.38	2158	511
		50	0.73	3.77	4.08	1.41	1.63	3.48	0.22	66.1	11.1
3	0.1	100	0.83	2.71	2.11	1.71	4.35	2.15	1.84	64.2	9.67
		500	0.85	2.09	1.98	1.44	1.78	2.40	1.45	43	1.71
		50	0.97	13.74	2.43	1.00	2.00	2.21	0.96	38.8	5.96
3	0.5	100	0.88	4.22	6.79	0.90	2.32	51.62	0.88	30.7	1.43
		500	0.85	51.91	9.39	0.87	2.44	213.13	0.85	26.1	0.58
		50	0.07	3.70	4.59	0.07	0.27	0.13	0.06	105	37.1
5	0.5	100	0.35	27.98	10.56	0.37	1.51	0.74	0.35	19.5	4.22
		500	0.71	171.51	71.09	0.77	3.24	2.47	0.71	8.67	0.13
3-*component mixture*		50	—	1.18	1.97	0.77	0.79	0.82	0.98	635	205
$\mu_1=-1, \mu_2=0, \mu_3=1,$		100	—	1.52	1.09	1.11	1.06	0.59	0.95	1719	612
$p_1=p_2=p_3=1/3, \sigma=1$		500	—	2.29	2.15	1.87	3.33	2.20	0.92	1799	447

3-component mixture

Δ	(p_1, p_2, p_3)		Random	Best of ten	Finch	Moment matching	Well separated	New	TRUE mean	TRUE st. error
		50	0.20	1.94	0.63	10.14	0.33	0.96	526.09	94
1	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	100	0.31	2.70	1.05	24.23	0.32	0.62	1185.94	208.2
		500	0.44	1.67	1.02	—	0.35	0.73	2540.02	359.18
		50	—	1.46	0.32	—	—	0.14	1590.84	425.77
1	0.7,0.2,0.1	100	0.39	0.92	0.46	1.52	0.04	0.17	2063.69	441.09
		500	—	0.78	0.25	0.78	0.07	0.18	5866.92	1164.32
		50	—	6.80	1.05	—	—	1.02	80.58	7.42
3	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	100	—	12.72	1.01	—	—	1.00	75.33	4.53
		500	—	64.40	1.02	—	—	1.01	55.17	1.48
		50	—	12.65	1.13	—	—	1.15	187.42	99.48
3	0.7,0.2,0.1	100	—	1.87	1.70	—	—	1.62	105.67	13.68
		500	—	1.18	—	—	—	1.70	69.12	3.39

Thus, the results are rather indicative and they cannot prove the superiority of one method upon each other with respect to the number of iterations.

As far as the three-component normal mixture case is concerned, some methods have a very poor performance. The 'true' values method had the best performance, while the new method described in (h) and the method proposed by Finch et al. (1989)

can be considered as alternative choices. It is worth mentioning that in the case of three-component mixtures more local maxima were located and, in general, the degree of complexity of the situation calls for the need of trying out more than one set of initial values. This holds for mixtures with more components.

4. A simulation comparison—finite Poisson mixtures

Consider the case of the Poisson distribution. Due to the discrete nature of the distribution and the restriction on its parameter space, the case of local maxima is not so common, while some methods for selecting initial values may lead to inadmissible initial estimates.

In this simulation experiment, the following methods for choosing initial values were considered: (a) The 'true' values of the parameters; (b) Moment estimates of the parameters. Note that moment estimates may not be obtainable in certain cases (e.g. when the sample size is small and/or the components are close together); (c) Finch et al.'s method as described for the normal case (but of course there is no need to specify an initial value for the variance); (d) A variant of Finch et al.'s method. This amounts to finding the initial value for p_1 as in the previous method. Then, the initial values for λ_1 and λ_2 are calculated as: $\lambda_1 = \bar{x} - [(s^2 - \bar{x})(1 - p_1)/p_1]^{1/2}$, $\lambda_2 = \bar{x} + [(s^2 - \bar{x})p_1/(1 - p_1)]^{1/2}$, where $\bar{x}$ is the sample mean and s^2 is the sample variance. The motivation for this algorithm is the fact that the mean and the variance calculated via the initial values are equal to the mean and the variance of the sample, respectively. If $\lambda_1 \leqslant 0$, we set $\lambda_1 = 0.01$; (e) Setting $p_1 = 0.5$ and $\lambda_1 = \bar{x} - s$, $\lambda_2 = \bar{x} + s$. This initial guess is symmetric, and satisfies the first two moment equations of the observed dataset. The choice of the value 0.5 as an initial value for the mixing proportion is expected to work reasonably well only when the mixing proportion is near 0.5. This method, however, can be easily extended to the case of more than two components; (f) A bootstrap root search type method along the lines discussed by Markatou et al. (1998), differing slightly from theirs in that moment estimation was applied to a bootstrap sample from the original sample to create the initial values. This method may overcome the shortcomings of the method of moments, which often fails to yield moment estimates. Two variants are considered. The first uses a bootstrap sample size of 15 (referred to as B-15) as proposed by Markatou (2000). The second variant uses a bootstrap sample of half the size of the original sample. If the bootstrap sample obtained fails to lead to the moment estimates, it is replaced by another bootstrap sample and this process is repeated until the moment estimates become available. This approach is referred to as B-n. Note that Markatou (2000) proposed the bootstrap root search method not as a method of choosing initial values (as the case is in this paper), but as a method for finding all the roots of the estimating equations with a high probability (in cases where there are more than one roots).

Other methods were also included in the simulations, but, since their performance was inferior to that of the methods discussed above, the obtained results are not reported. Again, for each sample, 12 different initial sets of initial values were considered.

The criteria used in order to assess the behavior of the initial values were the same as those considered in Section 3 in the case of normal mixtures. In addition, the proportion of times the method failed to provide initial estimates in the admissible range was used as a further criterion.

Our aim was to examine the behavior of all the methods in two distinct situations: when the model is correct and when the data have been generated from other models. The sampling distributions used were two-component Poisson mixtures (correct model), and the alternative mixed Poisson distributions considered were the negative binomial and various three-component Poisson mixtures. In the latter cases, the model to be estimated was incorrect. Finally, for each of the three distributions used, the parameter values were selected so as to allow for various shapes, thus covering a variety of different cases. The sample sizes considered were $n = 50$, 100 and 500. Complete results can be found in Karlis and Xekalaki (2001).

Table 3 summarizes the results of the simulation study. The entries in the table are the mean number of iterations for each of the methods considered relative to that of the moment method, while the mean number of iterations for the moment method, together with its standard error is reported in the last column. In most of the cases, the moment method turns out to be the method, which requires fewer iterations. Only the 'true' values can compete on this issue in the case where we sample from the two-component Poisson mixture whence the parameter values can be regarded as known. Unfortunately, the moment method has a high probability of failing to provide initial estimates, especially in the cases of small sample sizes and/or not well separated components.

The B-15 method does not seem to work well in terms of the number of iterations till convergence; this is not true for the B-n method, which is the 'best' method when moment estimates are non-obtainable. However, if the sample size is large, the B-15 method requires less computing time because calculations and resampling are performed only for samples of size 15. As far as the remaining methods are concerned, Finch et al.'s method seems to perform better as it never fails to provide initial estimates and requires fewer iterations in comparison to the modified method of Finch et al. (method (d)). A rather interesting finding is that all the methods performed very well in locating the global maximum when the data were generated from a two-component Poisson mixture (Table 4). When the model was not correct, the moment estimates had the best performance for locating the global maximum. The above implies that, for the bootstrap likelihood ratio test, when one samples from the underlying null distribution, the global maximum can be obtained without the need for starting from several initial values.

When the components are well separated, the algorithm terminates quite quickly for all the initial values and no appreciable differences exist between the methods. On the other hand, the number of iterations needed to meet the convergence criterion depends on the 'information' available.

Let us now turn to the case where one might try to estimate the parameters of a two-component Poisson mixture on the basis of data that do not come from a two-component Poisson mixture. In this case, the moment method is again very attractive because of the smaller number of iterations that are usually needed. However,

Table 3
Poisson mixture case—ratios of the mean numbers of iterations until convergence for each method relative to those for a reference method. The mean number of iterations and its standard error for the reference method are reported in the last two columns

Distribution	Sample size	True values	Finch	Momdified Finch	Symmetric around mean	B-15	B-*n*	Moment method mean	Moment method st. error
2-*component Poisson mixture*									
0.1, 1, 8	50	0.57	1.61	1.62	1.55	1.24	1.21	42.1	0.36
	100	0.77	1.85	1.83	1.78	1.67	1.50	25.69	0.20
	500	0.82	1.91	1.90	1.84	1.56	1.07	19.67	0.10
0.5, 1, 8	50	0.94	1.03	1.02	1.00	1.09	1.07	13.21	0.10
	100	0.93	1.06	1.02	1.04	1.13	1.06	12.57	0.08
	500	0.94	1.17	1.03	1.15	1.23	1.08	11.17	0.06
0.5, 1, 2	50	1.39	1.77	1.81	1.36	1.59	1.47	374.81	1.17
	100	1.73	2.03	2.05	1.66	1.79	1.56	493.73	1.09
	500	1.52	1.51	1.54	1.49	1.58	1.53	1099.07	1.79
Negative Binomial									
Mean = 1	50	—	1.61	2.29	1.56	1.64	1.59	108.98	0.55
Variance = 2	100	—	1.32	1.82	1.27	1.29	1.23	130.71	0.73
	500	—	1.19	1.57	1.07	1.11	1.05	103.39	0.31
Mean = 6	50	—	1.02	1.04	1.02	1.04	1.00	170.31	0.51
Variance = 24	100	—	1.00	1.02	0.99	1.02	1.00	185.54	0.48
	500	—	0.98	1.01	0.97	1.00	0.99	202.53	0.36
3-*component Poisson mixture*									
0.4, 0.3, 1, 5, 7	50	—	1.07	1.04	1.05	1.10	1.07	29.19	0.20
	100	—	1.12	1.05	1.09	1.13	1.06	26.07	0.16
	500	—	1.20	1.12	1.18	1.21	1.05	24.34	0.10
0.3, 0.4, 1, 2, 3	50	—	1.34	1.38	1.20	1.17	1.16	397.3	1.42
	100	—	1.35	1.35	1.27	1.43	1.22	443.09	1.11
	500	—	1.23	1.20	1.23	1.34	1.22	481.94	0.97
0.7, 0.2, 1, 5, 10	50	—	1.02	1.37	1.02	1.03	1.02	87.19	0.40
	100	—	1.03	1.26	1.02	1.04	1.01	86.97	0.29
	500	—	1.02	1.17	1.02	1.02	0.99	82.93	0.17

Finch et al.'s method is an interesting competitor. This is so since, for distributions with a high overdispersion, it requires a few iterations. Moreover, the initial parameter estimates are always obtainable. Note that, now, the proportion of times each method failed to obtain the global maximum is higher, thus suggesting that the use of a single initial value is not a good strategy.

Table 4
Poisson mixture case—proportions of times (based on 100 replications) the various methods succeeded in locating the "global maximum"

		Method						
	Sample size	True values	Modified Finch	Finch	Moment matching	Symmetric around mean	B-15	B-*n*
2-component mixture								
0.8, 1, 5	50	1.00	1.00	1.00	1.00	1.00	0.99	0.99
	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.1, 1, 8	50	0.98	0.96	0.98	0.98	0.96	0.96	0.97
	100	1.00	1.00	1.00	1.00	1.00	0.99	1.00
	500	1.00	1.00	1.00	1.00	1.00	0.99	1.00
0.5, 1, 8	50	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.5, 1, 2	50	0.75	0.94	0.95	0.95	0.77	0.76	0.77
	100	0.82	0.96	0.95	0.95	0.81	0.85	0.87
	500	0.90	0.95	0.90	0.91	0.89	0.91	0.91
Negative Binomial								
Mean = 1	50	—	0.99	0.99	1.00	0.99	0.97	0.98
Variance = 2	100	—	1.00	1.00	1.00	1.00	0.98	1.00
	500	—	1.00	1.00	1.00	1.00	0.98	1.00
Mean = 6	50	—	0.99	0.99	0.99	0.99	0.98	0.99
Variance = 24	100	—	1.00	1.00	1.00	1.00	0.99	0.99
	500	—	1.00	1.00	1.00	1.00	1.00	1.00
3-component mixture								
0.4, 0.3, 1, ,5, 7	50	—	0.96	0.97	0.97	0.90	0.89	0.89
	100	—	0.98	0.98	0.98	0.94	0.93	0.94
	500	—	0.99	0.99	0.99	0.99	0.98	0.99
0.3, 0.4, 1, 2, 3	50	—	0.99	0.99	0.98	0.99	0.98	0.99
	100	—	0.99	0.99	0.99	0.99	0.99	0.99
	500	—	1.00	1.00	1.00	1.00	1.00	1.00
0.7, 0.2, 1, 5, 10	50	—	1.00	1.00	1.00	1.00	1.00	1.00
	100	—	1.00	1.00	1.00	1.00	1.00	1.00
	500	—	1.00	1.00	1.00	1.00	1.00	1.00
0.4, 0.4, 1, 2, 5	50	—	0.99	0.99	1.00	0.99	0.99	0.98
	100	—	0.99	0.99	1.00	0.99	0.99	0.98
	500	—	1.00	1.00	1.00	1.00	0.99	1.00

Table 5
Proportions of times the various methods succeeded in locating the 'global maximum' using different stopping criteria. The reported values are averages over all the two-component Poisson mixture models considered

	50		100		500	
	$< 10^{-12}$	$< 10^{-6}$	$< 10^{-12}$	$< 10^{-6}$	$< 10^{-12}$	$< 10^{-6}$
True values	0.91	0.14	0.92	0.15	0.96	0.10
Moment estimates	0.97	0.56	0.97	0.66	0.96	0.50
Modified Finch	0.96	0.20	0.95	0.11	0.96	0.09
Symmetric around the mean	0.96	0.17	0.96	0.13	0.96	0.10
Finch	0.93	0.16	0.91	0.13	0.96	0.05
B-15	0.94	0.23	0.91	0.19	0.94	0.10
B-*n*	0.93	0.24	0.91	0.21	0.96	0.17

Table 5 provides the proportions of times at which each of the examined methods was considered to have failed to locate the 'global' maximum. Two different tolerance levels have been used for the criterion. Iterating was stopped when the relative change of the loglikelihood was smaller than 10^{-6} and 10^{-12}, respectively. The entries of Table 5 reveal the importance of the stopping criterion. If the criterion is not strict (as in the case of the first tolerance level) it is very likely that the resulting estimates will be far from the global maximum.

A smaller simulation comparison with four-component Poisson mixtures was also carried out. The obtained results were quite similar to those reported in the two-component case. Of course, the number of iterations required was much greater, as was expected, due to the greater amount of missing information. In the case of four-component Poisson mixtures, the moment estimates are not obtainable with a high probability due to the high order sample moments involved. Thus, in practice, one needs an alternative method. As in the case of normal mixtures, the performance of all the methods becomes worse in the case of more than two-components, pointing to the need of considering several sets of starting values in order to ensure more reliable results.

5. Concluding remarks

In this paper, a simulation comparison of several methods for choosing initial values was carried out. The results clearly show the dependence of the method on the choice of the initial values. In addition, the algorithm may run a lot of iterations trapped in areas away from the global maximum. This implies that it would be advisable to use a mixed strategy, by starting from several different initial values, making a small number of iterations without necessarily examining convergence and then, running until convergence from the point with the largest likelihood after these initial iterations, using a strict criterion. Such an approach helps in reducing the amount of time spent in areas of a flat likelihood, away from the global maximum.

For the purposes of a bootstrap likelihood ratio test, the 'true' values can be a successful choice of the initial values as they can lead to the location of the global

maximum with a high probability. Some other choices for particular finite mixtures (normal and Poisson) were also discussed. In concluding, attention should be paid to the often problematic behavior of the stopping criteria that fails to give a clear indication of whether the algorithm converged to the global maximum or it was trapped in an area of a flat likelihood.

Acknowledgements

The authors would like to thank Prof. Marianthi Markatou, the referees and an Associate Editor for their constructive comments.

References

Atwood, L.D., Wilson, A.F., Elston, R.C., Bailey-Wilson, J.E., 1992. Computational aspects of fitting a mixture of two normal distributions using maximum likelihood. Comm. Statist. Simulation Comput. 21, 769–781.

Böhning, D., 1999. Computer Assisted Analysis of Mixtures & Applications in Meta-analysis, Disease Mapping & Others. CRC Press, Boca Raton, FL.

Böhning, D., Dietz, E., Schaub, R., Schlattman, P., Lindsay, B., 1994. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Ann. Inst. Statist. Math. 46, 373–388.

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

Finch, S., Mendell, N., Thode, H., 1989. Probabilistic measures of adequacy of a numerical search for a global maximum. J. Amer. Statist. Assoc. 84, 1020–1023.

Fowlkes, E., 1979. Some methods for studying the mixture of two normal (lognormal) distributions. J. Amer. Statist. Assoc. 74, 561–575.

Furman, W.D., Lindsay, B.G., 1994a. Testing for the number of components in a mixture of normal distributions using moment estimators. Comput. Statist. Data. Anal. 17, 473–492.

Furman, W.D., Lindsay, B.G., 1994b. Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. Comput. Statist. and Data. Anal. 17, 493–508.

Karlis, D., Xekalaki, E., 1999. On testing for the number of components in finite Poisson mixtures. Ann. Inst. Statist. Math. 51, 149–162.

Karlis, D., Xekalaki, E., 2001. On implementing the EM algorithm in finite Poisson mixtures. Department of Statistics, Athens University of Economics, p. 132.

Laird, N., 1978. Nonparametric maximum likelihood estimation of a mixing distribution. J. Amer. Statist. Assoc. 73, 805–811.

Leroux, B.G., 1992. Consistent estimation of a mixing distribution. Ann. Statist. 20, 1350–1360.

Lindsay, B.G., 1983. The geometry of mixture likelihood. a general theory. Ann. Statist. 11, 86–94.

Lindsay, B.G., 1995. Mixture Models: Theory, Geometry and Applications. . NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5. Institute of Mathematical Statistics, California.

Lindsay, B.G., Basak, P., 1993. Multivariate normal mixtures: a fast consistent method of moments J. Amer. Statist. Assoc. 88, 468–475.

Lindstrom, M.J., Bates, D.M., 1988. Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. J. Amer. Statist. Assoc. 83, 1014–1022.

Markatou, M., 2000. Mixture models, robustness and the weighted likelihood methodology. Biometrics 56, 483–486.

Markatou, M., Basu, A., Lindsay, B.G., 1998. Weighted likelihood estimating equations with a bootstrap root search. J. Amer. Statist. Assoc. 93, 740–750.

McLachlan, G.J., 1987. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Appl. Statist. 36, 318–324.

McLachlan, G.J., 1988. On the choice of initial values for the EM algorithm in fitting mixture models. The Statistician 37, 417–425.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

Pilla, R.S., Lindsay, B.G., 2001. Alternative EM methods for nonparametric finite mixture models. Biometrika 88, 535–550.

Pilla, R.S., Kamarthi, S.V., Lindsay, B.G., 2001. Aitken-based acceleration methods for assessing convergence of multilayer neural networks. IEEE Trans. Neur. Net. 12, 998–1012.

Seidel, W., Mosler, K., Alker, M., 2000a. A cautionary note on likelihood ratio tests in mixture models. Ann. Inst. Statist. Math. 52, 481–487.

Seidel, W., Mosler, K., Alker, M., 2000b. Likelihood ratio tests based on subglobal optimisation: a power comparison in exponential mixture models Statist. Hefte 41, 85–98.

Seidel, W., Sevcikova, H., Alker, M., 2000c. On the power of different versions of the likelihood ratio test for homogeneity in an exponential mixture model. Dept. Stat. und Quantit. Ökon. Universität der Bundeswehr Hamburg, Report 92.

Woodward, W., Parr, W., Schucany, R., Lindsey, H., 1984. A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. J. Amer. Statist. Assoc. 79, 590–598.