Biometrika (2001), **88**, 3, *pp.* 767–778

Printed in Great Britain

Testing the number of components in a normal mixture

BY YUNGTAI LO

Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.
ytlo@hustat.harvard.edu

NANCY R. MENDELL

Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, New York 11794, U.S.A.
nmendell@notes.cc.sunysb.edu

AND DONALD B. RUBIN

Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.
rubin@hustat.harvard.edu

SUMMARY

We demonstrate that, under a theorem proposed by Vuong, the likelihood ratio statistic based on the Kullback–Leibler information criterion of the null hypothesis that a random sample is drawn from a k_0-component normal mixture distribution against the alternative hypothesis that the sample is drawn from a k_1-component normal mixture distribution is asymptotically distributed as a weighted sum of independent chi-squared random variables with one degree of freedom, under general regularity conditions. We report simulation studies of two cases where we are testing a single normal versus a two-component normal mixture and a two-component normal mixture versus a three-component normal mixture. An empirical adjustment to the likelihood ratio statistic is proposed that appears to improve the rate of convergence to the limiting distribution.

Some key words: Kullback–Leibler information criterion; Likelihood ratio test; Normal mixture; Weighted sum of chi-squared random variables.

1. INTRODUCTION

Mixture distributions are important in a wide range of applications; see Everitt & Hand (1981), Titterington et al. (1985), McLachlan & Basford (1988) and McLachlan & Peel (2000). The most widely used finite mixture distributions are those with normal distributions as components, and in many practical situations the number of components in a mixture model is unknown.

Determining the number of components in a mixture distribution is an important but difficult problem. One approach is to do a formal test of the null hypothesis that a random sample is from a k_0-component mixture versus the alternative hypothesis that the sample is from a k_1-component mixture, where $k_1 > k_0$. It is tempting to use the likelihood ratio test with an asymptotic chi-squared null reference distribution. However, the necessary regularity conditions are not met in mixture model problems since, under the null hypoth-

esis, the general mixing proportions are on the boundary of the parameter space and the parameters are not identifiable under the null model.

Extending the results of Redner (1981) and Self & Liang (1987), Feng & McCulloch (1996) showed that the maximum likelihood estimator is consistent when the true parameter is on the boundary of the parameter space and in a nonidentifiable subset. Hartigan (1985) showed that the likelihood ratio test statistic, for testing a standard normal against a two-component normal mixture with a shift in mean, tends to infinity in probability at a very low rate in the case of an unbounded parameter space. Ghosh & Sen (1985) considered the case of testing homogeneity against location-contaminated normal mixtures with σ known. They showed that, under a separation condition imposed on the parameter space, the likelihood ratio test statistic is asymptotically distributed as a certain functional $\{\sup W(\mu_2)\}^2 I_{\{\sup W(\mu_2)>0\}}$, where $W(.)$ is a Gaussian process with mean zero and covariance kernel that depends on the true value of μ_1 under the null hypothesis; μ_1 and μ_2 are the component means. Berdai & Garel (1996) proved that Ghosh & Sen's result holds for the case of normal mixtures with σ unknown. Following Chernoff & Lander (1995), Lemdani & Pons (1999) reparameterised so as to solve the problems of nonidentifiability of the parameters, and showed that the limiting distribution of the likelihood ratio test statistic, for testing various mixtures of distributions from the same parametric family, is the supremum of a squared truncated Gaussian process. Dacunha-Castelle & Gassiat (1997, 1999) obtained the same result for the asymptotic null distribution of the likelihood ratio statistic using a locally conic parameterisation.

Goffinet et al. (1992) considered the case where the mixing proportions of a mixture distribution are known a priori. Their results suggest that the asymptotic distribution of the likelihood ratio test statistic depends on the value of the mixing proportion, the dimensionality of the problem and whether or not the within component variance is known. Windham & Cutler (1992) proposed an approach called minimum information ratio estimation and validation, based on the ratio of the Fisher information matrix obtained given the origins of individual observations unknown to the Fisher information matrix obtained given the origins known. The magnitude of the smallest eigenvalue of the information matrix ratio is used as an indicator of the number of components. A normalised entropy criterion for determining the number of clusters in the mixture context was discussed by Celeux & Soromenho (1996). Richardson & Green (1997) proposed a reversible jump Markov chain Monte Carlo method for simulating the posterior distribution of the number of components.

Following Aitkin et al. (1981), McLachlan (1987) proposed a bootstrap test for testing the number of components in a normal mixture. Soromenho (1994) compared the performance of five different approaches, Orlov's approach, described in a 1987 thesis by G. Celeux from L'Université de Paris IX Dauphine, Wolfe's (1970) approach, Aitkin & Rubin's (1985) approach, McLachlan's (1987) bootstrap approach and a procedure based on a stochastic version of the EM algorithm (Celeux & Diebolt, 1985), in the context of detecting the presence of mixtures. He concluded that the bootstrap approach and the stochastic EM procedure yield higher percentages of correct identification of the true model and yield higher empirical power.

In this paper, we propose a test procedure based on an extension of a theorem by Vuong (1989). Extending the work of White (1982), Vuong (1989) derived a likelihood ratio test for model selection based on the Kullback & Leibler (1951) information criterion. He showed that, under certain regularity conditions, the limiting distribution of the likelihood ratio statistic is a weighted sum of independent χ_1^2 random variables when the

competing models are nested or overlapping, and a normal distribution when the competing models are nonnested. These results do not require that either of the competing models be correctly specified. Vuong's test procedure was developed mainly for selecting models in multiple regression. Section 2 extends Vuong's theorem to cover the proposed likelihood ratio test of a k_0-component normal mixture versus a k_1-component normal mixture. The performance and statistical properties of the test are assessed in § 3. Suggestions for further study are given in § 4.

2. Development of the test

Let $X_1, \ldots, X_n$ be a random sample of size n from a finite normal mixture distribution denoted by $H(x; \vartheta)$ and with probability density function

$$h(x; \vartheta) = \sum_{i=1}^{k} \pi_i f_i(x; \mu_i, \sigma_i^2), \tag{1}$$

where $f_i(.)$ is a normal density with mean μ_i and variance σ_i^2,

$$\vartheta = (\pi_1, \mu_1, \sigma_1^2, \ldots, \pi_k, \mu_k, \sigma_k^2)$$

is the vector of unknown parameters, the π_i's are the mixing proportions, summing to 1, and k is the number of components. We assume throughout that the parameter k is unknown. Since all the component distributions are from the same parametric family, the mixture density in (1) is invariant under permutation of the component labels in ϑ. To deal with this lack of identifiability of ϑ, we assume without loss of generality that $\mu_1 < \mu_2 < \ldots < \mu_k$. If $\pi_i^{(1)} \neq 0$ or 1 for $i = 1, \ldots, k$, then this assumption ensures that $h(x; \vartheta^{(1)}) \equiv h(x; \vartheta^{(2)})$ implies $\vartheta^{(1)} = \vartheta^{(2)}$. The classic asymptotic chi-squared theory does not hold for the likelihood ratio test in the context of mixtures, so we consider here a likelihood ratio test procedure based on the Kullback–Leibler information criterion.

We postulate that the data arise either from a k_0-component normal mixture or from a k_1-component normal mixture, where k_0 and k_1 are known constants with $k_0 < k_1$. Let

$$F_\theta \equiv \{F(x; \theta); \theta \in \Theta \subset \mathscr{R}^p\}, \quad G_\gamma \equiv \{G(x; \gamma); \gamma \in \Gamma \subset \mathscr{R}^q\}$$

be respectively the family of k_1-component normal mixtures and the family of k_0-component normal mixtures, where

$$\theta = (\pi_1, \mu_1, \sigma_1^2, \ldots, \pi_{k_1}, \mu_{k_1}, \sigma_{k_1}^2)$$

with dimension $p = 3k_1 - 1$ and $\mu_1 < \mu_2 < \ldots < \mu_{k_1}$, and

$$\gamma = (\pi_1, \mu_1, \sigma_1^2, \ldots, \pi_{k_0}, \mu_{k_0}, \sigma_{k_0}^2)$$

with dimension $q = 3k_0 - 1$ and $\mu_1 < \mu_2 < \ldots < \mu_{k_0}$. Suppose further that the two competing models satisfy the assumptions A2, A3, A4 and A5 of Vuong (1989). Note that a k_0-component normal mixture is nested within a k_1-component normal mixture since it can be obtained by imposing appropriate restrictions on the parameter vector θ under a k_1-component normal mixture framework. That is, there exists a mapping, ϕ say, defined on Γ such that $G(x; \gamma) = F\{x; \phi(\gamma)\}$ for $\gamma \in \Gamma$. The two competing families may or may not contain the true distribution $H(x; \vartheta)$.

Define

$$I(h:f;\theta) = E_h\left\{\log\frac{h(X;\vartheta)}{f(X;\theta)}\right\}$$
$$= \int_{-\infty}^{\infty} \log h(x;\vartheta)\,dH(x;\vartheta) - \int_{-\infty}^{\infty} \log f(x;\theta)\,dH(x;\vartheta), \tag{2}$$

where $f(.)$ is the k_1-component normal mixture density, $h(.)$ given in (1) is the true underlying density and E_h denotes expectation with respect to $h(.)$. It is well known that $I(h:f;\theta) \geqslant 0$, with equality if and only if $f(x;\theta) = h(x;\vartheta)$ for every x in the support, that is if and only if the density is specified correctly by $f(.)$. Similarly, we measure the distance between the assumed k_0-component normal mixture density $g(.)$ and the true density by

$$I(h:g;\gamma) = E_h\left\{\log\frac{h(X;\vartheta)}{g(X;\gamma)}\right\}. \tag{3}$$

Following Sawa (1978), we say that the k_1-component normal mixture is a better approximation to the true distribution $H(.)$ than the k_0-component normal mixture if and only if

$$\inf_{\theta} I(h:f;\theta) < \inf_{\gamma} I(h:g;\gamma),$$

that is

$$\sup_{\theta} E_h\{\log f(X;\theta)\} > \sup_{\gamma} E_h\{\log g(X;\gamma)\}.$$

Therefore $E_h\{\log f(X;\theta^*)\}$ and $E_h\{\log g(X;\gamma^*)\}$ can be used for model selection, where θ^* and γ^* minimise (2) and (3).

Of course $E_h\{\log f(X;\theta^*)\}$ and $E_h\{\log g(X;\gamma^*)\}$ depend on the unknown true distribution but, from White (1982, Theorem 2.2), we know that, under some regularity conditions and as $n\to\infty$,

$$\frac{1}{n} L_f(\hat{\theta};x) \to E_h\{\log f(X;\theta^*)\}, \tag{4}$$

$$\frac{1}{n} L_g(\hat{\gamma};x) \to E_h\{\log g(X;\gamma^*)\}, \tag{5}$$

almost surely, where $\hat{\theta}$ and $\hat{\gamma}$ are the maximum likelihood estimators of θ^* and γ^*, respectively. Regularity assumptions, based on Assumptions A1, A2 and A3 of White (1982) or Vuong (1989), under which (4) and (5) hold are the following.

Assumption 1. The random variables $X_1,\ldots,X_n$ are independent and identically distributed with the density function $h(.)$, which is strictly positive for almost all x.

Assumption 2. (a) For every θ in Θ, $f(x;\theta)$ is strictly positive for almost all x.

(b) The parameter space Θ is a compact subset of $\mathscr{R}^p$ and $f(x;\theta)$ is continuous in θ for almost all x.

(c) For every γ in Γ, $g(x;\gamma)$ is strictly positive for almost all x.

(d) The parameter space Γ is a compact subset of $\mathscr{R}^q$ and $g(x;\gamma)$ is continuous in γ for almost all x.

Assumption 3. (a) For almost all x, $|\log f(x;\theta)|$ is bounded above by a function of x integrable with respect to H.

(b) The function $E_h\{\log f(x;\theta)\}$ has a unique maximum at θ^* in Θ.

(c) For almost all x, $|\log g(x;\gamma)|$ is bounded above by a function of x integrable with respect to H.

(d) The function $E_h\{\log g(x;\gamma)\}$ has a unique maximum at γ^* in Γ.

Assumptions 3(a) and 3(c) ensure that $E_h\{\log f(X;\theta)\}$ and $E_h\{\log g(X;\gamma)\}$ respectively are well defined. It can be shown that Assumption 3(a) is satisfied whenever the true within-component variances are finite and Θ does not contain $\sigma_i^2 = 0$ for $i = 1, \ldots, k_1$. Similarly, Assumption 3(c) is satisfied whenever the true within-component variances are finite and Γ does not contain $\sigma_i^2 = 0$ for $i = 1, \ldots, k_0$. Assumption 3(b) ensures that θ^* is identifiable. That is, there exists no other θ in Θ such that $E_h\{\log f(X;\theta)\} = E_h\{\log f(X;\theta^*)\}$. Similarly, Assumption 3(d) ensures that γ^* is identifiable. It is known that $\hat{\theta}$ or $\hat{\gamma}$ may not converge to a particular parameter vector of interest if the true model is not a k_1- or k_0-component normal mixture. Under Assumptions 1–3, $\hat{\theta}$ is a strongly consistent estimator for θ^* and $\hat{\gamma}$ is a strongly consistent estimator for γ^* (White, 1982). If the true model is a k_1-component normal mixture, that is $h(x;\vartheta) \equiv f(x;\theta_0)$ for some θ_0 in Θ, then it follows from Assumption 3(b) and Jensen's inequality that $\theta^* = \theta_0$ and $\hat{\theta}$ converges to the true parameter vector θ_0. Similarly, if the true model is a k_0-component mixture, by Assumption 3(d) and Jensen's inequality, $\hat{\gamma}$ is consistent for the true parameter vector. From (4) and (5), it can be shown that $1/n$ times the loglikelihood ratio is strongly consistent for the difference in the two quantities $E_h\{\log f(X;\theta^*)\} - E_h\{\log g(X;\gamma^*)\}$. That is, as $n \to \infty$,

$$\frac{1}{n}\,\mathrm{LR} \to E_h\left\{\log \frac{f(X;\theta^*)}{g(X;\gamma^*)}\right\}, \tag{6}$$

almost surely, where

$$\mathrm{LR} = \mathrm{LR}(\hat{\theta}, \hat{\gamma}; x) = L_f(\hat{\theta}; x) - L_g(\hat{\gamma}; x) = \sum_{j=1}^{n} \log \frac{f(X_j;\hat{\theta})}{g(X_j;\hat{\gamma})}. \tag{7}$$

Consider a test based on (6) for testing the null hypothesis that the k_0-component normal mixture and the k_1-component normal mixture are equally close to the true underlying distribution $H(x;\vartheta)$, that is

$$E_h\{\log f(x;\theta^*)\} = E_h\{\log g(x;\gamma^*)\},$$

against the alternative hypothesis that the k_1-component mixture is better than the k_0-component mixture, that is

$$E_h\{\log f(x;\theta^*)\} > E_h\{\log g(x;\gamma^*)\}.$$

The test will reject the null hypothesis that the data arise from the k_0-component normal mixture $G(x;\gamma^*)$, if the test statistic 2LR as defined in (7) is greater than or equal to some constant determined by the size of the test. Define the matrices

$$A_f(\theta^*) = E_h\left\{\frac{\partial^2 \log f(X;\theta^*)}{\partial\theta\,\partial\theta'}\right\}, \tag{8}$$

$$A_g(\gamma^*) = E_h\left\{\frac{\partial^2 \log g(X;\gamma^*)}{\partial\gamma\,\partial\gamma'}\right\}, \tag{9}$$

$$B_f(\theta^*) = E_h\left\{\frac{\partial \log f(X;\theta^*)}{\partial\theta}\,\frac{\partial \log f(X;\theta^*)}{\partial\theta'}\right\}, \tag{10}$$

$$B_g(\gamma^*) = E_h\left\{\frac{\partial \log g(X;\gamma^*)}{\partial \gamma}\,\frac{\partial \log g(X;\gamma^*)}{\partial \gamma'}\right\}, \tag{11}$$

$$B_{fg}(\theta^*,\gamma^*) = B'_{gf}(\gamma^*,\theta^*) = E_h\left\{\frac{\partial \log f(X;\theta^*)}{\partial \theta}\,\frac{\partial \log g(X;\gamma^*)}{\partial \gamma'}\right\}. \tag{12}$$

THEOREM 1. *Under Assumptions* 1–3 *and some additional regularity assumptions, the asymptotic distribution of* 2LR *is a weighted sum of* $p+q$ *independent* χ_1^2 *random variables under the null hypothesis; that is, as* $n\to\infty$,

$$\mathrm{pr}(2\mathrm{LR}\leqslant y)\to M_{p+q}(y;\lambda)\quad (y\geqslant 0),$$

where $M_{p+q}(.)$ *is the distribution function of a weighted sum of* χ_1^2 *variables and* $\lambda = (\lambda_1,\ldots,\lambda_{p+q})$ *is the vector of* $p+q$ *eigenvalues of*

$$W = \begin{pmatrix} -B_f(\theta^*)A_f^{-1}(\theta^*) & -B_{fg}(\theta^*,\gamma^*)A_g^{-1}(\gamma^*) \\ B_{gf}(\gamma^*,\theta^*)A_f^{-1}(\theta^*) & B_g(\gamma^*)A_g^{-1}(\gamma^*) \end{pmatrix}.$$

The proof of Theorem 1 is given in Vuong (1989, Theorem 3.3(i)). The additional regularity assumptions are analogues of Assumptions A4, A5 and A8 of Vuong (1989).

Assumption 4. (a) For almost all x, $\log f(x;\theta)$ is twice continuously differentiable with respect to θ and $\log g(x;\gamma)$ is twice continuously differentiable with respect to γ.

(b) For almost all x,

$$\left|\frac{\partial^2 \log f(x;\theta)}{\partial\theta\,\partial\theta'}\right|,\quad \left|\left\{\frac{\partial \log f(x;\theta)}{\partial\theta}\right\}\left\{\frac{\partial \log f(x;\theta)}{\partial\theta'}\right\}\right|,\quad \left|\frac{\partial^2 \log g(x;\gamma)}{\partial\gamma\,\partial\gamma'}\right|,$$

$$\left|\left\{\frac{\partial \log g(x;\gamma)}{\partial\gamma}\right\}\left\{\frac{\partial \log g(x;\theta)}{\partial\gamma'}\right\}\right|$$

are bounded above by functions of x integrable with respect to H.

Assumption 5. (a) The value θ^* is an interior point of Θ and a regular point of $A_f(\theta)$.

(b) The value γ^* is an interior point of Γ and a regular point of $A_g(\gamma)$.

Assumption 6. There exists a function $\phi(.)$ from Γ to Θ such that, for almost all x, $g(x;\gamma) = f\{x;\phi(\gamma)\}$ for γ in Γ.

Assumption 4 ensures the existence of the matrices as defined in (8), (9), (10), (11) and (12). It can be shown that Assumption 4 is satisfied here. A regular point of the matrix $A_f(\theta)$ is defined as a value of θ such that $A_f(\theta)$ has constant rank in an open neighbourhood of θ. Under Assumptions 1–4, Assumption 5 ensures that $A_f(\theta^*)$ and $A_g(\gamma^*)$ are negative definite, and hence both are nonsingular (White, 1982). Given Assumptions 1–3 and 6, $E_h\{\log f(x;\theta^*)\} = E_h\{\log g(x;\gamma^*)\}$ implies that $f(x;\theta^*)\equiv g(x;\gamma^*)$. The weights λ_i are all real and some of them may be negative. If λ_i only takes the value 0 or 1, for $i=1,\ldots,p+q$, then 2LR has the regular chi-squared distribution with degrees of freedom equal to the number of nonzero λ_i's. In practice, λ can be consistently estimated by the eigenvalues, $\hat\lambda = (\hat\lambda_1,\ldots,\hat\lambda_{p+1})$, of the matrix $\hat W$ which is an estimator of W obtained by replacing (8), (9), (10), (11) and (12) by sample averages evaluated at $\hat\theta$ and $\hat\gamma$; for example $A_f(\theta)$ is

replaced by

$$A_{fn}(\hat{\theta}) = \left(\frac{1}{n} \sum_{j=1}^{n} \frac{\partial^2 \log f(x_j; \hat{\theta})}{\partial \theta \, \partial \theta'} \right),$$

and so on.

The distribution function of 2LR can be derived by the use of the inversion formula (Roussas, 1997, p. 141) and the uniqueness theorem of characteristic functions (Gnedenko & Kolmogorov, 1968, p. 50). It follows from Imhof (1961) that the distribution function of 2LR is

$$M_{p+q}(y; \lambda) = \text{pr}(2\text{LR} < y) = \frac{1}{2} - \frac{1}{\pi} \int_0^{\infty} \frac{\sin \delta(u)}{u\rho(u)} \, du, \tag{13}$$

where

$$\delta(u) = \frac{1}{2} \sum_{i=1}^{p+q} \arctan(\lambda_i u) - \frac{1}{2} yu, \quad \rho(u) = \prod_{i=1}^{p+q} (1 + \lambda_i^2 u^2)^{1/4}.$$

The integral in (13) is not available in closed form and will be evaluated through quadrature formulae such as Simpson's rule or the trapezoidal rule. Since the integrand $\sin \delta(u)/\{u\rho(u)\}$ in (13) converges to zero as $u\rho(u)$ tends to infinity, the numerical integration is carried out over a truncated range of the infinite interval $[0, \infty)$ and ignores the tail of the integrand which is defined as the truncation error. Since $\sin \delta(u)/\{u\rho(u)\}$ is undefined at $u = 0$, the integrand at $u = 0$ is replaced by its limit, obtained from L'Hôpital's rule as

$$\lim_{u \to 0} \frac{\sin \delta(u)}{u\rho(u)} = \frac{1}{2} \sum_{i=1}^{p+q} \lambda_i - \frac{1}{2} y.$$

Thus (13) can be approximated by

$$M_{p+q}(y; \lambda) = \text{pr}(2\text{LR} < y) \simeq \frac{1}{2} - \frac{1}{\pi} \int_0^{U} I(u) \, du, \tag{14}$$

where

$$I(u) = \begin{cases} \{\sin \delta(u)\}/\{u\rho(u)\} & (u \neq 0), \\ \frac{1}{2} \sum_{i=1}^{p+q} \lambda_i - \frac{1}{2} y & (u = 0). \end{cases}$$

In (14), U is chosen so that the truncation error is less than some small number.

3. Simulation results

Simulation studies were conducted to investigate the finite sample properties of the test. The maximum likelihood estimates of the parameters throughout were obtained by the EM algorithm (Dempster et al., 1977) with 100 sets of random starting values for the parameters. The p-values were computed from (14) with both the approximation error, which is defined as the error that arises from using quadrature formulae for computing (14), and the truncation error less than 10^{-7}. Two cases were considered. In the first case, the null hypothesis was that a random sample has been drawn from a normal distribution, $k = 1$, and the alternative hypothesis was that the sample has been drawn from a mixture of two normal distributions, $k = 2$, with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$. Sample sizes $n = 50$, 75, 100,

150, 200, 300, 400, 500 and 1000 were considered. For each sample size, 1000 samples were generated from the standard normal distribution. Under the null hypothesis, the test statistic 2LR defined in (7) is asymptotically distributed as a weighted sum of $p+q=6$ independent χ_1^2 random variables. Table 1(a) shows the simulated significance levels for 2LR at $\alpha=0{\cdot}01$ and $0{\cdot}05$ based on 1000 null samples for each sample size. Comparisons of the actual levels to the nominal levels for each sample size indicate that the rate of convergence of 2LR toward its limiting distribution is very slow. To improve the accuracy of the approximation, we considered the ad hoc adjusted test statistic

$$2\text{LR}^* = \frac{2\text{LR}}{1+\{(p-q)\log n\}^{-1}}. \tag{15}$$

Note that the correction term $\{(p-q)\log n\}^{-1}$ goes to zero as n goes to infinity. The results in Table 1(b) show that the approximation is reasonably satisfactory for most sample sizes.

Table 1. *Simulated significance levels for (a) the unadjusted and (b) the adjusted tests for testing a single normal versus a mixture of two normals based on* 1000 *replications for each sample size*

(*a*) *Unadjusted test*

Nominal level	Sample size								
	50	75	100	150	200	300	400	500	1000
0·01	0·009	0·014	0·013	0·010	0·012	0·012	0·015	0·015	0·019
0·05	0·078	0·076	0·072	0·075	0·061	0·070	0·067	0·069	0·061

(*b*) *Adjusted test*

Nominal level	Sample size								
	50	75	100	150	200	300	400	500	1000
0·01	0·005	0·009	0·006	0·008	0·008	0·009	0·008	0·010	0·016
0·05	0·057	0·054	0·049	0·047	0·045	0·050	0·052	0·051	0·053

The observed power of the adjusted and unadjusted likelihood ratio tests was estimated from 1000 samples drawn from a two-component normal mixture alternative. We used 27 configurations, based on three values of the mixing proportion π, 0·5, 0·7 and 0·9, three sample sizes, 50, 100 and 200, and three values of D, 1, 2 and 3, where $D=(\mu_2-\mu_1)/\sigma$ measures the distance between the two components; μ_1 was set to 0 and σ was set to 1. The empirical powers of the adjusted and unadjusted tests at $\alpha=0{\cdot}01$ and $0{\cdot}05$ are shown in Table 2. The power is seen to be very low for $n<200$ when the two components are not well separated, that is $D=1$ or 2, and a sample size of 100 or more is required to have reasonable power when the two components are well separated, $D=3$. There is no strong evidence that the power depends on the mixing proportion, although the power for $\pi=0{\cdot}5$ and $\pi=0{\cdot}7$ is somewhat higher than that for $\pi=0{\cdot}9$ when $D=3$. The power of the unadjusted test is inflated, because the approximation of the nominal levels of 0·01 and 0·05 to the actual levels of the unadjusted test is not accurate. As a consequence, the unadjusted test tends to reject the null hypothesis more often than does the adjusted test. For $\alpha=0{\cdot}05$, our results are consistent with those of Mendell et al. (1991). They reported that the power depends on the spacing between the components, sample sizes and the

Table 2. *Simulated powers, in percentages, of the adjusted and unadjusted tests for testing a single normal versus a two-component normal mixture based on* 1000 *replications under each alternative*

Mixing proportion	Sample size	Nominal level	$D=1$		$D=2$		$D=3$	
			2LR	2LR*	2LR	2LR*	2LR	2LR*
0·5	50	0·01	1·8	0·9	3·2	2·5	35·9	28·7
		0·05	7·8	5·4	13·4	9·4	64·2	57·4
	100	0·01	1·2	0·8	7·3	5·0	75·2	69·8
		0·05	7·3	5·7	22·4	17·5	91·0	88·4
	200	0·01	1·8	1·4	20·1	16·1	98·2	97·7
		0·05	8·4	7·0	43·2	37·9	99·5	99·3
0·7	50	0·01	2·2	1·4	5·8	3·5	40·8	33·0
		0·05	8·3	6·0	17·2	13·4	67·9	61·6
	100	0·01	1·5	0·8	11·0	8·8	80·7	75·5
		0·05	8·0	5·4	29·5	25·1	93·3	91·1
	200	0·01	1·8	0·7	29·1	22·9	98·9	98·7
		0·05	8·7	6·1	51·5	46·5	99·8	99·8
0·9	50	0·01	2·0	1·3	7·0	5·0	36·0	30·7
		0·05	9·6	7·1	18·6	14·3	56·7	51·3
	100	0·01	1·4	1·2	11·6	9·0	69·5	64·9
		0·05	8·0	6·1	29·1	25·0	85·7	83·3
	200	0·01	1·7	1·2	25·1	21·6	95·9	94·3
		0·05	9·1	7·1	47·9	43·5	99·1	98·8

mixing proportion; they used as the critical value the observed 95th percentage point based on 2500 replications (Thode et al., 1988).

We next considered the case of a null hypothesis that a random sample has been drawn from a mixture of two normal distributions, $k=2$, with $\mu_1<\mu_2$ and $\sigma_1=\sigma_2$, and an alternative hypothesis that the sample has been drawn from a mixture of three normal distributions, $k=3$, with $\mu_1<\mu_2<\mu_3$ and $\sigma_1=\sigma_2=\sigma_3$. Under the null hypothesis, 2LR is asymptotically distributed as a weighted sum of $p+q=10$ independent χ^2_1 random variables. Three values of π, 0·5, 0·7 and 0·9, six sample sizes, 50, 75, 100, 150, 200 and 300, and three values of D, 1, 2 and 3, were used to generate null samples. The simulated significance levels of the adjusted and unadjusted tests are shown for $\alpha=0{\cdot}01$ and 0·05 in Table 3; since the results vary little with π, only results for $\pi=0{\cdot}7$ are given. Table 3 shows that the approximation of the nominal sizes to the actual sizes of the unadjusted test is inadequate for $D=2$ and 3. The results also indicate that the rate of convergence of the adjusted test statistic toward the asymptotic distribution depends on the spacing between the components and the mixing proportion. Regardless of the mixing proportion, the distribution of 2LR* converges rather well for $D=1$ and 2. When $D=3$, the approximation is not very accurate but is acceptable for most sample sizes.

An empirical power study was conducted to examine the effect of the spacing between components on the power of the adjusted test for testing a two-component mixture versus a three-component mixture. A total of 96 configurations of the alternative were used, involving two sets of values of the mixing proportions, $\pi_1=0{\cdot}3$, $\pi_2=0{\cdot}4$ and $\pi_3=0{\cdot}3$, and $\pi_1=0{\cdot}5$, $\pi_2=0{\cdot}4$ and $\pi_3=0{\cdot}1$, three sample sizes, 50, 100 and 200, four values of D_1, 1, 2, 3 and 4, and four values of D_2, 1, 2, 3 and 4, where $D_1=(\mu_2-\mu_1)/\sigma$, $D_2=(\mu_3-\mu_2)/\sigma$, and

Table 3. *Simulated significance levels, in percentages, of the adjusted and unadjusted tests for testing a mixture of two normals versus a mixture of three normals based on* 1000 *replications under each null hypothesis*

Mixing proportion	Sample size	Nominal level	$D=1$		$D=2$		$D=3$	
			2LR	2LR*	2LR	2LR*	2LR	2LR*
0·7	50	0·01	1·4	0·9	1·9	1·2	2·3	1·0
		0·05	8·5	6·0	8·1	5·7	8·4	6·3
	75	0·01	1·5	1·1	1·0	0·6	1·7	1·0
		0·05	6·2	4·5	5·7	4·1	7·2	5·1
	100	0·01	1·0	0·6	2·0	1·4	2·1	1·4
		0·05	6·2	4·1	8·0	6·0	9·1	7·0
	150	0·01	1·4	0·7	1·5	1·2	2·2	1·8
		0·05	5·1	3·9	7·7	5·6	7·9	6·1
	200	0·01	0·6	0·2	2·0	1·5	2·4	1·3
		0·05	5·6	4·3	7·4	5·8	8·5	6·8
	300	0·01	1·0	0·8	1·5	1·0	2·2	0·9
		0·05	6·1	4·7	8·6	6·1	7·7	6·3

μ_1 and σ were set equal to 0 and 1, respectively. For each configuration, 1000 samples were generated. Table 4 gives the observed rejection rates of the null hypothesis at $\alpha = 0{\cdot}01$ and $0{\cdot}05$ for the first set of mixing weights. The results indicate that the factors with most influence on power are the spacings D_1 and D_2. The adjusted test exhibits low power for sample sizes less than 200 when $D_1 \leqslant 4$ and $D_2 \leqslant 2$, or $D_1 \leqslant 2$ and $D_2 \leqslant 4$. As in the first example, a sample of size 100 or more is required for reasonable power with mixtures with $D_1 \geqslant 3$ and $D_2 \geqslant 3$. Qualitatively similar results were obtained with the second set of mixing weights.

4. DISCUSSION

Recall that the result given in Theorem 1 for the asymptotic null distribution of 2LR imposes no explicit restriction on the parameters, except that needed to ensure that the parameters are identifiable. As noted earlier, the likelihood ratio statistic converges in probability to infinity in the case of an unbounded parameter space, and hence the asymptotic result fails. However, if a separation condition is imposed on the parameter space, for example if the Mahalanobis distance between the two normal components for $k=2$ is restricted to a closed interval, then 2LR has the same but nondegenerate limiting distribution. Of course, for finite n, 2LR has a proper distribution and the result given in this paper with the suggested adjustment for the asymptotic null distribution of 2LR can be a useful approximation in practice.

The simulation studies suggest that the test works well for the case of homoscedastic normal mixtures. A priority in future research will be to investigate the performance of the test for assessing the number of components in a normal mixture with unequal variances. In particular, we will examine how the rate of convergence of 2LR to the limiting distribution depends on the choice of restrictions imposed on the component variances to deal with the problem of unboundedness of the likelihood. It would be of interest to compare this method with the parametric bootstrap method (McLachlan, 1987), the Bayesian model checking procedure called posterior predictive checks (Rubin, 1984) and

Table 4. *Simulated powers, in percentages, of the adjusted and unadjusted tests for testing a mixture of two normals versus a mixture of three normals based on* 1000 *replications under each alternative with the mixing proportions* $\pi_1 = 0{\cdot}3$, $\pi_2 = 0{\cdot}4$

	Sample	Nominal	$D_2 = 1$		$D_2 = 2$		$D_2 = 3$		$D_2 = 4$	
D_1	size	level	2LR	2LR*	2LR	2LR*	2LR	2LR*	2LR	2LR*
1	50	0·01	2·5	1·5	2·3	1·5	1·9	1·2	2·6	1·4
		0·05	9·1	6·8	8·6	5·9	9·4	6·8	8·5	6·2
	100	0·01	1·7	1·1	2·0	1·1	1·4	1·0	2·0	1·5
		0·05	6·2	4·9	7·6	5·3	9·0	6·5	10·0	8·7
	200	0·01	0·8	0·5	2·3	1·1	2·0	1·3	2·7	2·2
		0·05	6·3	4·1	8·2	6·9	8·6	6·8	10·4	8·4
2	50	0·01	2·2	1·4	2·6	1·2	4·9	3·5	8·0	5·8
		0·05	9·1	7·5	9·6	6·6	15·6	10·9	23·5	17·2
	100	0·01	1·5	0·9	2·7	1·8	6·8	3·9	15·7	11·4
		0·05	9·6	6·7	9·8	6·8	20·5	15·5	38·4	33·1
	200	0·01	1·2	0·8	3·6	2·6	18·6	13·8	44·5	38·6
		0·05	8·1	6·2	12·7	9·9	40·9	35·9	68·3	64·0
3	50	0·01	1·5	0·8	4·3	2·4	15·5	10·4	37·3	29·4
		0·05	8·1	6·2	12·8	10·2	36·2	28·8	62·9	56·3
	100	0·01	2·4	1·8	7·3	4·5	39·3	32·5	78·1	71·9
		0·05	9·8	7·0	24·2	19·5	65·3	59·8	91·6	89·6
	200	0·01	1·9	1·1	18·8	15·1	80·7	75·5	98·5	98·1
		0·05	10·0	7·9	39·3	33·3	91·6	89·8	99·9	99·9
4	50	0·01	2·3	1·2	6·3	3·6	34·6	27·0	77·2	69·8
		0·05	8·9	6·0	18·4	13·0	61·3	54·0	91·6	88·9
	100	0·01	2·4	1·5	18·0	12·0	78·5	72·7	99·1	98·7
		0·05	11·4	9·2	38·1	32·3	92·0	90·2	99·9	99·8
	200	0·01	2·7	1·9	41·9	35·0	99·5	98·8	100·0	100·0
		0·05	12·2	10·2	67·5	61·9	100·0	100·0	100·0	100·0

the procedure developed by Lemdani & Pons (1999) or Dacunha-Castelle & Gassiat (1999). It would also be interesting to extend Vuong's theorem to the problems of determining the number of components in a gamma mixture, an exponential mixture, a binomial mixture or a Poisson mixture.

Acknowledgement

The authors wish to thank Stephen Finch for inspiring this work and the editor and referees for very helpful comments and suggestions that improved the presentation of this paper. This work was partially supported by grants from the National Institutes of Health and from the Roy Hunt Foundation.

References

Aitkin, M., Anderson, D. & Hinde, J. (1981). Statistical modelling of data on teaching styles (with Discussion). *J. R. Statist. Soc.* A **144**, 419–61.

Aitkin, M. & Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *J. R. Statist. Soc.* B **47**, 67–75.

BERDAI, A. & GAREL, B. (1996). Detecting a univariate normal mixture with two components. *Statist. Decis.* **14**, 35–51.

CELEUX, G. & DIEBOLT, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Statist. Quart.* **2**, 73–82.

CELEUX, G. & SOROMENHO, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *J. Classif.* **13**, 195–212.

CHERNOFF, H. & LANDER, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *J. Statist. Plan. Infer.* **43**, 19–40.

DACUNHA-CASTELLE, D. & GASSIAT, E. (1997). Testing in locally conic models and application to mixture models. *Eur. Ser. Appl. Indust. Math.: Prob. Statist.* **1**, 285–317.

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**, 1178–209.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum-likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1–38.

EVERITT, B. S. & HAND, D. J. (1981). *Finite Mixture Distributions.* London: Chapman and Hall.

FENG, Z. D. & MCCULLOCH, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *J. R. Statist. Soc.* B **58**, 609–17.

GHOSH, J. H. & SEN, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, **2**, Ed. L. M. Le Cam and R. A. Olshen, pp. 789–806. Monterey, CA: Wadsworth.

GNEDENKO, B. V. & KOLMOGOROV, A. N. (1968). *Limit Distributions for Sums of Independent Random Variables.* Translated by K. L. Chung. Reading, MA: Addison-Wesley.

GOFFINET, B., LOISEL, P. & LAURENT, B. (1992). Testing in normal mixture models when the proportions are known. *Biometrika* **79**, 842–6.

HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, **2**, Ed. L. M. Le Cam and R. A. Olshen, pp. 807–10. Monterey, CA: Wadsworth.

IMHOF, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48**, 419–26.

KULLBACK, S. & LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

LEMDANI, M. & PONS, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli* **5**, 705–19.

MCLACHLAN, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.* **36**, 318–24.

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

MCLACHLAN, G. J. & PEEL, D. (2000). *Finite Mixture Models.* New York: Wiley.

MENDELL, N. R., THODE, H. C. & FINCH, S. J. (1991). The likelihood ratio test for the two-component normal mixture problem: power and sample size analysis. *Biometrics* **47**, 1143–8.

REDNER, R. A. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Ann. Statist.* **9**, 225–8.

RICHARDSON, S. & GREEN, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *J. R. Statist. Soc.* B **59**, 731–92.

ROUSSAS, G. G. (1997). *A Course in Mathematical Statistics.* New York: Academic Press.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151–72.

SAWA, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica* **46**, 1273–91.

SELF, S. G. & LIANG, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Statist. Assoc.* **82**, 605–10.

SOROMENHO, G. (1994). Comparing approaches for testing the number of components in a finite mixture model. *Comp. Statist.* **9**, 65–78.

THODE, H. C., FINCH, S. J. & MENDELL, N. R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test for the mixture of two normals. *Biometrics* **44**, 1195–201.

TITTERINGTON, D. M., SMITH, A. F. M. & MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: Wiley.

VUONG, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57**, 307–33.

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

WINDHAM, M. P. & CUTLER, A. (1992). Information ratio for validating mixture analyses. *J. Am. Statist. Assoc.* **87**, 1188–92.

WOLFE, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivar. Behav. Res.* **5**, 329–50.

[*Received April* 2000. *Revised January* 2001]