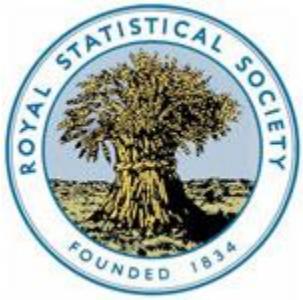



Estimation and Hypothesis Testing in Finite Mixture Models
Author(s): Murray Aitkin and Donald B. Rubin
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 47, No. 1 (1985), pp. 67-75
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2345545
Accessed: 12-02-2016 02:07 UTC



*J. R. Statist. Soc.* B (1985)
**47**, *No.* 1, *pp.* 67–75

# Estimation and Hypothesis Testing in Finite Mixture Models

By MURRAY AITKIN† and DONALD B. RUBIN‡

*University of Lancaster, UK* *University of Chicago, USA*

[Received September 1983. Revision January 1984]

SUMMARY

Finite mixture models are a useful class of models for application to data. When sample sizes are not large and the number of underlying densities is in question, likelihood ratio tests based on joint maximum likelihood estimation of the mixing parameter, $\lambda$, and the parameter of the underlying densities, $\theta$, are problematical. Our approach places a prior distribution on $\lambda$ and estimates $\theta$ by maximizing the likelihood of the data given $\theta$ with $\lambda$ integrated out. Advantages of this approach, computational issues using the *EM* algorithm and directions for further work are discussed. The technique is applied to two examples.

*Keywords*: MIXTURE MODEL; *EM* ALGORITHM; PRIOR DISTRIBUTION; LIKELIHOOD RATIO TEST

## 1. INTRODUCTION

Finite mixture models form a large class of interesting statistical models which can be used to address such practical problems as inference in the presence of outliers and clustering units by latent traits. The common idea underlying these models is that the data $X = (X_1, \ldots, X_n)^T$, where $X_i$ is generally a vector and $n$ is the sample size, arise from two or more underlying groups with common distributional form but different parameters. Our objectives are (1) to estimate the vector parameter of the density of $X$ in the $g$ groups, $\theta = (\theta_1, \ldots, \theta_g)$ where $\theta_j$ is the vector parameter of the distribution of $X$ in the $j$th group, and (2) to summarize evidence for the number of different groups (i.e. the number of different $\theta_j$) needed to fit the data. The relative sizes of the $g$ groups are specified by the mixing parameter $\lambda = (\lambda_1, \ldots, \lambda_g)$ where $\Sigma_1^g \lambda_j = 1$ and $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_g$.

The *EM* algorithm (Dempster, Laird and Rubin, 1977), is a basic tool for the estimation of parameters by maximum likelihood (m.l.) in mixture models. However, there can be some problems with the use of results based on the maximum likelihood estimate of $(\theta, \lambda)$, especially when the number of groups is in question and the sample sizes are not large. Since summarizing evidence about the existence of outliers and studying the number of latent classes are frequent practical questions, potential problems with the m.l. estimate of $(\theta, \lambda)$ are of practical concern.

Our approach is to place a prior distribution on the mixing parameter and to estimate the distributional parameter $\theta$ by m.l. from the likelihood $L(\theta \mid X)$ which is found by integrating the likelihood of $L(\theta, \lambda \mid X)$ over the prior distribution of $\lambda$. Tests for the number of groups are made relative to the one-group model; relative likelihoods and standard $\chi^2$ $p$-values can be used to evaluate the need for extra groups. The advantages of our approach over the standard approach treating $\lambda$ as a parameter to be jointly estimated with $\theta$ by m.l. (i.e. by maximizing the joint likelihood of $\theta, \lambda$) are:

(a) sensible (realistic) prior information about the likely values of $\lambda$ can be incorporated into the estimation (e.g. the prior distribution of $\lambda$ when looking for outliers should differ from the prior distribution of $\lambda$ when looking for latent groups);

† *Present address*: Centre for Applied Statistics, The University, Bailrigg, Lancaster LA1 4YL, UK.

‡ *Present address*: Department of Statistics, Harvard University

 0035–9246/85/47067 $2.00

(b) the resulting likelihood function for $\theta$ is better-behaved, and consequently relative likelihoods can be more easily interpreted than when $\lambda$ and $\theta$ are jointly estimated by m.l.

Our approach does not fully solve the problem of testing for the number of groups since the likelihood ratio statistics testing for $g$ vs. $g-1$ groups do not generally follow standard $\chi^2$ distributions. An analogous problem occurs within the multiple comparison context: with $g$ means, the likelihood ratio test of $g$ different means vs. one common mean is simple, whereas the test of $g$ different means vs. $g-1$ different means is not so simple in general, and in fact is rarely made in practice.

Section 2 fixes notation and presents the *EM* algorithm for finding the m.l.e. of $\theta, \lambda$. Section 3 describes the modified *EM* algorithm for finding the m.l.e. of $\theta$ having integrated over $\lambda$. Section 4 describes and compares the likelihood ratio tests arising from the approaches of Sections 2 and 3. Section 5 investigates outlier detection in Darwin's data. Section 6 examines a latent class model applied to data on teaching styles.

Other recent work on mixture models include the book by Everitt and Hand (1981), Ganesalingam and McLachlan (1980), Aitkin (1980), Shaked (1980), Lachenbruch and Broffitt (1980), Walker (1980), Olsson (1979), Quandt and Ramsey (1979), Chang (1979), Fowlkes (1979) and the International Encyclopaedia of Statistics article by Blischke (1977).

## 2. THE MIXTURE MODEL AND M.L. ESTIMATION OF $(\theta, \lambda)$ USING *EM*

In addition to the notation introduced in Section 1, let $Z_i$ indicate group membership for the $i$th unit, $Z_i = (Z_{i1}, \ldots, Z_{ig})$, where $Z_{ij} = 1$ indicates $X_i$ arose from the $j$th group, $\Sigma Z_{ij} = 1$, $j = 1, \ldots, g$. The basic mixture model specifies that for $i = 1, \ldots, n$:

$(X_i \mid Z_i, \theta, \lambda)$ are independent with densities

$$\sum_{j=1}^{g} Z_{ij} f(X_i \mid \theta_j) = \prod_{j=1}^{g} f(X_i \mid \theta_j)^{Z_{ij}} \tag{1}$$

and

$(Z_i \mid \theta, \lambda)$ are i.i.d. with distribution

$$\prod_{j=1}^{g} \lambda_j^{Z_{ij}}, \quad \sum_{1}^{g} \lambda_j = 1, \ \lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_g \geqslant 0. \tag{2}$$

The ordering of $\lambda_j$ is simply in order to identify the $g$ groups: group 1 is largest, ..., group $g$ is smallest. From specifications (1) and (2), we have that

$$(X_i \mid \theta, \lambda) \overset{\text{i.i.d.}}{\sim} \sum_{j=1}^{g} \lambda_j f(X_i \mid \theta_j), \quad \Sigma \lambda_j = 1, \quad \lambda_1 \geqslant \ldots \geqslant \lambda_g \geqslant 0. \tag{3}$$

Consequently, if we consider $\theta$ and $\lambda$ to be parameters to be estimated by m.l. from observed data $X = (X_1, \ldots, X_n)^{\mathrm{T}}$, we are led to the likelihood function

$$L(\theta, \lambda \mid X) = \prod_{j=1}^{n} \left[ \sum_{j=1}^{g} \lambda_j f(X_i \mid \theta_j) \right], \ \Sigma \lambda_j = 1, \ \lambda_1 \geqslant \ldots \geqslant \lambda_g \geqslant 0. \tag{4}$$

This function can be maximized quite easily using *EM* by treating $Z = (Z_1, \ldots, Z_n)$ as missing data. From (1) and (2), if $Z$ were observed, the log likelihood could be written

$$\log L(\theta, \lambda \mid X, Z) = \sum_{i=1}^{n} \sum_{j=1}^{g} Z_{ij} \log f(X_i \mid \theta_j) + \sum_{i=1}^{n} \sum_{j=1}^{g} Z_{ij} \log \lambda_j, \tag{5}$$

$$\Sigma \lambda_i = 1, \ \lambda_1 \geqslant \ldots \geqslant \lambda_g \geqslant 0.$$

The *E*-step of the *EM* algorithm requires us to calculate the expected value of the log likelihood over the conditional distribution of the missing data, $Z$, given (a) the observed data, $X$, and (b) current estimates of parameters $\theta$ and $\lambda$.

The expected value of the log likelihood (5) given the observed data, $X$, and parameter values $\theta = \theta^{(o)}$ and $\lambda = \lambda^{(o)}$ is

$$\sum_{i=1}^{n} \sum_{j=1}^{g} \log f(X_i \mid \theta_j) \, E\,(Z_{ij} \mid X, \theta^{(o)}, \lambda^{(o)}) \tag{6}$$

$$+ \sum_{j=1}^{n} \sum_{j=1}^{g} \log \lambda_j \; E\,(Z_{ij} \mid X, \theta^{(o)}, \lambda^{(o)})$$

$$\Sigma \lambda_j = 1, \ \lambda_1 \geqslant \ldots \geqslant \lambda_g \geqslant 0.$$

From (1) and (2), and Bayes' Theorem we have that

$$(Z_i \mid X_i, \theta, \lambda) \propto \prod_{j=1}^{g} [\lambda_j f(X_i \mid \theta_j)]^{Z_{ij}} \text{ and are i.i.d.} \tag{7}$$

Thus

$$E(Z_{ij} \mid X, \theta^{(o)}, \lambda^{(o)}) = \lambda_j^{(o)} f(X_i \mid \theta_j^{(o)}) / \sum_{j=1}^{g} \lambda_j^{(o)} f(X_i \mid \theta_j^{(o)}), \tag{8}$$

and consequently, the *E*-step of the *EM* algorithm is computationally simple.

Substitute the right hand side of equation (8), say $Z_{ij}^{(o)}$, into expression (6) to obtain expression (9):

$$\sum_{i=1}^{n} \sum_{j=1}^{g} \log f(X_{ij} \mid \theta_j) Z_{ij}^{(o)} + \sum_{i=1}^{n} \sum_{j=1}^{n} \log \lambda_j \, Z_{ij}^{(o)} \ \Sigma \lambda_j = 1, \lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_g \geqslant 0. \tag{9}$$

Having completed the *E*-step, we need in the *M*-step of the *EM* algorithm to maximize the expected log likelihood, expression (9), over $\theta$ and $\lambda$, in order to find the next values of $\theta$ and $\lambda$, say $\theta^{(1)}$ and $\lambda^{(1)}$. Ignore for the moment the restriction that the $\lambda_j$ are ordered. The maximization over $\theta$ is accomplished by maximizing the first term in (9), and thus is equivalent to finding the complete data maximum likelihood estimate of $\theta$ with the modification that the $Z_{ij}$, instead of being 0 or 1, are between 0 and 1. The maximization over $\lambda$ is accomplished by maximizing the second term in (9), and immediately gives

$$\lambda_j^{(1)} = \sum_{i=1}^{n} Z_{ij}^{(o)} / n, \quad j = 1, \ldots, g. \tag{10}$$

Repeated *EM* steps maximize the likelihood over $\theta$, $\lambda$ ignoring the order restriction on the $\lambda$. Assuming convergence to an m.l.e. $(\hat{\theta}, \hat{\lambda})$, we simply permute the group index $j$ on $\theta$ and $\lambda$ so that $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \ldots \hat{\lambda}_g$. This permutation will give the m.l.e. of $\theta, \lambda$ satisfying the ordering restriction. Note that the likelihood of $\theta, \lambda$ without the ordering restriction has $g!$ modes corresponding to the $g!$ orderings of the $\lambda_j$.

The precise conditions under which *EM* converges to an m.l.e. need more study. Some experience suggests that in common examples *EM* will converge. Models exist, however, with unbounded likelihoods at the boundary of the parameter space, and in these cases *EM* may not converge (e.g. $f(X_i \mid \theta_j)$ is the normal density with mean $\mu_j$ and variance $\sigma_j^2$). Recent work on this problem includes Wu (1983), which considers convergence properties of *EM* rather generally, Redner and Walker (1984), which concerns *EM* in mixture models, and Meester (1983), which presents an extensive Monte Carlo study of m.l. estimation in mixture models. Here, we restrict attention to cases in which *EM* has converged to an m.l.e.

## 3. A MODIFIED APPROACH INTEGRATING OVER THE MIXING PARAMETERS

Although it is computationally straightforward to find the m.l.e. of $(\theta, \lambda)$, there remain some serious statistical problems. The source of these problems is that unless the m.l.e. of $\lambda_g$, $\hat{\lambda}_g$, is well away from the boundary value 0 (in the sense that $L(\hat{\theta}, \hat{\lambda}_1, \ldots, \hat{\lambda}_{g-1}, \hat{\lambda}_g)$ is much larger than $L(\theta, \lambda_1, \ldots, \lambda_{g-1}, 0)$ for all $(\theta, \lambda_1, \ldots, \lambda_{g-1})$, the joint likelihood of $(\theta, \lambda)$ cannot be approximately normal. Since instances when it is desirable to test for the number of groups usually arise when $\hat{\lambda}_g$ is *not* well away from zero, the usual likelihood ratio tests relying on asymptotic normality for probabilistic interpretations may be deceptive.

More precisely, consider two ways of specifying the null hypothesis. If the null hypothesis that there are only $g-1$ groups is specified by $\lambda_g = 0$, then this null value is on the boundary of the parameter space, and under the null hypothesis, the m.l.e. $\hat{\lambda}_g$ does not have an asymptotic normal distribution about 0 since $\lambda_g \geqslant 0$; furthermore, at $\lambda_g = 0$, the likelihood is flat in $\theta_g$. If the null hypothesis that there are only $g-1$ groups is specified by $\theta_{g-1} = \theta_g$, then under this null hypothesis, the likelihood is the same for an infinity of values with the same value of $(\lambda_{g-1} + \lambda_g)$, and again is not asymptotically full-rank normal. Consequently, under the null hypothesis of fewer than $g$ groups, the likelihood will not be asymptotically full rank normal, and hence, the standard likelihood ratio tests (l.r.t's) will not generally follow standard $\chi^2$ distributions, or in direct likelihood terms, may not give an appropriate indication of how far the null hypothesis is from the region of high likelihood. For example, the maximum value of the likelihood may drop precipitously as $\lambda_g \to 0$, but the drop will be accompanied by a violent flattening of the likelihood in $\theta$: this flattening implies that comparing the maxima of the likelihood at $\lambda_g > 0$ and $\lambda_g = 0$ is not a good summary of the evidence for $\lambda_g > 0$ vs. $\lambda_g = 0$. The behaviour of the likelihood with real data is shown in Section 5.

Failings of the standard l.r.t's have been noted in practice by Wolfe (1970) and Everitt and Hand (1981). Wolfe (1971) performed a Monte Carlo study and conjectured another distribution for the standard l.r.t's which we do not believe holds in general. A more familiar example of testing on the boundary of the parameter space is the one-way variance components model: in this case the asymptotic distribution of the l.r.t. for the between variance equalling zero is not $\chi_1^2$ but more nearly $\chi_{g-1}^2$ where $g$ is the number of groups. Kendall and Stuart (1966, Chapter 36) provide details.

Our proposal for avoiding this problem with standard l.r.t's is to place a prior distribution on $\lambda$, integrate over $\lambda$, and find the value of $\theta$ that maximizes the likelihood $L(\theta \mid X)$. The advantages of our approach are first, that reasonable prior information about the expected sizes of the underlying groups can be incorporated, and second, that any null hypothesis about $\theta$ is specified in the interior of the parameter space with an expected asymptotically full-rank normal likelihood. Consequently, not only do we expect slightly improved estimates since some prior information is being incorporated, but also we anticipate that the inferences resulting from standard asymptotic $\chi^2$ distributions will be more reasonable.

The likelihood of $\theta$, $L(\theta \mid X)$, given observed data $X$ with prior distribution $p(\lambda)$ on $\lambda$ is:

$$L(\theta \mid X) = \int \prod_{i=1}^{n} \left[ \sum_{j=1}^{g} \lambda_j f(X_i \mid \theta_j) \right] p(\lambda)\, d\lambda. \tag{11}$$

The likelihood can be maximized by *EM* by treating the $Z_i$ *and* $\lambda$ as missing data. From (1), the likelihood of $\theta$ given $X, Z$ and $\lambda$ is proportional to

$$\prod_{j=1}^{g} f(X_i \mid \theta_j)^{Z_{ij}}. \tag{12}$$

*EM* maximizes $L(\theta \mid X)$ by iteratively maximizing the expected value of the logarithm of (12). This expected value over the distribution of $(\lambda, Z)$ given $X$ and $\theta = \theta^{(o)}$ is

$$\sum_{i=1}^{n} \sum_{j=1}^{g} \log f(X_i \mid \theta_j)\, E(Z_{ij} \mid X_i, \theta^{(o)}), \tag{13}$$

where

$$E(Z_{ij} \mid X, \theta^{(o)}) = \int \left[ \lambda_j f(X_i \mid \theta_j) / \sum_{j=1}^{g} \lambda_j f(X_i \mid \theta_j) \right] p(\lambda \mid X, \theta^{(o)})\, d\lambda \tag{14}$$

and $p(\lambda \mid X, \theta^{(o)})$ is the posterior density of $\lambda$ given $X$ and $\theta = \theta^{(o)}$:

$$p(\lambda \mid X, \theta^{(o)}) \propto p(\lambda) \prod_{i=1}^{n} \left[ \sum_{j=1}^{g} \lambda_j f(X_i \mid \theta_j^{(o)}) \right]. \tag{15}$$

In general, the evaluation of (14) requires a $g-1$ dimensional numerical integration. Once we have evaluated (14), the m.l.e. of $\theta$ is found as before in the *M*-step, i.e. using weighted sufficient statistics.

The need to perform a numerical integration at the *E*-step is a clear disadvantage of our approach. However, the value of $\theta$ that maximizes the joint likelihood of $\theta, \lambda$, $L(\theta, \lambda \mid X)$ is often quite close to the value of $\theta$ that maximizes the likelihood of $\theta$, $L(\theta \mid X) = \int L(\theta, \lambda \mid X)\, p(\lambda)\, d\lambda$. Although there are simple examples in statistics in which the marginal mode is quite different from the integrated mode, these examples typically involve integrating over a high-dimensional parameter to obtain the marginal mode of a low-dimensional parameter, such as with linear regression when integrating over the regression coefficient to obtain the modal estimate of the residual variance in the regression. The integration called for in our approach involves integrating over the $(p-1)$-dimensional $\lambda$ to obtain the marginal mode of $\theta$, which is at least $p$-dimensional and with multivariate $X_i$ substantially higher-dimensional. Consequently, we suggest maximizing $L(\theta \mid X)$ by (a) first maximizing $L(\hat{\theta}, \hat{\lambda} \mid X)$ using the simple *EM* algorithm described in Section 2 to find $\theta$, $\lambda$, and (b) second, starting at $\hat{\theta}$, maximizing $L(\theta \mid X)$ using the more complicated *EM* algorithm presented in this section. Examples suggest that only one, or possibly no, complicated iteration is needed. Thus our suggestion is to perform *EM* as in Section 2 to find $\hat{\theta}$, and perform the numerical integration called for in (11) to find $L(\hat{\theta} \mid X)$ for testing purposes.

## 4. TESTING FOR THE NUMBER OF GROUPS

Let $\tilde{\theta}$ be the value of $\theta$ that maximizes $L(\theta \mid X)$, i.e. the maximum likelihood estimate of $\theta$ under the specification (11) where $\lambda \sim p(\lambda)$. Suppose that we wish to test the null model that there is only one underlying distribution of $X$ rather than $g$, i.e. $\theta_1 = \theta_2 = \ldots = \theta_g$. It does not follow immediately that the l.r.t.

$$-2 \log \left[ L(\tilde{\theta} \mid X) / \max_{\theta_1} \left\{ \prod_{i=1}^{n} f(X_i \mid \theta_1) \right\} \right]$$

has a sampling distribution which is asymptotically $\chi^2_{\dim(\theta)-\dim(\theta_1)}$, since the introduction of the prior distribution for $\lambda$ produces a correlation between the observations $X_i$. However we show in the Appendix that this correlation is usually small in the two-group case, and that therefore if the sample size is not small, the usual asymptotic distribution can be expected to apply. See also Table 1.

However, we do not perform direct l.r.t.'s for intermediate numbers of groups, for example, for $g = 3$ vs. $g = 2$, for the following reason. In general, the two group model is not nested in the three-group model, since no prior information is incorporated in the models indicating that (say) the second group in the two-group model is itself a mixture of two sub-groups. Even when such prior information exists, it is not simple computationally to retain the identity of group 1 in the two-group solution as the first group in the three-group solution. Furthermore, in the absence of such prior information, the two-group prior will not be the marginal distribution from the three-group prior.

The essential differences between the l.r.t.'s with a prior distribution on $\lambda$ and the l.r.t.'s without such a distribution are easily seen in the simple case with $g = 2$ where we wish to test $\theta_1 = \theta_2$. Under the null hypothesis, the prior distribution on $\lambda$ is irrelevant, in the sense that the value of $\theta_1$ that maximizes the likelihood is the same, and the value of the maximized likelihood is the same, say $L_0$, whether or not $\lambda$ has a prior distribution. However when $\theta_2$ is not forced to equal $\theta_1$, the maximized values of the likelihoods with and withour prior distribution on $\lambda$ differ; in fact, the maximized likelihood without the prior, i.e. $L(\hat{\theta}_1, \hat{\theta}_2, \hat{\lambda} \mid X)$ is always larger than the maximized likelihood with a uniform prior on $\lambda$, i.e. $L(\tilde{\theta}_1, \tilde{\theta}_2 \mid X)$; that is

$$\sup_{\theta,\,\lambda} L(\theta, \lambda \mid X) \geqslant \sup_{\theta} \int L(\theta, \lambda \mid X)\, p(\lambda)\, d\lambda.$$

For notational simplicity, we define $\lambda$ in the context of this section so that $\lambda = \lambda_1$ and $1 - \lambda = \lambda_2$. For the maximum likelihood estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\lambda})$, the maximized likelihood equals

$$L(\hat{\theta}_1, \hat{\theta}_2, \hat{\lambda} \mid X) = \prod_{i=1}^{n} [\hat{\lambda} f(X_i \mid \hat{\theta}_1) + (1 - \hat{\lambda}) f(X_i \mid \hat{\theta}_2)] = \sum_{k=0}^{n} M(k, \hat{\theta}_1, \hat{\theta}_2)\, \hat{\lambda}^{(n-k)} (1 - \hat{\lambda})^k$$

where

$$M(k, \hat{\theta}_1, \hat{\theta}_2) = \sum_{S \in S_k} \Big[ \prod_{i \in S} f(X_i \mid \hat{\theta}_1) \prod_{i \in \bar{S}} f(X_i \mid \hat{\theta}_2) \Big]$$

and $S_k$ = set of all subsets of exactly $k$ of the integers from 1 to $n$. $M(k, \theta_1, \theta_2)$ is the relative likelihood of $k$ observations from group 1 and $n - k$ from group 2, for all partitions of the $n$ observations. In contrast, for the maximum likelihood estimates $\tilde{\theta}_1$ and $\tilde{\theta}_2$ corresponding to a prior $p(\lambda)$ on $\lambda$, the maximized likelihood of $\theta_1, \theta_2$ equals

$$L(\tilde{\theta}_1, \tilde{\theta}_2 \mid X) = \int \prod_{i=1}^{n} [\lambda f(X_i \mid \tilde{\theta}_1) + (1 - \lambda) f(X_i \mid \tilde{\theta}_2)]\, p(\lambda)\, d\lambda$$

$$= \sum_{k=0}^{n} M(k, \tilde{\theta}_1, \tilde{\theta}_2) \int \lambda^{n-k} (1 - \lambda)^k\, p(\lambda)\, d\lambda$$

If $\tilde{\theta}_1 = \hat{\theta}_1$ and $\tilde{\theta}_2 = \hat{\theta}_2$ (which will usually nearly hold), the difference between the values of the maximized likelihoods is due to the multipliers of the terms $M(k, \hat{\theta}_1, \hat{\theta}_2)$. With a prior distribution $\lambda$, these multipliers are constant and do not depend on the data, whereas when $\lambda$ has no prior distribution, the $(n + 1)$ multipliers are estimated from the data and are chosen to maximize the joint likelihood.

We consider now two practical examples.

## 5. OUTLIERS IN DARWIN'S DATA

The existence of outliers in Darwin's data was examined by Aitkin and Tunnicliffe Wilson (1980) using a normal mixture model. A normal probability plot of the 15 observations indicates that the two smallest observations are discrepant, and a two-component (common variance) normal mixture model isolates these two observations in one component with (posterior) probability 1.000, all the remaining observations belonging to the second component with probabilities at least 0.999. The reduction in $-2\log L$, compared with the single normal model, is 6.9. Figure 1 shows a contour plot in the space of $\delta\ (=\mu_2-\mu_1)$ and $\lambda$ of the profile likelihood, i.e. the likelihood maximized over the other parameters. The parameter space is restricted to $\delta \geqslant 0$.

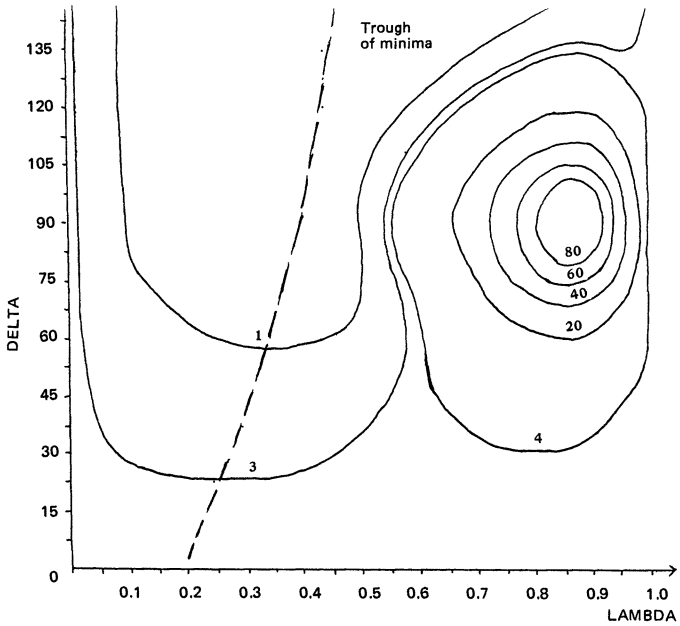


Fig. 1. Profile likelihood.

The likelihood is flat around the boundaries $\lambda = 0$, $\lambda = 1, \delta = 0$, and is noticeably skewed in $\lambda$ (restricting to $\delta \geqslant 0$ with $\lambda$ unrestricted is equivalent to restricting to $\lambda \geqslant \frac{1}{2}$ with $\delta$ unrestricted). A "trough" of minima of the likelihood over $\lambda$ for fixed $\delta$ is also visible, and is shown by the dotted line in Fig. 1. Starting values for $(\delta, \lambda)$ on the "wrong" side of this trough lead to very slow convergence of the *EM* algorithm to a point on the boundary of the parameter space (with a likelihood equal to that for the single normal model), instead of to the maximum.

The prior distribution of $\lambda$ on $(\frac{1}{2}, 1)$ was taken as a Beta distribution, with

$$\beta(\lambda) \propto (\lambda - \tfrac{1}{2})^p\ (1-\lambda)^q.$$

The maximized value of the likelihood $L(\theta \mid X)$ was found as described in Section 3 for a range of values of $(p, q)$, starting in each case from the m.l.e. $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma})$ found by ordinary *EM*. Adequate convergence occurred in one iteration in each case, the parameter estimates hardly changing. The difference between $\sup_\theta \log L(\theta \mid X)$ and the log likelihood for the one-sample normal model is given in Table 1 for various $(p, q)$. The estimated correlations between pairs of observations, calculated from the Appendix at $(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\sigma})$ are also shown.

TABLE 1

*LRTs and correlations between observations for $B(p, q)$ prior on $\lambda$*

| | | $p$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | -0.5 | 0 | 0.5 | 1 | 1.5 | 2 |
| $q$ | -0.5 | 4.8<br>0.133 | 4.5<br>0.119 | 4.1<br>0.100 | 3.6<br>0.083 | 3.1<br>0.070 | 2.7<br>0.060 |
| | 0 | 5.2<br>0.082 | 5.2<br>0.089 | 4.9<br>0.083 | 4.5<br>0.075 | 4.1<br>0.067 | 3.7<br>0.060 |
| | 0.5 | 5.3<br>0.055 | 5.5<br>0.067 | 5.4<br>0.067 | 5.1<br>0.063 | 4.8<br>0.059 | 4.5<br>0.054 |
| | 1 | 5.3<br>0.040 | 5.7<br>0.052 | 5.7<br>0.054 | 5.5<br>0.053 | 5.3<br>0.051 | 5.0<br>0.048 |
| | 1.5 | 5.3<br>0.030 | 5.8<br>0.041 | 5.9<br>0.045 | 5.8<br>0.045 | 5.6<br>0.044 | 5.4<br>0.043 |
| | 2 | 5.2<br>0.023 | 5.8<br>0.033 | 6.0<br>0.037 | 6.0<br>0.039 | 5.8<br>0.039 | 5.6<br>0.038 |

$\chi^2_{1,0.01} = 6.64$ $\chi^2_{1,0.05} = 3.84$ $\chi^2_{1,0.1} = 2.71$

LRT above, correlation below

Assuming that $\chi^2_1$ provides an adequate asymptotic distribution for the l.r.t, as seems reasonable from Fig. 1 (the relative likelihood for $\delta$ is close to normal for any fixed $\lambda$), the conclusion about the existence of outliers—i.e. of a two-component mixture—depends on the prior information. If *a priori* outliers are very unlikely ($p$ large, $q$ small), the evidence in the data is discounted, and the l.r.t. is not "significant" at a conventional 5 per cent level. If *a priori* outliers are likely, then the data point clearly to their existence. Of course we are not proposing that the prior distribution should be chosen after its effect on the likelihood $L(\theta \mid X)$ is observed. Our point is that the conclusion about outliers depends *strongly* on the prior distribution because of the small sample size.

## 6. LATENT CLASSES IN DATA ON TEACHING STYLES

Latent class models were fitted to data on 468 teachers measured on 38 binary variables describing teacher behaviour. A detailed discussion is given in Aitkin, Anderson and Hinde (1981). The two-latent-class model, when compared with a model of complete independence, gave an l.r.t. of 775.8, the proportion of the teacher population in the first ("formal style") class being estimated as 0.54.

Because of the extensive computations required, we examined only four prior distributions. Three Beta priors gave l.r.t.'s of 770.8 ($p = q = \frac{1}{2}$), 771.6 ($p = q = 0$) and 773.6 ($p = -\frac{1}{2}$, $q = \frac{1}{2}$). An extremely concentrated prior ($p(\lambda) = 1 - e^{-50}$ at $\lambda = 0.5125$ and $= e^{-50}$ elsewhere) gave an l.r.t. of 670.2, but corresponds to impossibly precise prior information. As expected, the data overwhelmed any realistic prior assumption. The existence of two components is beyond question, given the model.

## APPENDIX

$$(X_i \mid \lambda) \sim \lambda f_1 + (1-\lambda) f_2 \qquad \lambda \sim p(\lambda)$$

$$E(X_i \mid \lambda) = \lambda\mu_1 + (1-\lambda)\mu_2 = \lambda(\mu_1 - \mu_2) + \mu_2$$

$$\begin{aligned} V(X_i \mid \lambda) &= \lambda(\mu_1^2 + \sigma_1^2) + (1-\lambda)(\mu_2^2 + \sigma_2^2) - \lambda^2(\mu_1 - \mu_2)^2 - \mu_2^2 - 2\lambda\mu_2(\mu_1 - \mu_2) \\ &= -\lambda^2(\mu_1 - \mu_2)^2 + \lambda(\mu_1^2 - \mu_2^2 + \sigma_1^2 - \sigma_2^2 - 2\mu_1\mu_2 + 2\mu_2^2) + \sigma_2^2 \\ &= \lambda(1-\lambda)(\mu_1 - \mu_2)^2 + \lambda\sigma_1^2 + (1-\lambda)\sigma_2^2 \end{aligned}$$

$$\operatorname{cov}(X_i, X_i' \mid \lambda) = 0$$

$$\text{corr}\,(X_i, X_i') = \frac{\text{cov}\,[E(X_i \mid \lambda), E(X_i' \mid \lambda)]}{V\,[E(X_i \mid \lambda)] + E\,[V(X_i \mid \lambda)]}$$

$$= \frac{(\mu_1 - \mu_2)^2\ V(\lambda)}{(\mu_1 - \mu_2)^2\ V(\lambda) + (\mu_1 - \mu_2)^2\ E[\lambda\,(1-\lambda)] + \bar{\lambda}\,\sigma_1^2 + (1-\bar{\lambda})\,\sigma_2^2}$$

$$= \left[1 + \frac{E[\lambda\,(1-\lambda)]}{V(\lambda)} + \frac{\bar{\lambda}\,\sigma_1^2 + (1-\bar{\lambda})\,\sigma_2^2}{(\mu_1 - \mu_2)^2\ V(\lambda)}\right]^{-1}$$

This correlation is bounded below 1, and the first term is determined solely by the prior distribution on $\lambda$.

$E$, $V$ refer to the prior distribution of $\lambda$ (given $\theta$), and $\bar{\lambda} = E(\lambda)$.

ACKNOWLEDGEMENTS

This research was carried out with the support of SSRC (UK) grant HR6132 while D. B. Rubin was a Senior Visiting Fellow at the University of Lancaster in 1979 and 1980. We are grateful to Ian Pate and Brian Francis for help with some of the computing.

REFERENCES

Aitkin, M. (1980) Mixture applications of the *EM* algorithm in GLIM. In *COMPSTAT 1980*, pp. 537–541.

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 Tunnicliffe Wilson, G. (1980) Mixture models, outliers, and the *EM* algorithm. *Technometrics*, **22**, 325–331.

Blischke, W. R. (1977) Distributions, statistical: IV. Mixtures of distributions. *Int. Encyc. Statist*, **1**.

Chang, W.-C. (1979) Confidence interval estimation and transformation of data in a mixture of two multivariate normal distributions with any given large dimension. *Technometrics*, **21**, 351–356.

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.

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

Fowlkes, E. B. (1979) Some methods for studying the mixture of two normal (lognormal) distributions. *J. Amer. Statist. Ass.*, **74**, 561–575.

Ganesalingam, S. and McLachlan, G. J. (1980) A comparison of the mixture and classification approaches to cluster analysis. *Comm. in Statist. A*, **9**, 923–933.

Kendall, M. G. and Stuart, A. (1966) *The Advanced Theory of Statistics*, Vol. 3, New York: Hafner.

Lachenbruch, P. A. and Broffitt, B. (1980) On classifying observations when one population is a mixture of normals. *Biometric J.*, **22**, 295–301.

Meester, L. E. (1983) A simulation study of the small sample behavior of parameter estimation for mixtures of distributions. Delft University of Technology, The Netherlands.

Olsson, D. M. (1979) Estimation for mixtures of distributions by direct maximization of the likelihood function. *J. Qual. Technol.*, **11**, 153–159.

Quandt, R. E. and Ramsey, J. B. (1978) Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Ass.*, **73**, 730–752 (with Discussion).

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the *EM* algorithm. *SIAM Review*, to appear.

Shaked, M. (1980) On mixtures from exponential families. *J. R. Statist. Soc.* B, **42**, 192–198.

Walker, H. F. (1980) Estimating the proportions of two populations in a mixture using linear maps. *Commun. in Statist. A*, **9**, 837–849.

Wolfe, J. H. (1970) Pattern clustering by multivariate mixture analysis. *Multiv. Behav. Res.*, **5**, 329–350.

——— (1971) A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Naval Personnel and Training Research Laboratory. *Technical Bull., STB 72-2*, San Diego, California.

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