On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture

By G. J. McLACHLAN†

University of Queensland, Australia

SUMMARY
An important but difficult problem in practice is assessing the number of components *g* in a mixture. An obvious way of proceeding is to use the likelihood ratio test statistic λ to test for the smallest value of *g* consistent with the data. Unfortunately with mixture models, regularity conditions do not hold for $-2 \log \lambda$ to have it usual asymptotic null distribution of chi-squared. In this paper the role of the bootstrap is highlighted for the assessment of the null distribution of $-2 \log \lambda$ for the test of a single normal density versus a mixture of two normal densities in the univariate case.

Keywords: Mixture distributions; Normal components; EM algorithm; Likelihood ratio test; Bootstrap procedures

1. Introduction

Finite mixtures of distributions, in particular normal mixtures, have been used widely in practice over the years; see Do and McLachlan (1984) for a recent case study. Let $x_1, \ldots, x_n$ denote the *p*-dimensional observed values of a random sample of size *n*. In fitting a finite mixture of *g* components to these data, it is assumed that the probability density function (p.d.f.) of an observation **x** can be represented in the form

$$f(\mathbf{x}; \boldsymbol{\phi}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}; \boldsymbol{\theta}), \qquad (1.1)$$

where $f_i(\mathbf{x}; \boldsymbol{\theta})$ is the p.d.f. corresponding to the *i*th component and $\boldsymbol{\theta}$ denotes the vector of all unknown parameters associated with the parametric forms adopted for these *g* component densities. We let

$$\boldsymbol{\phi} = (\boldsymbol{\pi}', \boldsymbol{\theta}')',$$

be the vector of all unknown parameters, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)'$ is the vector of mixing proportions. An estimate of $\boldsymbol{\phi}$ can be obtained as a solution of the likelihood equation

$$\partial L(\boldsymbol{\phi})/\partial \boldsymbol{\phi} = \mathbf{0}, \qquad (1.2)$$

where $L(\boldsymbol{\phi})$ denotes the log likelihood function for $\boldsymbol{\phi}$ formed under the mixture model (1.1). We shall refer to our chosen solution of (1.2), $\hat{\boldsymbol{\phi}}$, as the likelihood estimate of $\boldsymbol{\phi}$. For an observed sample $\hat{\boldsymbol{\phi}}$ is usually taken to be the root of (1.2) corresponding to the largest of the local maxima located. That is, in those cases where $L(\boldsymbol{\phi})$ has a global maximum in the interior of the parameter space, $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimate of $\boldsymbol{\phi}$, assuming that the global

---

† *Address for correspondence*: Department of Mathematics, University of Queensland, St. Lucia 4067, Australia.

maximum has been located. An account of likelihood estimation under a mixture model may be found in Basford and McLachlan (1985b), who discuss the computation of $\hat{\phi}$ using the *EM* algorithm of Dempster, Laird and Rubin (1977), and who give the form of the likelihood equation for a mixture of normal densities with arbitrary covariance matrices. For additional references on finite mixture distributions the reader is referred to the review articles by McLachlan (1982) and Redner and Walker (1984), and to the monographs devoted entirely to the topic by Everitt and Hand (1981), Titterington, Smith and Makov (1985) and, more recently, by McLachlan and Basford (1987).

In many situations in practice, for example in cluster analysis where the number of components g correspond to the number of non-overlapping clusters to be produced for the data, there is no *a priori* knowledge of g. An obvious way of approaching this problem is to use the likelihood ratio test statistic λ to test for the smallest value of g compatible with the data. Unfortunately with mixture models, it is well known that regularity conditions do not hold for $-2 \log \lambda$ to have its usual asymptotic null distribution of chi-squared with degrees of freedom equal to the difference between the number of parameters under the null and alternative hypotheses.

In the context of latent class models, Aitkin, Anderson and Hinde (1981) assessed the null distribution of $-2 \log \lambda$ using a resampling method which can be viewed as a particular application of the general bootstrap approach of Efron (1979, 1982). The assessment is made on the basis of a number, say K, of bootstrap replications of $-2 \log \lambda$ obtained by appropriate resampling. More recently, Aitkin and Rubin (1985) adopted an approach which places a prior distribution on the vector of mixing proportions over which the likelihood is integrated before the evaluation of $-2 \log \lambda$.

In this paper we shall examine further the bootstrapping of the log likelihood ratio statistic for testing the number of components g in the simplest situation of $g = 1$ under the null hypothesis H_0 versus $g = 2$ under the alternative hypothesis H_1. Attention is to be focussed on H_0: a single univariate normal distribution *versus* H_1: a mixture of two normal distributions with a common variance. We shall report the results of simulations performed in this case to demonstrate the effect of the size of K on the estimate of the P-value of the test based on $-2 \log \lambda$, and also on the power of the test constructed from $-2 \log \lambda$ and its K subsequent bootstrap replications so as to have a nominal size. An example involving a real data set is to be discussed.
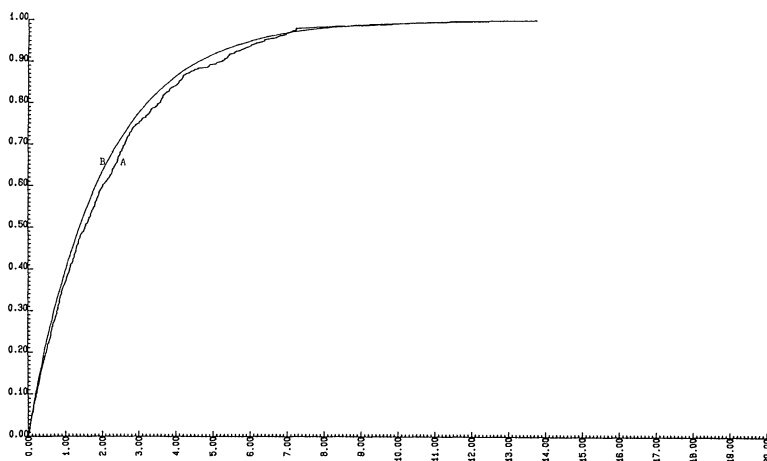

Fig. 1. Plots of distribution functions: A, simulated null distribution of $-2 \log \lambda$ for the test of $g = 1$ versus $g = 2$ normal component densities with equal variances for $n = 100$; B, χ_2^2 distribution.

Before we proceed to discuss the bootstrapping of the log likelihood ratio statistic in testing for g, we briefly report the results of a simulation experiment performed for assessing the null distribution of $-2 \log \lambda$ for H_0 versus H_1 with univariate normal component densities; see also Everitt (1981). Since $g = 1$ under H_0, the distribution of λ, and hence of $-2 \log \lambda$, clearly depends only on the sample size n. For the homoscedastic case of equal component variances, 500 samples of size $n = 25$, 50 and 100 respectively, were generated under H_0. The mean (variance) of the simulated null distribution of $-2 \log \lambda$ is equal to 2.47(5.66), 2.36(5.06), and 2.16(4.30) for $n = 25$, 50 and 100 respectively. The empirical distribution function of $-2 \log \lambda$, formed from the 500 simulated values of this statistic for $n = 100$, is plotted in Fig. 1, along with the χ_2^2 distribution function. The latter corresponds to the approximation of Wolfe (1971), where the degrees of freedom of the chi-squared distribution is taken to be twice the difference in the number of parameters under H_0 and H_1, not including the mixing proportions. Concerning the approximation of the null distribution of $-2 \log \lambda$, in particular its upper tail, by the χ_2^2 distribution, it would appear that it is reasonable to use the latter for the determination of P-values at least for $n \geqslant 100$.

It is worth noting in passing that Wolfe's approximation to the null distribution of $-2 \log \lambda$ may not be applicable in the heteroscedastic case. This can be seen from Fig. 2, which gives the empirical distribution function of $-2 \log \lambda$ constructed from 500 samples of size $n = 100$ generated under H_0. On this occasion in evaluating $-2 \log \lambda$ for each of the simulated samples, the normal component densities were taken to have unequal variances under H_1. Also plotted in Fig. 2 are the χ_4^2 and χ_6^2 distribution functions, the former corresponding to Wolfe's approximation. However, it can be seen that the latter distribution provides a better fit. The simulated null distribution of $-2 \log \lambda$ has mean and variance equal to 5.96 and 13.86 respectively.

2. Bootstrapping the Log Likelihood Test Statistic

The log likelihood ratio statistic for the test of the null hypothesis of $g = g_1$ versus the alternative of $g = g_2$ can be bootstrapped as follows. Proceeding under the null hypothesis, a so-called bootstrap sample is generated from the mixture density $f(x; \hat{\boldsymbol{\phi}})$, where $\hat{\boldsymbol{\phi}}$ is taken to be the likelihood estimate of $\boldsymbol{\phi}$ formed under the null hypothesis from the original sample. The value of $-2 \log \lambda$ is computed for the bootstrap sample after fitting mixture models for $g = g_1$ and g_2 in turn to it. This process is repeated independently a number of times K, and
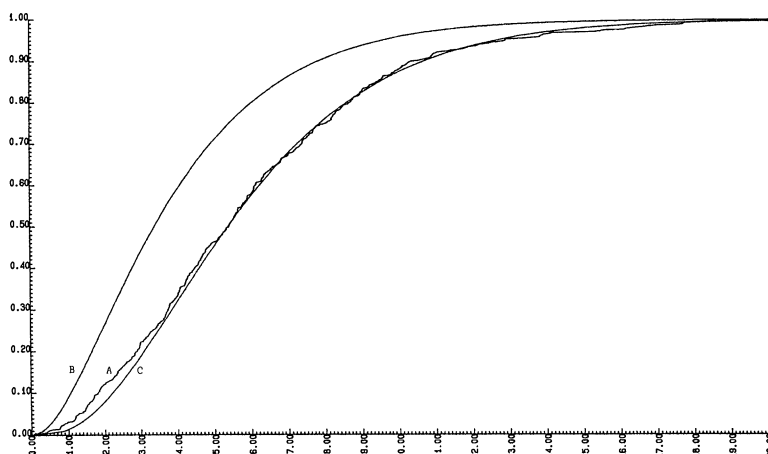

Fig. 2. Plots of distribution functions: A, simulated null distribution of $-2 \log \lambda$ for the test of $g = 1$ versus $g = 2$ normal component densities with unequal variances for $n = 100$; B, χ_4^2 distribution; C, χ_6^2 distribution.

the replicated values of $-2 \log \lambda$ evaluated from the successive bootstrap samples can be used to assess the bootstrap, and hence the true, null distribution of $-2 \log \lambda$. In particular, it enables an approximation to be made to the achieved level of significance P corresponding to the value of $-2 \log \lambda$ evaluated from the original sample. The value of the jth order statistic of the K replications can be taken as an estimate of the quantile of order $j/(K+1)$, and the P-value can be assessed by reference with respect to the ordered bootstrap replications of $-2 \log \lambda$.

We focus again on the simplest situation of $H_0: g = 1$ versus $H_1: g = 2$, where the component distributions are normal with a common variance. Let q_α denote the αth quantile of the null distribution of $-2 \log \lambda$. We consider $q_{0.95}$ corresponding to a conventional 0.05 level test. An estimate of this quantile, $\hat{q}_{0.95}$, is provided by the 19th, 38th, and 95th order statistics of the K replications of $-2 \log \lambda$, corresponding to $K = 19$, 39, and 99 respectively. These sizes of K have been so chosen as in the next section the K bootstrap replications are to be used in conjunction with the original value of $-2 \log \lambda$ to provide a test of size 0.05.

For a sample size of $n = 100$, the sample mean (standard deviation) of $\hat{q}_{0.95}$ taken over 100 simulation trials executed under the null hypothesis was 7.62(2.48), 7.15(1.78), and 6.62(1.03) for $K = 19$, 39, and 99 respectively. As noted earlier for the present situation, the null distribution of λ depends only on n, and so the bootstrap and true null distributions of $-2 \log \lambda$ are the same. The true value of $q_{0.95}$ is unknown but, as seen in the previous section, it can be well approximated for $n = 100$ by $\chi^2_{2,0.95} = 5.99$, the 95th percentile of the χ^2_2 distribution. As anticipated the variance of $\hat{q}_{0.95}$ is quite large for $K = 19$, but drops off sharply as K is increased to 99. With the null distribution F of $-2 \log \lambda$ approximated by the χ^2_2 distribution, we computed also for the same simulated data the sample mean (standard deviation) of $F(\hat{q}_{0.95})$, obtaining 0.961(0.036), 0.961(0.030), and 0.959(0.018) for $K = 19$, 39, and 99 respectively. If a very accurate estimate of a P-value were required, then K would have to be very large. Indeed, for less complicated models than mixtures, Efron and Tibshirani (1986) have shown that whereas 50 to 100 bootstrap replications may be sufficient for standard error and bias estimation, a larger number, say 350, are needed to give a useful estimate of a percentile or P-value, and many more for a highly accurate assessment. Usually, however, there is no interest in estimating a P-value with high precision.

Given the occurrence of local maxima in the log likelihood $L(\phi)$ for mixture models, there is no guarantee that the largest of the local maxima will be located for any observed sample. Hence $-2 \log \lambda$ as simulated above for the test of H_0 versus H_1 may be biased downwards. In an attempt to limit this bias, on each simulation trial and each subsequent bootstrap replication, a systematic search for all local maxima was undertaken in fitting the mixture model under H_1 via the *EM* algorithm. As the two-component normal mixture model fitted under H_1 has both components belonging to the same parametric family, the log likelihood function will have two local maxima of the same value corresponding to an interchange of the component labels. This was avoided here by the restriction that the mean μ_1 of the first component be less than the mean μ_2 of the second component.

3. Bootstrap Test of Specified Size

In the narrower sense where the decision to be made concerns solely the rejection or retention of the null hypothesis at a specified significance level α, Aitkin, Anderson and Hinde (1981) noted how analogous to the Monte Carlo test procedure of Hope (1968), the bootstrap replications can be used to provide a test of approximate size α. The test which rejects H_0 if $-2 \log \lambda$ for the original data is greater than the jth smallest of its K bootstrap replications, has size

$$\alpha = 1 - j/(K+1) \tag{3.1}$$

approximately. For if the bootstrap and true null distributions of $-2 \log \lambda$ were the same (as in the case here of H_0 versus H_1), then the original and subsequent bootstrap values of

$-2 \log \lambda$ can be treated as the realizations of a random sample of size $K + 1$, and the probability that a specified member is greater than j of the others is $1 - j/(K + 1)$. Note that the result (3.1) applies to the unconditional size of the test and not to its size conditional on the K bootstrap values of $-2 \log \lambda$. For a specified significance level α, the values of j and K can be appropriately chosen according to (3.1). For example, for $\alpha = 0.05$, the smallest value of K needed is 19 with $j = 19$. As cautioned above on the estimation of the P-value for the likelihood ratio test, K needs to be very large to ensure an accurate assessment. In the present context the size of K manifests itself in the power of the test; see Hope (1968). Although the test may have essentially the prescribed size for small K, its power may be well below its limiting value as $K \to \infty$. Some simulations were performed to demonstrate the power of this test as a function of K for the test of H_0 versus H_1. The sample size was $n = 100$ and the size of the test was specified to be $\alpha = 0.05$ for each of three levels of $K(K = 19, 39, 99)$. On a given simulation trial, a sample of size $n = 100$ was generated from a mixture in equal proportions of two univariate normal distributions with the same variance σ^2 for various levels of the Mahalanobis distance

$$\Delta = |\mu_1 - \mu_2|/\sigma$$

between the two components. The entries in Table 1 for $K = \infty$ correspond to the power of the likelihood ratio test where, in the simulations, the 95th percentile of the null distribution of $-2 \log \lambda$ was approximated by $\chi^2_{2,0.95}$.

It can be seen for each of the two values of Δ taken for $n = 100$ that, as K increases from 19 through 39 to 99, the simulated power also increases, so that at $K = 99$, for $\Delta = 3$, it is the same as that of the likelihood ratio test of size 0.05. The power of the latter test is poor unless the components of the mixture are widely separated, corresponding to $\Delta = 3$ in Table 1. This was noted too by Everitt (1981) in his simulations.

4. Example

We consider an application of the likelihood ratio statistic for testing for g in the context of fitting a normal mixture model to cluster treatment means. The data set concerns seven varieties of barley grown in $r = 6$ blocks as used by Duncan (1955). The sample means in bushels per acre are listed below:

1	2	3	4	5	6	7
49.6	58.1	61.0	61.5	67.6	71.2	71.3

An analysis of variance gave $s^2 = 79.64$ with $v = 30$ degrees of freedom as an estimate of $\sigma^2(r$ times their common variance), and a P-value of 0.002 for the F test of homogeneity of all the means.

TABLE 1
Simulated power of the test of size
$\alpha = 0.05$ for $n = 100$ using K replica-
tions

	Simulated power ($\alpha = 0.05$)	
K	$\Delta = 2$	$\Delta = 3$
19	0.13	0.82
39	0.15	0.86
99	0.17	0.90
∞	0.21	0.90

Previous results on the grouping of these means tend to agree on the partitions into g groups (1 to 4 and 5 to 7 for $g = 2$ and 1, 2 to 4, and 5 to 7 for $g = 3$), but there is no clear cut choice as to $g = 2$ or 3. With the hierarchical splitting method of Scott and Knott (1974), the split from $g = 2$ to $g = 3$ groups was found to be on the borderline of significance. Binder (1978, 1981) with his Bayesian approach did not identify an obvious solution, while Cox and Spjøtvoll (1982) found there were five possible partitions into $g = 2$ groups with a probability greater than 0.05 of being consistent with the data. They also listed some partitions for $g = 3$ "partly because none of the partitions into two groups gives an excellent fit". The hierarchical agglomerative method of Caliński and Corsten (1985) led to $g = 2$ or 3, depending on the criterion used to terminate it. Their other proposed method of clustering pointed to $g = 2$.

In the light of this uncertainty over $g = 2$ or 3 as the smallest value of g compatible with the data, it is of interest to consider the choice between these two values of g on the basis of the log likelihood ratio statistic for the number of components in a normal mixture model fitted to the data.

A mixture of $g = 2$ normal densities

$$f(x; \boldsymbol{\phi}) = \pi_1 h(x; \mu_1, \sigma^2/r) + \pi_2 h(x; \mu_2, \sigma^2/r) \tag{4.1}$$

was fitted to the seven sample means $\bar{x}_1$ to $\bar{x}_7$, where $h(x; \mu_i, \sigma^2/r)$ denotes the normal density with mean μ_i and variance σ^2/r. The maximum likelihood estimate of

$$\boldsymbol{\phi} = (\pi_1, \pi_2, \mu_1, \mu_2, \sigma^2)'$$

was computed on the basis of $\bar{x}_1$ to $\bar{x}_7$ and also s^2, and is

$$\hat{\boldsymbol{\phi}} = (0.54, 0.46, 57.40, 69.35, 83.77)'.$$

The estimated standard errors of $\hat{\pi}_1$, $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\sigma}^2$ are equal to 0.24, 1.61, 4.72 and 18.87, respectively. Their estimated correlation matrix is

$$\begin{bmatrix} 1 & & & \\ 0.20 & 1 & & \\ 0.44 & 0.31 & 1 & \\ -0.28 & 0.09 & 0.07 & 1 \end{bmatrix}$$

The likelihood equation for a randomized complete block design with fixed block effects in Basford and McLachlan (1985a) can be easily manipulated to handle this situation where the individual observations are not available and the data are presented in the above summarized form; see Aitkin (1980). A FORTRAN listing of a program for the fitting of normal mixture models is given in McLachlan and Bashford (1987).

A mixture of $g = 3$ normal densities was fitted also, and twice the increase in the log likelihood for $g = 3$ over $g = 2$ was found to be 2.40. The P-value of 0.31 computed according to the χ_2^2 approximation to the null distribution of $-2 \log \lambda$ suggests $g = 2$. In view of the questionable reliablity of P-values so approximated for small sample sizes (although the data are replicated here), the null hypothesis of $g = 2$ versus the alternative of $g = 3$ is considered further using $K = 19$ replications of $-2 \log \lambda$ to construct a test with a significance level of 0.05 in mind. Each bootstrap sample consists of observations on $\bar{x}_1^*$ to $\bar{x}_7^*$ and s^{*2}, Before being ordered, $\bar{x}_1^*$ to $\bar{x}_7^*$ are distributed independently according to the null density (4.1) for the treatment means but with $\hat{\boldsymbol{\phi}}$ in place of $\boldsymbol{\phi}$, each independent of s^{*2} distributed as

$$vs^{*2}/\hat{\sigma}^2 \sim \chi_v^2.$$

As the value of 2.40 for $-2 \log \lambda$ formed from the original data is less than two of the bootstrap values of 2.53 and 2.81 obtained for this statistic, the null hypothesis of $g = 2$ groups is not rejected at the 0.1 level. Regarding the actual grouping of the means for $g = 2$ and also $g = 3$, the estimated posterior probabilities of group membership which are not listed here, clearly suggest the same partitions as obtained with the aforementioned methods of partitioning.

References

Aitkin, M. (1980) Mixture applications of the *EM* algorithm in *GLIM*. *In Compstat 1980, Proc. Computational Statistics*, pp. 537–541. Vienna: Physica-Verlag.

Aitkin, M., Anderson, D. and Hinde, J. (1981) Statistical modelling of data on teaching styles (with discussion). *J. R. Statist. Soc.* A, **144**, 419–461.

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

Binder, D. A. (1978) Bayesian cluster analysis. *Biometrika*, **65**, 31–38.

Binder, D. A. (1981) Approximations to Bayesian clustering rules. *Biometrika*, **68**, 275–285.

Basford, K. E. and McLachlan, G. J. (1985a) Cluster analysis in a randomized complete block design. *Commun. Statist.-Theor. Meth.*, **14**, 451–464.

Basford, K. E. and McLachlan, G. J. (1985b) Likelihood estimation with normal mixture models. *Appl. Statist.*, **34**, 282–289.

Caliński, T. and Corsten, L. C. A. (1985) Clustering means in ANOVA by simultaneous testing. *Biometrics*, **41**, 39–48.

Cox, D. R. and Spjøtvoll, E. (1982) On partitioning means into groups. *Scan. J. Statist.*, **9**, 147–152.

Day, N. E. (1969) Estimating the components of a mixture of two normal distributions. *Biometrika*, **56**, 463–474.

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

Do, K. and McLachlan, G. J. (1984) Estimation of mixing proportions: a case study. *Appl. Statist.*, **33**, 134–140.

Duncan, D. B. (1955) Multiple range and multiple *F* tests. *Biometrics*, **11**, 1–42.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: SIAM.

Efron, B. and Tibshirani, R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion). *Statist. Science*, **1**, 54–77.

Everitt, B. S. (1981) A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivar. Behav. Res.*, **16**, 171–180.

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

Hope, A. C. A. (1968) A simplified Monte Carlo significance test procedure. *J. R. Statist. Soc.* B, **30**, 582–598.

McLachlan, G. J. (1982) The classification and mixture maximum likelihood approaches to cluster analysis. In *Handbook of Statistics* Vol. 2 (P. R. Krishnaiah and L. N. Kanal, eds) pp. 199–208. Amsterdam: North-Holland.

McLachlan, G. J. and Basford, K. E. (1987) *Mixture Models: Inference and Applications to Clustering.* To be published by Marcel Dekker, New York.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the *EM* algorithm. *SIAM Review*, **26**, 195–239.

Scott, A. J. and Knott, M. (1974) A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, **30**, 507–512.

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

Wolfe, J. H. (1971) A monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. *Technical Bulletin STB 72-2.* San Diego: U.S. Naval Personnel and Training Research Laboratory.