Ann. Inst. Statist. Math.
Vol. 46, No. 2, 373-388 (1994)

THE DISTRIBUTION OF THE LIKELIHOOD RATIO FOR MIXTURES OF DENSITIES FROM THE ONE-PARAMETER EXPONENTIAL FAMILY

DANKMAR BÖHNING[1], EKKEHART DIETZ[1], RAINER SCHAUB[1], PETER SCHLATTMANN[1] AND BRUCE G. LINDSAY[2]

[1] *Department of Epidemiology, Free University Berlin, Augustastr. 37, 12 203 Berlin, Germany*
[2] *Department of Statistics, The Pennsylvania State University, 422 Classroom Building, University Park, PA 16802, U.S.A.*

(Received November 16, 1992; revised November 29, 1993)

Abstract. We here consider testing the hypothesis of *homogeneity* against the alternative of a two-component mixture of densities. The paper focuses on the asymptotic null distribution of $2\log\lambda_n$, where λ_n is the likelihood ratio statistic. The main result, obtained by simulation, is that its limiting distribution appears pivotal (in the sense of constant percentiles over the unknown parameter), but model specific (differs if the model is changed from Poisson to normal, say), and is not at all well approximated by the conventional $\chi^2_{(2)}$-distribution obtained by counting parameters. In Section 3, the binomial with sample size parameter 2 is considered. Via a simple geometric characterization the case for which the likelihood ratio is 1 can easily be identified and the corresponding probability is found. Closed form expressions for the likelihood ratio λ_n are possible and the asymptotic distribution of $2\log\lambda_n$ is shown to be the mixture giving equal weights to the one point distribution with all its mass equal to zero and the χ^2-distribution with 1 degree of freedom. A similar result is reached in Section 4 for the Poisson with a small parameter value ($\theta \leq 0.1$), although the geometric characterization is different. In Section 5 we consider the Poisson case in full generality. There is still a positive asymptotic probability that the likelihood ratio is 1. The upper precentiles of the null distribution of $2\log\lambda_n$ are found by simulation for various populations and shown to be nearly independent of the population parameter, and approximately equal to the $(1-2\alpha)100$ percentiles of $\chi^2_{(1)}$. In Sections 6 and 7, we close with a study of two continuous densities, the *exponential* and the *normal with known variance*. In these models the asymptotic distribution of $2\log\lambda_n$ is pivotal. Selected $(1-\alpha)100$ percentiles are presented and shown to differ between the two models.

Key words and phrases: Aitken acceleration, boundary problem, mixtures, asymptotic distribution of likelihood ratio.

1. Introduction

Mixture models arise as a simple and natural way to model population heterogeneity. Suppose that the population consists of k homogeneous subgroups or component populations (which we will simply call *components*). A simple parametric model, such as the Poisson or Binomial, is then assumed to hold in each component. For a general introduction into mixture models see Titterington *et al.* (1985) or McLachlan and Basford (1988). Mixture models can be viewed as the *semiparametric* compromise between a fully parametric model such as the homogeneous Poisson, which often forces too much structure to the data leading to problems such as *overdispersion*, and a *nonparametric* model, which—though avoiding strong structural assumptions—experience other disadvantages including high data dependency of model estimates.

Formally, let $f(x, \theta_j)$ be the probability density function for observation X, when sampled from the j-th component. Suppose further that the j-th component is a fraction p_j of the total population, with $p_1 + \cdots + p_k = 1$. Assuming that one samples from the entire population, without knowledge of component membership, then the observation X has the *mixture density*

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

where the unknown parameter vector P consists of k component parameters $\theta_1, \ldots, \theta_k$ and k component proportions $p_1, \ldots, p_k$ which is written as:

$$P = \begin{bmatrix} \theta_1, \ldots, \theta_k \\ p_1, \ldots, p_k \end{bmatrix}.$$

Typically, P is estimated by the method of maximum likelihood. A maximum likelihood estimator $\hat{P}$ of P is defined as the probability measure $\hat{P}$ (assigning mass p_j to support point θ_j) which maximizes the log-likelihood function

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

This paper focuses on the asymptotic distribution of the likelihood-ratio statistic

$$2 \log \lambda_n = 2[l(\hat{P}) - l(\hat{\theta})]$$

where $\hat{\theta}$ is the maximum likelihood estimator under the null hypothesis of homogeneity $H_0 : k = 1$ whereas $\hat{P}$ is the maximum likelihood estimator under the alternative of a two-component model: $H_1 : k = 2$.

2. Some fallacies

The log-likelihood ratio test statistic $2\log\lambda_n$ generally has an asymptotic $\chi^2_{(d)}$-distribution, where the degrees of freedom, d, equal the difference between the number of parameters under the alternative and null hypothesis (Cox and Hinkley (1974), p. 323). In the case of univariate θ, this would imply a $\chi^2_{(2)}$-distribution for the test of H_0 and H_1 above. However, this theory is known to fail for the mixture problem (Titterington *et al.* (1985), p. 154). The explanation is that the null hypothesis does not lie in the interior of the parameter space. In fact, our simulation results will show that the large sample distribution in a number of models appears more like a mixture of $\chi^2_{(2)}$, $\chi^2_{(1)}$ and $\chi^2_{(0)}$ (degenerate at 0) distributions, where the proportions depend on the model under consideration.

Recently, Goffinet *et al.* (1992) found the exact limiting distribution in several problems involving the normal distribution, but under the assumption that p_1 and p_2 are known under the alternative hypothesis. In general, however, there is little known other than that the standard theory does not apply.

In statistical applications of mixture models the problem is still often ignored. An example is given by Gibbons *et al.* (1990) who use a mixture of Poissons in modelling suicide surveillance. They argue that a χ^2 distribution with 1 df for $2\log\lambda_n$ can be used for testing a one component against a two component model if n is beyond 20 or 30, although they point out in the same paper that the boundary condition is violated.

A systematic investigation *by simulation* of the distribution of $2\log\lambda_n$ for various densities has not been done until very recently. Thode *et al.* (1988) study the case of a normal with an *additional* free and common variance parameter. They conclude that the distribution of the likelihood ratio statistic is asymptotically χ^2 with 2 df, although convergence is rather slow. In Mendell *et al.* (1991) the asymptotic distribution of $2\log\lambda_n$ is studied *under the alternative hypothesis*. It is conjectured that the asymptotic distribution could be noncentral χ^2, possibly with 2 df.

In this paper we will concentrate on examples from the one parameter exponential family.

3. The binomial Bi(m,θ) for $m = 2$

Suppose that $X_1,\ldots,X_n$ are a random sample in which each X_i is binomial Bi$(2,\theta)$. We consider this simple example since by a geometric analysis we are able to characterize exactly the structure of the likelihood ratio test. If we record $Y_0 = \#$ zeros, $Y_1 = \#$ ones, and $Y_2 = \#$ twos, then $(Y_0, Y_1, Y_2)^T$ has a multinomial distribution with probabilities $(\alpha_0,\alpha_1,\alpha_2)^T = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)^T$. This vector is in the probability simplex $\{(\alpha_0,\alpha_1,\alpha_2)^T \mid \alpha_i \geq 0, \alpha_0+\alpha_1+\alpha_2 = 1\}$. We can graphically reproduce this simplex in two dimensions by omitting the inessential last coordinate, giving us $\Sigma_2 = \{(\alpha_0,\alpha_1)^T \mid \alpha_0 \geq 0, \alpha_1 \geq 0, \alpha_0+\alpha_1 \leq 1\}$. See Fig. 1. The set of binomial probabilities $\Gamma = \{(\theta^2, 2\theta(1-\theta))^T \mid \theta \in (0,1)\}$ form a curve which connects the vertices $(0,0)^T$ and $(1,0)^T$. *This curve represents the multinomial probabilities allowable under the null hypothesis of* $k = 1$. We can identify each value of θ with a point on this curve.

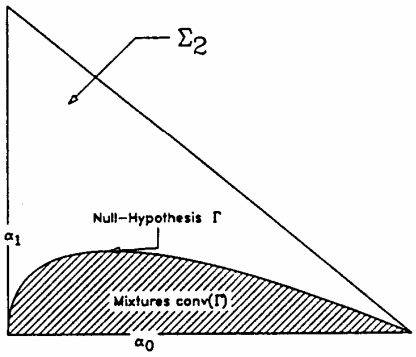

Fig. 1. Null-hypothesis and alternative in the case of the binomial $(\theta, 2)$.

The alternative hypothesis consists of densities having the form $p_1 f(x, \theta_1) + p_2 f(x, \theta_2)$, with $p_1 + p_2 = 1$. In our plot (Fig. 1), such a density corresponds to a convex combination of the two points on Γ corresponding to θ_1 and θ_2, with weights p_1 and p_2. Thus, it is clear that in this case, the alternative hypothesis yields as multinomial probabilities the entire convex hull of Γ, the shaded portion of Fig. 1. As Fig. 1 nicely shows, *the null hypothesis is part of the boundary of the alternative*. In this problem, we can *analytically* describe $2 \log \lambda_n$ as follows. If we place no restrictions on $(\alpha_0, \alpha_1, \alpha_2)^T$, then the maximum likelihood estimator is $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)^T = (Y_0, Y_1, Y_2)^T/n$, and so corresponds to a point $(\hat{\alpha}_0, \hat{\alpha}_1)^T$ in Σ_2. The maximum likelihood estimator under $H_1 : k = 2$ must correspond to a point in the convex hull conv(Γ) of Γ. First, if $(\hat{\alpha}_0, \hat{\alpha}_1)^T$ is in conv(Γ), then this point must also be the maximum likelihood point under H_1. On the other hand, if $(\hat{\alpha}_0, \hat{\alpha}_1)^T \notin \text{conv}(\Gamma)$, then it can be shown that the maximum likelihood estimator under H_1 is on the boundary of conv(Γ), hence in Γ, and so corresponds also to the maximum likelihood estimator under H_0, $\hat{\theta} = (Y_1 + 2Y_2)/n$. In this case, $2 \log \lambda_n$ is zero.

What still remains to be answered is the question: what is the probability that $\lambda_n = 1$? According to the above remarks this is equivalent to $\hat{\alpha}$ lying above Γ or, $\hat{\alpha}_1 \geq 2\sqrt{\hat{\alpha}_0}(1 - \sqrt{\hat{\alpha}_0})$ or, $y_1 \geq 2\sqrt{y_0}(\sqrt{n} - \sqrt{y_0})$. Since $n\hat{\alpha}$ has the multinomial density

$$\binom{n}{y_0 y_1 y_2} \theta^{2y_0} (2\theta(1-\theta))^{y_1} (1-\theta)^{2y_2},$$

this probability can be computed as

$$\begin{aligned} (3.1) \qquad \xi_n(\theta) &:= \Pr(\lambda_n = 1) \\ &= \sum_{y_1 \geq 2\sqrt{y_0}(\sqrt{n} - \sqrt{y_0})} \binom{n}{y_0 y_1 y_2} \theta^{2y_0} (2\theta(1-\theta))^{y_1} (1-\theta)^{2y_2}. \end{aligned}$$

Because of the asymptotic normality of $\hat{\alpha}$ we expect $\xi_n(\theta)$ to converge to $\frac{1}{2}$. But we can even say more about our expectations: because of the convex curvature

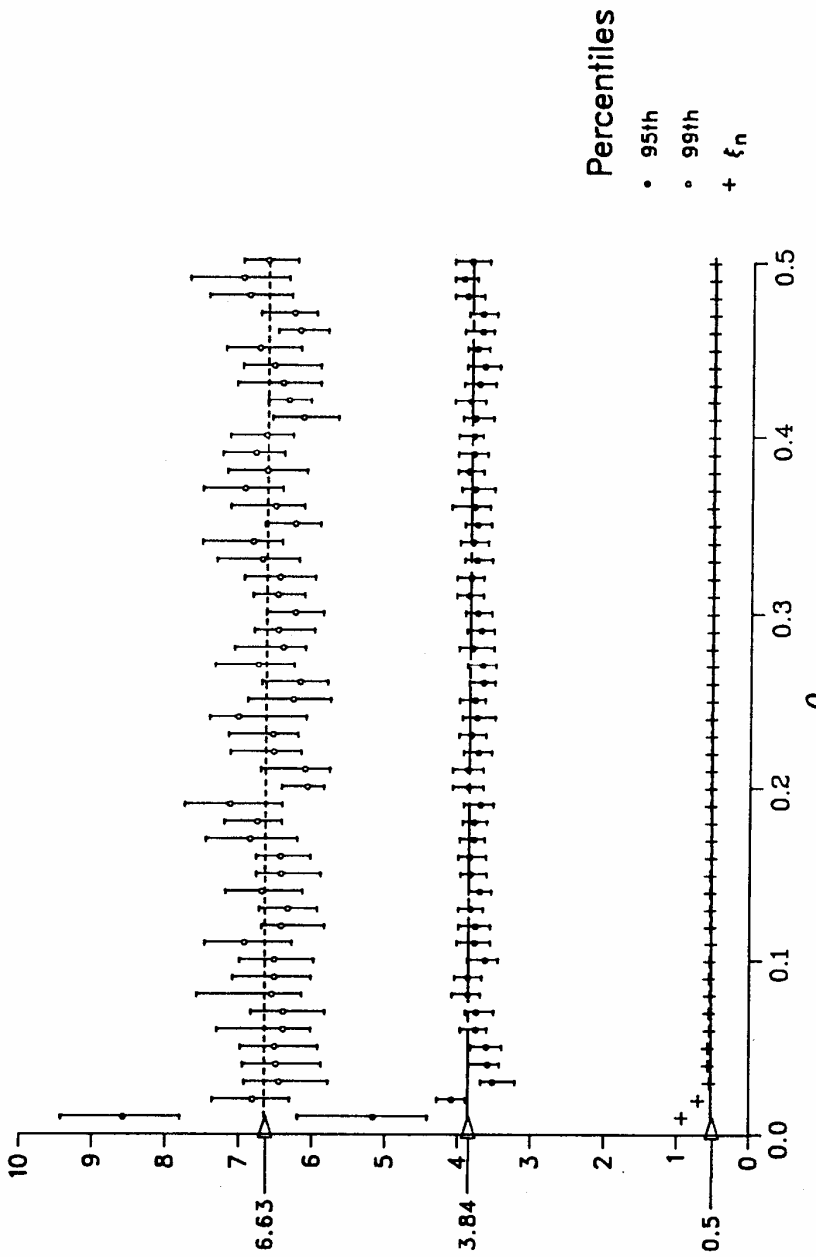

Fig. 2. Binomial $(2, \theta)$: ξ_n and percentiles of $2 \log \lambda_n$ given that $\lambda_n > 1$ for $n = 1000$.

of Γ we might expect that convergence is from above. In principle, the distribution of $2\log\lambda_n$ conditional that $\lambda_n \neq 1$ can be found in a similar way via the multinomial distribution. However, since for large n the multinomial coefficients are expensive to compute we have simulated the conditional distribution of $2\log\lambda_n$. In Fig. 2, an estimate of $\xi_n(\theta)$ is shown for $n = 1000$, and a nonparametric estimate with 95%-confidence interval of $\Phi_{\lambda_n}^{-1}(.95)$ and $\Phi_{\lambda_n}^{-1}(.99)$ is presented. Φ_{λ_n} is the conditional distribution function of $2\log\lambda_n$: $\Phi_{\lambda_n}(x) = \Pr\{2\log\lambda_n \leq x \mid \lambda_n > 1\}$. The confidence intervals were constructed in the usual manner using the normal approximation to the binomial. The estimate is based on a replication size of 10000. The solid lines in Fig. 2 correspond to the 95th and 99th percentile of the χ^2-distribution with 1 df. For θ-values larger than 0.05, the χ^2-approximation appears to be rather satisfactory. Therefore, our analysis suggests that in this case the (unconditional) asymptotic distribution of $2\log\lambda_n$ is:

$$0.5\chi^2_{(0)} + 0.5\chi^2_{(1)},$$

where $\chi^2_{(0)}$ is the distribution with all its mass at zero. If we use the above geometric description and apply the asymptotic theory of Self and Liang (1987), then this example fits exactly in their "Case 5".

4. The Poisson Po(θ) for small θ

Here we consider $f(x,\theta) = \exp(-\theta)\theta^x/x!$, the Poisson distribution. However, we restrict θ to small values, in the interval $[0, 0.1]$, say. Motivation for this assumption lies in the fact that then $\Pr(X > 2) \approx 0$. Therefore, we can undertake an analysis similar to Section 3, since the nonzero Poisson probabilities $[f(0,\theta), f(1,\theta), f(2,\theta)]^T$ define a curve Γ in the two-dimensional simplex $\{\alpha \mid \alpha_i \geq 0$, for $i = 0,1,2$, and $\alpha_0 + \alpha_1 + \alpha_2 = 1\}$. Again, we consider only the first two coordinates: $\Gamma = \{(f(0,\theta), f(1,\theta))^T \mid \theta \in [0, 0.1]\} \subset \Sigma_2$. Figure 3 demonstrates the geometry of the situation. Again, the null hypothesis is part of the boundary of the alternative. The complication is very similiar to Section 3. What is different here is the way the event "$\hat{\alpha}$ is above Γ" is determined. We can write $\Gamma = \{\exp(-\theta)(1,\theta)^T \mid \theta \in [0, 0.1]\}$. Thus the event "$\hat{\alpha}$ is above Γ" is equivalent to $\hat{\alpha}_1 \geq \hat{\alpha}_0[-\log\hat{\alpha}_0]$ or, $y_1/y_0 \geq -\log(y_0/n)$ and can be computed via

$$\begin{aligned}(4.1)\qquad \xi_n(\theta) &:= \Pr(\lambda_n = 1)\\ &= \sum_{y_1/y_0 \geq -\log(y_0/n)} \binom{n}{y_0 y_1 y_2} f(0,\theta)^{y_0} f(1,\theta)^{y_1} f(2,\theta)^{y_2},\end{aligned}$$

the multinomial distribution. However, since the multinomial coefficients become expensive to compute for large n, we use simulation to find $\xi_n(\theta)$ and $\Phi_{\lambda_n}(x)$ for $\theta = 0.01, 0.02, \ldots, 0.10$. The replication size is again 10000. Estimates of $\xi_n(\theta)$ and $\Phi_{\lambda_n}^{-1}(.95)$, $\Phi_{\lambda_n}^{-1}(.99)$ are shown in Figs. 4 and 5 for $n = 1000$ and $n = 10000$. The asymptotic distribution of $2\log\lambda_n$ appears to be the mixture $0.5\chi^2_{(0)} + 0.5\chi^2_{(1)}$ again, although convergence is slow for small θ-values. Just as in the previous

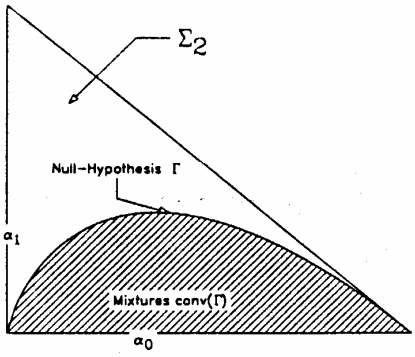

Fig. 3. Null-hypothesis and alternative in the case of the Poisson with small θ.

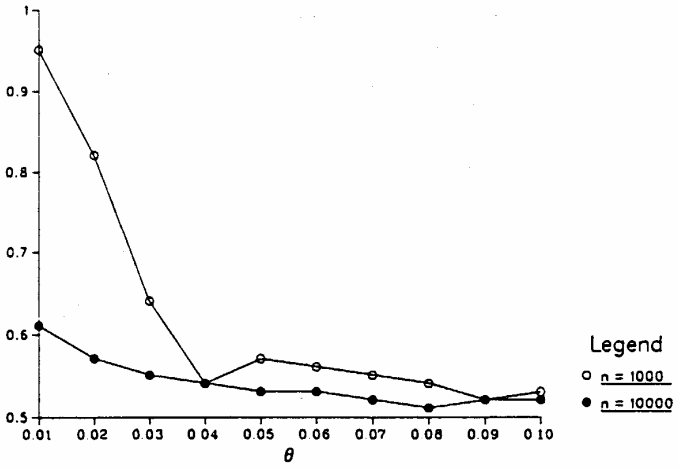

Fig. 4. ξ_n: Proportion of $\lambda_n = 1$ for $n = 1000$ and 10000.

example, this is the correct asymptotic result if the probability of getting 3 or greater is small enough to be negligible.

5. The Poisson, general case

We now consider $f(x,\theta) = \exp(-\theta)\theta^x/x!$, $\theta > 0$, $x = 0,1,2,\ldots$. Unfortunately, the result of Section 4 does *not* generalize. A simple geometric characterization of the maximum likelihood estimator is no longer possible. However, with the tool of the *general mixture maximum likelihood theorem* (Böhning (1982, 1989), Böhning and Hoffmann (1982), Lindsay (1983)) it is possible to identify the cases for which $\lambda_n = 1$ rather easily. Recall that $\lambda_n = 1$ is equivalent to

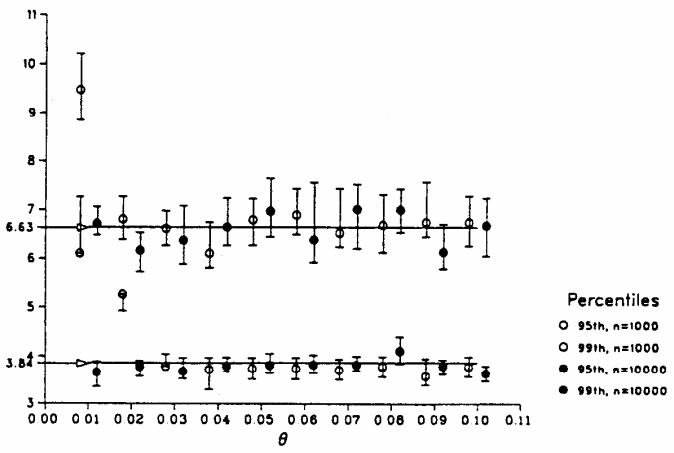

Fig. 5. Poisson with small θ: Percentiles of $2\log\lambda_n$ given that $\lambda_n > 1$ for $n = 1000$ and 10000.

$l(\hat{P}) = \sup_P l(P) = \sup_\theta l(\theta) = l(\hat{\theta})$. Here, $\hat{P}$ and $\hat{\theta}$ are the maximum likelihood estimators under the alternative and null hypothesis, respectively. Define $D(P,\theta) = \lim_{\beta\to 0+}[l((1-\beta)P + \beta Q_\theta) - l(P)]/\beta$ as the *directional derivative at P in the direction of the one point measure Q_θ*. Then the *general mixture maximum likelihood* theorem states that P^* is the unrestricted nonparametric maximum likelihood estimator if and only if $\sup_\theta D(P^*,\theta) = 0$, which can easily be deduced from the following inequality chain:

$$\sup_\theta D(P,\theta) \geq l(P^*) - l(P) \geq 0.$$

Now, in particular this inequality chain implies (with the convention $D(\mu,\theta) = D(Q_\mu,\theta)$) that

$$\sup_\theta D(\hat{\theta},\theta) \geq l(P^*) - l(\hat{\theta}) \geq 0.$$

From here we have:

$$\sup_\theta D(\hat{\theta},\theta) = 0 \Leftrightarrow l(P^*) - l(\hat{\theta}) = 0. \tag{5.1}$$

Noting that $l(P^*) \geq l(\hat{P}) \geq l(\hat{\theta})$, it is clear that $\sup_\theta D(\hat{\theta},\theta) = 0$ implies $l(\hat{P}) = l(\hat{\theta})$, and so $\lambda_n = 1$. On the other hand, it can be shown that if $l(\hat{P}) > l(\hat{\theta})$, then $\sup_\theta D(\hat{\theta},\theta) > 0$. Thus our first step in the calculation of λ_n is to check if $\sup_\theta D(\hat{\theta},\theta) \leq 0$. For the maximization of $D(\hat{\theta},\theta)$ we use a *global* maximization algorithm on the interval $(x_{(1)}, x_{(n)})$, in the sense that $D(\hat{\theta},\theta)$ is computed on a grid of 101 equally spaced points from $x_{(1)}$ to $x_{(n)}$, and the maximum was taken as the initial value for a Newton-Raphson iteration. Note that $D(\hat{\theta},\theta)$ is simply

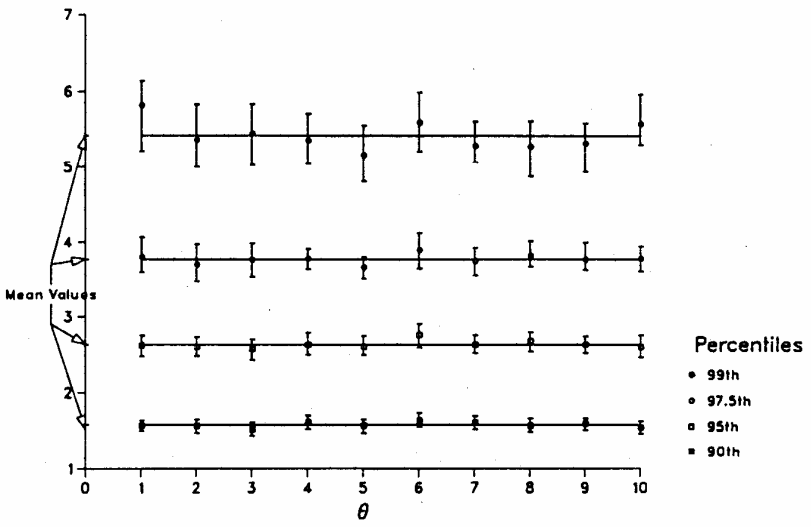

Fig. 6. Poisson: Selected percentiles of unconditional distribution of $2 \log \lambda_n$ for $n = 10000$.

$\sum_i \left\{ \frac{f(x_i,\theta)}{f(x_i,\hat{\theta})} - 1 \right\}$. If the inequality holds, we set $\lambda_n = 1$ (in the computer program $\sup_\theta D(\hat{\theta},\theta) \leq 0$ was implemented as $\sup_\theta D(\hat{\theta},\theta) \leq 10^{-4}$). This procedure can be carried out quickly at low computational expense.

For $\sup_\theta D(\hat{\theta},\theta) > 0$ the computation of $2 \log \lambda_n$ is more difficult, and the use of algorithmic methodology cannot be avoided. Note that (5.1) holds independently of the form of the density $f(x,\theta)$.

Although many algorithms have been developed for computing the nonparametric maximum likelihood estimate of the mixing distribution (see Böhning *et al.* (1992), Lesperance and Kalbfleisch (1992)), the EM algorithm (Dempster *et al.* (1977)) is still the simplest technique to compute the maximum likelihood estimate when the support size is fixed (as it is here with $k = 2$). We here used the EM algorithm together with a *stable acceleration device* that can save about 60–90% of the number of iterations *without* increasing the numerical complexity. This device is discussed in the appendix. Moreover, a key difficulty in mixture computations with fixed support size is the possible existence of multiple modes. In our simulations, we used as starting values under the alternative $p_1 = p_2$ and $\theta_1 = x_{(1)} + 1/2$, $\theta_2 = x_{(n)} - 1/2$, since well separated values have often turned out to be a good strategy for avoiding local maxima which are not global ones—at least in univariate problems (Böhning *et al.* (1992)). We note that computational error in the sense of not finding the global maximum would have the effect of biasing our percentiles downward, but we believe the effect to be negligible.

The (unconditional) distribution of $2 \log \lambda_n$ is found by simulation. The replication size was set to 10000. The parameters were chosen from the grid $1, 2, \ldots, 10$. Figure 6 shows estimates of the 90th, 95th, 97.5th, and 99th percentile of the dis-

Table 1. Selected percentiles of $2\log\lambda_n$ for Poisson distribution (averaged over parameter).

$1-\alpha$	$n=100$	$n=1000$	$n=10000$	P-value for $\chi^2_{(1)}$
90th	2.65	2.22	1.58	0.209
95th	4.01	3.86	2.62	0.105
97.5th	5.30	5.30	3.75	0.053
99th	7.15	7.15	5.44	0.020

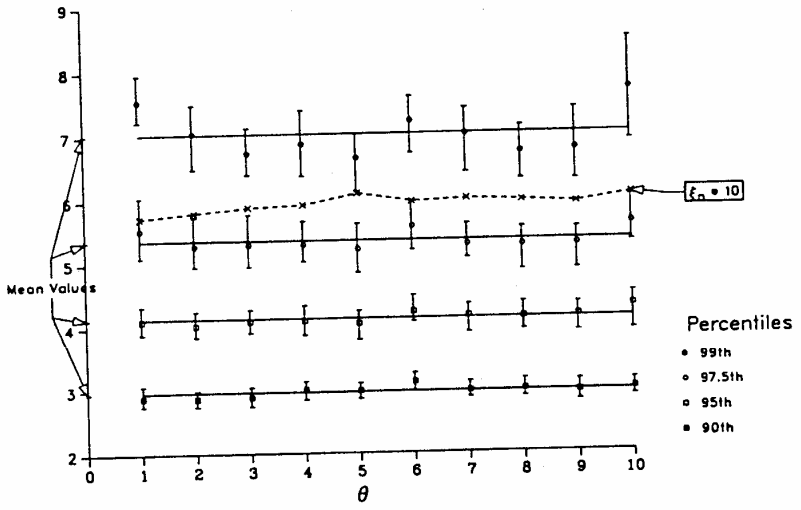

Fig. 7. Poisson: Selected percentiles of conditional distribution of $2\log\lambda_n$, conditional that $\lambda_n > 1$ for $n = 10000$.

tribution of $2\log\lambda_n$, $n = 10000$ with 95% confidence intervals. Plots for $n = 100$ and $n = 1000$ have similiar stability across θ. It appears that the upper percentiles become independent of θ, as is desired for testing. Table 1 presents the four percentiles averaged over the 10 parameters under consideration ($\theta = 1, 2, \ldots, 10$). Column 4 gives the p-value for the value of $2\log\lambda_n$ in column 3 *under the* χ^2 *with* 1 *df.* It can be seen that the $(1-\alpha)100$ percentile of $2\log\lambda_n$ corresponds to the $(1-2\alpha)100$ percentile of the χ^2 with 1 df.

Figure 7 shows estimates of the 90th, 95th, 97.5th, and 99th percentile of the conditional distribution of $2\log\lambda_n$, conditional that $\lambda_n > 1$, as well as the estimated proportion $\xi_n = \xi_n(\theta)$ of $\lambda_n = 1$. It appears as well that the percentiles and ξ_n become independent of θ.

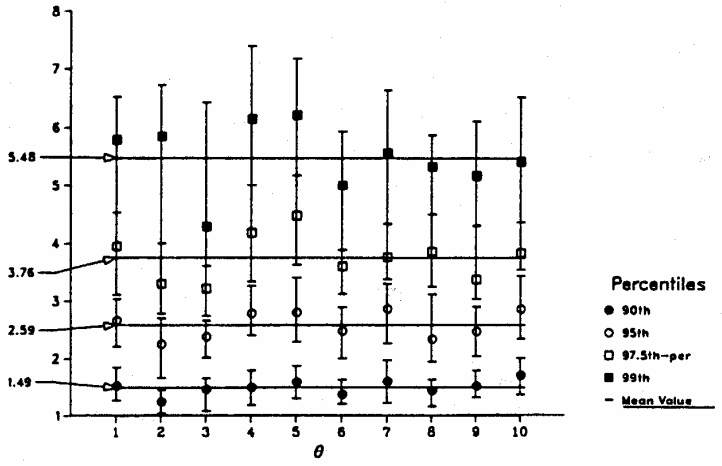

Fig. 8. Exponential: Upper percentiles of simulated distribution of $2 \log \lambda_n$ for $n = 1000$.

Table 2. Selected percentiles of $2 \log \lambda_n$ for exponential distribution (averaged over parameter).

$1 - \alpha$	$n = 100$	$n = 1000$	$n = 10000$	P-value for $\chi^2_{(1)}$
90th	1.69	1.49	0.50	0.479
95th	3.26	2.59	1.86	0.172
97.5th	4.67	3.76	3.19	0.074
99th	6.33	5.48	4.94	0.026

6. The exponential

Next, we study—as a first example with a continuous sample space—the *exponential* density $f(x, \theta) = \frac{1}{\theta} \exp\{-\frac{x}{\theta}\}$. Figure 8 shows four selected percentiles of the distribution of $2 \log \lambda_n$ for $n = 1000$. In this case, the homogeneity of the distribution over θ is not surprising, as λ_n is invariant under scale transformations of the data, and so under the null hypothesis has a distribution *not* depending on θ. If we average over the 10 parameters under consideration ($\theta = 1, 2, \ldots, 10$), we obtain the values of Table 2. Here, the asymptotic distribution of $2 \log \lambda_n$ is shifted even more to the left. Its $(1 - \alpha)100$-percentile corresponds more to the $(1 - 3\alpha)100$-percentile of a χ^2 with 1 df than to its $(1 - 2\alpha)100$-percentile.

7. The normal with known variance

We consider the normal density with known variance. By a location invariance argument the distribution of λ_n does not depend on parameter θ under H_0. This time, to find the percentiles of the limiting distribution, we use a technique suggested in Thode *et al.* (1988). The selected percentiles are computed by simulation

Table 3. Selected percentiles of $2\log\lambda_n$ for normal distribution with known variance.

n	90th	95th	97.5th	99th
100	2.13	3.50	5.14	6.63
200	1.87	3.35	4.67	6.23
300	1.56	2.95	4.59	6.79
400	1.52	2.92	4.66	6.51
500	1.49	2.65	4.29	6.52
600	1.43	2.47	3.65	5.67
700	1.52	2.49	3.77	5.59
800	1.55	2.46	3.56	5.22
900	1.57	2.57	3.90	5.32
1000	1.38	2.37	3.58	5.14
$\infty(10000)^{\ddagger}$	1.10 (0.968)	1.89 (2.23)	3.02 (3.44)	4.90 (5.06)
P-value for $\chi^2_{(1)}$	0.29 (0.32)	0.16 (0.14)	0.08 (0.06)	0.03 (0.03)

‡The number in brackets refer to the sample size $n = 10000$.

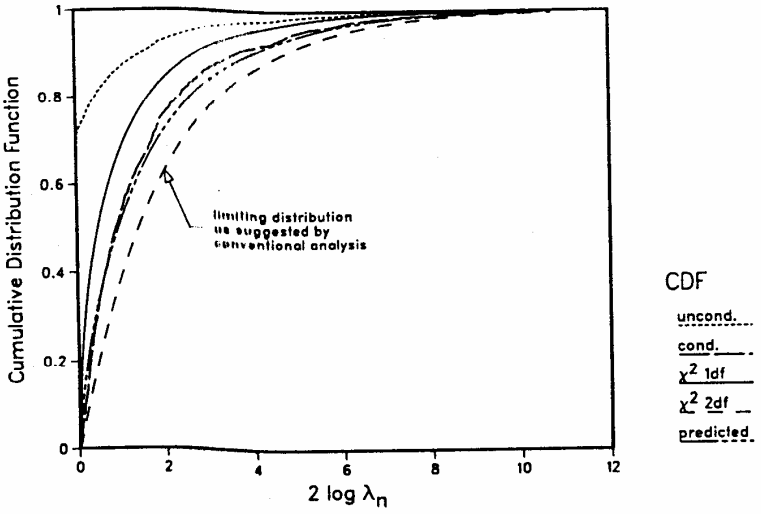

Fig. 9. Simulated distribution function of $2\log\lambda_n$, simulated distribution function of $2\log\lambda_n$ conditional that $\lambda_n > 1$, and distribution function of χ^2 with 1 and 2 df for sample size $n = 10000$.

(replication size 5000) for small n, from 100 to 1000. A limiting 95th percentile $\beta_\infty = 1.89$ can be obtained by extrapolation along a simple regression line. The regression model used is percentile $(2\log\lambda_n) = \beta_\infty + \beta/\sqrt{n}$. The line fits the percentiles well, and agrees with the percentile for $n = 10000$, which was not used

in fitting the line. See Table 3.

We close with a comparison of the simulated distribution function of $2\log\lambda_n$ for $n = 10000$ with the χ^2-distribution function of 1 degree of freedom and 2 degrees of freedom, the latter being suggested by conventional analysis. The difference from the χ^2-distribution with 1 degree of freedom is evident from Fig. 9. In addition, if we look at the simulated distribution function of $2\log\lambda_n$ conditional on $\lambda_n > 1$, we observe that it lies between the χ^2-distributions (with 1 df and 2 df). From the latter result it is reasonable to model the limiting conditional distribution Φ_{λ_n} of the log-likelihood ratio statistic conditional on $\lambda_n > 1$ as $(1-\alpha)\Phi_{\chi^2_{(1)}} + \alpha\Phi_{\chi^2_{(2)}}$. In our case we find $\hat{\alpha} = 0.56$ (by regressing $\Phi_{\lambda_n} - \Phi_{\chi^2_{(1)}}$ on $\Phi_{\chi^2_{(2)}} - \Phi_{\chi^2_{(1)}}$) and the corresponding mixture of χ^2-distribution functions is found under "*predicted*" in Fig. 9. *This illustrates the non-standard asymptotic result.*

8. Concluding remarks

A major achievement of this work can be seen in the finding that the upper percentiles (which are of interest of testing) appear to become independent of θ, although they vary from model to model under consideration.

It might be of interest to see whether the asymptotic result for testing $H_0 : k = 1$ against $H_1 : k = 2$ generalizes to the nested hypothesis case $H_0 : k = k^*$ against $H_1 : k = k^* + 1$ for $k^* = 1, 2, 3, \ldots$. The potential value of such an investigation lies in it's application in identifying the number of components by a selection procedure such as the backward elimination with the natural starting value taken as the *nonparametric maximum likelihood estimator* (Böhning (1982), Lindsay (1983)). However, such an investigation will require extensive simulation studies.

Acknowledgements

The research of Böhning, Dietz, Schaub, and Schlattmann is under current support of the German Research Foundation. Lindsay received support from the Humboldt-Foundation and the National Science Foundation.

Appendix

Aitken acceleration for computing the likelihood-ratio statistic

The EM algorithm (Dempster *et al.* (1977), Wu (1983)) is a well-known algorithmic concept for finding the maximum likelihood estimate in incomplete data problems, which is attractive because it often leads to simple iteration formulas with guaranteed stepwise increase of the likelihood function. Unfortunately, the EM algorithm converges only *linearly*, and in practice the rate is often very slow. This implies that many iterations will be necessary to achieve parameter estimates of reasonable accuracy. In a simulation study, where the algorithm has to be used many thousand of times, a speeding device can be very useful.

In addition, there is another *severe* problem: It is often not easy to say when reasonable accuracy is *satisfied*. Let (l_i) be the sequence of log-likelihoods created

by the EM algorithm. Often it is suggested to stop the iteration when $|l_{i+1} - l_i| \leq tol$ is met, with *tol* small. See for example Agha and Ibrahim (1984). However, this serves more as a *lack of progress* criterion than a useful stopping criterion.

We will discuss the possibilities of using Aitken acceleration to develop a *useful* stopping rule and then explore further the possibilities of using Aitken acceleration on the log-likelihood estimates. Aitken acceleration on the parameter estimates itself has been suggested by various authors (Louis (1982), Laird *et al.* (1985), Gediga and Holling (1988) and more recently by Meilijson (1989)). According to Section 1 the mixture likelihood becomes

$$\prod_{i=1}^{n} \sum_{j=1}^{k} f(x_i, \theta_j) p_j.$$

and the sequence (l_ν) is defined by $l_\nu = \log[\prod_{i=1}^{n} \sum_{j=1}^{k} f(x_i, \hat{\theta}_j^{(\nu)}) \hat{p}_j^{(\nu)}]$ where $\hat{\theta}^{(\nu)}$ and $\hat{p}^{(\nu)}$ are the iterates at the ν-th step of the EM algorithm.

Suppose we have an arbitrary sequence (l_i) converging *linearly* to $\hat{l}$, that is

$$\text{(A.1)} \qquad l_{i+1} - \hat{l} \cong c(l_i - \hat{l}) \quad \text{for all } i\text{'s and some } c,\ 0 < c < 1.$$

Here "$\cong$" means that the equality in (A.1) is valid in the sense $\lim_{i \to \infty} \frac{(l_{i+1} - \hat{l})}{(l_i - \hat{l})} = c$. Equation (A.1) can be rearranged algebraically to give the equation

$$l_{i+1} - l_i \cong (1 - c)(\hat{l} - l_i).$$

From this it is clear that if c is very close to 1, then a small increment in the log-likelihood, $l_{i+1} - l_i$, need not mean that l_i is close to the maximum, $\hat{l}$.

Aitken acceleration is a device to exploit the regularity of the convergence process. Because of (A.1) we find that

$$l_{i+1} - l_i \cong c(l_i - l_{i-1}) \quad \text{for all } i$$

implying that

$$l_{i+1} - l_i \cong c^i(l_1 - l_0),$$

and we get the *geometric series*

$$\text{(A.2)} \qquad \hat{l} = \lim_{i \to \infty} l_i \cong l_0 + \left(\sum_{i=0}^{\infty} c^i \right) (l_1 - l_0) = l_0 + \frac{1}{1 - c}(l_1 - l_0).$$

Since c is *unknown* we have to estimate it; this can be done with two *consecutive errors* $\epsilon_1 = (l_2 - l_1)$ and $\epsilon_0 = (l_1 - l_0)$ as

$$c_1 = \frac{(l_2 - l_1)}{(l_1 - l_0)} \quad \text{or its general form} \quad c_i = \frac{(l_{i+1} - l_i)}{(l_i - l_{i-1})}.$$

leading to the *Aitken accelerated estimate* of $\hat{l}$

$$l_i^{\infty} = l_{i-1} + \frac{1}{1-c_i}(l_i - l_{i-1}). \tag{A.3}$$

For c_i in $(0,1)$ we notice the nice *monotonicity property* $l_i^{\infty} \geq l_i$. In fact, in many cases l_i^{∞} is much bigger than l_i, as c_i is nearly one, corresponding to a slow linear rate of convergence.

The principle idea in applying (A.3) is to leave the sequence of parameter EM estimates unchanged (since they are *not* of primary interest) but instead of considering (l_i), we use (l_i^{∞}). We stop the EM algorithm if $|l_i^{\infty} - l_{i-1}^{\infty}| < tol$ and use l_i^{∞} as a prediction of $l(\hat{P})$. Note that this acceleration device can be used for any log-likelihood-sequence that is linearly convergent.

REFERENCES

Agha, M. and Ibrahim, M. T. (1984). Maximum likelihood estimation of mixtures of distributions, *J. Roy. Statist. Soc. Ser. C*, **33**, 327–332.

Böhning, D. (1982). Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process, *Ann. Statist.*, **10**, 1006–1008.

Böhning, D. (1989). Likelihood inference for mixtures: geometrical and other constructions of monotone step-length algorithms, *Biometrika*, **76**, 375–383.

Böhning, D. and Hoffmann, K.-H. (1982). Numerical techniques for estimating probabilities, *J. Statist. Comput. Simulation*, **14**, 283–293.

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

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

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

Gediga, G. and Holling, H. (1988). On the convergence of the EM algorithm including different methods of Aitken acceleration for finite mixture models, *Proceedings of the COMPSTAT Congress*, Short Communications and Posters, 31–32.

Gibbons, R. D., Clark, D. C. and Fawcett, J. (1990). A statistical method for evaluating suicide clusters and implementing cluster surveillance, *American Journal of Epidemiology*, **132**, 183–191.

Goffinet, B., Loisel, P. and Laurent, B. (1992). Testing in normal mixture models when the proportions are known, *Biometrika*, **79**, 842–846.

Laird, N., Lange, N. and Stram, D. (1985). Maximum likelihood computation with repeated measures: application of the EM algorithm, *Proceedings of the Statistical Computing Section*, 34–44.

Lesperance, M. and Kalbfleisch, J. D. (1992). An algorithm for computing the nonparametric MLE of a mixing distribution, *J. Amer. Statist. Assoc.*, **87**, 120–126.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory, *Ann. Statist.*, **11**, 86–94.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm, *J. Roy. Statist. Soc. Ser. B*, **44**, 226–233.

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

Meilijson, I. (1989). A fast improvement to the EM algorithm in its own terms, *J. Roy. Statist. Soc. Ser. B*, **51**, 127–138.

Mendell, N. R., Thode, H. C. and Finch, S. J. (1991). The likelihood ratio test for the two-component normal mixture problem: power and sample size analysis, *Biometrics*, **47**, 1143–1148.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *J. Amer. Statist. Assoc.*, **82**, 605–610.

Thode, H. C., Finch, S. J. and Mendell, N. R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test, *Biometrics*, **44**, 1195–1201.

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

Wu, C. F. (1983). On the convergence properties of the EM algorithm, *Ann. Statist.*, **11**, 95–103.