Ann. Inst. Statist. Math.
Vol. 52, No. 3, 481–487 (2000)

A CAUTIONARY NOTE ON LIKELIHOOD RATIO TESTS IN MIXTURE MODELS

WILFRIED SEIDEL[1], KARL MOSLER[2] AND MANFRED ALKER[1]*

[1]*Fachbereich Wirtschafts- und Organisationswissenschaften, Universität der Bundeswehr Hamburg, D-22039 Hamburg, Germany*

[2]*Seminar für Wirtschafts- und Sozialstatistik, Universität zu Köln, D-50923 Köln, Germany*

(Received December 1, 1997; revised April 5, 1999)

Abstract. We show that iterative methods for maximizing the likelihood in a mixture of exponentials model depend strongly on their particular implementation. Different starting strategies and stopping rules yield completely different estimators of the parameters. This is demonstrated for the likelihood ratio test of homogeneity against two-component exponential mixtures, when the test statistic is calculated by the EM algorithm.

Key words and phrases: EM algorithm, exponential mixture models, initial values, stopping criteria, maximum likelihood estimation, likelihood ratio test.

1. Introduction

To estimate the parameters of a finite mixture model by maximum likelihood, iterative methods are commonly used. The likelihood function in mixture models tends to have multiple maxima. The expectation-maximization (EM) algorithm converges monotonically to a local maximum or a saddle point. Its convergence can be rather slow, which has given rise to many accelerated variants in the literature. Frequently it is argued that the behaviour of the EM algorithm (and of these variants as well) does not depend on its special implementation and, in particular, is independent of the rule by which initial values are selected.

In this note we shall demonstrate that estimates and tests which are based on iterative maximization of the likelihood may well depend on the starting strategy. Moreover, they may depend on the stopping criterion. We show that these effects really occur and that their consequences cannot be neglected in practice.

As the effects are particularly severe when the data come from the "wrong" model, likelihood ratio tests for the number of mixing components are heavily affected by them. Therefore we argue that it is not feasible to perform such a test on the basis of quantiles that have been obtained by Monte Carlo simulations of the test statistic and published in the literature, unless the algorithm employed by the authors is fully known and reimplemented.

In particular, we investigate the likelihood ratio test for exponential homogeneity against finite mixtures of exponentials (Böhning *et al.* (1994)). We show that the details of the maximization algorithm strongly influence the simulated quantiles. So, when applying the test, not only are the functional form of the test statistic and the critical values decisive but also the specific implementation of the algorithm. A required level

* This research has been partially sponsored (M. Alker) by a grant from the German Research Foundation.

of the test can only be guaranteed if, in every application, the test statistic is calculated with precisely the same algorithm that has been used in the former simulation of the quantiles.

For a general introduction into the analysis of mixture models and their applications, the reader is referred to the monographs by Everitt and Hand (1981), Titterington *et al.* (1985) and Lindsay (1995). McLachlan and Basford (1988) present statistical tools including a resampling approach for assessing the null distribution of the likelihood ratio test for homogeneity. For the EM algorithm see Dempster *et al.* (1977) and, for recent developments, McLachlan and Krishnan (1997).

Section 2 introduces the exponential mixture model. We discuss different implementations of the EM algorithm, while in Section 3 we present the likelihood ratio test for exponential homogeneity and its quantiles and sizes obtained by different implementations. The Appendix collects some details about random number generation and the level of accuracy required. This level must be very high in order to obtain stable quantiles.

2. Exponential mixture model and EM algorithm

For $x, \theta > 0$, let $f(x,\theta) = \frac{1}{\theta}\exp(-\frac{x}{\theta})$ denote the density of the exponential distribution with expectation θ. A mixture of two exponential distributions with parameters θ_1 and θ_2 and with mixing weights p and $1-p$, $0 \le p \le 1$, has density $f(x,P) = pf(x,\theta_1) + (1-p)f(x,\theta_2)$, where

$$P = \begin{bmatrix} \theta_1 & \theta_2 \\ p & 1-p \end{bmatrix}$$

denotes the parameter of the mixture. A maximum likelihood (ML) estimator $\hat{P}$ of P is defined as a parameter value $\hat{P}$ that maximizes the log-likelihood function $l(P) = \sum_{i=1}^{n} \ln f(x_i,P)$. Starting from an initial value P^0, the EM algorithm generates a sequence

$$P^k = \begin{bmatrix} \theta_1^k & \theta_2^k \\ p^k & 1-p^k \end{bmatrix}, \qquad k \in \mathbb{N} \cup \{0\},$$

according to an iterative scheme which has a simple form in our situation; see Böhning and Schlattmann ((1992), p. 293). The sequence $l(P^k)_{k\in\mathbb{N}}$ is nondecreasing.

2.1 *Stopping criteria*

Let $acc > 0$ be a given level of accuracy. In our investigation we employ two criteria to stop the EM algorithm:

Stopping criterion 1 (**"function values"**) is based on the size of change in the log-likelihood. From Böhning and Schlattmann (1997) we borrow the following version of it: Choose $\hat{P} = P^k$, if $l(P^k) - l(P^{k-2}) < n \cdot acc$ for some $k = 3\nu + 2$, $\nu \ge 0$.

However, as Lindstrom and Bates (1988) write, this first criterion "is a measure of lack of progress but not of actual convergence". The following stopping criterion has a better theoretical foundation; see Böhning *et al.* (1994).

Stopping criterion 2 (**"directional derivatives"**) uses the expression

$$D_P(\theta) = \sum_{i=1}^{n} \frac{f(x_i,\theta) - f(x_i,P)}{f(x_i,P)}.$$

It can be shown that $D_P(\theta)$ has the properties of a directional derivative of $l(P)$ in the direction of θ. In Böhning and Schlattmann (1992) and below the iterations are stopped at $\hat{P} = P^k$ if $\max\{D_{P^k}(\theta_1^k), D_{P^k}(\theta_2^k)\} < n \cdot acc$ and $k \ge 3$.

Observe that both stopping criteria do not depend on data scale, i.e., they are scale invariant in the probability law.

2.2 *Scale Invariance*

If X is exponentially distributed with parameter θ and $a > 0$ then aX is exponentially distributed with parameter $a\theta$. Let $(P^k)_{k\in\mathbb{N}}$ be the sequence of estimates obtained from a sample $x_1, \ldots, x_n$ by the EM algorithm. Then the sample $ax_1, \ldots, ax_n$ results in a sequence $(P_a^k)_{k\in\mathbb{N}}$ with

$$P_a^k = \begin{bmatrix} a\theta_1^k & a\theta_2^k \\ p^k & 1-p^k \end{bmatrix}.$$

Therefore, if the starting values θ_1^0 and θ_2^0 are scale equivariant and p^0 is scale invariant in the underlying probability law, the ML estimator calculated by the EM algorithm is scale equivariant with respect to θ_1 and θ_2 and scale invariant with respect to p.

2.3 *Starting Values*

There are some indications for good starting values of the EM algorithm, at least for mixtures of univariate distributions. Let $x_{\min} = \min\{x_1, \ldots, x_n\}$ and $x_{\max} = \max\{x_1, \ldots, x_n\}$. In Böhning and Schlattmann ((1992), p. 296) it is shown for a particular example that an initial parameter with $p = 0.5$ and well separated values of θ_1 and θ_2 yields the global maximum in a normal mixture model, and in Böhning *et al.* ((1994), Section 5) $p = 0.5$, $\theta_1 = x_{\min} + 0.5$ and $\theta_2 = x_{\max} - 0.5$ are chosen as initial values in a Poisson mixture model. So our first strategy, abbreviated as $x_{\min}/x_{\max}$, is

$$x_{\min}/x_{\max}: \qquad P^0 = \begin{bmatrix} x_{\min} & x_{\max} \\ .5 & .5 \end{bmatrix}.$$

We contrast this with a second strategy,

$$\bar{x} \pm .5\theta: \qquad P^0 = \begin{bmatrix} \bar{x} - .5\theta & \bar{x} + .5\theta \\ .5 & .5 \end{bmatrix}.$$

Observe that both strategies start with parameters θ_1^0 and θ_2^0 that are scale equivariant in the data generating probability law, while p^0 does not depend on it.

3. Application to quantiles of a likelihood ratio test

Let us consider the likelihood ratio test for the null hypothesis of *exponential homogeneity*, H_0: $X \sim f(x, \theta)$ for some θ, against the alternative that the distribution of X is a *mixture of exponential distributions*, H_1: $X \sim f(x, P)$ with $\theta_1 \neq \theta_2$ and $0 < p < 1$. Under H_0, the ML estimator $\hat{\theta}$ of θ is given by $\hat{\theta} = \bar{x}$. The ML estimator under the alternative has to be calculated numerically. The test statistic of the likelihood ratio test, $2 \ln \lambda_n = 2[l(\hat{P}) - l(\hat{\theta})]$, has a null distribution that does not depend on θ (see Subsection 2.2).

Table 1. Quantiles under different maximization strategies.

	Start: $x_{\min}/x_{\max}$		Start: $\bar{x} \pm .5\theta$		published in
	Stop: funct.	Stop: deriv.	Stop: funct.	Stop: deriv.	Böhning *et al.* (1994)
$n = 100$					
$\alpha = 0.1$	3.36	3.36	2.22	2.44	1.69
$\alpha = 0.05$	4.73	4.76	3.64	3.80	3.26
$\alpha = 0.025$	6.13	6.19	5.04	5.23	4.67
$\alpha = 0.01$	8.15	8.01	6.94	7.09	6.33
$n = 1000$					
$\alpha = 0.1$	3.49	3.52	1.50	2.60	1.49
$\alpha = 0.05$	4.95	4.99	2.55	4.09	2.59
$\alpha = 0.025$	6.42	6.42	3.71	5.52	3.76
$\alpha = 0.01$	8.30	8.40	5.34	7.41	5.48
$n = 10000$					
$\alpha = 0.1$	3.23	3.41	1.09	1.57	0.50
$\alpha = 0.05$	4.65	4.84	2.22	2.73	1.86
$\alpha = 0.025$	6.06	6.19	3.39	4.01	3.19
$\alpha = 0.01$	7.93	8.16	5.10	5.73	4.94

3.1 *Quantiles*

In Böhning *et al.* ((1994), p. 383), selected simulated quantiles of the null distribution of $2 \ln \lambda_n$ are published. Details of the algorithm for maximizing the likelihood function (starting values, accuracy) are not given. Another program (Böhning and Schlattmann (1997)) uses $\bar{x} \pm .5\theta$ as a starting strategy, which is possible by the scale invariance of the test problem, and a stopping criterion of type "function values". A stopping criterion of type "directional derivative" is built into the C.A.MAN algorithms (Böhning and Schlattmann (1992)). To calculate quantiles of the likelihood ratio tests based on different implementations of the EM algorithm, we simulate the null distribution of $2 \ln \lambda_n$ for $n = 100, 1000, 10000$ and the four possible combinations of starting strategies and stopping criteria described above, using $acc = 10^{-5}$ as level of accuracy. Each distribution is simulated for $\theta = 1, 2, \ldots, 10$, with 10000 replications for each parameter. (There is an exception at $n = 10000$: With the second stopping rule the distribution is simulated only for $\theta = 1, 2, \ldots, 5$, with 2000 replications for each parameter.) Table 1 presents the $1 - \alpha$ quantiles, $\alpha = 0.1, 0.05, 0.025$ and 0.01, averaged over the 10 (5) parameters under consideration, and contrasts them with the quantiles published in Böhning *et al.* (1994).

As it is seen from Table 1, different implementations of the EM algorithm result in different tests. The distribution of the test statistic depends heavily on the starting and stopping strategies chosen.

While the published quantiles are the smallest, the quantiles based on $x_{\min}/x_{\max}$ are much larger and the quantiles based on $\bar{x} \pm .5\theta$ lie in between. Whereas the published quantiles are decreasing as n increases, the quantiles based on $x_{\min}/x_{\max}$ seem, for every α, to be independent of n.

The influence of the stopping rule is obvious when we look at $n = 1000$ and starting value $\bar{x} \pm .5\theta$.

The only set of quantiles which seems to match the published quantiles fairly well occurs at $n = 1000$ with $\bar{x} \pm .5\theta$ and the function values stopping rule. However, we recalculate this set using the accuracy $acc = 10^{-8}$. The result is shown in Table 2. The

Table 2. Quantiles under accuracy change.

α	.1	.05	.025	.01
$acc = 10^{-5}$	1.50	2.55	3.71	5.34
$acc = 10^{-8}$	1.66	3.17	4.85	6.93

Table 3. Size when different quantiles are used.

	$x_{\min}/x_{\max}$ deriv.	published in Böhning *et al.* (1994)
$n = 100$		
$\alpha = 0.1$	.1041	.2366
$\alpha = 0.05$	.0520	.1089
$\alpha = 0.025$	.0264	.0550
$\alpha = 0.01$	.0108	.0252
$n = 1000$		
$\alpha = 0.1$	.0976	.2640
$\alpha = 0.05$	.0486	.1536
$\alpha = 0.025$	.0239	.0866
$\alpha = 0.01$	.0090	.0394
$n = 10000$		
$\alpha = 0.1$	0.0998	0.4166
$\alpha = 0.05$	0.0496	0.2091
$\alpha = 0.025$	0.0248	0.1123
$\alpha = 0.01$	0.0105	0.0466

quantiles based on $acc = 10^{-8}$ are larger, yet still much smaller than the quantiles based on the directional derivatives stopping rule. This situation is analyzed more closely in Seidel *et al.* (1997).

3.2 *Size*

Let us now investigate the size of the test if the "wrong" quantiles are used. Of course, the implementation of the EM algorithm that produces the largest quantiles is the best maximizer of the likelihood. As the largest quantiles are the ones based on $x_{\min}/x_{\max}$ and on a derivatives stopping rule, these are regarded to be "closest to global maximization". (Recall that the denominator of the likelihood ratio test statistic is given explicitly.) Therefore the corresponding test statistic, which maximizes the likelihood with initial value $x_{\min}/x_{\max}$ and a derivatives stopping rule, is considered as the best one. What happens if it is used together with the quantiles obtained from another inplementation of the EM algorithm? To demonstrate this with an example, we use the quantiles published in Böhning *et al.* (1994). The size of this test, i.e. the probability of rejecting H_0, given H_0, is simulated for the published quantiles as well as for the quantiles calculated with the above "best" implementation of the test statistic. The results of the simulation, based on $\theta = 1$ and 10000 replications, are exhibited in Table 3. It is seen that the size of the test can drastically exceed α if the published quantiles are used and the test statistic is calculated with a different implementation of the EM algorithm.

4. Conclusions

When maximizing the likelihood by an iterative algorithm, different starting and stopping strategies yield different ML estimators. In particular, when the test statistic of a likelihood ratio test is determined with an iterative algorithm, each implementation of it defines a different test. Using the example of the EM algorithm in exponential mixture testing, we have demonstrated that these differences cannot be neglected.

In practice, our observation has far-reaching consequences: If simulated quantiles are used, the test statistic must, in every application, be calculated with precisely the same algorithm that has been used in the simulation. Simulated quantiles from the literature cannot be used unless they come with a complete specification of all aspects of the algorithm by which the test statistic has been evaluated.

Even if one believes that the "true" likelihood ratio test is the one based on global maximization (which is in principle possible here since, for a finite mixture of exponentials, the likelihood function is bounded), one has to be aware that the distribution of the test statistic under H_0 depends heavily on the accuracy of the approximation to the global maximum. The reason is that, with a large probability, the difference between the log-likelihoods under the null hypothesis and under the alternative is small; see the Appendix.

In our opinion, one has to decide on the basis of power comparisons which version of the likelihood ratio test is the appropriate one. Preliminary results (Seidel *et al.* (1997)) suggest that global optimization is not necessarily the best strategy: A simple starting strategy, which often fails to find the global maximum under the null hypothesis, results in a rather powerful test. On the other hand, a multiple starting strategy that comes close to global maximization under both the null and the alternative hypotheses leads to inferior power.

Acknowledgements

We thank Dankmar Böhning and Peter Schlattmann who kindly gave us the source codes of their C.A.MAN program and of another unpublished program for simulating quantiles of the likelihood ratio test and provided details of their calculations. Our investigations build on their work. We are also grateful to the referees for their helpful comments concerning the presentation.

Appendix: Random number generation and level of accuracy required

As usual, exponentially distributed random numbers are generated by the inversion method from uniformly distributed random numbers. Let (x_k) and (y_k) be sequences of random numbers from exponential distributions with expectation 1 and θ, respectively. If the same seed is used, $y_k = \theta x_k$ should hold for each k and, according to Subsection 2.2, all simulated quantiles obtained with the EM algorithm should be the same. A trivial consequence is that another seed has to be used if the simulation study is repeated for another value of θ. Consider, however, the following example: Using the same seed, we generate $x_1, \ldots, x_{100}$ for $\theta = 1$ and $y_1, \ldots, y_{100}$ for $\theta = 10$. To compute a value λ^x_{100} and λ^y_{100} of the likelihood ratio test statistic, the EM algorithm is performed on the basis of the second stopping criterion and of the starting values $x_{\min}/x_{\max}$ and $y_{\min}/y_{\max}$, respectively. We obtain $2 \ln \lambda^x_{100} = .18954$ and $2 \ln \lambda^y_{100} = .18964$. These values are slightly different, although they should be exactly equal.

Let us analyze the situation more closely. Denote by $\hat{\theta}_x$, $\hat{P}_x$, $\hat{\theta}_y$, $\hat{P}_y$ the maximum likelihood estimators based on the x-values and the y-values. Observe that $2 \ln \lambda_{100}^{x/y} = 2[l(\hat{P}_{x/y}) - l(\hat{\theta}_{x/y})]$. Then the following values are obtained:

$l(\hat{P}_x) = -121.80106$	$l(\hat{\theta}_x) = -121.89584$
$\hat{P}_x = \begin{bmatrix} .61178 & 1.34480 \\ .13644 & .86356 \end{bmatrix}$	$\hat{\theta}_x = 1.24478$
$l(\hat{P}_y) = -352.05954$	$l(\hat{\theta}_y) = -352.15436$
$\hat{P}_y = \begin{bmatrix} 6.11788 & 13.44798 \\ .13645 & .86355 \end{bmatrix}$	$\hat{\theta}_y = 12.4478$

Obviously the values of $l(\hat{P})$ and $l(\hat{\theta})$ are nearly equal, so that as a result of cancellation small relative errors in $l(\hat{P})$ produce large relative errors in $2 \ln \lambda_{100}$. This situation is very likely to occur under the null hypothesis. Therefore $l(P)$ has to be calculated with a very high accuracy.

REFERENCES

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

Böhning, D. and Schlattmann, P. (1997). Personal communication.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. and Lindsay, B. G. (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. and Rubin, D. B. (1997). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. B*, **39**, 1–38.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*, Chapman and Hall, London.

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

Lindstrom, M. J. and 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.

McLachlan, G. J. and Basford, K. E., (1988). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*, Wiley, New York.

Seidel, W., Mosler, K., Alker, M. and Ruck. A. (1997). Size and power of likelihood ratio tests in exponential mixture models based on different implementations of the EM algorithm, *Discussion Papers in Statistics and Quantitative Economics*, **79**, Universität der Bundeswehr Hamburg.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, Chichester.