Statistical Papers **41**, 85–98 (2000)

Statistical Papers

Likelihood ratio tests based on subglobal optimization: A power comparison in exponential mixture models

Wilfried Seidel, Karl Mosler, and Manfred Alker

Received: December 9, 1997; revised version: November 2, 1998

The paper compares several versions of the likelihood ratio test for exponential homogeneity against mixtures of two exponentials. They are based on different implementations of the likelihood maximization algorithm. We show that global maximization of the likelihood is not appropriate to obtain a good power of the LR test. A simple starting strategy for the EM algorithm, which under the null hypothesis often fails to find the global maximum, 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.

Key words: Maximum likelihood, likelihood ratio tests, mixtures of exponentials, homogeneity tests, EM algorithm.

AMS 1991 *subject classifications.* Primary 62H05; secondary 52A22, 60F05.

1 Introduction

Analysis of mixture distributions is still a challenge for application of numerical methods in statistics. Finite mixtures are important tools for modelling heterogeneous populations that split up into several homogeneous components, where "homogeneity" means that a simple parametric model holds in each component.

Testing for the number of components in a mixture model is an important but difficult problem. Often a likelihood ratio (LR) test is proposed. Its application, however, is not straightforward for two main reasons: The first one is that in mixture models the null distribution of the likelihood ratio test statistic $2 \ln \lambda$ is not easily obtained. The second reason is that in most cases no closed expression for the maximum likelihood estimator exists, so it has to be calculated numerically. Moreover, in mixture models the likelihood function often has multiple local maxima.

In this paper we shall argue that global maximization of the likelihood function is not of primary importance for constructing a powerful test. In general, every maximizing algorithm approaches a local maximum value which is more or less below the global maximum value of the likelihood function. We show that a test based on an algorithm that operates rather badly under the null hypothesis may have much better power than another test based on an algorithm that seems to come close to global optimization.

It follows, first, that an LR test should not be defined in terms of (global) maxima of likelihood functions. Rather the test statistic should be defined in terms of some maximization algorithm and the critical quantiles be calculated by Monte Carlo simulation of the statistic. Second, it is not necessary that the algorithm be a good global optimizer. By construction, any such test guarantees a pregiven level of significance while its power on relevant alternatives has to be evaluated and compared.

In particular, we analyze several likelihood ratio tests for homogeneity in an exponential mixture model, which are based on different versions of the EM algorithm for maximizing the likelihood function. The null hypothesis is that the data follow an exponential distribution with unknown mean. As departures from the null hypothesis of homogeneity we consider mixtures of two exponential distributions.

Under standard assumptions LR tests have nice asymptotic properties. But in the exponential mixture model the usual regularity conditions do not hold to ensure that the LR statistic has an asymptotic chi-squared distribution. An asymptotic distribution of the LR statistic is given by Dacunha-Castelle and Gassiat (1997a). See Maller and Zhou (1996, pp 176 ff), for the special situation when one of two mixing distributions is degenerate.

In this paper we do not persue the asymptotic approach. We rather study the behaviour of LR tests for finite samples and argue that every LR test depends on the specific implementation of the likelihood maximizing algorithm. It comes out that from different starting strategies of the EM algorithm, even for $n = 10\ 000$, very different distributions of the test statistic arise. Therefore it is not feasible to employ the LR asymptotics, (taken for granted that they hold in each of these cases) if n equals 10 000 or less.

For finite samples we will demonstrate that the classic LR test, which is defined by global maximization of the likelihood, has poor power. Subglobal maximization of the likelihood yields a test that is more powerful on a large range of alternatives.

For a general introduction into mixture models, applications and statistical tools for their analysis, the reader is referred to the monographs by Everitt and Hand (1981), Lindsay (1995), McLachlan and Basford (1988) and Titterington, Smith and Makov (1985). The maximum likelihood estimator of the mixing distribution has been investigated by various authors, among them Jewell (1982), and algorithms for calculating maximum likelihood estimates of the parameters of exponential mixtures are discussed by Hasselblad (1982), Kaylan and Harris (1981) and McLachlan (1995). Besides LR testing, there are several other approaches to determine the number of components in a finite mixture model. See Richardson and Green (1997) for Bayesian analysis, Robert (1996) for MCMC inference, Dacunha-Castelle and Gassiat (1997b) for algebraic methods, and Bozdogan (1993) for informational criteria.

2 Testing for exponential mixtures

The mixture of two exponential distributions in the mean value parameterization has density

$$f(x, P) = p\frac{1}{\theta_1}e^{-\frac{x}{\theta_1}} + (1-p)\frac{1}{\theta_2}e^{-\frac{x}{\theta_2}} \quad \text{for } x > 0.$$

The parameter of the mixture is denoted by

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

$\theta_1 > 0$ and $\theta_2 > 0$ are the expectations of the two mixing exponential distributions and p and $1-p$ are the *mixing weights*, $0 \leq p \leq 1$. Let

$x_1, \ldots, x_n$ be the outcome of a random sample of size n with respect to the variable X. We consider the likelihood ratio test for the null hypothesis of *homogeneity*,

$$H_0 : X \text{ has density } \frac{1}{\theta} e^{-\frac{x}{\theta}} \text{ for all } x > 0 \text{ and some } \theta \,,$$

against the alternative that the distribution of X is a proper mixture of two exponential distributions,

$$H_1 : X \sim f(x, P) \text{ with } \theta_1 \neq \theta_2 \text{ and } 0 < p < 1 \,.$$

Let $f(x_1, \ldots, x_n, \theta)$ and $f(x_1, \ldots, x_n, P)$ denote the likelihood of the sample under the null hypothesis and under the alternative.

Under H_0, the maximum likelihood estimate $\hat{\theta}$ of θ is given by $\hat{\theta} = \overline{x}$. A maximum likelihood estimate $\hat{P}$ of P is defined as a parameter value $\hat{P}$ which maximizes $f(x_1, \ldots, x_n, P)$ or, equivalently, the log-likelihood-function

$$l(P) = \ln f(x_1, \ldots, x_n, P) = \sum_{i=1}^{n} \ln f(x_i, P).$$

The likelihood ratio test is based on the test statistic

$$\lambda_n = \frac{f(x_1, \ldots, x_n, \hat{P})}{f(x_1, \ldots, x_n, \hat{\theta})} \,. \tag{2.1}$$

For better comparison with the literature, we consider the logarithm of the test statistic,

$$2 \ln \lambda_n = 2\,[l(\hat{P}) - l(\hat{\theta})]. \tag{2.2}$$

Note that the denominator of the likelihood ratio test statistic is explicitly given while, in every concrete application, the numerator has to be evaluated by some numerical procedure.

Under H_0 the distribution of $2 \ln \lambda_n$ is independent of θ. Therefore it can be simulated once in order to provide tables of the corresponding critical values.

3 Maximizing the likelihood function

To calculate the LR statistic (2.1) or its logarithm (2.2), the likelihood has to be maximized numerically under the alternative hypothesis. This is often done by the EM algorithm (Dempster, Laird and Rubin, 1977, McLachlan and Krishnan, 1997), which in our situation takes the following form: Choose some initial parameter P^0 and calculate the sequence

$$P^k = \begin{bmatrix} \theta_1^k & \theta_2^k \\ p^k & 1-p^k \end{bmatrix}, \qquad k = 1, 2, 3, \dots,$$

by

$$p^{k+1} = p^k \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i, \theta_1^k)}{f(x_i, P^k)}, \qquad \theta_j^{k+1} = \frac{\sum_{i=1}^{n} \frac{x_i f(x_i, \theta_j^k)}{f(x_i, P^k)}}{\sum_{i=1}^{n} \frac{f(x_i, \theta_j^k)}{f(x_i, P^k)}}.$$

As we shall see below, the likelihood function under H_1, which depends on the three parameters θ_1, θ_2 and p, is bounded, but rather flat, so that its maximization comes out to be a nontrivial numerical problem. In addition, the behaviour of the EM algorithm depends, contrary to most other situations reported in the literature, on the choice of initial values. Also the stopping criterion has to be carefully selected in order not to stop early. We shall conclude that different implementations of the EM algorithm, regarding the starting and stopping strategy, yield different test statistics, which have different distributions under the null hypothesis as well as under the alternative.

In the sequel we consider three such LR tests that are based on different implementations of the EM algorithm.

In order to describe our stopping criterion, let

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

It can be shown that $D_P(\theta)$ has the properties of a *directional derivative* of $l(P)$ in the direction of θ. In all three cases the EM algorithm is stopped at

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

if

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

We note that this stopping criterion is scale invariant in the probability law.

In the first test a starting strategy, abbreviated x_{min}/x_{max}, is employed,

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

The second test uses a starting strategy, which is abbreviated by $.5\overline{x}/1.5\overline{x}$,

$$.5\overline{x}/1.5\overline{x}: \qquad P^0 = \begin{bmatrix} .5\,\overline{x} & 1.5\overline{x} \\ .5 & .5 \end{bmatrix}.$$

Observe that these strategies start the algorithm 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.

These two implementations of the EM algorithm often produce different local maxima of the likelihood and, therefore, different test statistics. In another implementation, which yields our third test statistic, we use a multi-start strategy.

We start the EM algorithm at 55 initial values with equal mixing weights as follows: Let $u_0 = x_{(1)}$, $u_1 = x_{(20)}$, $u_2 = x_{(40)}$, $\ldots$, $u_{10} = x_{(200)}$, where $x_{(i)}$ denotes the i-th order statistic. Then the initial values are given by

$$P_{i,j}^0 = \begin{bmatrix} u_i & u_j \\ .5 & .5 \end{bmatrix}, \; 0 \leq i < j \leq 10\,.$$

These are 55 different values. The initial mixing weights are kept at the common value .5, since their choice seems to be less crucial than the choice of the parameters θ_i. Note that the log-likelihood is strictly concave in the mixing weight p. (Strictly speaking, concavity in p only implies that, if two local maxima have the same values of θ_1 and θ_2, they also have the same value of p. It does *not* imply that, if two starting values differ only in p, the algorithm converges to the same local maximum.)

It is easy to see that, for each of the three implementations of the EM algorithm, the null distribution of the test statistic $2\ln\lambda_n$ does not depend on the parameter θ.

4 Test quantiles

We simulate the null distribution of $2\ln\lambda_n$, for the two starting strategies x_{min}/x_{max} and $.5\overline{x}/1.5\overline{x}$, using $acc = 10^{-5}$ as level of accuracy and 100 000 replications. The $1-\alpha$ quantiles, $\alpha = 0.1, 0.05, 0.025$ and 0.01, for $n = 200$ are found in Table 1.

Table 1. Quantiles and starting strategies ($n = 200$)				
α	0.1	0.05	0.025	0.01
$.5\,\overline{x}/1.5\,\overline{x}$	2.51	3.90	5.22	7.06
x_{min}/x_{max}	3.48	4.90	6.35	8.24
multi-start	3.98	5.42	6.80	8.73

From Table 1 it is obvious that the different implementations of the EM algorithm result in different tests. Observe that the quantiles based on x_{min}/x_{max} are usually larger than the quantiles based on $.5\overline{x}/1.5\overline{x}$. Of course, the larger the quantiles, the closer comes the likelihood maximization to the global maximum.

We also calculate these quantiles for other samples sizes n between 100 and 10 000 and observe similar results. Moreover, the quantiles based on x_{min}/x_{max} seem, for every α, to be independent of n.

For the multi-start implementation of the EM algorithm we simulate the quantiles of the null distribution of $2\ln\lambda_n$ for n = 200, using 20 000 replications. This multi-start implementation is expected to come close to global maximization of the likelihood and therefore to yield larger quantiles for the test statistic than the two previous implementations, which obviously are no global maximizers. The results are exhibited in the last row of Table 1. As expected, these quantiles are larger than the quantiles of the first two tests.

Let us try to explain the surprising difference between x_{min}/x_{max} and $.5\overline{x}/1.5\overline{x}$. Given a sample $x_1, \ldots, x_n$, consider $\hat{\theta} = \bar{x}$ as above and the log-likelihood difference

$$ldiff(P) = 2\,[l(P) - l(\hat{\theta})]$$

depending on P. Let $\hat{P}(x_{min}/x_{max})$ and $\hat{P}(.5\overline{x}/1.5\overline{x})$ denote the estimates found by the starting strategies x_{min}/x_{max} and $.5\overline{x}/1.5\overline{x}$, respectively. If the sample is simulated under the null hypothesis, we often observe that $ldiff(\hat{P}(x_{min}/x_{max}))$ is larger than $ldiff(\hat{P}(.5\overline{x}/1.5\overline{x}))$.

For a particular (simulated) sample we get the parameter estimates $\hat{\theta} = \overline{x} = 0.9118$ and

$$\hat{P}(.5\overline{x}/1.5\overline{x}) = \begin{bmatrix} 0.8916 & 0.9285 \\ 0.4527 & 0.5473 \end{bmatrix}$$

with the log-likelihood difference $ldiff(\hat{P}(.5\overline{x}/1.5\overline{x})) = -0.0104$. The log-likelihood difference of the other starting strategy amounts to

$ldiff(\hat{P}(x_{min}/x_{max})) = 4.6602$. Figure 1 shows the graph of the function $2[l^*(\theta_1,\theta_2) - l(\hat{\theta})]$, where

$$l^*(\theta_1,\theta_2) = \max_p \, l\left(\begin{bmatrix}\theta_1 & \theta_2 \\ p & 1-p\end{bmatrix}\right).$$

The large figure exhibits the graph for $0.0 \le \theta_j \le 2.0$, $j = 1,2$. If we look at this graph, the function seems to be well behaved, with a global maximum near $\theta_1 = \theta_2 = \hat{\theta}$. Note that $2[l^*(\hat{\theta},\hat{\theta}) - l(\hat{\theta})] = 0$. The reason why the EM algorithm does not find this maximum appears to be the flatness of the likelihood function rather than its multimodality.

The graph in the large figure is evaluated at an equidistant grid of 101^2 points in the square $[0.0, 2.0] \times [0.0, 2.0]$. But this grid is too coarse to provide the correct picture. The small figure enlarges a small section of the graph, which now is evaluated at an equidistant grid of 101^2 points in the rectangle $[0.0000, .0048] \times [0.8, 1.1]$. The rectangle is so small that it has no intersection with the original grid. The enlargement reveals a much larger maximum. By the x_{min}/x_{max} starting strategy this maximum is found at

$$\hat{P}(x_{min}/x_{max}) = \begin{bmatrix}0.0004 & 0.9197 \\ 0.009 & 0.991\end{bmatrix}.$$

To analyze the situation more closely, let us compare the contributions of single observations to the likelihood of the sample. The ordered sample of our numerical example reads $(x_{(1)}, x_{(2)}, \ldots, x_{(200)}) = (0.000256, 0.000640, 0.008298, 0.008725, \ldots, 4.7614)$. We observe that $f(x_{(i)}, \hat{P}(x_{min}/x_{max}))$ is much larger than $f(x_{(i)}, \hat{P}(.5\overline{x}/1.5\overline{x}))$ for $i = 1$ and $i = 2$, whereas for $i \ge 3$ the two contributions to the likelihood have similar size.

Often, under the null hypothesis, the lowest observations are very small. In these cases an estimate that builds a cluster from a few very small observations obtains a relatively large likelihood. Such an estimate corresponds to a small value of θ_1 and a small value of p. It is found, usually, by starting at x_{min}/x_{max} but not at $.5\overline{x}/1.5\overline{x}$.

We finally note that the observed effects are not bound to the strategy x_{min}/x_{max}, which might be considered as extreme or non-robust. In the present example with the starting strategy $x_{(5)}/x_{(196)}$ in place of x_{min}/x_{max} the same maximum is found, whereas with the starting strategy $x_{(10)}/x_{(191)}$ a local maximum near $\hat{P}(.5\overline{x}/1.5\overline{x}))$ is obtained.

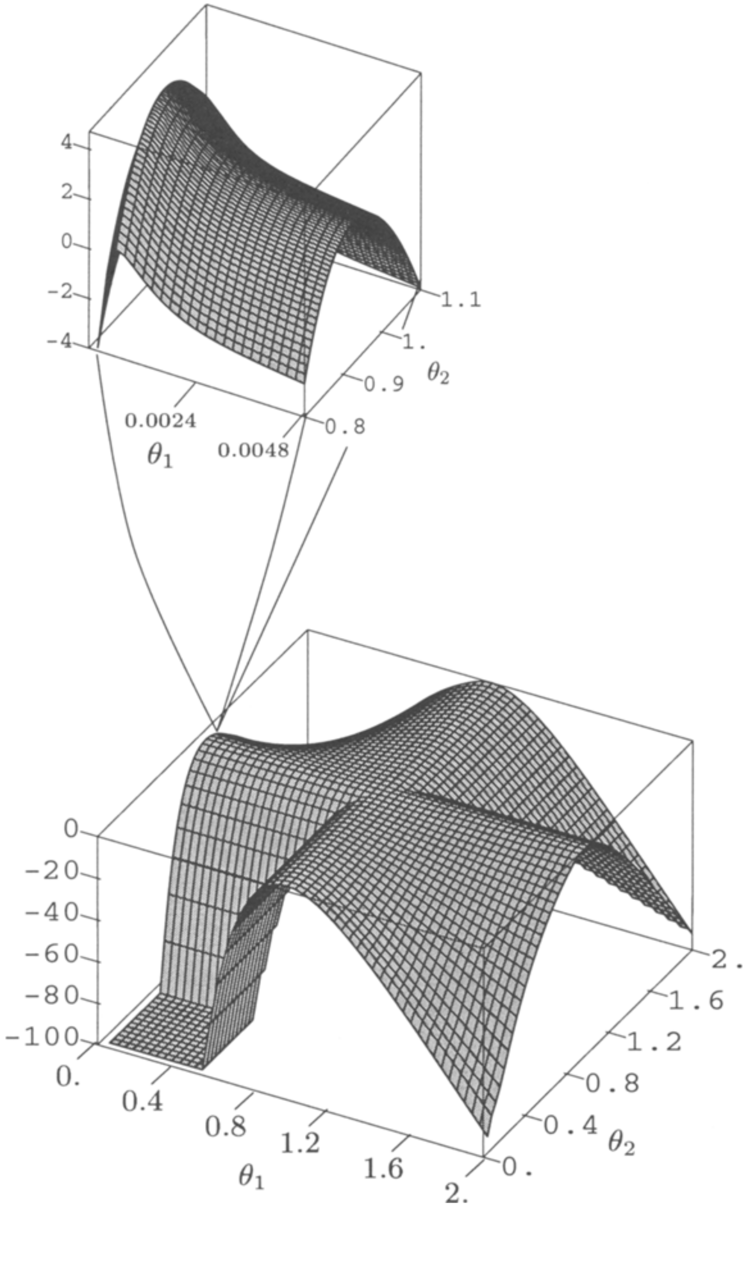

Figure 1 : Log-likelihood difference

5 Power of the three tests

In the power study, for a sample size of n = 200, we compare the three tests under various two-point mixture alternatives. In Table 3, the simulated quantiles for n = 200 are listed for the three tests. The power of the first two tests is simulated, using 10 000 replications, for each of the parameters

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

with $\theta_1 = 1$, p = 0.1, 0.2, ..., 0.9 and $\theta_2 = 0.1, 0.14, 0.25, 0.33, 0.5, 0.67, 0.8$, and significance levels $\alpha = 0.1, 0.05, 0.01$. Figures 2 – 6 show the power depending on θ_2 for significance level $\alpha = 0.05$ and selected values of the mixing proportion p. The power functions for other significance levels are very similar.

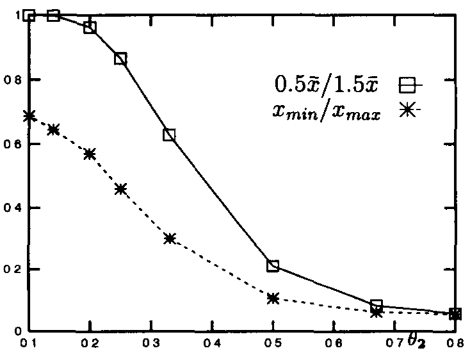

Figure 2: Power at 0.05 level of significance, $\theta_1 = 1$, mixing proportion $p = 0.1$

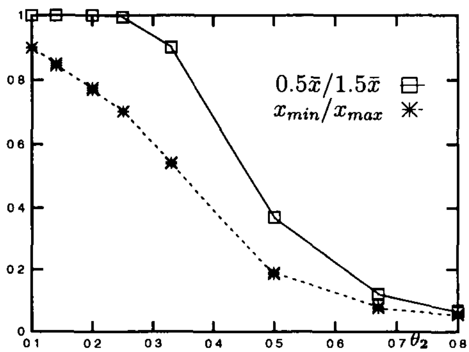

Figure 3: Power at 0.05 level of significance, $\theta_1 = 1$, mixing proportion $p = 0.3$

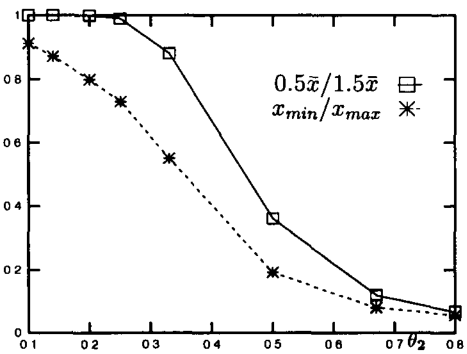

Figure 4: Power at 0.05 level of significance, $\theta_1 = 1$, mixing proportion $p = 0.5$

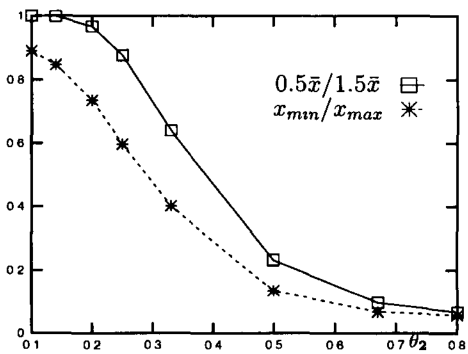

Figure 5: Power at 0.05 level of significance, $\theta_1 = 1$, mixing proportion $p = 0.7$

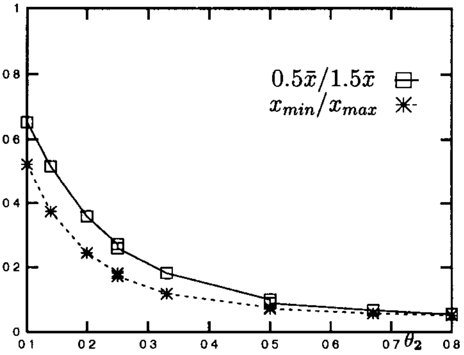

Figure 6: Power at 0.05 level of significance, $\theta_1 = 1$, mixing proportion $p = 0.9$

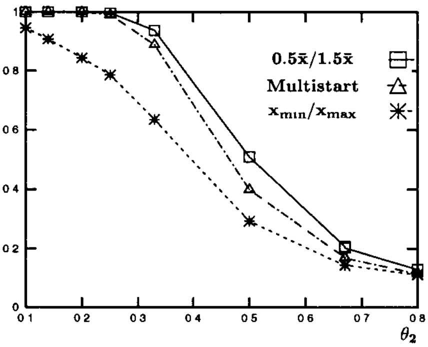

Figure 7: Power, $\alpha = 0.1$

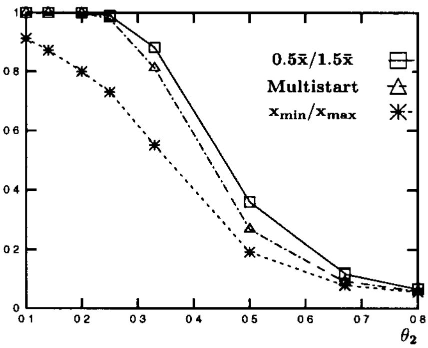

Figure 8: Power, $\alpha = 0.05$

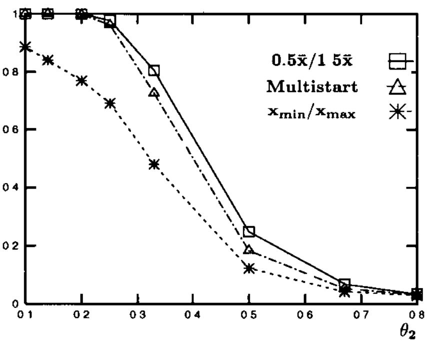

Figure 9: Power, $\alpha = 0.025$

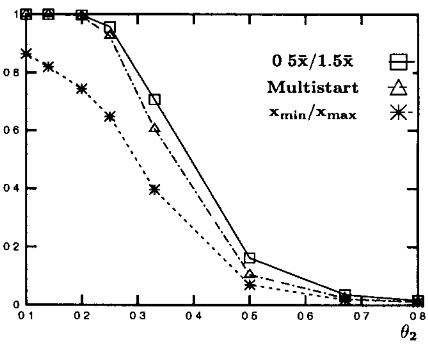

Figure 10: Power, $\alpha = \alpha = 0.01$

We also simulate the power of the third test based on the multi-start implementation for $\theta_1 = 1.0$, $p_1 = p_2 = .5$ and several values of θ_2. Figures 7 – 10 show the results, compared to the power of the first two tests. Surprisingly the power of "global optimization" is smaller than the power of the test based on $.5\,\overline{x}/1.5\,\overline{x}$.

A closer analysis reveals the following reason for the observed behaviour: Under H_0, the strategy $.5\,\overline{x}/1.5\,\overline{x}$ is inferior to the other two strategies in maximizing the likelihood (yet *not* inferior in estimating the parameter!), therefore the quantiles of the test statistic based on $.5\,\overline{x}/1.5\,\overline{x}$ are rather low. Under the alternative, however, $.5\,\overline{x}/1.5\,\overline{x}$ is not much worse than multi-start, and in a considerable number of cases it is (much) better than x_{min}/x_{max}. In several cases, the latter estimates a mixing component $\theta_1 \approx x_{min}$ with a very small mixing weight, which may give a higher likelihood than other parameter estimates under H_0, but a smaller one under H_1.

6 Conclusions

We have shown that the performance of the LR test for exponential homogeneity against mixtures of two exponentials depends heavily on the particularities of the likelihood maximization algorithm employed.

Moreover an algorithm that is rather poor in terms of approaching the global maximum under the null hypothesis, has proven to produce better empirical power of the test, on a large range of alternatives, than the algorithm that comes close to global maximization. This holds even for very large samples (n= 10 000). Subglobal maximization of the likelihood appears to yield more power than the classic LR test, which is defined by global maximization of the likelihood. Our results are based on an extensive computational experience and have strong consequences on the practice of LR testing. But still a theory is missing that could explain these effects in a satisfactory way.

Although we cannot be sure that our multi-start strategy comes sufficiently close to the global maximum in all situations, the following suggestion may be justified. When constructing an LR test, the primary goal should not consist in global maximization of the likelihood but rather in having a simple and well-defined subglobal maximization strategy that produces good power on the relevant alternatives. The latter has to be secured by proper simulation studies.

The simulation load may be lightened by using proper accelerated versions of the EM algorithm. As the parameter estimates are of no interest for the test, we tried a simple version of the Aitken acceleration that approximates only the function value of the local maximum of the likelihood, see e.g. Böhning *et al.* (1994, pp 386 ff). Unfortunately it turned out that this approach leads to instabilities; the quantiles often seem to be drastically overestimated. Future work has to show whether more sophisticated approaches (Meilijson, 1989, and others) yield better results.

Clearly, for every test problem, a lot of work has still to be done to identify likelihood maximization strategies that yield satisfactory power of the test against relevant alternatives and to eliminate strategies that produce poor power.

Acknowledgements

The authors would like to thank the referees for their constructive suggestions. This research has been partially sponsored (M. Alker) by a grant from the German Research Foundation.

References

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., Lindsay, B. G. (1994). The Distribution of the Likelihood Ratio for Mixtures of Densities from the One-Parameter Exponential Family. *Ann. Institute Statistical Mathematics 46*, 373-388.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse Fisher information matrix, pp 40–54 in *Information and Classification*, (ed. O. Opitz, B. Lausen, R. Klar). Berlin: Springer Verlag.

Dacunha-Castelle, D., Gassiat, E. (1997a). The estimation of the order of a mixture model, *Bernoulli 3*, 279–299.

Dacunha-Castelle, D., Gassiat, E. (1997b). Testing in locally conic models, and application to mixture models, *ESAIM: Probability and Statistics 1*, 285–317.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Royal Statistical Soc. B 39*, 1–38.

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

Hasselblad, V. (1982). Estimation of finite mixtures of distributions from the exponential family. *J. American Statistical Association* *64*, 1459–1471.

Jewell, N.P. (1982). Mixtures of exponential distributions. *Ann. Statistics* *10*, 479–484.

Kaylan, A.R., Harris, C.M. (1981). Efficient algorithms to derive maximum likelihood estimates for finite exponential and Weibull distributions. *Computers and Operations Research* *8*, 97–104.

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

Maller, R.A., Zhou, X. (1996). *Survival Analysis with Long-Time Survivors.* Chichester: J. Wiley.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal Royal Statistical Soc. B 51*, 127–138.

McLachlan, G. J. (1995). Mixtures - Models and Applications, chapter 19 in *The Exponential Distribution* (eds. N. Balakrishnan, A.P. Basu). Amsterdam: Gordon and Breach.

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

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

Richardson, S., Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statistical Soc. B 59*, 731–792.

Robert, C. (1996). Mixtures of distributions: Inference and estimation, pp 441–464 in *Practical Markov Chain Monte Carlo,* (eds. W.R. Gilks, S. Richardson, D.J. Spiegelhalter). London: Chapman and Hall.

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

Wilfried Seidel
Manfred Alker
Universität der Bundeswehr Hamburg
22039 Hamburg, Germany

Karl Mosler
Universität zu Köln
50923 Köln, Germany