Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Statistical Planning and Inference 129 (2005) 93 – 107

journal of statistical planning and inference

www.elsevier.com/locate/jspi

# Modified likelihood ratio test in finite mixture models with a structural parameter

Jiahua Chen[a,*], John D. Kalbfleisch[b]

[a] Department of Statistics and Actuarial Science, University of Waterloo, 200 University Ave West, Waterloo, Ontario Canada N2L 3G1

[b] Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109–2029, USA

Available online 11 September 2004

## Abstract

The finite mixture model is an example of a non-regular parametric family, and most classical asymptotic results cannot be directly applied. In particular, the asymptotic properties of likelihood ratio statistics for testing for the number of subpopulations are complicated and difficult to establish. One approach that has been found to simplify the asymptotic results while preserving the power of the test is to modify the likelihood function by incorporating a penalty term to avoid boundary problems. The asymptotic properties and the use of likelihood ratio results are even more difficult when an unknown structural parameter is involved in the model. In this paper, we study an application of the modified likelihood approach to finite normal mixture models with a common and unknown variance in the mixing components and consider a test of the hypothesis of a homogeneous model versus a mixture on two or more components. We show that the $\chi^2_2$ distribution is a stochastic lower bound to the limiting distribution of the likelihood ratio statistic. This same distribution is also shown to be a stochastic upper bound to the limiting distribution of the modified likelihood ratio statistic. A small simulation study suggests that both bounds are relatively tight and practically useful. An example from genetics is used to illustrate the technique.

*MSC:* 62F03; 62F05

*Keywords:* EM-algorithm; Hardy–Weinberg law; Normal mixture; Stochastic bounds

[*] Corresponding author. Tel.: +519-888-4567; fax: +519-746-1875.
*E-mail address:* jhchen@uwaterloo.ca (J. Chen).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.06.041

## 1. Introduction

Finite mixture models are often used to study data from a population that is suspected to be composed of a number of homogeneous subpopulations. For example, when a disease has a simple genetic cause, the population may be divided into two or three homogeneous groups. In the initial stage of these investigations, it is important to have a sensitive test for the number $k$ of subpopulations included in the data. The construction of such a test, however, is often more challenging than might be expected.

Finite mixture models belong to a class of non-regular models and, as a consequence, many classical asymptotic results do not apply. Many researchers have tried to understand the large sample properties related to the analysis of finite mixture models. Hartigan (1985) first demonstrated the peculiar behavior of the likelihood ratio statistic for mixture models. Ghosh and Sen (1985) obtained the limiting distribution under a separation condition. The separation condition turned out to be unnecessary, which was shown by Chernoff and Lander (1995) for binomial mixtures, and in general by Chen and Chen (2001, 2002), Dacunha-Castelle and Gassiat (1999), and others.

Even though the large sample behavior of the likelihood ratio statistic under a mixture model is now better understood, its implementation still poses a challenge. The main difficulty involves determining the critical value based on a limiting distribution that involves the supremum of a Gaussian process. Techniques given in Adler (1990) and Sun (1993) may be useful in this respect. An alternative, discussed in McLachlan (1987), Chen and Chen (2001) and elsewhere, is to use resampling methods. Bayesian methods can also be applied in this context (Richardson and Green, 1997). Additional recent work can be found in McLachlan and Peel (2000), Lo et al. (2001), Garel (2001) and Garel and Goussanou (2002).

Chen and Kalbfleisch (1996), Chen (1998) and Chen et al. (2001, 2002) suggest a modification of the likelihood by incorporating a penalty term that forces certain estimates away from the boundary of the parameter space. The likelihood ratio statistic based on the modified estimators is shown, in many instances, to yield relatively simpler limiting distributions and hence simpler tests.

These results focus on finite mixtures of kernel distributions belonging to a one-parameter family and are applicable to Poisson and Binomial mixtures, for example, and to normal mixture when the variance (or the mean) is known. In many applications, however, one wishes to consider finite mixture models where, in addition to the mixing parameter, the kernel has a common but unknown structural parameter common to all components. For example, if a quantitative trait is largely determined by a simple Mendelian gene, then its distribution is often modeled as a mixture of normal distributions with common (but typically unknown) variance.

In this paper, we study application of the modified likelihood approach to finite normal mixture models with a common variance, which is a typical example of location and scale distribution family. The results on normal mixture are likely applicable to general location scale families. However, we believe that a lucid discussion on a specific model can sometimes be more meaningful than on a class of models under a long list of conditions. Specifically, we consider a test of a homogeneous normal model with $k=1$ mixture components versus a general alternative with $k \geqslant 1$ components. We find that the related modified likelihood ratio statistic does not have a simple limiting distribution. Nonetheless, we find a simple chi-squared upper bound for the limiting distribution and carry out simulations to assess its tightness. As expected, the ordinary or unmodified likelihood ratio statistic also does not have a simple limiting distribution and it is interesting that the same chi-squared distribution serves as a lower bound. Simulations indicate that the bounds are very tight in both cases, at least for sample sizes between 200 and 500. We illustrate our method using data from Wilson et al. (1988, 1990) and Zabetian et al. (2001).

## 2. Modified likelihood ratio test

### 2.1. The inference procedure

We consider the normal mixture model with density function

$$f(x;\sigma,G)=\int \sigma^{-1}\phi\{\sigma^{-1}(x-\theta)\}\,\mathrm{d}G(\theta). \tag{1}$$

where $\phi(x)$ is the density function of the standard normal distribution and $G$ is a cumulative distribution function on the compact parameter space $\Theta\subset R$. Other location-scale mixture models can be considered in the same manner.

The mixing distribution $G$ is assumed to have a finite number of support points. It is convenient to define

$$\mathscr{M}_k=\{G: G \text{ has at most } k \text{ support points}\}.$$

for $k=1,2,\ldots$ . We consider first a test of the null hypothesis $k=1$ versus the alternative $k=2$; or more precisely, we consider a test of the hypothesis $G\in\mathscr{M}_1$ versus $G\in\mathscr{M}_2$.

If $X_1,X_2,\ldots,X_n$ is a random sample from density (1), the log likelihood function is

$$l_n(\sigma,G)=\sum_{i=1}^{n}\log f(X_i;\sigma,G). \tag{2}$$

If $G\in\mathscr{M}_2$, then we can write $G(\theta)=(1-\pi)I(\theta_1\leqslant\theta)+\pi I(\theta_2\leqslant\theta)$ where $\pi\in(0,1)$ and $\theta_1\leqslant\theta_2$.

The ordinary likelihood ratio statistic for testing $G\in\mathscr{M}_1$ against $G\in\mathscr{M}_2$ is given by

$$\tilde{R}_n=2\left\{\sup_{\sigma,G\in\mathscr{M}_2} l_n(\sigma,G)-\sup_{\sigma,G\in\mathscr{M}_1} l_n(\sigma,G)\right\}.$$

Due to non-regularity of the finite mixture models, $\tilde{R}_n$ does not have a usual chi-squared limiting distribution.

In the modified likelihood approach, we define

$$pl_n(\sigma,G)=l_n(\sigma,G)+C\,\log\{4\pi(1-\pi)\}, \tag{3}$$

where $C$ is a positive constant. When $\theta_1=\theta_2$, we let $\pi=0.5$. The purpose of the "penalty term", $C\log\{4\pi(1-\pi)\}$ in (3) is to restore regularity to the problem by avoiding estimates of $\pi$ on or near the boundary.

Let $\hat{\sigma}_1$ and $\hat{G}_1$ maximize $pl_n(\sigma, G)$ for $G \in \mathscr{M}_1$, and $\hat{\sigma}_2$ and $\hat{G}_2$ maximize $pl_n(\sigma, G)$ for $G \in \mathscr{M}_2$. The modified likelihood ratio statistic is

$$R_n = 2\{l_n(\hat{\sigma}_2, \hat{G}_2) - l_n(\hat{\sigma}_1, \hat{G}_1)\}. \tag{4}$$

We consider a test in which the null hypothesis $k = 1$ is rejected for values of $R_n$ that are sufficiently large.

### 2.2. Stochastic bounds of modified and ordinary likelihood ratio statistics

The modified likelihood ratio test cannot be implemented in applications unless the null distribution of $R_n$ can be computed or estimated in some way. When the structural parameter $\sigma$ is known, the modified likelihood ratio statistic has a very simple limiting distribution of $0.5\chi_1^2 + 0.5\chi_0^2$ under the null hypothesis $k = 1$ (Chen et al., 2001). The analysis of a similar statistic for testing the hypothesis $k = 2$ versus $k > 2$ is much more complicated, but also yields a relatively simple limiting null distribution (Chen et al., 2002). However, with $\sigma$ unknown and estimated, the exact limiting distribution of $R_n$ defined by (4) does not have a simple form.

One could consider various strategies to address this situation. For example, one possibility is to revise the penalty term $\log\{4\pi(1-\pi)\}$ to achieve a simpler limiting distribution (Qin and Smith, personal communication). A second possibility is to allow $C = C_n$ in (3) to approach zero at an appropriate rate as $n \to \infty$ which we conjecture would yield a simpler limiting distribution. In this paper, we take an alternative approach and provide a simple upper bound to the asymptotic distribution of $R_n$. At the same time, we also derive a lower bound for the ordinary likelihood ratio statistic $\tilde{R}_n$.

For the normal mixture model with $\Theta = [-M, M]$, Chen and Chen (2002) show that under the null hypothesis

$$\tilde{R}_n \overset{\mathscr{D}}{\to} \max\left[\sup_{|\theta| \leqslant M} \{\varsigma^+(\theta)\}^2, \varsigma(0)^2 + Z^2\right] \tag{5}$$

as $n \to \infty$. In this expression, $\varsigma(\theta)$, $|\theta| \leqslant M$ is a Gaussian process with mean 0, variance 1 and autocorrelation function (for $st \neq 0$)

$$\rho(s, t) = \mathrm{sgn}(st)\frac{b(st)}{\sqrt{b(s^2)b(t^2)}},$$

where $b(x) = e^x - 1 - x - x^2/2$, and $\rho(0, t) = |t|^3/\sqrt{6b(t^2)}$. Also, $\varsigma(0)$ and $Z \sim N(0, 1)$ are independent, and for $s \neq 0$,

$$\mathrm{Cov}\{\varsigma(s), Z\} = \frac{s^4}{\sqrt{24b(s^2)}}.$$

Since $\varsigma(0)^2 + Z^2$ has a $\chi_2^2$ distribution, it is obvious that the $\chi_2^2$ distribution is a stochastic lower bound to the limiting distribution of $\tilde{R}_n$. When $M$ goes to infinity, this limiting distribution is un-bounded. When $M$ is moderate, however, the chance for the second term in the square brackets of (5) to exceed extreme values of the $\chi_2^2$ distribution is small. Thus, this bound is likely to be a good one at least in the upper tail of the distribution. Simulations presented later suggest this is true.

Our second task is to derive an upper bound for the limiting distribution of $R_n$. Note from (5) that

$$\tilde{R}_n = \mathrm{O_p}(1).$$

As a consequence $C\ \log\{\hat{\pi}(1-\hat{\pi})\} \geqslant -\tilde{R}_n = \mathrm{O_p}(1)$ when the likelihood is modified as in (3). Consequently, with the modified likelihood, we can investigate the asymptotic properties of $R_n$ while restricting attention to those mixing distributions $G \in \mathscr{M}_2$ such that $\varepsilon < \pi < 1-\varepsilon$ for an arbitrarily small positive constant $\varepsilon$. Under this restriction on $\pi$, the modified likelihood ratio statistic has the expansion given by $R(\varepsilon; II)$ in Chen and Chen (2002). The derivation of $R(\varepsilon; II)$ is long and tedious, and we will not repeat it here. It is interesting that the expansion itself does not depend on the choice of $\varepsilon$ which enables us to obtain an upper bound for the limiting distribution of $R_n$. We now present the mathematical derivation based on the result in Chen and Chen (2002).

We use the same notation as introduced and motivated in Chen and Chen (2002). Specifically, let

$$Y_i = X_i, \quad Y_i' = U_i = (X_i^2 - 1)/2,$$

$$Y_i'' = (X_i^3 - 3X_i)/3, \quad Y_i''' = 2U_i' = (X_i^4 - 6X_i^2 + 3)/4$$

and denote the first four moments of $G$ as $m_j$, $j = 1, 2, 3, 4$. Let

$$s_1 = m_1, \quad s_2 = \sigma^2 - 1 + m_2, \quad s_3 = m_3/2, \quad s_4 = (m_4 - 3m_2^2)/6.$$

Directly from Chen and Chen (2002), we have

$$\begin{aligned} R_n \leqslant \sup_{s_1,s_2,s_3,s_4} \Bigg[ & 2\left\{ s_1 \sum_{i=1}^n Y_i + s_2 \sum_{i=1}^n Y_i' + s_3 \sum_{i=1}^n Y_i'' + s_4 \sum_{i=1}^n Y_i''' \right\} \\ & - \left\{ s_1^2 \sum_{i=1}^n Y_i^2 + s_2^2 \sum_{i=1}^n (Y_i')^2 + s_3^2 \sum_{i=1}^n (Y_i'')^2 + s_4^2 \sum_{i=1}^n (Y_i''')^2 \right\} \Bigg] \\ & + C\ \log\{4\pi(1-\pi)\}] - n\bar{X}^2 + (2/n)\left\{ \sum_{i=1}^n U_i \right\}^2 + \mathrm{o_p}(1). \end{aligned}$$

Note that the leading term is a quadratic function in the first four moments of $G$. Hence, using the fact that $2ax - x^2 \leqslant a^2$ for all $x$, it is easy to show

$$\begin{aligned} R_n \leqslant & \sup_{s_1,s_2,s_3,s_4} \left[ 2\left\{ s_1 \sum_{i=1}^n Y_i + s_2 \sum_{i=1}^n Y_i' + s_3 \sum_{i=1}^n Y_i'' + s_4 \sum_{i=1}^n Y_i''' \right\} \right. \\ & \left. - \left\{ s_1^2 \sum_{i=1}^n Y_i^2 + s_2^2 \sum_{i=1}^n (Y_i')^2 + s_3^2 \sum_{i=1}^n (Y_i'')^2 + s_4^2 \sum_{i=1}^n (Y_i''')^2 \right\} \right] \\ & - n\bar{X}^2 + (2/n)\left\{ \sum_{i=1}^n U_i \right\}^2 + \mathrm{o_p}(1). \\ \leqslant & \frac{(\sum_{i=1}^n Y_i'')^2}{\sum_{i=1}^n (Y_i'')^2} + \frac{(\sum_{i=1}^n Y_i''')^2}{\sum_{i=1}^n (Y_i''')^2} + \mathrm{o_p}(1) \xrightarrow{\mathscr{D}} \chi_2^2. \end{aligned}$$

The first inequality follows since $\log\{4\pi(1-\pi)\} \leqslant 0$ and the second is the result of a simple computation. To achieve the upper bound for $R_n$ it is clear that $\pi = 0.5$ is necessary. But if $\pi = 0.5$, the range of $s_1, s_2, s_3, s_4$ will be restricted to a three-dimensional manifold. Thus, limiting distribution of $R_n$ may not achieve the upper bound.

When the sample size is moderate, say in the range of 200–500, many other factors can affect the accuracy of the above approximation. In particular, the higher order term represented by $\mathrm{o_p}(1)$ may be relatively large with high probability and the distribution of $R_n$ could even be stochastically greater than $\chi_2^2$. In Section 4, we evaluate the accuracy of the upper bound as an approximation and find it to be accurate in the cases considered.

The accuracy of $\chi_2^2$ as an approximation can perhaps be motivated in another way. If $C$ decreases as $n$ increases, the limiting distribution of $R_n$ will increase. We conjecture that when $C = C_n$ decreases at a rate of $(\log\, n)^{1/2}$ or so, the limiting distribution of $R_n$ is indeed given by $\chi_2^2$.

### 2.3. A genetic model with $k = 3$

Suppose that a quantitative trait is strongly affected by a gene with two alleles, $A$ and $a$ say, whose effects are additive. In this case, a normal mixture model with $k = 3$ is often used. According to the Hardy–Weinberg law, the mixture probabilities are $\pi_1 = p^2$, $\pi_2 = 2pq$ and $\pi_3 = q^2$ where $p \in [0, 1]$ is the population frequency of the $A$ allele and $q = 1 - p$. We call this the H–W mixture model. The theory in Chen and Chen (2002) does not apply directly to the H–W mixture model and that generalization is left for future work. However, we discover through simulation that the ordinary and modified likelihood ratio statistics can also be approximated by chi-squared distribution with 3 degrees of freedom.

### 2.4. Computations

The EM-algorithm is often used for computational problems in finite mixture models with given number of components (Böhning, 1999). It turns out that it needs only slight modification for the computational problems of the modified likelihood ratio test. We illustrate this result with the H–W mixture model.

Note that if the class of each observation $X_i$ is known, the complete likelihood function can be easily found. Let $I_{ij}$ the indicator that the $i$th observation is from the $j$th subpopulation, $i=1,2,\ldots,n,\ j=1,2,3$. The complete log-likelihood function is then

$$l_n(\theta_1,\theta_2,\theta_3,p,\sigma)=-n\ \log\ \sigma-\sum_{i=1}^{n}\sum_{j=1}^{3}I_{ij}(X_i-\theta_j)^2/(2\sigma^2)+\sum_{j=1}^{3}I_{.j}\ \log(\pi_j),$$

where $I_{.j}=\sum_{i=1}^{n}I_{ij}$. Given $X_1,\ldots,X_n$ and the parameters $\sigma,\ p,\ \theta_j,\ \ j=1,2,3$, the E-step involves the conditional expectations

$$\hat{\pi}_{ij}=E\{I_{ij}\}=\phi\{\sigma^{-1}(X_i-\theta_j)\}/\left[\sum_{\ell=1}^{3}\pi_j\phi\{\sigma^{-1}(X_i-\theta_\ell)\}\right].$$

In the M-step, we find

$$\hat{\theta}_j=\sum_{i=1}^{n}\hat{\pi}_{ij}X_i/\sum_{i=1}^{n}\hat{\pi}_{ij},$$

$$\hat{\sigma}^2=n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{3}\hat{\pi}_{ij}(X_i-\hat{\theta}_j)^2,$$

$$\hat{p}=\sum_{i=1}^{n}(2\hat{\pi}_{i1}+\hat{\pi}_{i2})/(2n)$$

and $\hat{\pi}_1=\hat{p}^2$. The E- and M-steps are then continued to convergence.

Let the modified log-likelihood function be defined as

$$pl_n(\theta_1,\theta_2,\theta_3,p,\sigma)=l_n(\theta_1,\theta_2,\theta_3,p,\sigma)+C\ \log(\pi_1\pi_2\pi_3).$$

To maximize this modified likelihood, we need only revise the estimate of $p$ in the M-step above to

$$\hat{p}=\left[\sum_{i=1}^{n}(2\hat{\pi}_{i1}+\hat{\pi}_{i2})+3C\right]/(2n+6C).$$

We implemented the above EM-algorithm in our simulation experiment. In order to increase the likelihood of finding the global maximum, we used three initial values for each simulated sample. We found that the EM-algorithm is much more reliable when $C=1$ than when $C=0$. In fact, in about 20% of the trials, the algorithm led to a modified LRT statistic ($C=1$) that was larger than the ordinary LRT statistic ($C=0$). This is not possible and indicates the failure of the algorithm to find the global maximum at least when $C=0$. More details are given in the simulation section.

When $\sigma^2$ is known, the modified likelihood ratio statistic for testing $G\in\mathscr{M}_1$ versus $G\in\mathscr{M}_r$ has the same asymptotic distribution for any $r>1$. There is also a similar property for testing $G\in\mathscr{M}_2$. See Chen (1998), Chen et al. (2001, 2002) for details. This is a nice property that, for example, enables a single test of the hypothesis $k=1$ versus $k>1$. When $\sigma^2$ is unknown, these results do not appear to generalize.

In the next section, we consider simulations of the null distribution of the ordinary and the modified likelihood ratio statistics with $\sigma^2$ unknown. The results are compared with the $\chi_2^2$ bound for the simple case of testing $k=1$ versus $k=2$ as described in Section 2.2, and the comparison is found to be quite close. For alternatives in the H–W mixture model, we find that the modified LR statistics can be better approximated by the $\chi_3^2$ distribution. Some theoretical argument would be useful in explaining this.

## 3. Simulation

We use simulation to investigate the precision of the upper bound for the limiting distribution of $R_n$.

In the first case, we generate 20,000 samples of size $n=200$ and 500 from a standard normal distribution and compare the sample quantiles of $R_n$ with the quantiles of the $\chi_2^2$ distribution. The modification constant $C$ is set to be 0.0 and 1.0. When $C=0.0$, $R_n$ becomes the ordinary likelihood ratio statistic, whereas when $C=1.0$, $R_n$ is a modified likelihood ratio statistic. The theory of Section 2 implies that for large enough $n$ the quantiles of $R_n$ when $C=1$ are bounded above by those of $\chi_2^2$ and the quantiles of $R_n=\tilde{R}_n$ when $C=0$ are bounded below by those of $\chi_2^2$.

To assess the power of the tests, we generated samples from six alternative models with $k=2$ as summarized in Table 1. We used the simulated quantiles for the null distribution to obtain estimates of the power.

We used the EM-algorithm in our simulation. It is well known that EM-algorithms can yield a local rather than the global maximum and we took measures to ensure the validity of the simulation. First, we used several different initial values to increase the chance of locating the global maximum. Second, we used two different strategies for picking initial values and examined the difference between implementations. Although for some samples, different implementations gave different values of $R_n$, the overall rejection rates were very similar. Table 2 summarizes the results, based on 20,000 repetitions, in which the null rates are with reference to the critical values of a $\chi_2^2$ distribution. Fig. 1 displays plots of the simulation results under the null model with sample size $n=200$.

As seen in Table 2, the null rejection rates of both ordinary and modified LR tests based on critical values from the $\chi_2^2$ distribution are very close to the nominal values. The first row of Fig. 1 gives Q–Q plots of the simulated LR statistics under the null model versus the $\chi_2^2$ for $n=200$ and indicates good agreement over a much broader region. The rejection rates under the alternative models in Table 2 represent the empirical powers of the methods

Table 1
The alternative models considered for the normal mixtures with $k=2$ components

| Model | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|
| $\mu_1$ | −1.0 | −1.0 | −1.0 | −1.0 | −1.0 | −1.0 |
| $\mu_2$ | 1.5 | 1.5 | 1.5 | 2.0 | 2.0 | 2.0 |
| $\pi$ | 0.5 | 0.3 | 0.1 | 0.5 | 0.3 | 0.1 |

Table 2
The rejection rates of the ordinary ($C = 0$) and modified ($C = 1$) likelihood ratio tests for $k = 1$ versus $k = 2$

| | $C = 0$ | | | $C = 1$ | | |
|---|---|---|---|---|---|---|
| $n = 200$ | | | | | | |
| Nominal | 0.100 | 0.050 | 0.010 | 0.100 | 0.050 | 0.010 |
| Null | 0.117 | 0.061 | 0.013 | 0.101 | 0.054 | 0.012 |
| A1 | 0.935 | 0.878 | 0.704 | 0.944 | 0.888 | 0.717 |
| A2 | 0.964 | 0.927 | 0.792 | 0.969 | 0.932 | 0.802 |
| A3 | 0.910 | 0.853 | 0.683 | 0.909 | 0.851 | 0.683 |
| A4 | 0.999 | 0.997 | 0.985 | 0.999 | 0.998 | 0.986 |
| A5 | 1.000 | 0.999 | 0.995 | 1.000 | 0.999 | 0.995 |
| A6 | 0.995 | 0.990 | 0.962 | 0.994 | 0.989 | 0.963 |
| $n = 500$ | | | | | | |
| Null | 0.108 | 0.055 | 0.012 | 0.088 | 0.047 | 0.011 |
| A1 | 1.000 | 0.999 | 0.995 | 1.000 | 0.999 | 0.995 |
| A2 | 1.000 | 1.000 | 0.999 | 1.000 | 1.000 | 0.999 |
| A3 | 1.000 | 0.998 | 0.991 | 1.000 | 0.998 | 0.991 |

using the simulated quantiles to control the size of the test. The two methods, ordinary and modified LR tests, have very similar powers. Note that some alternative models under $n=500$ are omitted since the rejection rates are all nearly 100%. The second row of plots in Fig. 1 illustrate the effect of the modification on estimation of the "penalty" term $\log\{4\pi_1(1-\pi_1)\}$ under the null hypothesis. The size of this term is much larger for the modified LR test which provides some illustration that the modification has the effect of restoring regularity of finite mixture models by pushing the estimates away from the boundary.

Chen and Chen (2002) show that the MLE of $\sigma^2$ is consistent, but the rate of convergence is $n^{-1/4}$ instead of the usual $n^{-1/2}$. Histograms of the ordinary and modified variance estimates under the null are given in the third row of the Fig. 1, and the mean, variance, and mean squared error of the variance estimators are given in Table 3. Under the null hypothesis with $n = 200$, the common variance parameter is seriously under estimated. Some decrease in the bias is seen for $n=500$, but the improvement is relatively small. Under the alternative models, however, the biases are small and even negligible. Histograms of the variance estimates are given in row 3 of Fig. 1. These histograms tend to have two modes corresponding to situations where the data are best fitted with $k = 1$ or with $k = 2$ components. When the null model is true, fitting a mixture model with $k = 2$ and smaller common variance can be viewed as a type of over-fitting. Such models allow more flexibility to fit some spurious observations in one of the tails that may appear by chance. One approach to prevent such over-fitting and so reduce the bias of the variance estimator might be to incorporate a prior distribution for the common variance. Further investigation of the estimation of $\sigma$ is needed.

Based on the above simulation, we recommend the modified likelihood approach using the $\chi^2_2$ approximation. This test tends to be conservative and is somewhat more accurate. It is also easier to implement since, as we noted, the EM-algorithm has better convergence properties when $C = 1$.

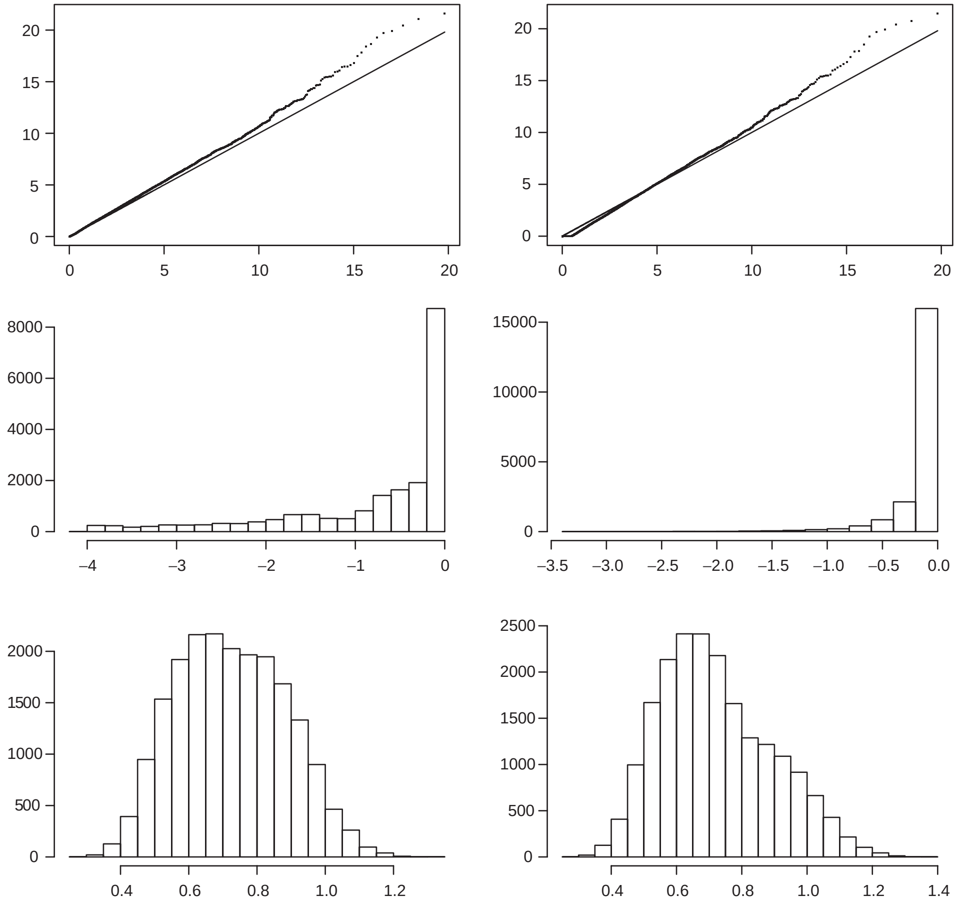

Fig. 1. Null distributions of estimates and tests in the normal mixture with $k=2$ using modified likelihood ($C=0$, left column; $C=1$, right column). Row 1: Q–Q plots of $R_n$ against $\chi_2^2$. Row 2: histograms of $\log\{4\pi_1(1-\pi_1)\}$. Row 3: histograms of $\hat{\sigma}^2$.

We conducted a similar simulation for the H–W mixture model as specified in Table 4. Table 5 gives the null rejection rates of both ordinary and modified LR tests using a $\chi_3^2$ distribution to define critical values. The null rejection rates are very close to the nominal values. Fig. 2 gives Q–Q plots for $C=0$ and 1 with $n=200$. The agreement between the simulated quantiles and those of the $\chi_3^2$ is very good. At this stage, the $\chi_3^2$ approximation is entirely empirical and perhaps provides motivation to a future theoretical investigation. Table 6 summarizes the variance estimator. Under the null model, the common variance is again seriously under estimated, and the bias is reduced when the sample size increases. As before, but more apparently, the histograms of the variance estimators in Fig. 2 have two modes again corresponding to fits with $k=1$ and $k>1$ components.

Table 3
The mean, variance, and MSE of $\hat{\sigma}^2$ under the ordinary and modified likelihood

| | $C = 0$ | | | $C = 1$ | | |
|---|---|---|---|---|---|---|
| $n = 200$ | | | | | | |
| Null | 0.730 | 0.026 | 0.101 | 0.720 | 0.029 | 0.107 |
| A1 | 0.990 | 0.025 | 0.025 | 0.990 | 0.024 | 0.024 |
| A2 | 0.988 | 0.024 | 0.024 | 0.983 | 0.022 | 0.023 |
| A3 | 0.974 | 0.021 | 0.021 | 0.950 | 0.018 | 0.021 |
| A4 | 0.991 | 0.018 | 0.018 | 0.991 | 0.018 | 0.018 |
| A5 | 0.991 | 0.017 | 0.018 | 0.989 | 0.017 | 0.017 |
| A6 | 0.983 | 0.016 | 0.016 | 0.972 | 0.015 | 0.015 |
| $n = 500$ | | | | | | |
| Null | 0.790 | 0.017 | 0.060 | 0.780 | 0.019 | 0.066 |
| A1 | 0.995 | 0.009 | 0.009 | 0.995 | 0.009 | 0.009 |
| A2 | 0.995 | 0.009 | 0.009 | 0.993 | 0.009 | 0.009 |
| A3 | 0.989 | 0.009 | 0.009 | 0.978 | 0.008 | 0.008 |

Table 4
The alternative models considered for the H–W mixture

| Model | A1 | A2 | A3 | A4 | A5 | A6 |
|---|---|---|---|---|---|---|
| $\mu_1$ | −1.0 | −1.0 | −1.0 | −1.0 | −1.0 | −1.0 |
| $\mu_2$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| $\mu_3$ | 1.0 | 1.0 | 2.0 | 2.0 | 3.0 | 3.0 |
| $p$ | 0.5 | 0.3 | 0.5 | 0.3 | 0.5 | 0.3 |

Table 5
The rejection rates of tests for $k = 1$ versus the H–W mixture model based on 20,000 simulations

| | $C = 0$ | | | $C = 1$ | | |
|---|---|---|---|---|---|---|
| $n = 200$ | | | | | | |
| Nominal | 0.100 | 0.050 | 0.010 | 0.100 | 0.050 | 0.010 |
| Null | 0.100 | 0.051 | 0.010 | 0.098 | 0.054 | 0.011 |
| A1 | 0.116 | 0.062 | 0.015 | 0.119 | 0.063 | 0.014 |
| A2 | 0.144 | 0.079 | 0.021 | 0.146 | 0.078 | 0.021 |
| A3 | 0.586 | 0.442 | 0.207 | 0.626 | 0.482 | 0.237 |
| A4 | 0.575 | 0.438 | 0.202 | 0.586 | 0.443 | 0.202 |
| A5 | 0.999 | 0.998 | 0.988 | 1.000 | 0.999 | 0.990 |
| A6 | 0.999 | 0.998 | 0.986 | 0.999 | 0.997 | 0.982 |
| $n = 500$ | | | | | | |
| Null | 0.077 | 0.039 | 0.008 | 0.073 | 0.039 | 0.008 |
| A1 | 0.147 | 0.083 | 0.019 | 0.153 | 0.085 | 0.020 |
| A2 | 0.220 | 0.130 | 0.035 | 0.220 | 0.129 | 0.037 |
| A3 | 0.947 | 0.889 | 0.711 | 0.956 | 0.907 | 0.746 |

The rejection rates under the null distribution are with reference to critical values from a $\chi^2_3$ distribution.

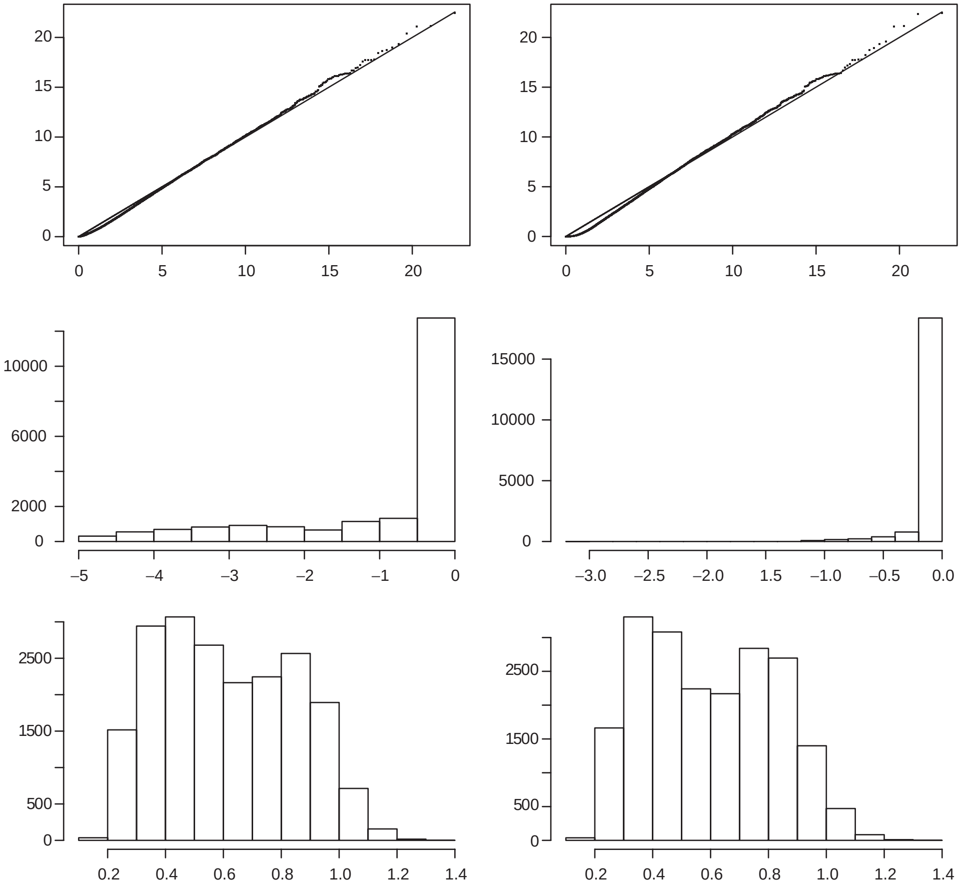

Fig. 2. Null distributions of estimates and tests in the H–W mixture model using modified likelihood ($C = 0$, left column; $C = 1$, right column). Row 1: Q–Q plots of $R_n$ against $\chi^2_3$. Row 2: histograms of $\log\{4p(1-p)\}$. Row 3: histograms of $\hat{\sigma}^2$.

## 4. A genetic example

Wilson et al. (1988, 1990) and Zabetian et al. (2001), give analyses of family data on dopamine-$\beta$-hydroxylase (D$\beta$H), a chemical that catalyzes the conversion of dopamine to norepinephrine. Altered plasma-D$\beta$H activity has been reported in a variety of psychiatric and neurological disorders. A large study of European Americans (EAs) identified a subgroup consisting of 3–4% of the population with very low levels of plasma-D$\beta$H activity. It is postulated that there exists a functional low-activity allele, D$\beta$H$^L$, with a frequency of about 20% in the EA population. Further, this gene is believed to be linked with the ABO blood-group gene. Their analyses support these findings in general.

Table 6
The mean, variance, and MSE of $\hat{\sigma}^2$ under the H–W mixture model based on 20,000 simulations

| | $C = 0$ | | | $C = 1$ | | |
|---|---|---|---|---|---|---|
| $n = 200$ | | | | | | |
| Null | 0.61 | 0.054 | 0.206 | 0.60 | 0.051 | 0.211 |
| A1 | 0.833 | 0.112 | 0.140 | 0.812 | 0.106 | 0.141 |
| A2 | 0.807 | 0.099 | 0.136 | 0.790 | 0.093 | 0.137 |
| A3 | 1.025 | 0.078 | 0.079 | 0.897 | 0.078 | 0.089 |
| A4 | 0.882 | 0.062 | 0.076 | 0.852 | 0.058 | 0.080 |
| $n = 500$ | | | | | | |
| Null | 0.72 | 0.40 | 0.478 | 0.70 | 0.35 | 0.440 |
| A1 | 0.958 | 0.086 | 0.088 | 0.930 | 0.077 | 0.082 |
| A2 | 0.929 | 0.074 | 0.079 | 0.912 | 0.067 | 0.075 |
| A3 | 1.141 | 0.030 | 0.049 | 1.020 | 0.045 | 0.046 |
| A4 | 0.963 | 0.031 | 0.032 | 0.951 | 0.028 | 0.030 |

Table 7
The maximum modified likelihood estimates with $C = 1$ of mixture models for the family data on D$\beta$H

| Family | $\pi_1$ | $\mu_1$ | $\mu_2$ | $\sigma$ | MLRT | $p$-value |
|---|---|---|---|---|---|---|
| HGAR 6 | 0.255 | 2.73 | 4.75 | 0.95 | 2.78 | 0.249 |
| HGAR 7 | 0.163 | 3.03 | 6.05 | 1.26 | 13.65 | 0.001 |
| HGAR 9 | 0.438 | 3.82 | 6.56 | 1.14 | 2.73 | 0.255 |
| HGAR 10 | 0.265 | 3.20 | 5.21 | 1.34 | 0.99 | 0.610 |

The data set contains four families with a total of 923 individuals. The number of individuals with D$\beta$H values determined in the four families are 56, 204, 48 and 191 to give a total of 499. Due to the complexity of the family structure, it is very difficult to use the complete likelihood function based on the finite normal mixture model. Following Wilson et al. (1988) and others, we fit the mixture model with common but unknown variance to the square root of the D$\beta$H reading, and treat the observations within each family as independent and identically distributed. We consider tests of the null hypothesis of a homogeneous model in favor of a mixture model with $k = 2$ or an H–W mixture model with $k = 3$ components.

Since the families are not random samples from the population, their corresponding gene-frequencies may vary, but the subpopulation means should be the same. In fitting models, however, we introduced no such restrictions on the parameters. The fitted models based on the modified likelihood with $C = 1$ are reported in Tables 7 and 8.

From Table 7, only the data in family HGAR 7 lead to rejection of the null homogeneous normal model.

The modified likelihood approach produced similar estimates to those in Wilson et al. (1988) except that the estimate of the gene frequency for HGAR 7 is only about half of that which they obtained. Using the $\chi^2_3$ distribution as a benchmark, the corresponding $p$-values are given and again there is significant evidence that the HGAR 7 family can be better

Table 8
The maximum modified likelihood estimates with $C = 1$ of the H–W mixture models for the family data on D$\beta$H

| Family | $p$ | $\mu_1$ | $\mu_2$ | $\mu_3$ | $\sigma$ | MLRT | $p$-value |
|---|---|---|---|---|---|---|---|
| HGAR 6 | 0.334 | 1.85 | 3.76 | 5.27 | 0.76 | 5.17 | 0.160 |
| HGAR 7 | 0.179 | 1.23 | 4.17 | 6.37 | 1.09 | 15.32 | 0.0016 |
| HGAR 9 | 0.376 | 2.76 | 4.71 | 7.02 | 0.91 | 2.83 | 0.419 |
| HGAR 10 | 0.453 | 2.90 | 4.82 | 5.66 | 1.28 | 1.22 | 0.748 |

modeled with the H–W mixture than the homogeneous case. Comparing H–W mixture to normal mixture with $k = 2$, the MLRT values are increased but only moderately so. There is no statistical evidence that the H–W mixture should be favored over the simpler two component normal mixture.

## Acknowledgements

We would like to thank Professor Robert Elston for providing us with the D$\beta$H data used for illustration in Section 4, and also for his helpful comments on general analysis strategies. This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada to Chen and by a grant from the University of Michigan to Kalbfleisch.

## References

Adler, R.J., 1990. An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. Institute of Mathematical Statistics, Lecture Notes, vol. 12, Hayward, CA

Böhning, D., 1999. Computer-Assisted Analysis of Mixtures and Applications: Meta-analysis, Disease Mapping, and Others, Chapman & Hall, London.

Chen, H., Chen, J., 2001. Large sample distribution of the likelihood ratio test for normal mixtures. Canad. J. Statist. 29, 201–216.

Chen, H., Chen, J., 2002. Tests for homogeneity in normal mixtures with presence of a structural parameter. Preprint.

Chen, J., 1998. Penalized likelihood-ratio test for finite mixture models with multinomial observations. Canad. J. Statist. 26, 583–599.

Chen, J., Kalbfleisch, J.D., 1996. Penalized minimum-distance estimates in finite mixture models. Canad. J. Statist. 24, 167–175.

Chen, H., Chen, J., Kalbfleisch, J.D., 2001. A modified likelihood ratio test for homogeneity in finite mixture models. J. Roy. Statist. Soc. B 63, 19–29.

Chen, H., Chen, J., Kalbfleisch, J.D., 2002. Testing for a finite mixture model with two components. Statistics and Actuarial Science Technical Report #2001–02, University of Waterloo.

Chernoff, H., Lander, E., 1995. Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. J. Statist. Plann. Inference 43, 19–40.

Dacunha-Castelle, D., Gassiat, E., 1999. Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes. Ann. Statist. 27, 1178–1209.

Garel, B., 2001. Likelihood ratio test for univariate Gaussian mixture. J. Statist. Plann. Inference 96, 325–350.

Garel, B., Goussanou, F., 2002. Removing separation conditions in a 1 against 3-components Baussian mixture problem. In: Jajuga, K., Sokolowski, A., Bock, H.H. (Eds.), Classification, Clustering, and Data Analysis. Springer, Berlin, pp. 61–73.

Ghosh, J.K., Sen, P.K., 1985. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In: LeCam, L., Olshen, R.A. (Eds.), Proceedings of the Berk. Conference in Honor of J. Neyman and J. Kiefer, vol. 2. pp. 799–806.

Hartigan, J.A., 1985. A failure of likelihood asymptotics for normal mixtures. In: LeCam, L., Olshen, R.A. (Eds.), Proceedings of the Berk Conference in Honor of J. Neyman and J. Kiefer, vol. 2. pp. 807–810.

Lo, Y., Mendell, N.R., Rubin, D.B., 2001. Testing for the number of components in a normal mixture. Biometrika 88, 767–778.

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

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

Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures of with an unknown number of components. J. Roy. Statist. Soc. B 59, 731–792.

Sun, J., 1993. Tail probabilities of the maxima of Gaussian random fields. Ann. Probab. 21, 34–71.

Wilson, A.F., Elston, R.C., Siervogel, R.M., Tran, L.D., 1988. Linkage of a gene regulating dopamine-$\beta$-hydroxylase activity and the ABO blood group locus. Amer. J. Hum. Genet. 42, 160–166.

Wilson, A.F., Elston, R.C., Sellers, T.A., Bailey-Wilson, J.E., Gersting, J.M., Deen, D.K., Sorant, A.J.M., Tran, L.D., Amox, C.I., Siervogel, R.M., 1990. Stepwise oligogenic segregation and linkage analysis illustrated with dopamine-$\beta$-hydrozylase activity. Amer. J. Medical Genet. 35, 425–432.

Zabetian, C.P., Anderson, G.M., Buxbaum, S.G., Elston, R.C., Ichinose, H., Nagatsu, T., Kim, K.S., Kim, C.H., Malison, R.T., Gelernter, J., Cubells, J.F., 2001. A quantitative-trait analysis of human plasma-dopamine $\beta$-hydroxylase activity: evidence for a major functional polymorphism at the DBH locus. Amer. J. Hum. Genet. 68, 515–522.