An introduction to finite mixture distributions

BS Everitt

Stat Methods Med Res 1996 5: 107
DOI: 10.1177/096228029600500202

Statistical Methods in Medical Research 1996; **5**: 107–127

An introduction to finite mixture distributions

BS Everitt Biostatistics and Computing Department, Institute of Psychiatry, London, UK

Finite mixture densities can be used to model data from populations known or suspected to contain a number of separate subpopulations. Most commonly used are mixture densities with Gaussian (univariate or multivariate) components, but mixtures with other types of component are also increasingly used to model, for example, survival times. This paper gives a general introduction to the topic which should help when considering the other more specialized papers in this issue.

1 Introduction

Finite mixture densities are a family of probability density functions of the form

$$f(\mathbf{x};\, \mathbf{p},\boldsymbol{\theta}) = \sum_{i=1}^{c} p_i g_i(\mathbf{x};\boldsymbol{\theta}_i) \tag{1.1}$$

where $\mathbf{x}$ is a d-dimensional random variable, $\mathbf{p}' = [p_1, p_2, \ldots, p_{c-1}]$ and $\boldsymbol{\theta}' = [\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \ldots, \boldsymbol{\theta}_c']$, with the p_i being the *mixing proportions* and the $g_i, i = 1,2 \ldots, c$ the *component densities*, with density g_i parameterized by $\boldsymbol{\theta}_i$. The mixing proportions are non-negative and are such that $\sum_{i=1}^{c} p_i = 1$. The number of components forming the mixture is c.

In most applications of mixture densities, the g_i are assumed to take the same *specified* form, for example, univariate Gaussian, although in some applications it is appropriate to allow different forms as i varies. A particular case of the latter, called a *nonstandard mixture*, is that for which $c = 2$ and one of the component densities is concentrated on a single value; an example will be given in Section 5. (Dealing with mixtures where the form of the component densities is *not* explicitly specified will not be considered in this paper; a recent account of this aspect of such densities is given in Tarter and Lock.[1])

Finite mixture densities are most often used in one or other of the following contexts:

- where the population whose distribution is to be modelled is *known* to consist of well-defined subpopulations but the individual class memberships are unavailable or too expensive to obtain. A simple example would occur if clinical measurements were available for patients, but disease classifications were not.
- where subpopulations are only *suspected* and finite mixture models are used to 'explore' the data for any potentially informative groupings. Here finite mixture densities act as a relatively sound model for *cluster analysis* (see McLachlan and Basford[2] and the article by McClaren later in this issue).

Mixtures in which the component densities are Gaussian (either univariate or multivariate) are most commonly encountered in practice. McLachlan and Basford,[2] for example, fit a mixture of bivariate normal densities to haemophilia data collected from a sample of women, in a bid to identify normal women and haemophilia A

Address for correspondence: Professor Brian S Everitt, Biostatistics and Computing Department, Institute of Psychiatry, de Crespigny Park, Denmark Hill, London SE5 8AF, UK.

carriers. In addition, however, mixtures of exponential, Weibull or Gompertz densities are increasingly used to model the distribution of survival times in situations where deaths occur from more than one cause (see, for example, McGiffin *et al.*[3] and Blackstone *et al.*[4]), and mixtures of multivariate Bernoulli densities can be used to examine the structure of categorical data; Pickering and Forbes,[5] for example, use such an approach to study how to allocate neonatal resources throughout Scotland. A number of other medical applications of mixture densities will be described in detail in Section 5.

The main problem to be faced when applying finite mixture densities is the estimation of the parameters of the mixture (which may or may not include the number of components, c). Before considering estimation, however, a brief review of the historical background will be given.

2 Brief history of mixture densities

> When a series of measurements gives rise to a normal curve, we may probably assume something approaching a stable condition; there is production and destruction impartially round the mean. In the case of certain biological, sociological and economic measurements there is, however, a well-marked deviation from this normal shape, and it becomes important to determine the direction and amount of such deviation. The asymmetry may arise from the fact that the units grouped together in the measured material are not really homogeneous. It may happen that we have a mixture of 2,3, . . ., n homogeneous groups, each of which deviates about its own mean symmetrically and in a manner represented with sufficient accuracy by the normal curve. Thus an abnormal frequency-curve may be really built up of normal curves having parallel but not necessarily coincident axes and different parameters . . . . The object of the present paper is to discuss the dissection of abnormal frequency curves into normal curves, . . . the analytical difficulties, even for the case $n = 2$ are so considerable, that it may be questioned whether the general theory could ever be applied in practice to any numerical case.

Thus wrote Karl Pearson in the introduction of his classic, century-old paper,[6] that contained a method of moments approach to the estimation of the five parameters in a two component, univariate Gaussian mixture, i.e. the density, $f(x)$, given by

$$f(x) = p\frac{1}{\sigma_1\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} + (1-p)\frac{1}{\sigma_2\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2} \qquad (2.1)$$

Despite the estimation procedure involving the solution of a ninth degree polynomial (see Section 3), Pearson managed to apply it to a set of measurements on the ratio of forehead to body length of 1000 crabs, supplied by Professor WR Weldon. Figure 1 shows a frequency polygon of the original data together with the two component Gaussian mixture estimated by Pearson. In addition, Figure 1 shows the mixture density as estimated by maximum likelihood (see Section 3) and a single normal density having the sample mean and variance of the 1000 observations. Here the method of moments and maximum likelihood solutions are very close, although it is known in general that maximum likelihood is far more efficient (see Tan and Chan[7]).

The amount of arithmetic involved in solving a ninth degree polynomial would not have been an attractive proposition for statisticians in the early part of the twentieth century, and there were several attempts to simplify Pearson's original method. Charlier,[8] for example, devised a somewhat simpler solution, although great computational difficulties remained. A comment made by Charlier would probably have been echoed by most statisticians of the day:

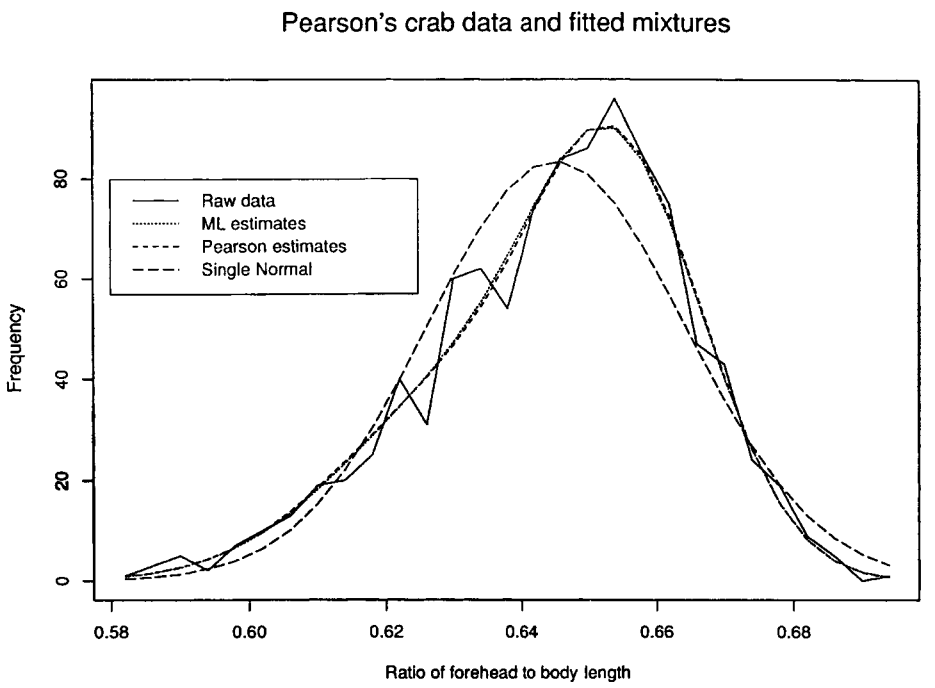

Figure 1 Frequency polygon of ratio of forehead to body length in 1000 crabs and two component normal mixtures fitted by moments and maximum likelihood

> The solution of an equation of ninth degree, where almost all powers, to the ninth, of the unknown quantity are existing, is, however, a very laborious task. Mr. Pearson has indeed possessed the energy to perform this heroic task in some instances in his first memoir on these topics from the year 1894. But I fear that he will have few successors, if the dissection of the frequency curve into two components is not very urgent.

During the next 30 years there were a number of other attempts to simplify Pearson's proposed method. These included the use of cumulants rather than moments by Strömgren[9] and the use of k-statistics by Rao.[10] Despite the computational problems associated with the two component, univariate Gaussian mixture, Charlier and Wicksell[11] attempted the estimation of the parameters in a two component mixture of *bivariate* normals, and Doetsch[12] considered the problem of normal mixtures with more than two components. In each case the method of moments was the estimation procedure used.

Because of the daunting amount of calculation involved in the estimation procedures arising from the method of moments, it is no surprise to find the development of other *ad hoc*, generally graphical, approaches to the problem. The essence of many suggested methods of this kind was to identify parts of the x-axis where the influence of all but one of the components is negligible. The estimation process is then temporarily restricted to one of these sections and the parameters of the corresponding component estimated. The same procedure is then repeated in a new part of the x-axis, often after subtraction of the previously estimated components from the observed distribution. Specific examples of this type of procedure are given in.[13–17]

Maximum likelihood estimation for the normal mixture problem was first suggested by Rao[10] (although see Newcombe[18]), who developed an iterative solution for the

case of two components, with equal standard deviations, based on Fisher's method of scoring. Hasselblad[19,20] attacked the more general case, allowing for more than two components and for components to have unequal variances. Wolfe[21–23] extended maximum likelihood estimation to the situation involving mixtures of *multivariate* normal densities, and also developed the first computer program that allowed the *almost* routine application of such mixture models. The procedure suggested by Wolfe for solving the maximum likelihood equations is an early example of the EM algorithm, formulated later in more general terms by Dempster, Laird and Rubin.[24]

Although finite mixtures with Gaussian components have always been those most widely discussed and applied, several other mixture densities have found application in particular areas. Medgyessi,[25] for example, used a mixture of binomial distributions in the context of identifying the constituents of small amounts of organic chemicals. Blishke[26] gives an example of a mixture of two trinomial distributions applied to the sex distributions of twin pairs. Thomas[27] uses a mixture of exponentials to model the distribution of the time to discharge of nerve cells and Joffe[28] employed a mixture with components of the form

$$g(x) = r \exp(-s\sqrt{x}) \tag{2.2}$$

to model the frequency distribution of the size of dust particles in mines.

Mixtures with other than normal components are also important in modelling *failure time* or *survival time* data. Here the observations are the times to failure or death of a sample of items or patients. Since death can be due to a variety of causes, each with its own particular survival time distribution, the overall survival time distribution will be a mixture. The components of such a mixture may be negative exponential, Weibull, etc., and early attempts to fit such distributions are described by Mendenhall and Hader[29] and by Kao.[30] More recent applications of mixture distributions to survival times are described by McGiffin *et al.*,[13] who deal with the distribution of time to death following major cardiac surgery, and Lui, Darrow and Rutherford,[31] who use a particular class of mixture models to estimate the probability of developing AIDS after HIV infection and the distribution of the *incubation time*, the time interval from infection with HIV to the date of diagnosis with AIDS. A review of the use of mixture distributions for survival data is given in McLachlan and McGiffin.[32]

Mixture models with multivariate Bernoulli components are the basis of *latent class analysis* (see, for example,[33–35]). This method, which is often regarded as the categorical analogue of factor analysis, is described in detail in the paper by Formann and Kohlmann in this issue of the journal. A variation of the EM algorithm (see next section), can be used to find estimates of the parameters in the latent class model. Two examples of the application of the latent class model are those described by Aitkin *et al.*[36] and Pickering and Forbes.[5]

3 Estimating the parameters in finite mixture densities

The finite mixture density given in equation (1.1) is characterized by the parameters, $\mathbf{p}$ and $\boldsymbol{\theta}$, *and* by the number of components, c. In this section the problem of estimating the former is addressed. Determining the number of components is left until Section 4.

3.1 The method of moments applied to the two component normal mixture

Pearson approached the problem of estimating the parameters in a two component normal mixture by equating sample moments to their theoretical counterparts, an approach which leads to the following system of five nonlinear simultaneous equations,

$$\begin{aligned}
p\delta_1 + (1-p)\delta_2 &= 0 \\
p(\sigma_1^2 + \delta_1^2) + (1-p)(\sigma_2^2 + \delta_2^2) &= V_2 \\
p(3\delta_1\sigma_1^2 + \delta_1^3) + (1-p)(3\delta_2\sigma_2^2 + \delta_2^3) &= V_3 \\
p(3\sigma_1^4 + 6\sigma_1^2\delta_1^2 + \delta_1^4) + (1-p)(3\sigma_2^4 + 6\sigma_2^2\delta_2^2 + \delta_2^4) &= V_4 \\
p(15\sigma_1^4\delta_1 + 10\sigma_1^2\delta_1^3 + \delta_1^5) + (1-p)(15\sigma_2^4\delta_2 + 10\sigma_2^2\delta_2^3 + \delta_2^5) &= V_5\,,
\end{aligned} \tag{3.1}$$

where $V_r = (1/n)\ \Sigma_{i=1}^{n}(x_i - \bar{x})^r$ are the sample moments of the sample observations $x_1, x_2, \ldots, x_n$ with arithmetic mean $\bar{x}$; $\delta_k = (\mu_k - \mu)$, $k = 1,2$, with $\mu = p\mu_1 + (1-p)\mu_2$. The terms p, μ_1, μ_2, σ_1, σ_2 are the five parameters of the mixture density defined in equation (2.1).

By some fairly tedious algebra, these equations may be reduced to the 'fundamental nonic' originally derived by Pearson (details are given in Cohen[37] and Everitt and Hand[38]). The problem of multiple solutions of the ninth degree polynomial was dealt with by Pearson by choosing the set of estimates which resulted in closest agreement between the *sixth* central moment of the sample and the corresponding moment of the fitted mixture distribution (although in 1894, Pearson was troubled by 'the great labour' involved in the calculation of the sixth moment being sufficient to deter the practical statistician). Although this method of estimation might now appear to hold little more than historical interest, there have been a number of relatively recent proposals for estimating the parameters of mixture distributions which use, essentially, the same approach. Quandt and Ramsay,[39] for example, suggested a scheme based on the moment generating function which seemed to work well in some situations. Lindsay and Basak,[40] in a more recent paper, demonstrate that the parameters of even multivariate normal mixtures *can* be found quickly and efficiently using the method of moments, and, in particular, recommend their approach as a means of providing starting values for algorithms designed to search for roots of the likelihood equation (see Section 3.3). A simulation study described in the paper shows that the suggested moment estimates worked better (in the sense of giving higher initial likelihoods) than the *true values* as starting values for the EM algorithm (again see Section 3.3).

3.2 Graphical estimation procedures

As mentioned in Section 2, in the 1940s and 1950s several authors suggested graphical techniques for the estimation of the parameters in a mixture of univariate normals, in an attempt to overcome the computational burden of the more formal estimation methods that had been proposed. Harding,[14] for example, describes a graphical approach based on normal probability plots. Such plots of data from a *single* normal distribution result in an approximately straight line. When the data arise from a mixture of normals, sigmoidal type curves result, and these can be used to obtain estimates of the parameters of the mixture, after first identifying their points of inflexion. In many cases the identification process is not easy, and the examples given in Cassie[15] illustrate the degree of subjectivity that may be involved. Parameter esti-

mates obtained in this way are unlikely to be very accurate, although the plotting procedure (or some variant of it) may still be useful for indicating the presence of a mixture in a set of data (see Section 4).

A further graphical estimation technique is that suggested by Bhattacharya,[17] which involves a plot of $\log\phi_{i+1}/\phi_i$ against x_i, where ϕ_i and ϕ_{i+1} are the observed frequencies of adjacent classes i and $i+1$, in the empirical frequency distribution of the data and x_i is the mid-point of class i. Bhattacharya demonstrates that for a normal mixture distribution, such a plot should lead to a series of approximately straight lines with negative slopes, each line corresponding to an area where one particular component dominates. Estimates of the parameters, μ_k and σ_k^2, of the kth component normal distribution are obtained from using the following equations:

$$\hat{\mu}_k = \lambda_k + w/2 \tag{3.2}$$

$$\hat{\sigma}_k^2 = w\cot(\alpha_k) - w^2/12, \tag{3.3}$$

where α_k is the angle between the kth straight line and the negative direction of the x-axis, λ_k is the x intercept of the line, and w is the class width. After the means and variances have been calculated in this way, Bhattacharya suggests several methods which might be used to determine the mixing proportions. An example of an application of this method is given in Everitt and Hand.[38]

3.3 Maximum likelihood estimation: the EM algorithm

Given a sample of observations, $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ from the mixture density described in equation (1), the log-likelihood function, L, is

$$L = \sum_{i=1}^{n} \ln f(\mathbf{x}_i;\mathbf{p},\boldsymbol{\theta}). \tag{3.4}$$

Estimates of the parameters in the model can be obtained as a solution of the likelihood equation

$$\partial L(\boldsymbol{\phi})/\partial\boldsymbol{\phi} = 0 \tag{3.5}$$

where $\boldsymbol{\phi}' = [\mathbf{p}',\boldsymbol{\theta}']$

Various authors,[19–23,41] have shown that the likelihood equations can be so manipulated that the likelihood estimate of $\boldsymbol{\phi}$, $\hat{\boldsymbol{\phi}}$, satisfies the following equations

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{P}(j|\mathbf{x}_i) \tag{3.6}$$

$$\sum_{j=1}^{c}\sum_{i=1}^{n} \hat{P}(j|\mathbf{x}_i)\partial \ln g_j(\mathbf{x}_i;\boldsymbol{\theta}_j)/\partial\hat{\boldsymbol{\theta}} = 0 \tag{3.7}$$

where $\hat{P}(j|\mathbf{x})$ is the estimated posterior probability of an observation, $\mathbf{x}$ arising from component density, j. For a mixture in which the jth component density is multivariate normal with mean, $\boldsymbol{\mu}_j$, and covariance matrix, $\boldsymbol{\Sigma}_j$, equations (3.6) and (3.7) become

$$\hat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \hat{P}(j|\mathbf{x}_i) \tag{3.8}$$

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n\hat{p}_j} \sum_{i=1}^{n} \mathbf{x}_i \hat{P}(j|\mathbf{x}_i) \tag{3.9}$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{n\hat{p}_j} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j)' \hat{P}(j|\mathbf{x}_i) \tag{3.10}$$

Hasselblad,[19,20] Wolfe[21–23] and Day[41] all suggested an iterative scheme for solving the likelihood equations which involved finding initial estimates of the posterior probabilities from given initial values of the parameters of the mixture, and then evaluating the right-hand sides of equations (3.8), (3.9) and (3.10) to give revised estimates of the parameter values. From these, new estimates of the posterior probabilities are derived and the procedure is repeated until some suitable convergence criterion is satisfied.

This procedure is, in fact, a particular example of the EM algorithm described by Dempster, Laird and Rubin[24] in the context of likelihood estimation for incomplete data problems. (In the estimation of the parameters in a mixture it is the labels of the component density from which an observation arises that are missing.)

Details of the EM algorithm as applied to mixture distributions are given in McLachlan and Basford.[2] One of the problems of the algorithm noted by these authors is its generally slow convergence rate. (Jones and McLachlan[42] consider a number of procedures for improving the convergence rate.) Others have also pointed out that the algorithm need not converge to the global maximum. Alternative algorithms, for example, Quasi-Newton, have also been shown to have convergence problems and, again, not infrequently converge to a local maximum of the likelihood function.

A more general problem of likelihood estimation for mixture models is that examples can be found where the likelihood is unbounded, so that the maximum likelihood estimate does not exist. Kiefer and Wolfowitz[43] point out that such a situation arises in a mixture of two normal densities with unknown different means and unknown different variances. In this simple situation the problem can be overcome by constraining the two component variances to be equal, and in general the problem of singularities in the likelihood surface of mixtures of multivariate normals can be avoided by a similar constraint on the component covariance matrices. As pointed out by McLachlan and Basford, however, such constraints may often not be necessary, since the roots of the likelihood equation may correspond to local maxima in the interior of the parameter space which are consistent and asymptotically efficient (see also Hathaway[44]).

Having estimated the parameters in a mixture of multivariate normal densities, a *cluster analysis* (see Everitt[45]) of the observations can be carried out by considering the maximum values of the estimated posterior probabilities. McLachlan and Basford discuss the use of mixtures as models for cluster analysis in great detail. Banfield and Raftery[46] also discuss the use of mixture models as the basis for clustering, and present a framework which is sufficiently general to allow for multivariate normal mixtures with different covariance matrices, and for non-Gaussian mixtures; they also describe a model-based approximate Bayesian approach to choosing the number of clusters.

The asymptotic variance–covariance matrix of the maximum likelihood estimators

of the parameters in a normal mixture can be obtained from the inverse of Fisher's information matrix, that is, the inverse of a matrix whose *ij*th element is given by

$$E\left[\frac{\partial \ln f(\mathbf{x},\boldsymbol{\phi})}{\partial \phi_i}\,\frac{\partial \ln f(\mathbf{x},\boldsymbol{\phi})}{\partial \phi_j}\right] \tag{3.11}$$

For a two component univariate, Behboodian[47] shows that the elements of the information matrix are obtained by the numerical evaluation of integrals of the form

$$M_{mn}(g_i,g_j) = \int_{-\infty}^{\infty} \left(\frac{x-\mu_i}{\sigma_i}\right)^m \left(\frac{x-\mu_j}{\sigma_j}\right)^n \left[\frac{g_i(x)g_j(x)}{f(x)}\right] dx \tag{3.12}$$

Chang[48] considers the information matrix in the case of a two component multivariate normal mixture with assumed common variance–covariance matrix. He shows that by suitable reparameterization, the standard errors of the maximum likelihood estimates in *any* such mixture (i.e. one with any given value of d) may always be obtained from the information matrix for the three-dimensional case. Chang further shows that the three-dimensional integrals then involved can be simplified to functions of 14 one-dimensional integrals which can be evaluated numerically relatively simply. McLachlan and Basford[2] consider the evaluation of the *observed* information matrix, which can also be used to provide an estimate of the covariance matrix of the maximum likelihood estimates, but which is much simpler to compute than the expected Fisher information matrix.

4 Detecting finite mixture densities and determining the number of components in a mixture

Application of finite mixture densities is most convincing in situations where the existence of separate groups of observations with differing distributions is strongly implied by the nature of the application. In many cases considered in practice, however, the evidence for modelling data with a mixture distribution must be obtained empirically. It is in such circumstances that the questions of the appropriate value of c, the number of components in the mixture, and whether a mixture density is or is not an appropriate model for the data at all, become of considerable importance. The procedures that have been suggested for answering such questions involve a combination of graphical techniques and quasi-significance tests, both of which will be considered in the following sections.

4.1 Detecting mixtures

For univariate data, at least, perhaps the most natural candidate to consider when assessing whether a mixture distribution might be a suitable model for a sample of observations is the sample histogram, with clear multimodality being taken as strong evidence that a mixture distribution is appropriate. Figure 2, for example, taken from Holgerson and Jorner,[49] shows the fibre size distribution of myelinated lumbosacral ventral root fibres from a 96-day-old kitten. Two modes are clearly visible, corresponding here to axons of gamma neurones (first mode) and alpha neurones (second mode). Unfortunately the mixing of two or more unimodel densities will lead to a mixed density with more than a single mode only in particular circumstances (see Everitt and Hand[38]), so that cases as clear-cut as Figure 2 are likely to be the excep-

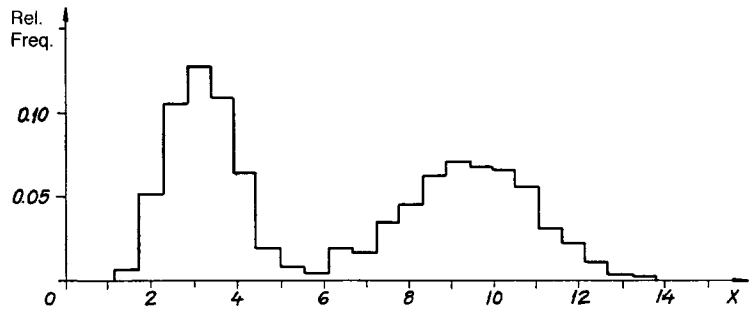

Figure 2 Fibre size distribution of myelinated lumbosacral ventral root fibres from a 96-day-old kitten

tion rather than the rule. Consequently, it becomes of importance to look beyond the sample histogram for evidence of a mixture. One possibility, particularly for mixtures with normal components, is to use a Q–Q plot of the data, or perhaps an adaptation of such a plot proposed by Fowlkes[50] which the author claims is particularly sensitive to the presence of mixtures.

Fowlkes' suggested procedure is to plot $(x_{(i)} - \bar{x})/s$ against $\Phi((x_{(i)} - \bar{x})/s) - p_i$ where $x_{(1)} \leq x_{(2)} \leq \ldots x_{(n)}$ represent the ordered sample values, which have mean $\bar{x}$ and standard deviation s; $p_i = (i - 0.5)/n)$ and Φ is the standard normal distribution function. When the observations arise from a single normal density, Fowlkes' proposed plot results in an approximately horizontal line at $y = 0$. Mixture densities lead to plots having a characteristic cyclical pattern about zero, which differs from the pattern given by other non-normal distributions.

As an illustration of the application of each of the procedures described above for detecting mixtures, Figure 3 shows the resulting histograms, probability plots and Fowlkes' plots, for 100 observations from a single normal distribution and from a two component mixture in which one component has mean zero and standard deviation unity, and the other has mean one, two, three or four, with again unit standard deviation. It appears that the component densities need to be relatively widely separated for *any* of the methods to give a clear indication of the presence of a mixture, with perhaps Fowlkes' plot being marginally more effective.

For the detection of multivariate normal mixtures, a possible procedure is a chi-squared probability plot of the Mahalanobis distance of each observation from the sample mean vector. If the data are from a single multivariate normal density these distances have, approximately, a chi-squared distribution with d degrees of freedom, where d is the number of variables; consequently a chi-squared probability plot of the ordered distances will result in an approximately straight line through the origin. Mixtures of multivariate normals will, however, tend to give plots that are 'S' shaped. An example of this type of plot is given in Section 5.

4.2 The likelihood ratio test for the number of components in a mixture density

A natural candidate for testing the hypothesis, $c = c_0$, against $c = c_1 (c_1 > c_0)$, where c is the number of components in a mixture density, is the likelihood ratio statistic, λ. Unfortunately this does not lead to a suitable significance test, since for mixture densities regularly conditions do not hold for $-2\ln\lambda$ to have its usual asymptotic null distribution i.e. a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters under the competing hypotheses. The prob-

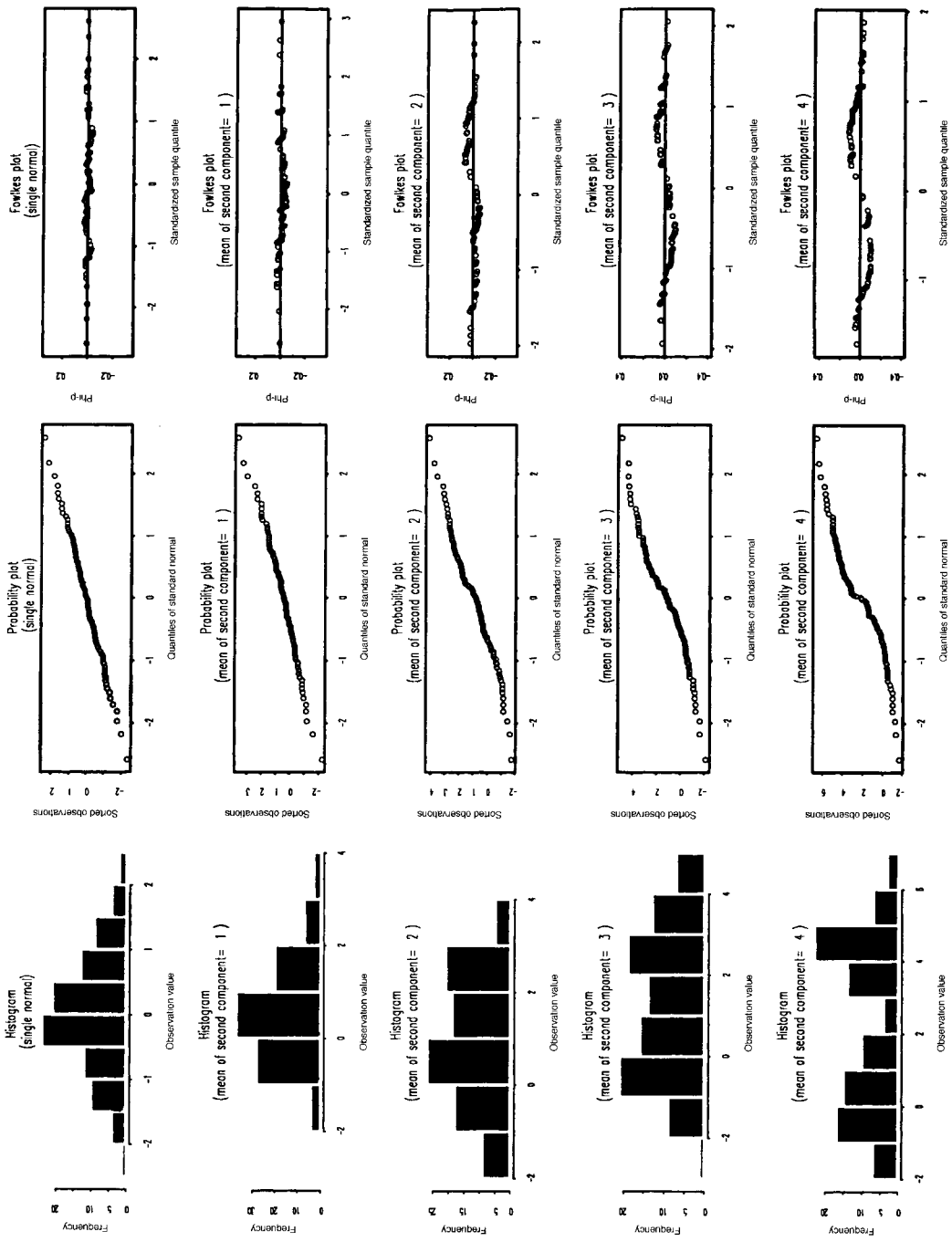

Figure 3 Histograms, probability plots and Fowlke's plots for two component normal mixtures with increasing separation between the two components

lem is that the null distribution is on the edge of the parameter space, in the sense that when two components coincide, their proportions become unidentifiable, or if a proportion tends to be zero, the corresponding parameters of the distribution become unidentifiable.

The problem has been considered by a number of authors. Wolfe,[51] for example, on the basis of a limited simulation study, suggested that the null distribution of the likelihood ratio statistic for testing $c+1$ against c components is chi-squared with $2\nu-2$ degrees of freedom, where ν is the number of extra parameters in the $c+1$ component mixture. Later authors, such as Hernandez-Avila,[52] Everitt[53] and Thode *et al.*[54] have shown that Wolfe's suggestion behaves reasonably in particular circumstances. McLachlan[55] describes a bootstrap approach to the problem. Large sample properties of likelihood ratio tests when the true parameter value may be on the boundary of the parameter space have been considered by Gosh and Sen[56] and by Self and Liang.[57]

Although no completely satisfactory method of using the likelihood ratio statistic as a formal significance test for number of components is available, McLachlan and Basford's suggestion that Wolfe's modified likelihood ratio test can be usefully employed as an *informal* guide to the appropriate number of components seems not unreasonable.

5 Some examples of the application of finite mixture densities in medical research

5.1 Age of onset of schizophrenia

A sex difference in the age of onset of schizophrenia was noted by Kraepelin.[58] Subsequently, it has been one of the most consistent findings in the epidemiology of the disorder. Lewine[59] collated the results of seven studies on the age of onset of the illness, and 13 studies on age at first admission, and showed that all these studies were consistent in reporting an earlier onset of schizophrenia in men than in women. Lewine suggested two competing models to explain these data:

> The timing model states that schizophrenia is essentially the same disorder in the two sexes but has an early onset in men and a late onset in women . . . In contrast with the timing model, the subtype model posits two types of schizophrenia. One is characterised by early onset, typical symptoms, and poor premorbid competence, and the other by late onset, atypical symptoms, and good premorbid competence . . . the early onset typical schizophrenia is highly a disorder of men, and late onset, atypical schizophrenia is largely a disorder in women.

The subtype model implies that the age of onset distribution for both male and female schizophrenics will consist of a mixture, with the mixing proportion for early onset schizophrenia being larger for men than for women. The proposal was investigated by fitting normal mixture distributions to the age of onset (determined as age on first admission) of 250 male and 250 female schizophrenics. Analysis on a logarithmic scale gave the results shown in Table 1. The characteristics of the two component solutions of men and women are very different, that for the women fitting very well into the subtype model. In fact, using a significance level of 0.01 to judge the likelihood ratio test for number of components, suggests that the data for men can be adequately described by a *single* normal distribution, but that for the women, a two component mixture is needed. Histograms of the data and the fitted normal densities are shown in Figures 4 and 5. These results suggest that the subtype model may be reasonable for women, but may be an unnecessary complication for men.

Table 1 Results from fitting normal mixtures to log age of onset of 250 male and 250 female schizophrenics

1) Males

Component	Log-likelihood	Chi square	df	*p*
1	295.46			
2	300.77	10.53(2v1)	4	0.0324
3	303.50	5.39(3v2)	4	0.2492

The sample mean is 1.5236 and the sample variance is 0.1734.

The estimated parameters of the two component solution and their standard errors are:

Component	Mixing proportion	Mean	Variance
1	0.919(0.03424)	1.5236(0.01477)	0.03007(0.00339)
2	0.081	1.8267(0.01462)	0.00138(0.00107)

2) Females

Component	Log-likelihood	Chi square	df	*p*
1	342.58			
2	350.41	15.52(2v1)	4	0.0037
3	351.00	1.17(3v2)	4	0.8835

The sample mean is 1.4932 and the sample variance is 0.1541.

The estimated parameters of the two component solution and their standard errors are:

Component	Mixing proportion	Mean	Variance
1	0.315(0.11683)	1.3337(0.02318)	0.00534(0.00199)
2	0.685	1.5663(0.03065)	0.01510(0.00419)

5.2 The variation of mortality rates between geographical areas

Current residence is widely used for the comparison of incidence or mortality rates between geographical areas. When studying diseases with long latency periods, migration between geographical areas reduces the sensitivity of this method. Beginning in 1979, mortality statistics by state or country of birth were published by the US Federal Government and made available in computerized files. An analysis of residence histories provided in the US 1958 current population survey found that 77.4% of people had not moved from their birthplace by the age of 19. Consequently, birthplace, which is listed on death certificates, provides a reasonably stable measure for geographical comparisons of potential early exposures for cohorts born before 1940, and Betemps and Buncher[60] investigated state of birth as a possible risk factor for motor neurone disease (MND), Parkinson's disease (PD), multiple sclerosis (MS) and cerebrovascular disease (CVA). Using proportional mortality rates for each of the diseases for each of the states in the USA except Hawaii, they found positive correlations between each disease rate and the latitudes and longitudes of the states, apart from CVA. Here mixtures of multivariate normal densities will be used to 'explore' the structure of the four mortality rates for 48 states (the District of Columbia is excluded). The relevant data are given in Table 1 of Betemps and Buncher's paper.

Before fitting mixture distributions to the data, a chi-squared plot of the generalized distances, as described in the previous section, was constructed, and is shown in

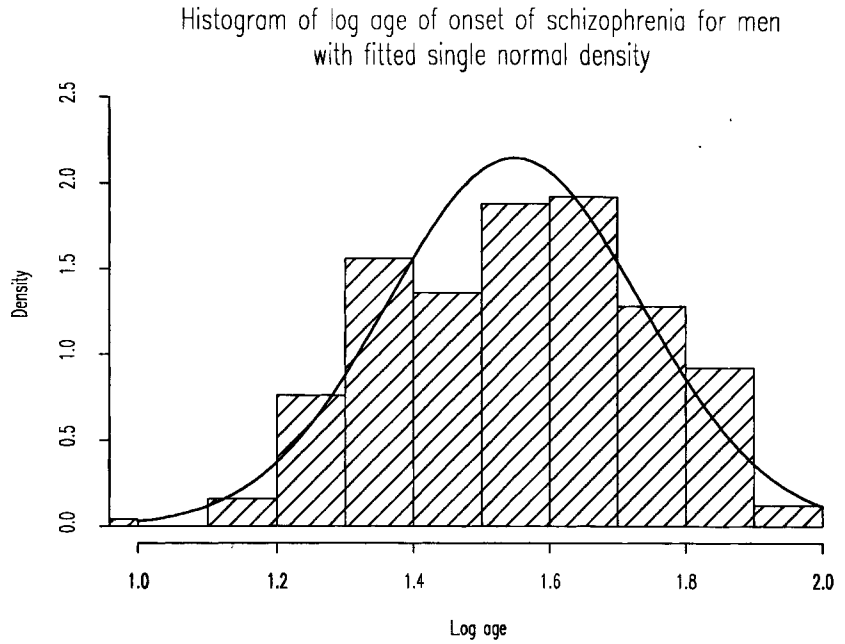

Figure 4 Age of onset data and fitted single normal density for male schizophrenics

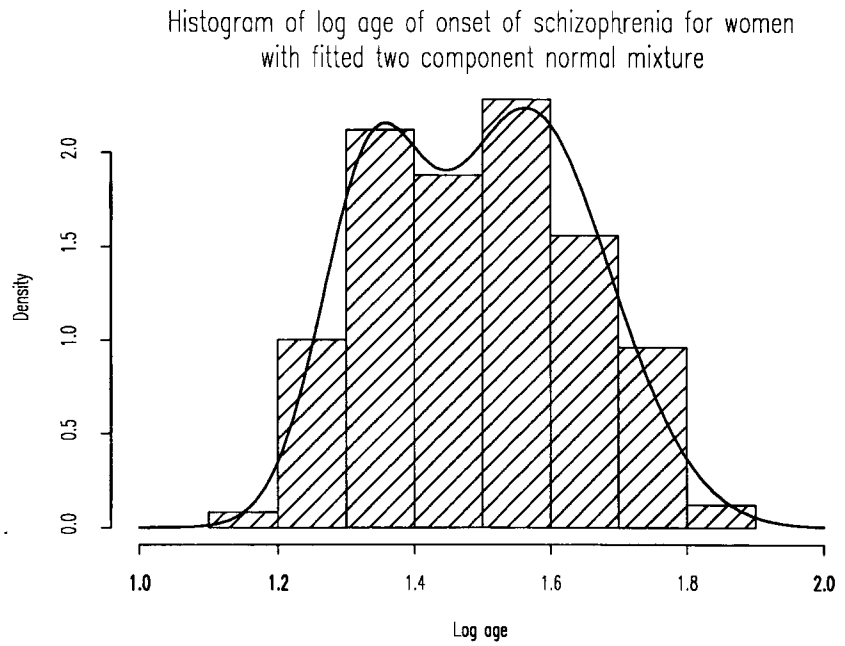

Figure 5 Age of onset data and fitted two component normal mixture for female schizophrenics

Figure 6. Some evidence of a departure from linearity is seen in the plot suggesting the presence of a mixture.

Two, three and four component normal mixtures were fitted to the data, with, in each case, the covariance matrices of the components constrained to be the same. With only 48 observations, allowing unequal covariance matrices leads to far too many parameters to estimate. Because of the small sample size, the likelihood ratio test for number of components should, perhaps, be taken even less seriously than is usual. The results, given in Table 2, suggest some rather weak evidence in favour of two components. The parameter estimates for the two component solution are shown in Table 3; this solution corresponds to an approximately equal division of the states

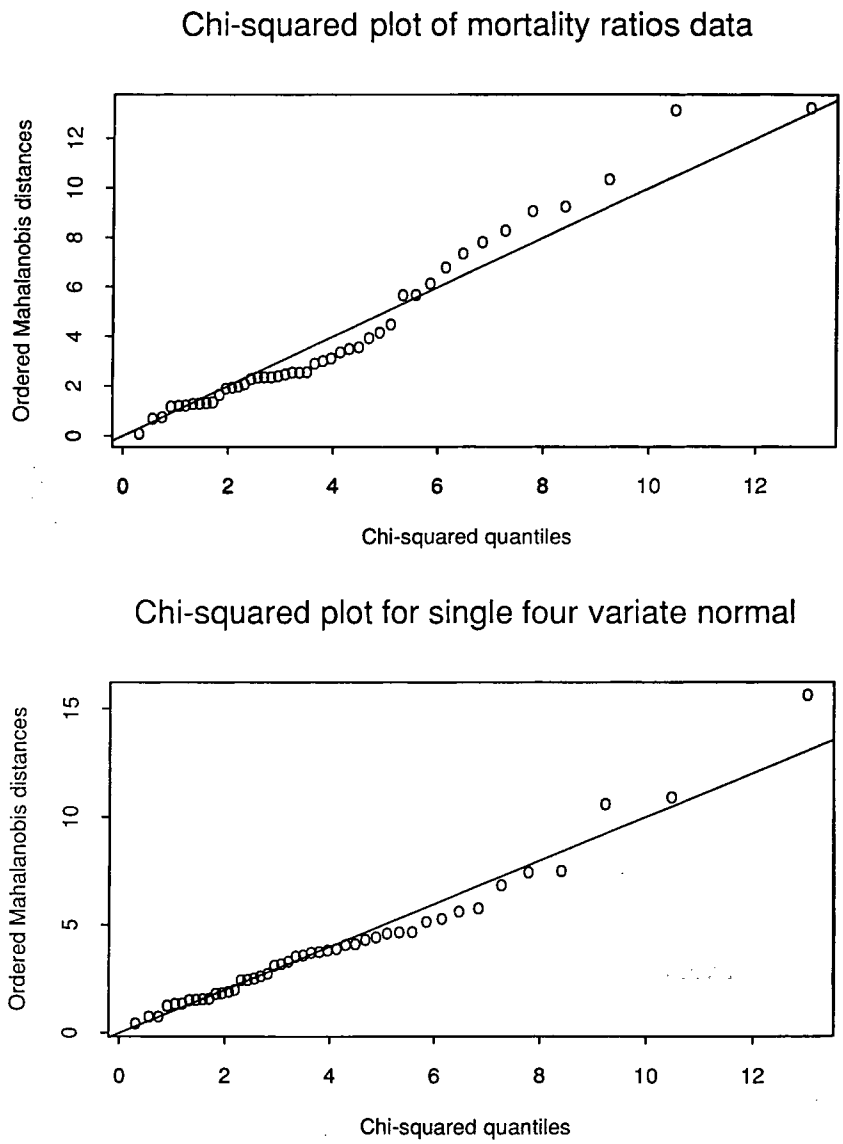

Figure 6 Chi-squared plots of the proportional mortality ratios for USA states

Table 2 Likelihood ratio tests for normal mixture densities fitted to mortality ratios data

Component	Log-likelihood	Chi square	df	*p*
1	−537.62			
2	−529.80	13.84(2v1)	8	0.09
3	−522.16	13.37(3v2)	8	0.10
4	−518.65	6.07(4v3)	8	0.64

Table 3 Parameter estimates for two component normal mixture fitted to mortality ratios data

Two component mixture

		Means			
Component	**Mixing proportion**	**MND**	**PD**	**MS**	**CVA**
1	0.42	15.74	25.37	6.60	945.45
2	0.58	17.31	22.51	11.93	765.23

Standard deviations			
MND	**PD**	**MS**	**CVA**
4.81	6.30	3.77	54.76

Correlation matrix

$$\mathbf{R} = \begin{array}{c} \\ \text{MND} \\ \text{PD} \\ \text{MS} \\ \text{CVA} \end{array} \begin{array}{c} \begin{array}{cccc} \textbf{MND} & \textbf{PD} & \textbf{MS} & \textbf{CVA} \end{array} \\ \begin{pmatrix} 1.00 & & & \\ 0.33 & 1.00 & & \\ 0.30 & 0.62 & 1.00 & \\ 0.10 & 0.24 & 0.17 & 1.00 \end{pmatrix} \end{array}$$

into those with high CVA and low MS rates, and those where the reverse is the case. An obvious way to display the solution in this case, is to identify the states belonging to each cluster formed on the basis of the estimated posterior probabilities, on a map of the USA. This is done in Figure 7. There is a clear general division of the states in terms of longitude, and, in addition, amongst those states on the East Coast, in terms of latitude. Florida differs from this overall pattern because its profile of mortality rates is somewhat between the mean vectors of the two clusters; its estimated posterior probabilities of belonging to cluster one (0.06) and cluster two (0.94), however, clearly place it in the second cluster.

5.3 Latent class analysis of mother and child reports of child smoking behaviour

Table 4 shows data collected by Fergusson and Horwood.[61] These data arise from mother and child reports of mother and child smoking behaviour. A latent class model with four classes was fitted to the data, where the four classes were assumed to be:

- *C*1: mother smoker, child smoker,
- *C*2: mother nonsmoker, child smoker,
- *C*3: mother smoker, child nonsmoker,
- *C*4: mother nonsmoker, child nonsmoker.

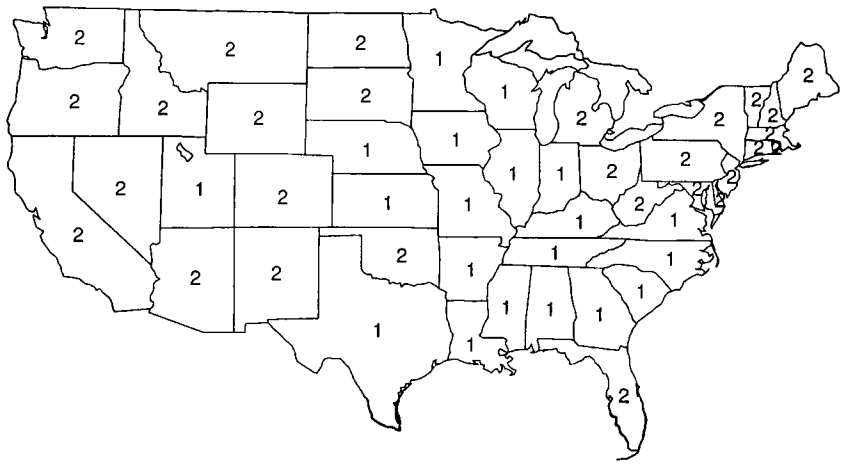

Figure 7 Two component normal mixture clusters for the mortality ratios data, showing states assigned to each cluster on the basis of the maximum values of the estimated posterior probabilities

Table 4 Response patterns for child and maternal reports of child and maternal smoking (1=smoker, 2=nonsmoker)

x_1	x_2	x_3	x_4	Frequency
1	1	1	1	35
2	1	1	1	31
1	2	1	1	25
2	2	1	1	105
1	1	2	1	2
2	1	2	1	2
1	2	2	1	2
2	2	2	1	8
1	1	1	2	5
2	1	1	2	3
1	2	1	2	8
2	2	1	2	23
1	1	2	2	51
2	1	2	2	28
1	2	2	2	52
2	2	2	2	429

x_1 = Child's report of child's smoking behaviour, x_2 = Mother's report of child's smoking behaviour, x_3 = Child's report of mother's smoking behaviour, x_4 = Mother's report of mother's smoking behaviour.

It was further assumed that the reporting accuracies of x_1, child's report of child's smoking behaviour and x_2, mother's report of child's smoking behaviour, do not depend on maternal smoking, and that the reporting accuracies of x_3, child's report of mother's smoking behaviour and x_4, mother's report of mother's smoking behaviour, do not depend on child smoking. These assumptions imply the following constraints on the class probabilities:

$$P(x_1 = 1|C1) = P(x_1 = 1|C2)$$
$$P(x_2 = 1|C1) = P(x_2 = 1|C2)$$

$$
\begin{aligned}
P(x_1 = 1|C3) &= P(x_1 = 1|C4) \\
P(x_2 = 1|C3) &= P(x_2 = 1|C4) \\
P(x_3 = 1|C1) &= P(x_3 = 1|C3) \\
P(x_4 = 1|C1) &= P(x_4 = 1|C3) \\
P(x_3 = 1|C2) &= P(x_3 = 1|C4) \\
P(x_4 = 1|C2) &= P(x_4 = 1|C4) \qquad (5.1)
\end{aligned}
$$

where a variable value of 1 indicates that the report involved identified the person involved as a smoker.

The estimated parameters in the model are shown in Table 5. The fitted model suggests that errors of measurements in reports of child smoking largely arise from false negative responses in which children who smoke describe themselves as non-smokers. Fergusson and Horwood show that one of the consequences of the high false negative rates is an underestimation of the prevalence of smoking behaviour and an underestimation of the strength of association between maternal and child smoking.

Several other examples of the application of latent class models are given in the paper by Formann and Kohlmann in this issue of the journal.

5.4 Mixture models for the time to reoperation for degeneration of xenograft values

As a further example, the use of finite mixture models in the analysis of survival data will be described, using the case study reported in McLachlan and McGiffin.[32] This study involved an investigation of the time to reoperation for degeneration of xenograft values implanted in the aortic position of some 1004 patients. Of these patients, 73 subsequently underwent reoperations for xenograft degeneration, while 212 died without requiring a reoperation for degeneration. The remaining 719 survival times were all censored, since at the end of the study the corresponding patients

Table 5 Estimated parameters for four class model for smoking behaviour data

1) **Estimates of mixing proportions**

Class 1(C1): Mother smoker, child smoker-$\hat{p}_1 = 0.147$,
Class 2(C2): Mother nonsmoker, child smoker-$\hat{p}_2 = 0.156$,
Class 3(C3): Mother smoker, child nonsmoker-$\hat{p}_3 = 0.140$,
Class 4(C4): Mother nonsmoker, child nonsmoker-$\hat{p}_4 = 0.557$.

2) **Estimates of class probabilities**

Parameter	Value	Interpretation
$p(x_1 = 1\|C1)$	0.593	True positive rate : child report of child
$p(x_2 = 1\|C1)$	0.639	True positive rate : mother report of child
$p(x_1 = 1\|C3)$	0.061	False positive rate : child report of child
$p(x_2 = 1\|C3)$	0.000	False positive rate : mother report of child
$p(x_3 = 1\|C1)$	0.950	True positive rate : child report of mother
$p(x_4 = 1\|C1)$	0.888	True positive rate : mother report of mother
$p(x_3 = 1\|C2)$	0.025	False positive rate : child report of mother
$p(x_4 = 1\|C2)$	0.008	False positive rate : mother report of mother

were either still living (without having undergone reoperation for xenograph degeneration), or had undergone a reoperation for some reason unrelated to xenograft degeneration.

For this problem, the failure time, T, is the time to the occurrence of the event *reoperation for degeneration.* Clearly, patients who die without a reoperation will never experience this event. Hence in order to model the survival function, $S(t)$ for this event, the patients are separated into two groups, G_1 and G_2, where G_1 refers to the group of patients who die without having undergone a reoperation for degeneration and G_2 refers to the group of patients who do undergo a reoperation in their lifetime. The survival function is then modelled by the following two component mixture:

$$S(t) = \pi_1 + \pi_2 S_2(t) \tag{5.2}$$

where π_1 is the probability that a patient belongs to G_1 and therefore has a survival function of unity, and $\pi_2 = 1 - \pi_1$ is the probability that a patient belongs to G_2. McLachlan and McGiffin model π_1 as a logistic function of the age of a patient at the time of the initial operation for replacement, and S_2 as a Gompertz distribution with a single covariate, again, age of patient at the time of the initial operation. The model was fitted by maximum likelihood using the EM algorithm. Results and further details are given in McLachlan and McGiffin.

6 Summary

The first 50 years of using finite mixture densities, following Pearson's pioneering work in 1894, involved, primarily, the development of techniques designed to simplify the daunting amount of arithmetic involved in applying the method of moments in estimating their parameters. During the last 20 years the wide availability of powerful computers has lessened the arithmetical problems and maximum likelihood estimation can now be applied routinely to find parameter estimates for mixtures of normals and of other densities. Mixtures continue, however, to provide a rich source of material for statisticians and recent work includes the development of models for data consisting of both continuous and categorical variables,[62,63] the linking of mixture models with *hidden Markov chain models* and with artificial neural networks,[64,65] the use of influence-based diagnostics for normal mixtures[66,67] and further consideration of the number of components problem.[68]

As in other areas of statistics, the ease with which finite mixture densities can now be applied is not without its problems and it seems appropriate to end with the following caveat issued in another context by Pearson in 1922: 'The imagination of man has always run riot, but to imagine a thing is not meritorious unless we demonstrate its reasonableness by the laborious process of studying how it fits experience'.

References

1 Tarter ME, Lock MD. *Model-free curve estimation.* London: Chapman and Hall, 1993.

2 McLachlan GJ, Basford KE. *Mixture models; inference and applications to clustering.* New York: Marcel Dekker, Inc., 1988.

3 McGiffin DC, Galbraith AJ, McLachlan GJ *et al.* Aortic valve infection-risk factors for death and recurrent endocarditis following aortic valve replacement. *Journal of Thoracic Cardiovascular Surgery* 1992; **104**: 511–20.

4 Blackstone EH, Naftel DC, Turner ME.

The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American Statistical Association* 1986; **81**: 615–24.
5 Pickering RM, Forbes JF. A classification of Scottish infants using latent class analysis. *Statistics in Medicine* 1984; **3**: 249–59.
6 Pearson K. Contribution to the mathematical theory of evolution. *Philosophical Transactions A* 1894; **185**: 71–110.
7 Tan WY, Chang WC. Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association* 1972; **67**: 702–708.
8 Charlier CVL. Researchers into the theory of probability. *Lunds Universities Årskrift, Ny foljd* 1906; **2.1**, No. 5.
9 Strömgren B. Tables and diagrams for dissecting a frequency curve into components by the half-invariant method. *Skand. Aktuarietidskr* 1934; **17**: 7–54.
10 Rao CR. The utilization of multiple measurements in the problems of biological classification. *Journal of the Royal Statistical Society, Series B* 1948; **10**: 159–203.
11 Charlier CVL, Wicksell SD. On the dissection of frequency functions. *Arkiv. for Matematik, Astronomi och Fysik* 1924; **18**: no. 6.
12 Doetsch G. Die Elimination des Dopplereffekts bei spektroskopischen feinstrukturen and exakte Bestimming der Kompronenten. *Zeitschrift für Physik* 1928; **49**: 705–30.
13 Tiselius A, Kabat EA. Electrophoretic study of immune sera and purified antibody preparations. *Journal of Experimental Medicine* 1939; **69**: 119–31.
14 Harding JP. The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biology Association of the UK* 1949; **28**: 141–53.
15 Hald A. *Statistical theory with engineering applications.* New York: Wiley, 1952.
16 Cassie RM. Some uses of probability paper in the analysis of size frequency distributions. *Australian Journal of Marine and Freshwater Research* 1954; **5**: 513–22.
17 Bhattacharya CG. A simple method of resolution of a distribution into Gaussian components. *Biometrics* 1967; **23**: 115–35.
18 Newcombe S. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics* 1886; **8**: 343–66.
19 Hasselblad V. Estimation of parameters for a mixture of normal distributions. *Technometrics* 1966; **8**: 431–44.
20 Hasselblad V. Estimation of finite mixtures of distribution from the exponential family. *Journal of the American Statistical Association* 1969; **64**: 1459–71.
21 Wolfe JH. A computer program for the maximum likelihood analysis of types. *Technical Bulletin, 65–15:* San Diego: US Naval Personnel Research Activity, 1965.
22 Wolfe JH. NORMIX: computational methods for estimating the parameters of multivariate normal mixtures of distributions. *Research Memorandum, SRM 68–2.* San Diego: US Naval Personnel Research Activity, 1967.
23 Wolfe JH. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioural Research* 1970; **5**: 329–50.
24 Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1977; **39**: 1–38.
25 Medgyessi P. *Decompositions of superpositions of distribution functions.* Budapest: Hungarian Academy of Sciences, 1961.
26 Blischke WR. Mixtures of distributions, In: *International encyclopedia of statistics,* Kruskal WH, Tanur JM eds. New York: The Free Press, 1979.
27 Thomas EAC. Mathematical models for the clustered firing of single cortical neurones. *British Journal of Mathematical and Statistical Psychology* 1966; **19**: 151–62.
28 Joffe AD. Mixed exponential estimation by the method of half moments. *Applied Statistics* 1964; **13**: 91–98.
29 Mendenhall W, Hader RJ. Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika* 1958; **45**: 504–20.
30 Kao JHK. A graphical estimation of mixed Weibull parameters in life-testing electron tubes. *Technometrics* 1959; **1**: 389–407.
31 Lui KJ, Darrow WW, Rutherford GW. A model-based estimate of the mean incubation period for AIDS in homosexual men. *Science* 1988; **20**: 1333–35.
32 McLachlan GJ, McGiffin DC. On the role of finite mixture models in survival analysis.

Statistical Methods in Medical Research 1994; **2**: 211–26.
33 Green BF. A general solution for the latent class model of latent structure analysis. *Psychometrika* 1951; **16**: 151–66.
34 Gibson WA. Three multivariate models: factor analysis, latent structure analysis and latent profile analysis. *Psychometrika* 1959; **24**: 229–52.
35 Lazarsfeld PF, Henry NW. *Latent structure analysis.* New York: Houghton Mifflin, 1968.
36 Aitkin M, Anderson D, Hinde J. Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, Series A* 1981; **144**: 419–48.
37 Cohen AC. Estimation in mixture of two normal distributions. *Technometrics* 1967; **9**: 15–28.
38 Everitt BS, Hand DJ. *Finite mixture distributions.* London: Chapman and Hall, 1981.
39 Quandt RE, Ramsey JB. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association* 1978; **73**: 730–38.
40 Lindsay BG, Basak P. Multivariate normal mixtures: a fast consistent method of moments. *Journal of the American Statistical Association* 1993; **88**: 468–76.
41 Day NE. Estimating the components of a mixture of normal distributions. *Biometrika* 1969; **56**: 463–74.
42 Jones PN, McLachlan GJ. Improving the convergence rate of the EM algorithm for a mixture model fitted to grouped truncated data. *Journal of Statistical Computation and Simulation* 1992; **43**: 31–44.
43 Keifer J, Wolfowitz J. Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* 1956; **27**: 887–906.
44 Hathaway RJ. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics* 1985; **13**: 795–800.
45 Everitt BS. *Cluster analysis.* London: Arnold, 1993.
46 Banfield JD, Raferty AE. Model-based Gaussian and non-Gaussian clustering. *Biometrics* 1993; **49**: 803–21.
47 Behboodian J. Information matrix for a mixture of two normal distributions. *Journal of Statistical Computation and Simulation* 1972; **1**: 295–314.
48 Chang WC. Confidence interval estimation and transformation of data in a mixture of two multivariate normal distributions with any given large dimension. *Technometrics* 1979; **21**: 351–55.
49 Holgersson N, Jorner U. *Decomposition of a mixture of two normal components.* Research Report 76-13, University of Uppsala, Sweden, 1976.
50 Fowlkes EB. Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association* 1979; **74**: 561–75.
51 Wolfe JH. A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. *Technical Bulletin STB 72–2.* San Diego: Naval Personnel and Training Research Laboratory, 1971.
52 Hernandez Avila A. *Problems in cluster analysis.* [Thesis]. Oxford, 1979.
53 Everitt BS. A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioural Research* 1981; **16**: 171–80.
54 Thode HC, Finch SJ, Mendell NR. Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals. *Biometrics* 1989; **44**: 1195–1201.
55 McLachlan GJ. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* 1987; **36**: 318–24.
56 Gosh JM, Sen PK. On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. In: *Proceedings of the Berkeley Conference in Honour of Jerzy Neyman and Jack Keifer, Volume II,* Le Cam LM, Olshen RA eds. Monterey: Wadsworth, 1985.
57 Self SG, Liang KY. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 1987; **82**: 605–10.
58 Kraeplin E. *Dementia praecox and paraphenia.* Edinburgh: Livingstone, 1919.
59 Lewine RRJ. Sex differences in schizophrenia: timing or subtypes? *Psychological Bulletin* 1981; **90**: 432–44.
60 Betemps EJ, Buncher CR. Birthplace as a risk factor in motor neurone disease and Parkinson's disease. *International Journal of Epidemiology* 1993; **22**: 898–904.
61 Fergusson DM, Horwood LJ. A latent class model of smoking experimentation in

children. *Journal of Child Psychology and Psychiatry* 1989; **30**: 761–73.
62 Everitt BS. A finite model for the clustering of mixed mode data. *Statistics and Probability Letters* 1988; **6**: 305–309.
63 Lawrence CJ, Krzanowski WJ. Mixture separation for mixed-mode data. *Statistics and Computing* 1995, **6**: 85–92.
64 Titterington DM. Mixture distributions (update). In: Kotz SM ed. *Encyclopedia of statistical science (update)*. New York: Wiley, 1996.
65 Ripley BD. Neural networks and related models for classification (with discussion). *Journal of the Royal Statistical Society, Series B* 1994; **56**: 409–56.
66 Jorgensen MA. Influence-based diagnostics for finite mixture models. *Biometrics* 1990; **46**: 1047–58.
67 Lindsay BG, Roeder K. Residual diagnostics for mixture models. *Journal of the American Statistical Association* 1992; **87**: 785–94.
68 Roeder K. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association* 1994; **89**: 487–95.