(2)

MRC Technical Summary Report #2369

SOME PROBLEMS WITH DATA FROM FINITE MIXTURE DISTRIBUTIONS

D. M. Titterington

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

April 1982

(Received February 15, 1982)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

SOME PROBLEMS WITH DATA FROM FINITE MIXTURE DISTRIBUTIONS

D. M. Titterington*

Technical Summary Report #2369
April 1982

ABSTRACT

Finite mixture distributions arise in many statistical applications. After the basic definition of mixture distributions, many of these applications are listed, sampling models are proposed and the basic statistical problems are described. More detailed study is then made of the use of the familiar statistical methodologies in mixture decomposition, of the incorporation of mixture data into discrimination procedures and of the problems that arise in hypothesis testing.

AMS (MOS) Subject Classifications: Primary: 62-02; Secondary: 62F05, 62F10, 62F12, 62G05, 62H30, 65D10

Key Words: finite mixture distributions, incomplete data, method of moments, graphical methods, maximum likelihood, Bayesian analysis, minimum distance, curve fitting, discrimination, likelihood ratio test, multimodality

Work Unit Number 4 (Statistics and Probability)

*Department of Statistics, University of Glasgow, Glasgow G12 8QW, Scotland.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

Finite mixture distributions are usually characterized by a probability density function of the form

$$p(x) = \sum_{j=1}^{k} \pi_j f_j(x) ,$$

where $\pi_1,\ldots,\pi_k$ are probabilities and $f_1(\cdot),\ldots,f_k(\cdot)$ are themselves probability density functions. It is often helpful to interpret the $\{\pi_j\}$ as prevalence rates of observations from k sources and the $\{f_j(\cdot)\}$ as the density functions for the observed random quantity, conditional on the source. In a typical application, in sedimentology, a sand sample if analyzed for grain-size (giving a frequency distribution of values of x). The k sources correspond to the constituent minerals of the sand. Mixtures find application in a very wide number of applied fields, such as geology, fisheries research, medicine, electrophoresis, economics, botany and communications. They are also useful as tools in some branches of statistical analysis.

The paper surveys the methods of solution of statistical problems which arise with data from a mixture, possibly supplemented by further data whose source identities are known.

The most detailed comments are related to mixture decomposition: given data, to estimate any unknown features of the model underlying the formula for $p(x)$. All the standard statistical estimation procedures are discussed and particular emphasis is placed on the points where difficulties arise that are peculiar to this problem.

The statistical discrimination problem usually involves the use of a "training set" of data, whose sources are unknown, to develop a procedure to aid the identification of the source of a future observation. The present paper investigates the extent to which mixture data can contribute to such a discriminant rule.

Finally the problem of testing for the number, k, of components is discussed. The interesting feature again is that, although a very familiar general technique may be considered, particular difficulties arise in the present context.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

SOME PROBLEMS WITH DATA FROM FINITE MIXTURE DISTRIBUTIONS

D. M. Titterington*

1. DEFINITION OF FINITE MIXTURE DISTRIBUTIONS

Suppose that a random variable, X, takes values in a sample space, S, and that its distribution is represented by a probability density function (p.d.f.) of the form

$$p(x) = \sum_{j=1}^{k} \pi_j f_j(x) , \qquad (x \in S) \tag{1}$$

where $\{\pi_j\}$ are a set of probabilities and $\{f_j(\cdot)\}$ are themselves p.d.f.'s on S. Then X is said to have a finite mixture distribution. The parameters $\{\pi_j\}$ are the mixing weights and the $\{f_j(\cdot)\}$ are the component densities. It is easy to check that $p(\cdot)$, as defined above, is indeed a p.d.f. on S.

Although equation (1) appears to be written as if X is meant to be a univariate continuous random variable, we shall subsume, under the same notation, the cases of random vectors and discrete data, interpreting $p(\cdot)$ and $\{f_j(\cdot)\}$ as probability mass functions in the latter case.

If the densities $\{f_j(\cdot)\}$ are of specified parametric forms, we shall write

$$p(x) = \sum_{j=1}^{k} \pi_j f_j(x|\theta_j) = p(x|\underline{\pi},\underline{\theta}) = p(\underline{\psi}) ,$$

in which θ_j denotes the parameters relevant to $f_j(\cdot)$, $\underline{\theta}$ denotes the aggregate of all distinct parameters in $\theta_1,\ldots,\theta_k$ and $\underline{\psi}$ denotes the set of all parameters in the model.

Although there are a few exceptions (see Davis, 1952, for instance) most applications of finite mixtures of parametric densities involve component densities of the same parametric type. In this case, $\theta_1,\ldots,\theta_k$ all belong to the same parameter space, Θ, say. We may then regard $\underline{\pi}$, as defining a probability distribution over Θ, and write

*Department of Statistics, University of Glasgow, Glasgow G12 8QW, Scotland

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

$$
\begin{aligned}
p(x|\underline{\psi}) &= \sum_{j=1}^{k} \pi_j f(x|\theta_j) \\
&= \int_{\Theta} f(x|\theta)\, dG_{\underline{\psi}}(\theta) \qquad (2) \\
&= E_{G_{\underline{\psi}}} f(x|\theta) \ ,
\end{aligned}
$$

where $G_{\underline{\psi}}(\cdot)$ denotes the probability measure on Θ defined by $\underline{\pi}$.

Finite mixtures correspond to finite discrete measures $G_{\underline{\psi}}(\cdot)$ and we shall be concentrating on these. The more general notation of (2) clearly suggests the generation of p.d.f.'s using more general probability measures on Θ. These may be called *general mixtures*. The formulation in (2) also clarifies the origin of the term *compound distribution*, which is sometimes used instead of mixture distribution. The distribution on S represented by $f(\cdot|\theta)$ is *compounded* with that on Θ given by $G_{\underline{\psi}}(\cdot)$. If, for instance, $f(\cdot|\theta)$ is a Poisson density, we obtain so-called *compound Poisson* distributions.

Another revealing feature of the basic p.d.f., as given in (1), is that mixture data can be regarded as *incomplete data*, in a certain sense. Suppose we have a pair of random variables (X,Y), where X has sample space S and Y is discrete, with sample space $\{1,\ldots,k\}$. Suppose also that the joint p.d.f. at $X = x$ and $Y = j$ is factorised as

$$
\begin{aligned}
p(x,j) &= p(j)p(x|j) \\
&= \pi_j f_j(x) \qquad (x \in S,\ j = 1,\ldots,k) \ .
\end{aligned}
$$

Then the mixture density (1) is the marginal p.d.f. for X. An observation from the mixture can therefore be regarded as a realization of (X,Y), but with the value of Y missing. As we shall see, not only does this interpretation have immediate meaning in many practical problems (in which we may have observations, each of which is known to come from one or other of a set of k source populations, but it is not known exactly which) but it also motivates some of the numerical methods required for parameter estimation, particularly with maximum likelihood (Section 4.4).

2. APPLICATIONS OF FINITE MIXTURE DISTRIBUTIONS

In this section we motivate the study of statistical methodology for dealing with data from mixture distributions by giving some indication of the number and variety of applications. These applications we shall divide into two categories to be called, somewhat arbitrarily, direct and indirect.

In direct applications there will be a belief in the existence, or possible existence, of k underlying sources from which the experimental unit generating X comes. The mixture model then appears directly, as built up in the final paragraph of Section 1. The following list of direct applications is therefore a list of applied fields. In each case there is a "physical" meaning for the sources or mixture components.

By an indirect application we mean a circumstance in which the mixture density is being used as a mathematical device, to facilitate the analysis in some way.

The following catalogue is intended to give a mere taste of the galaxy of applications that may be unearthed.

2.1 DIRECT APPLICATIONS

(i) Sedimentology. Samples of sand are often analyzed by measuring the frequency distribution of grain sizes. The sand may be known to be a (literal) mixture of several minerals. It is of interest to estimate the proportions of the different minerals in the sand. It may also be desired to estimate the grain size distributions for the different minerals, although these may already be "known" from extensive previous survey work.

(ii) Botany. In (i) above, if for "mineral type" we write "plant type" and for "sand grain size" we write "pollen grain size", "plant height" or "petal dimensions", then we account for a wealth of botanical applications.

(iii) Fisheries and marine biological research. Some characteristics of a fish are easy to measure once it has been landed. These include length but often do not include sex (only another fish can do this easily in some species!) or age. Data on, say, fish length, are often used for the estimation of sex proportions among a population of fish of the same age or of the age distribution of a mixture of several years' spawnings. Figure 1, taken from Hosmer (1973), shows a histogram of length data from a set of male and female

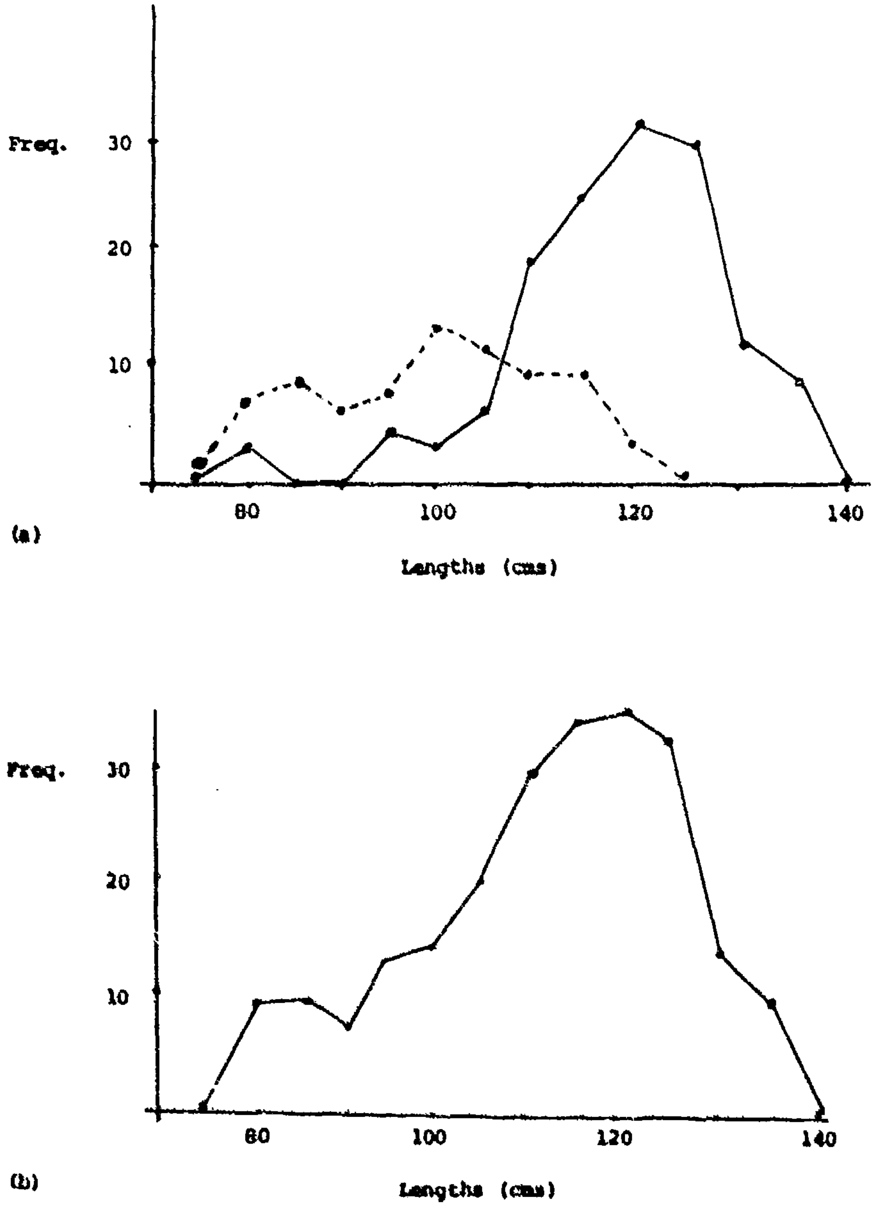


Figure 1. Frequency polygons for halibut lengths
(a) Males ------ ; Females ———
(b) All fish.

halibut. The separate male and female data are shown in *Figure 1(a)* and the mixture data in *Figure 1(b)*.

(iv) *Medicine*. Sometimes data may be available from clinical tests on a group of patients, each of whom is known to be suffering from one of two diseases. It is not,however, known exactly which disease is affecting each patient. A mixture model is sometimes used, with the particular aim of aiding diagnosis or prognosis; see Section 5.

(v) *Electrophoresis and gas chromatography*. Electrophoresis is used to estimate the relative concentrations of proteins in experimental samples and sometimes also to establish which proteins are actually present. *Figure 2*, adapted from Tiselius and Kabat (1939), shows a typical electrophoresis curve of concentration against the "migration position" achieved relative to a common initial position by the end of the experiment. Different proteins migrate at different rates, so the constituent proteins (in the example of *Figure 2* they are albumin and α-, β- and γ-globumin) may be identified. This differs from the other applications in that the "data" are themselves in the form of a smooth curve.

(vi) *Economics*. In one model for wage bargaining it is proposed (Quandt and Ramsey, 1978) that there are two possible phases, distinguished by some critical value of the cost of living index,characterized by two different regression models. In practice it may not be known, at any time at which data are gathered, which phase is in operation and this leads to a statistical model which is a "mixture" of the two regressions (*switching regressions*).

(vii) *Communications*. A sequence of messages is received, each one of which is either a signal or just noise. The proportion of signals and the signal and noise distributions may be of interest.

(viii) *Others*. Psychology, paleantology, geology, agriculture and zoology are a few of the many other fields of application.

2.2. INDIRECT APPLICATIONS.

(i) *Outlier models*. A mixture of $k = 2$ densities with one mixing weight close to one and the other close to zero is sometimes used to model outliers. The so-called *contaminated Normal* distributions form one such class. Their densities are of the form

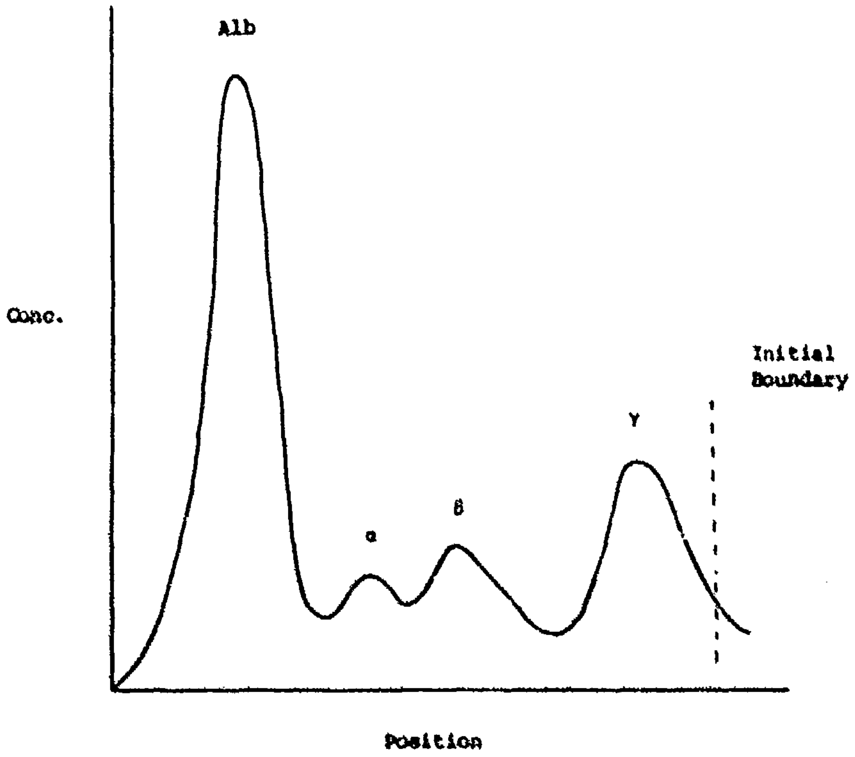


Figure 2. Electrophoresis curve from anti-egg albumin rabbit serum.

$$\pi_1\sigma_1^{-1}\phi((x - \mu_1)/\sigma_1) + (1 - \pi_1)\sigma_2^{-1}\phi((x - \mu_2)/\sigma_2) \ , \qquad (3)$$

where $\sigma_1, \sigma_2 > 0$, $\phi(u) = (2\pi)^{-1/2}\exp\{-\frac{1}{2}u^2\}$ and π_1, say, is close to one. For symmetric models $\mu_1 = \mu_2$ is imposed; see Barnett and Lewis (1978), Abraham and Box (1978).

(ii) *Heavy-tailed and multimodal densities*. The two-component Normal mixture (3), with $\mu_1 = \mu_2$, is one way of representing a symmetric heavy-tailed distribution. When the means are sufficiently well separated, relative to the variances, (3) represents a bimodal density; see Section 6.

(iii) *Cluster analysis and latent structure models*.

Multivariate mixture densities (Normal-based in particular) may be used as a basis for clustering techniques (Symons, 1981) and, in special cases, form latent structure models (Fielding, 1977). In the latter application the problem becomes that of finding a mixture model to fit the data. It is not essential that the components of the mixture that is chosen have meaning as physical sources, although some interpretation *may* be made, in the same spirit in which factors are interpreted in factor analysis.

(iv) *Nonparametric density estimation*. In the kernel method, a nonparametric estimate of a p.d.f. $f(\cdot)$ is obtained in the form

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} K((x - x_i)/h) \ .$$

Here h is a so-called smoothing parameter, $x_1, \ldots, x_n$ is a random sample from a population with the distribution which gives rise to $f(\cdot)$ and $K(\cdot)$, the kernel function, is itself a p.d.f.; see, for instance, Wegman (1972). The estimate $\hat{f}(\cdot)$ can obviously be described as an equally-weighted mixture of n component densities.

(v) *Modelling of prior densities*. Mixtures can provide rich families of conjugate prior densities in Bayesian analysis. If, for instance, each observation is distributed as $N(\theta, 1)$ and θ is given a Normal prior, then the posterior for θ is also Normal. The same conjugacy holds if the prior for θ is taken to be a k-component mixture of

Normals. If a "general" mixture is used we are led to hierarchical priors as in Lindley and Smith (1972).

(vi) Others. These include random number generation (Marsaglia, 1961), modelling of error distributions (Sorenson and Alspach, 1971), manifestation in empirical Bayes methods (Deely and Lindley, 1981), and as approximations to other distributions. Sometimes this last application is reversed. For instance, a lognormal density may be used to approximate to a skew mixture of two Normals; see also Smith and Naylor (1981).

3. SAMPLING STRUCTURES AND BASIC STATISTICAL PROBLEMS

Hosmer (1973) distinguishes, in a helpful way, among three sampling models. In the first one (M0) the data are realizations of n independent random variables distributed according to (1). The likelihood is, therefore,

$$L_0 = \prod_{i=1}^{n} \{ \sum_{j=1}^{k} \pi_j f_j(x_i) \} . \tag{4}$$

A more complicated likelihood, which results if the underlying sources for the observations follow Markov chain behaviour, has also been studied; see Baum et al. (1971) and Lindgren (1978).

Often supplementary data are available whose sources have been identified. If we denote these data by $\{x_{j\ell} : j = 1,\ldots,k,\ \ell = 1,\ldots,n_j\}$, then the new likelihood is

$$L_1 = L_0 \prod_{j=1}^{k} \{ \prod_{\ell=1}^{n_j} f_j(x_{j\ell}) \} .$$

This is model M1. It clearly provides extra data about the $\{f_j(\cdot)\}$. If the sampling rates of these categorized observations are equal to the mixing weights, then there is also more information about $\underline{\pi}$. This is model M2, for which the likelihood is

$$L_2 = L_1 \cdot \prod_{j=1}^{k} (\pi_j^{n_j}) .$$

Data of this type arise when a large set of data is available from the mixture and some of them, selected at random, have their sources identified by further "physical" examination.

The availability of some categorized ("complete") data usually boosts the statistical power of the data considerably, as Hosmer (1973) points out.

We shall consider in detail the treatment of three somewhat overlapping statistical problems.

3.1. Mixture decomposition.

Given data from M0, M1 or M2, to "estimate" the mixture density function. This will involve estimation of some or all of the following: the number of mixture components (for M0 data), the mixing weights and the component densities, or parameters thereof.

3.2. Discrimination (Pattern Recognition).

Given data from M0, M1 or M2, to use them for deriving discrimination procedures and to assess the worth of mixture data in this context.

3.3. Testing for the number of components.

Given data from M0, to find the model with the smallest number, k, of components but which is still compatible with the data. We may, for instance, wish to test whether the data come from a mixture of two univariate Normals as opposed to a single Normal. A related, but not equivalent, activity is that of testing for the modality of the p.d.f.

The rest of the paper discusses these objectives. Most of the space is devoted to mixture decomposition, on which there is the most voluminous literature. In general, the methodological principles that will be considered are very familiar and we shall be discussing what are just particular applications of these standard procedures. What makes the mixture problem special is that with many of the techniques there are snags, both theoretical and computational. We shall emphasize these particularly and point out that some of the complications remain unresolved.

4. MIXTURE DECOMPOSITION

Before launching into a catalogue of various estimation methodologies and their application to mixture data we go through some initial questions that have to be answered before calculations can begin.

4.1. *Preliminaries*

(i) *Which sampling structure is in operation: M0, M1, M2?* It is particularly important to decide correctly whether M1 or M2 obtains. With M0 data, estimation of the mixing weights is notoriously imprecise, so if the supplementary categorized data tell us more about $\underset{\sim}{\pi}$ it can be quite a bonus.

(ii) *What in the model is unknown?*

(a) *k*? This sets the problem up essentially as one in *cluster analysis*.

(b) *$\underset{\sim}{\pi}$ only*? In some problems, extensive previous experience may provide detailed knowledge about the component densities so that they may be treated as known. This occurs in some problems in sedimentology (Section 2.1) and in *remote sensing*, in which aerial photographs are analyzed to discover the relative concentrations of several crops in a geographical area.

(c) *$\{f_j(\cdot)\}$'s only*? Sometimes the mixing weights may be, for all practical purposes, known. A sex-ratio may sometimes be assumed to be unity, for instance.

(d) *$\underset{\sim}{\pi}$ and $\{f_j(\cdot)\}$'s*? Perhaps the most common case.

In cases (c) and (d) the $\{f_j(\cdot)\}$'s are unknown and we have the following dilemma.

(iii) *Can the $\{f_j(\cdot)\}$'s be assumed to have specified parametric forms, or not?* If the answer is yes then we may subsequently aspire that the parametric forms be simple ones, such as Normal!

(iv) *Is the class of mixtures we have chosen identifiable?* This is the first of the complications that may arise with mixture data, although it does not happen often in practical problems. For some classes of mixtures the members of the class are not uniquely defined and, if this is the case, estimation procedures are likely to run into difficulties. The main culprits are some discrete distributions on finite sample spaces and mixtures of uniform distributions.

(a) Consider mixtures of binomial distributions $B_1(N,\theta)$, with N known but θ variable. Then the class of k-component mixtures is identifiable only if $N \geq 2k - 1$ (Biischke, 1962).

(b) Let $U_x(a,b)$, as x varies, denote the p.d.f. for the uniform distribution on (a,b). Then the following two-component uniform mixture p.d.f.'s are identical

$$\alpha U_x(0,\alpha) + (1 - \alpha)U_x(\alpha,1), \quad \text{for all} \quad 0 < \alpha < 1 .$$

Theoretical work which reassures us that most classes of finite mixture densities of interest are identifiable is available in various papers, including Teicher (1963), Yakowitz and Spragins (1968) and Chandra (1977).

(v) Which method of estimation to use? The decision about what technique we shall use may well be based on our statistical philosophy but practical feasibility may also play a part, as we shall see. We now list, by subsection, methods that have been used for the mixture problem.

4.2. Graphical methods.

4.3. Method of moments.

4.4. Maximum likelihood.

4.5. Minimum distance methods.

4.6. Bayesian methods.

4.7. Sequential methods.

4.8. Curve fitting.

For purposes of illustration we shall restrict detailed attention to two simple examples.

Example 1. Mixture of two known densities.

$$p(x) = \pi_1 f_1(x) + (1 - \pi_1)f_2(x) \qquad (x \in S) , \qquad (5)$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are known and $0 \leq \pi_1 \leq 1$.

Example 2. Mixture of two univariate Normal densities.

The p.d.f. is given by (3), which we rewrite here, for convenience

$$p(x) = \pi_1\sigma_1^{-1}\phi((x - \mu_1)/\sigma_1) + (1 - \pi_1)\sigma_2^{-1}\phi((x - \mu_2)/\sigma_2) , \qquad (3)$$

where $\sigma_1 > 0$, $\sigma_2 > 0$ and $0 < \pi_1 < 1$.

As an indication of the flexibility of this as a model, we illustrate, in Figure 3, just 6 special examples.

4.2. Graphical methods

These have been used both in an exploratory way, for obtaining an informal assessment of the number, k, of components, along with quick, if crude, parameter estimates for subsequent numerical improvement, and also as the only method of analysis applied to the data. The latter was common in early work in applied fields and was stimulated to some extent by the numerical problems associated with the other methods.

The graphical methods are based on convenient plots, related to either the cumulative distribution function or the p.d.f. itself. The most familiar of the former is the use of Normal probability paper with Example 2. Figure 4 shows the theoretical plots for a particular Normal mixture and its components. The corresponding plot from a set of data can be used to assess whether the characteristic Normal mixture shape is apparent and, if so, to provide estimates of the means and variances (from the asymptotes) and for the mixing weight (roughly, from the point of inflexion); see Fowlkes (1979) for a useful survey and extension of this technique.

Other plots for Example 2 have been based on the p.d.f. and one of its data-based estimators, the histogram. First differences of the logarithms of the histogram frequencies give local approximations to the derivatives of the logarithms of the Normal component that is dominant at the given point. Furthermore, this derivative will be linear with negative slope which is inversely proportional to the variance of the dominant component. These facts form the basis of a graphical method of Bhattacharya (1967). The quadratic nature of the logarithm of a Normal p.d.f. also stimulated a semi-graphical

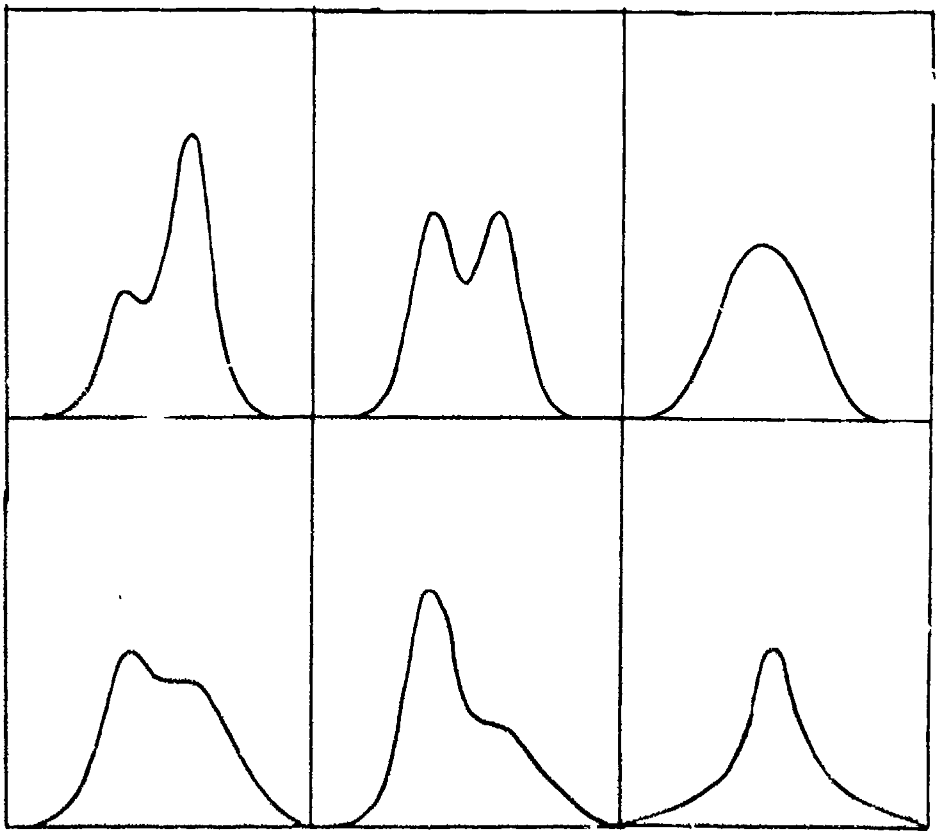
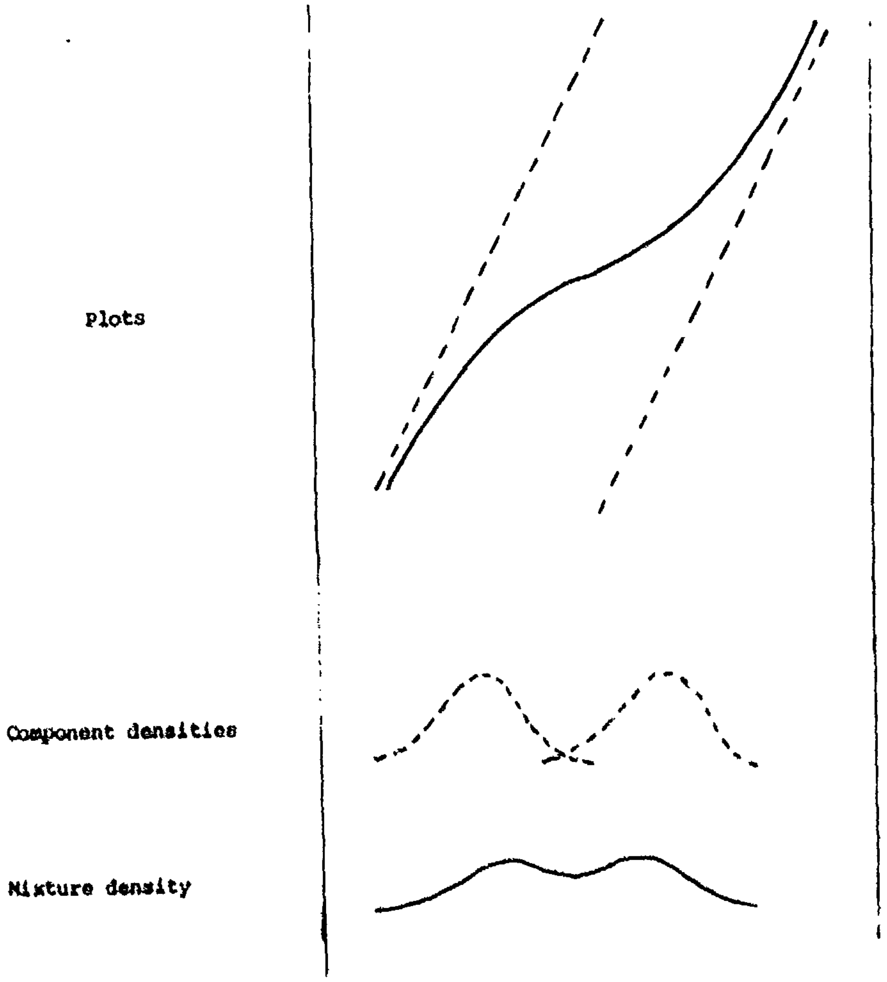


Figure 3. A selection of density functions for mixture of two univariate Normal densities.



<u>Figure 4.</u> Probability plots for a Normal mixture and its components.

method of Buchanan-Wollaston and Hodgson (1929). In general, success of the methods relies heavily on the mixture components being fairly well separated.

4.3. Method of moments

Suppose $\underline{\psi}$ contains s distinct parameters and that $m_1(X),\ldots,m_s(X)$ are s real-valued functions on the sample space such that their expected values exist as independent functions of $\underline{\psi}$. Write

$$\mu_j(\underline{\psi}) = \mathbb{E}m_j(X) , \qquad j = 1,\ldots,s .$$

Then, given $\underline{X} = (X_1,\ldots,X_n)$, a random sample of size n, a set of moment estimators for $\underline{\psi}$ can be obtained by solving

$$\underline{\mu}(\underline{\psi}) = \bar{\underline{m}}(\underline{x}) , \tag{6}$$

where $(\bar{\underline{m}})_j = n^{-1}\sum_{i=1}^{n} m_j(x_i), \qquad j = 1,\ldots,s .$

If the class of distributions under investigation is identifiable, consistent estimators of $\underline{\psi}$ are usually obtained, thanks to the laws of large numbers. Asymptotic Normality and the asymptotic covariance structure can usually be deduced from a first order Taylor Expansion of (6) into, approximately,

$$\underline{\mu}(\underline{\psi}) + D(\underline{\psi})(\hat{\underline{\psi}} - \underline{\psi}) = \bar{\underline{m}}(\underline{X}) ,$$

where $D(\cdot)$ denotes the matrix of first derivatives of $\underline{\mu}(\underline{\psi})$. Approximately, therefore,

$$\mathrm{cov}(\hat{\underline{\psi}}) = D(\underline{\psi})^{-1}\mathrm{cov}(\bar{\underline{m}})(D(\underline{\psi})^{T})^{-1} . \tag{7}$$

Although this is so far quite satisfying, problems do sometimes arise when attempts are made to solve the appropriate realization of (6). Firstly, explicit solution may not be possible and, secondly, there may be no solution in the parameter space, or more than one.

Example 1. Mixture of two known densities.

With only one unknown parameter, π_1, only one moment equation is required. Furthermore, the equation will be linear in π_1, giving

$$\hat{\pi}_1 = (\bar{m}_1 - \mu_{12})/(\mu_{11} - \mu_{12}) , \tag{8}$$

where $\mu_{1j} = \int m_1(x)f_j(x)dx$, $j = 1,2$. It is easy to check that $\hat{\pi}_1$ is unbiased for π_1

and

$$\mathrm{var}(\hat{\pi}_1) = \mathrm{var}(\bar{m}_1)/(\mu_{11} - \mu_{12})^2 . \tag{9}$$

Unfortunately there is no guarantee, except asymptotically, that $0 < \hat{\pi}_1 < 1$, although in this simple example this may not have great practical import. In principle, study of the right hand side of (9) may suggest a function $m_1(\cdot)$ for which $\mathrm{var}(\hat{\pi}_1)$ is small, or even minimal and which would therefore give an "optimal" moment estimator. Although achievement of this requires knowledge of π_1 itself, some practical guideline may well be possible in many examples.

The usual power moments are commonly used in (8), or in (6) for that matter. Another possibility in this example is to use an indicator function for $m_1(\cdot)$. That is, take

$$\begin{aligned} m_1(X) &= 1 \quad \text{if} \quad X < c \\ &= 0 \quad \text{otherwise} . \end{aligned}$$

Then $\bar{m}_1$ is the proportion of observations in the sample $< c$: see Johnson (1973) and James (1978).

Example 2. Mixture of two univariate Normals.

Possibly the earliest systematic look at mixtures was the application of the method of moments to this example by Pearson (1894) in a study of forehead measurements of a set of male and female crabs. We now have five parameters and Pearson used moment equations for the first five central moments. After a certain amount of elimination of variables the computation problem reduces (!?) to that of finding a negative root of a ninth degree polynomial, solution of which was no mean feat in the 1890's! Back-substitution then provides the parameter estimates. Sometimes, however, the nonic has no negative root and sometimes more than one. This is awkward, although in the former case a single Normal is often an adequate model and in the latter, either solution is usually satisfactory. Just how often these and other complications arise has been investigated by Bowman and Shenton (1973).

The number of papers that are directly derivative of that of Pearson (1894) runs into dozens, with many applications and modifications of the method of solution. For the special case of $\sigma_1 = \sigma_2$, the "nonic" is replaced by a cubic (Cohen, 1967) and a neat graphical aid for this case is given by Preston (1953). The bivariate case is mentioned by Charlier and Wicksell (1924) and the multivariate by Day (1969) and John (1970).

Mixtures of the other simple parametric distributions have also been given the method-of-moments treatment. Several of them, in which the component densities are one-parameter p.d.f.'s, lead to a general pattern in which there is a set of moment equations of the form

$$\sum_{j=1}^{k} \theta_j^{s-1} \pi_j = c_s, \qquad s = 1,\ldots,2k , \tag{10}$$

where θ_j is the (scalar) parameter associated with the jth component density, $c_0 = 1$ and the other $\{c_s\}$ are data-based.

They include mixtures of exponentials (based on ordinary power moments), binomials and negative binomials (both based on weighted factorial moments, with θ as the "success probability"), Poissons (factorial moments), one-parameter Weibulls and one-parameter gammas (both based on weighted power moments), and a generalized method of moments due to Kabir (1968). For an illustration of the standard method of solution of equations like (10), see Blischke (1964).

As in <u>Example 1</u>, (7) may, in principle, be used to select "optimal" moments for use in more general problems. Tallis and Light (1968) discuss the choice of fractional power moments so as to minimize det cov($\underset{\sim}{\hat{\theta}}$), as given by (7), for a mixture of two exponentials.

4.4. <u>Maximum likelihood</u>

For a given parametric mixture model, the method of maximum likelihood is available. That there are difficulties is immediately apparent if we look at the MO likelihood given by equation (4). Almost certainly the order statistic (in the case of univariate continuous data) will be minimal sufficient and explicit MLE's will not be available. Numerical optimization will be necessary although, in many cases, maximum likelihood analysis of the "complete" categorized version of the data may be very easy, as would be the case for both our special examples.

<u>Example 1</u>

$$L_0 = L_0(\pi_1) = \prod_{i=1}^{n} \{\pi_1(f_{i1} - f_{i2}) + 1\}$$

where $f_{ij} = f_j(x_i)$, $i = 1,\ldots,n$, $j = 1,2$. Thus

$$\partial \log L_0/\partial\pi_1 = \sum_{i=1}^{n} (f_{i1} - f_{i2})/p(x_i) \tag{11}$$

and

$$\partial^2 \log L_0/\partial\pi_1^2 = - \sum_{i=1}^{n} (f_{i1} - f_{i2})^2/p(x_i)^2 \ . \tag{12}$$

Although we see, from (11), that the likelihood equation is a polynomial equation of degree up to $(n - 1)$ in π_1, equation (12) shows that $\log L_0$ is strictly concave in π_1, so that there is at most one real root, π_1^*, of (11) and it gives a global maximum of L_0. It is easy to check whether $0 \leq \pi_1^* \leq 1$ and thus determine the maximum likelihood, $\hat{\pi}_1$, say. Peters and Coberly (1976) generalize this to a version of this example with more than two components.

Even with this simple problem, however, there is a complication, which arises in the asymptotic theory of maximum likelihood. It is fairly easy to discover that, if the <u>true</u> value of π_1 is 1 then, asymptotically, $\hat{\pi}_1 = 1$ with probability 1/2. Thus $\hat{\pi}_1$ is not asymptotically Normal, although it is consistent. The standard theory fails because the true π_1 is on the boundary of the parameter space.

<u>Example 2</u>

$$L_0 = L_0(\underline{\psi}) = \prod_{i=1}^{n} (\pi_1\sigma_1^{-1}\phi((x_i - \mu_1)/\sigma_1) + (1 - \pi_1)\sigma_2^{-1}\phi((x_i - \mu_2)/\sigma_2)) \ .$$

The two-component univariate Normal mixture is by far the most commonly researched or applied case and yet its likelihood surface is a potential disaster area. It is riddled with singularities. If we set, say, $\mu_1 = x_1$, then it is easy to see that, as $\sigma_1 \to 0$, $L_0 \to \infty$. Furthermore, there are many reported cases of weird features on the likelihood surfaces, quite apart from the problem of singularities: see for instance the related <u>Figure 1</u> of Hartigan (1977). They include multiple maxima, unusual troughs and unusual behaviour at the boundary of the parameter space. In spite of this, the method of maximum likelihood <u>is</u> used in practice for this problem and Kiefer (1978) has even established the existence of a local maximum of L_0 for which the usual asymptotic theory holds. To some extent the difficulties are lessened if there are supplementary categorized data of if the parameter space is restricted, by demanding that $\sigma_1 = \sigma_2$, for instance. As far as the computation of maximum likelihood estimates is concerned, we may employ traditional numerical methods, of which the most familiar to statisticians are the method of Newton Raphson and its derivative, the method of scoring, both of which calculate, automatically, via a Hessian matrix, an estimate of the asymptotic covariance matrix. Boes (1967) discusses a one-stage method of scoring for Example 1. If the initial estimator is consistent, then the first iterate is Best Asymptotically Normal.

It is also possible to exploit our interpretation of mixture data as being "incomplete" and use a version of the EM (Expectation-Maximization) algorithm of Dempster et al. (1977). The algorithm generates a sequence of estimates $\{\underline{\psi}^{(r)}\}$ of $\underline{\psi}$ for which the corresponding sequence of likelihoods is monotonic increasing. Although it can be slow to converge, the algorithm is usually very easy to program. Many of its manifestations, including those related to mixture problems, appeared in much earlier papers as appealing successive-approximations procedures, without the general structure or simple proof of monotonicity being spotted. In Section 1 we interpreted the observed mixture data $\underline{x}$ (M0 data) as originating from a complete data-set

$$\{(x_1,y_1),\ldots,(x_n,y_n)\} = (\underline{x},\underline{y}) ,$$

but with the source identifiers $y_1,\ldots,y_n$ missing. The two-step iterative stage of the EM algorithm is as follows, in which g denotes the p.d.f. for the complete data. We suppose that parameter estimates $\{\underline{\psi}^{(r)}\}$ are currently available, to be improved upon to give $\{\underline{\psi}^{(r+1)}\}$. Hopefully, as $r \to \infty$, $\underline{\psi}^{(r)} \to \hat{\underline{\psi}}$.

E-step: Evaluate $E\{\log g(\underline{x},\underline{y}|\underline{\psi})|\underline{x},\underline{\psi}^{(r)}\} = Q(\underline{\psi},\underline{\psi}^{(r)})$, say.

M-step: Find $\underline{\psi} = \underline{\psi}^{(r+1)}$ to maximize $Q(\underline{\psi},\underline{\psi}^{(r)})$.

Details of the general EM algorithm for finite mixtures are given Section 4.3 of Dempster et al. (1977), where it is found more convenient to express the source identifiers in terms of indicator vectors. Here we show the appealing forms for our two examples.

Example 1.

E-step: Given $\pi_1^{(r)}$, let $w_{ij}^{(r)} = \pi_j^{(r)} f_{ij}/p^{(r)}(x_i)$, $i = 1,\ldots,n$, $j = 1,2$, where $\pi_2^{(r)} = 1 - \pi_1^{(r)}$ and $p^{(r)}(x) = \pi_1^{(r)} f_1(x) + (1 - \pi_1^{(r)}) f_2(x)$.

M-step: $\pi_1^{(r+1)} = n^{-1} \sum_{i=1}^{n} w_{i1}^{(r)}$.

Note how, in the E-step, the n observations are "allocated" to the two components by fractions which are current estimates of predictive probabilities. That is

$$w_{ij}^{(r)} = \mathrm{Prob}(y_i = j | x_i, \pi_1^{(r)}) \ .$$

In the M-step, $\pi_1^{(r+1)}$ is obtained as a "relative frequency" based on aggregating these fractions. The categorized-data version would have all w_{ij}'s as zero or unity.

Example 2.

E-step: Given $\underline{\psi}^{(r)} = (\pi_1^{(r)}, \mu_1^{(r)}, \sigma_1^{(r)}, \mu_2^{(r)}, \sigma_2^{(r)})$, let

$$w_{ij}^{(r)} = \pi_j^{(r)} f_{ij}^{(r)}/p^{(r)}(x_i), \quad i = 1,\ldots,n, \ j = 1,2$$

where $p^{(r)}(x_i) = \sum_{j=1}^{2} \pi_j^{(r)} f_{ij}^{(r)}$, $i = 1,\ldots,n$, and

$$f_{ij}^{(r)} = (\sigma_j^{(r)})^{-1} \phi((x_i - \mu_j^{(r)})/\sigma_j^{(r)}), \text{ for each } i,j \ .$$

Again the $\{w_{ij}^{(r)}\}$ are current predictive probabilities.

M-step: For $j = 1,2$,

$$\pi_j^{(r+1)} = n^{-1} \sum_{i=1}^{n} w_{ij}^{(r)} \ ,$$

$$\mu_j^{(r+1)} = \sum_{i=1}^{n} w_{ij}^{(r)} x_i / \sum_{i=1}^{n} w_{ij}^{(r)} \ ,$$

and

$$(\sigma_j^2)^{(r+1)} = \sum_{i=1}^{n} w_{ij}^{(r)} (x_i - \mu_j^{(r+1)})^2 / \sum_{i=1}^{n} w_{ij}^{(r)} \ .$$

Note the similarity of M-step to the calculations for fully-categorized data.

Similar simple recursions are available for mixtures of other parametric distributions such exponentials, Poissons and their generalization, the exponential family. Wolfe (1970) and Day (1969) give the EM algorithm for multivariate Normal mixtures, Skene (1978) that for latent class analysis, Hartley (1978) that for the switching-regressions model and Baum et al. (1971) that for the Markov chain case referred to in Section 3.

Many other families of mixtures have had their maximum likelihood methodology dealt with by this or other algorithms. They include binomials and others (Hasselblad, 1969), truncated exponentials (Mendenhall and Hader, 1958), uniforms (Gupta and Miyawaki, 1978), von Mises (Mardia and Sutton, 1975), logistics (Anderson, 1979) and even the compound Poisson distribution (Simar, 1976).

A related approach is the so-called "cluster analysis" method. For Example 2 this amounts to the following. Consider all 2^n partitions of the data into two clusters. For each partition, maximize the likelihood and choose that partition and corresponding parameter estimates which give a global maximum. Symons (1981) emphasizes that the major usefulness of this method and its multivariate version is in cluster construction as opposed to parameter estimation in which obvious biases occur. In the univariate Normals case, Example 2, the optimal partition corresponds to some cut-off value c, say, such that all $x_i < c$ go into one component and the rest into the other. That the resulting variance estimates, say, are biased is quite clear.

4.5 Minimum distance estimation

A wide variety of estimation procedures may be envisaged which can be interpreted informally as the minimisation of

$$\delta \text{ (data, theoretical distribution)}$$

over the second argument, where δ is some measure of difference or distance. More formally, we may choose $\underline{\psi}$ to minimize

$$\delta(\tilde{F},F_{\underline{\psi}}) ,$$

where $F_{\underline{\psi}}$ is the theoretical cumulative distribution function and $\tilde{F}$ is some data-based version, the most natural being the empirical distribution function. All sorts of δ may be chosen, some of them metrics, some not, and indeed the previously mentioned methods of moments and maximum likelihood can be described in these terms. The latter corresponds to the Kullback-Leibler directed divergence

$$\delta_{KL}(F,G) = \int \log(dF(x)/dG(x))dF(x) ,$$

where $dF(x)/dG(x)$ is a ratio of "densities". Given data of type M0, the part of $\delta_{KL}(\tilde{F},F_{\underline{\psi}})$ depending on $\underline{\psi}$ is

$$- \int \log(p(x|\underline{\psi}))d\tilde{F}(x) = - \sum_{i=1}^{n} \log p(x_i|\underline{\psi}) = - \log L_0 .$$

Other special versions are

$$\delta_C(F,G) = \int \{f(x) - g(x)\}^2 g(x)^{-1}dx$$

and

$$\delta_{MC}(F,G) = \delta_C(G,F).$$

Discrete versions of these give the methods of minimum chi-squared and minimum modified chi-squared, the former of which was used by Fryer and Robertson (1972) for Normal mixtures using grouped data.

The quadratic distance function

$$\delta_Q(F,G) = \int \{F(x) - G(x)\}^2 dF(x)$$

is useful, particularly for our Example 1.

Example 1.

$$\delta_Q(\tilde{F},F_{\underline{\pi}}) = n^{-1} \sum_{i=1}^{n} \left(\sum_{j=1}^{2} \pi_j F_j(x) - i/n\right)^2 ,$$

where $F_j(\cdot)$ is the cumulative distribution function from $f_j(\cdot)$. We have to minimize, therefore, a quadratic function of π_1,π_2, subject to $\pi_1 + \pi_2 = 1$, $\pi_1 \geq 0$, $\pi_2 \geq 0$. If the nonnegativity constraints are ignored, explicit solution is possible for $\underline{\pi}$; see Macdonald and Pitcher (1979), for instance.

When explicit solution is not possible, numerical solution is required. A first order Taylor Expansion of the stationarity equations can be made the basis for asymptotic results, as in the method of moments or maximum likelihood. In particular, asymptotic covariance matrices may be obtained.

A modification of the basic technique is to minimize a distance measure between, not $\tilde{F}$ and $F_{\underline{\psi}}$, but $\tilde{\phi}_u$ and $\phi_u(\underline{\psi})$, say, where $\phi_u(\underline{\psi})$ is some transform of $F_{\underline{\psi}}$, with auxiliary variable u, and $\tilde{\phi}_u$ is the empirical version. The distance measure depends on u, clearly. One approach is to impose a weighting measure, W(u) on the range of u and to minimize

$$\Delta_W(\underline{\psi}) = \int \delta(\tilde{\phi}_u, \phi_u(\underline{\psi}))dW(u) .$$

Quandt and Ramsey (1978) use this method with

(i) δ quadratic;

(ii) $W(\cdot)$ a measure with finite support;

(iii) $\phi_u(\underline{\psi})$ the moment generating function.

They apply the technique to Normal mixtures and switching regressions. Kumar et al. (1979) use the characteristic funciton with a continuous measure for $W(\cdot)$. So far, little has been said about the obvious problem of choosing an "optimal" measure $W(\cdot)$, as far as the asymptctic covariance matrix, $\text{cov}(\underline{\hat{\psi}})$, say, is concerned. It corresponds to the choice of optimal moment equations in Section 4.3.

A slightly different use of distance functions is that of Hall (1981), for estimating mixing weights when there are date available from the mixture, providing empirical c.d.f. $\tilde{F}$, and from the k component distributions, giving empirical c.d.f.'s $\tilde{F}_1,\dots,\tilde{F}_k$. The $\hat{\pi}_1,\dots,\hat{\pi}_k$ are chosen to minimize

$$\delta(\tilde{F}, {}_j\sum \pi_j \tilde{F}_j) .$$

As in the treatment of Example 1 above, the use of a quadratic δ gives explicit minimization, if the nonnegativity constraints are ignored. For this essentially nonparametric technique, Hall (1981) derives asymptotic theory. Titterington (1983) looks at versions for discrete and smoothed continuous data.

4.6. Bayesian method

There is usually a strong similarity between the relative ease that is possible with likelihood inference and Bayesian methods. In principle the Bayesian approach promises to be the more amenable with mixture data. In practice we run into difficulty again, as illustrated below with M0 data.

Example 1

$$L_0 = \prod_{i=1}^{n} (\pi_1 f_{i1} + (1 - \pi_1) f_{i2}) = \sum_{2^n \text{ terms}} g(\underline{x}, \underline{y}) , \qquad (13)$$

where the summation is over all possible $\underline{y}$. L_0, therefore, is the sum of 2^n likelihoods each of which corresponds to categorized data. If categorized data are easy to deal with in Bayesian analysis (in other words, if there is a convenient, conjugate family of priors) then the same will be true for mixture data. In this example, if a Beta prior is available for π_1, then the posterior density for π_1 will be that of a calculable mixture of Betas. Unfortunately, the number of mixture components is 2^n, which quickly becomes large with n. If the number of mixture components were k, we would end up with a k^n-component mixture for the posterior for π_1.

Example 2

Here the natural prior structure is to have π_1, (μ_1,σ_1^2) and (μ_2,σ_2^2) mutually independent. Then "as usual" choose a Beta prior for π_1 and a Normal/inverse Gamma prior for each of (μ_1,σ_1^2) and (μ_2,σ_2^2). Again exact results may be written down in terms of 2^n-component mixtures for joint and marginal posterior p.d.f.'s.

Various ways of coping with this computational and storage problem have been considered.

(i) If only posterior expected values are of interest, use numerical integration based on (13). This may not, however, be very helpful in some circumstances. If, for instance, a posterior density is multimodal, then the posterior mean may be an unhelpful index of location. Numerical integration may however be the way to calculate predictive densities, as given by

$$q(z) = E\, p(z|\underline{\psi}) = \int p(z|\underline{\psi})t(\underline{\psi}|\underline{x})d\underline{\psi} \ ,$$

where $t(\cdot|\underline{x})$ denotes the posterior p.d.f. for $\underline{\psi}$.

(ii) Neglect terms in the posterior which are known to be small. When a contamination model is used for outliers, π_1 is considered to be close to 1. Only those terms in L_0 with small powers of $(1 - \pi_1)$ are retained and the posterior p.d.f. is renormalized appropriately (Box and Tiao, 1968, Abraham and Box, 1978).

(iii) Select a (comparatively) small number of the 2^n terms at random, evaluate them and renormalize (Leonard, 1982).

(iv) If only the predictive density, say, is of interest and not the parameters, $\underline{\psi}$, themselves, replace the mixture density by another with similar characteristics but which is more amenable to practical Bayesian analysis (Smith and Naylor, 1981).

(v) Use an approximate method based on sequential incorporation of the data (Section 4.7).

A Bayesian version of the "cluster analysis" approach (see end of Section 4.4) is discussed by Binder (1978).

4.7. Sequential methods

There is an important class of methods in which the data, $x_1,\ldots,x_n$, are treated sequentially and which lead to ways of decomposing the mixture approximately. Many of the procedures, particularly related to Example 1 (known component densities), were developed in the electrical engineering literature and, consequently, introduce a new jargon. The decomposition problem itself is called that of unsupervised learning, in that we have to process NO data without being told the whole story, namely, the identities of the sources. In the engineering context, the sequential nature of the analysis serves the need to process, on-line, data which become available sequentially. When such methods have been developed in the statistical literature there has also been the principle of trying to obviate the computational difficulties implicit in maximum likelihood and Bayesian analysis, as we shall see. We shall use Example 1 to illustrate four procedures.

(i) Decision directed (DD).

(ii) Learning with a probabilistic teacher (PT).

(ii) Quasi maximum likelihood (QML).

(iv) Quasi Bayes (QB).

Example 1

Suppose, after r observations have been dealt with, the "current" estimate of π_1 is $\pi_1^{(r)}$. For the next observation, x_{r+1}, we evaluate (cf. Section 4.4) weights

$$w_1^{(r+1)} = \pi_1^{(r)} f_1(x_{r+1})/p^{(r)}(x_{r+1}) ,$$

and

$$w_2^{(r+1)} = 1 - w_1^{(r+1)} .$$

These weights have possible application (see Section 5) in the classification of x_{r+1} into one or other of the two component populations. The procedures develop quite naturally from this interpretation, particularly the first three.

DD: Assign observation $r + 1$ to component 1 (resp. 2) if $w_1^{(r+1)} >$ (resp. $<$) $w_2^{(r+1)}$.

PT: With probability $w_j^{(r+1)}$, assign observation $r + 1$ to component j, $j = 1,2$.

QML: Assign a "fraction" $w_j^{(r+1)}$ of observation $(r + 1)$ to component j, $j = 1,2$.

This leads to the following recursive algorithms, stated here in forms which fit in with comments later on.

DD: If

$$w_1^{(r+1)} > w_2^{(r+1)}, \quad \pi_1^{(r+1)} = \pi_1^{(r)} - (r + 1)^{-1}(\pi_1^{(r)} - 1) \qquad (14)$$

If

$$w_1^{(r+1} < w_2^{(r+1)}, \quad \pi_1^{(r+1)} = \pi_1^{(r)} - (r + 1)^{-1}\pi_1^{(r)} \qquad (15)$$

PT: With probability $w_1^{(r+1)}$, (14) holds; otherwise, (15) holds.

QML: For $j = 1,2$, $\pi_j^{(r+1)} = \pi_j^{(r)} - (r +1)^{-1}(\pi_j^{(r)} - w_j^{(r+1)})$. (16)

In the QB approach the rationale is to maintain a Beta density for π_1 at each stage and a recursion is set up on the mean, for which we use the notation $\pi_1^{(r)}$. If, at stage r, $\pi_1 \sim Be(\alpha_r,\beta_r)$, so that $\pi_1^{(r)} = \alpha_r/(\alpha_r + \beta_r)$, then the distribution of π_1 at stage $r + 1$ ought to be a mixture of a $Be(\alpha_r + 1,\beta_r)$ and a $Be(\alpha_r,\beta_r + 1)$. Instead, we approximate to this mixture by a single Beta, with parameters $\alpha_r + w_1^{(r+1)}$ and $\beta_r + w_2^{(r+1)}$, with $(w_1^{(r+1)},w_2^{(r+1)})$ defined in terms of $\pi_1^{(r)}$ exactly as above. We obtain

$$\pi_j^{(r+1)} = \pi_j^{(r)} - (\alpha_r + \beta_r + 1)^{-1}(\pi_j^{(r)} - w_j^{(r)}), \qquad j = 1,2 .$$

Obviously the results will depend on the order in which the data are incorporated but the on-line facility may over-ride this criticism. The important theoretical question is whether convergence can be guaranteed of $\pi_1^{(r)}$ to the true π_1 as $r \to \infty$ $(n \to \infty)$. The recursions (14) - (17) have been written in forms which suggest that the key will lie in the theory of stochastic approximations (Wasan, 1964). For the DD method it is known that, sometimes, the sequence $\{\pi_1^{(r)}\}$ may "runaway" to a value other than the true π_1. For the other methods, consistency can be established. Similar sequential procedures may be set up for more complicated mixtures (Smith and Makov, 1978), Titterington, 1976, Titterington and Jiang, 1981). A useful survey is provided by Makov (1980).

4.8. Curve fitting.

So far we have given no indication of how to analyze (exactly the right word here!) the electrophoresis curve of Figure 2. Here the data are themsleves a smooth curve. In electrophoretic practice informal methods are sometimes used for estimating the relative concentrations of the proteins. The area under the curve is divided up in as fair a way as possible and the sub-areas are measured using a gadget called a planimeter.

For a more formal analysis, minimum distance methods may be used (Section 4.5) and a modified type of Fourier analysis is also available, thanks largely to Medgyessy (1977).

This approach is stimulated by the following obvious statement about curves like the p.d.f. corresponding to Example 2. Suppose we let

$$p(x|\underline{\psi},\lambda) = \pi_1\sigma_{1\lambda}^{-1}\phi((x - \mu_1)/\sigma_{1\lambda}) + (1 - \pi_1)\sigma_{2\lambda}^{-1}\phi((x - \mu_2)/\sigma_{2\lambda}) , \qquad (18)$$

where $\sigma_{j\lambda}^2 = \sigma_j^2 - \lambda$, $j = 1,2$ and $0 < \lambda < \min(\sigma_1^2,\sigma_2^2)$.

As λ increases from zero, the mixture becomes more and more clearly bimodal and the parameters become easier and easier to estimate from the curve. By operating mathematically in a specified way on the datum curve it is indeed possible to draw data-based versions of (18) and, thence, to decompose the mixture. Medgyessy (1977) gives details for both continuous and discrete data. Stanat (1968) gives multivariate versions. Gregor (1969) applies the procedure to histogram data and Tarter and Silvers (1975) decompose bivariate Normal mixtures in a rather similar manner.

5. DISCRIMINANT ANALYSIS

In usual discriminant analysis there are training sets of categorized observations from k sources. From these data a procedure is developed for assessing the possible source of a further observation, x, say, the source of which is unknown. Our questions here are whether further uncategorized data can be built into the discrimination procedure and whether the discriminatory performance is improved thereby. In the limiting case only uncategorized data are available from the start (M0 data). (In practice it may be expensive or, in some medical contexts, dangerc.s to obtain enough information to fully categorize an experimental unit. If therefore uncategorized data are useful as such, this could be very welcome.)

Whether or not uncategorized data are useful at all depends critically on the model chosen for the joint probability density

$$p(x,y)$$

of x and the source identifier y. We may write either

$$p(x,y) = p(x)p(y|x) \qquad (D)$$

or

$$p(x,y) = p(x|y)p(y) \qquad (S)$$

in which (D) recognises the diagnostic paradigm and (S) the sampling paradigm of Dawid (1976).

In discriminant analysis we are interested in using the training data to tell us about $p(y|x)$. If a parametric version of (D) is set up such that the parameters associated with the two factors on the right hand side are distinct, then no amount of data on uncategorized data give any information at all about $p(y|x)$. If (S) is used similarly, however and we obtain, by Bayes Theorem,

$$p(y|x) = p(y)p(x|y)/p(x) ,$$

where $p(x)$ is a mixture density, as in Section 1, then the uncategorized data will affect the discrimination procedure and its performance. In particular, as the amount of uncategorized data available increases, $p(y|x)$ should be estimated consistently.

Discriminant rules based on estimated likelihood ratios will tend to the optimal rule (in terms of misclassification rates, that is).

Example 2 (Restriction: $\sigma_1 = \sigma_2 = \sigma$)

In this case the likelihood ratio rule can be written in terms of a discriminant function that is linear in x and depends on the unknown parameters (Lachenbruch, 1975). These parameters may be estimated from mixture data, with or without supplementary categorized data sets, using, for instance, the EM algorithm of Section 4.4. Performance may be assessed either empirically or by considering the asymptotic expected rate of misclassification. O'Neill (1978) and Ganesalingam and McLachlan (1978), in almost simultaneous publications, showed that the mixture data can help in this context, although the two Normal components have to be rather well separated for the effect to be substantial. Let $\Delta = |\mu_1 - \mu_2|/\sigma$ and suppose $\pi_1 = 1/2$. Then, relative to the case in which all data are categorized, the asymptotic efficiencies for M0 data and for M2 data with 50% categorized data are, respectively, 10% and 50% (for $\Delta = 2$); 65% and 83% (for $\Delta = 4$).

Empirical evidence of the gains from an approximate Bayesian version of the multivariate Normal case is given by Titterington (1976) and Anderson (1979) combines the paradigms (D) and (S) by parametrising according to the factorization (S) and then analysing the data by logistic methods, which are diagnostic in spirit. Silverman (1978) estimates likelihood ratio's nonparametrically using data some of which are uncategorized.

It is clearly disturbing that the two parametrisations based on (D) and (S) lead to qualitatively different results in the present context. If (D) is used wrongly then potentially useful information is not being used; if (S) is used wrongly then the bonus the mixture data appear to offer is misleading. It makes it important to find the right model in any given application and, needless to say, it has led to considerable controversy about principle.

6. HYPOTHESIS TESTING AND MULTIMODALITY

In Section 4 we mentioned cluster analysis as a means of establishing the number of components present in a mixture when we only have uncategorized data. Alternatively, we may use likelihood criteria with added penalties for the number of parameters involved (Akaike, 1973, Schwarz, 1978). Another possibility is to seek the mixture with fewest components which is still compatible with the data. In particular, we may want to ask whether there really is a mixture or whether there is just a single underlying component. This could be just the sort of question we want to ask in practice. We seem to be on well-trodden ground, if parametric models may be assumed, because the problem can be formulated as one of testing between two nested hypotheses, for which the generalized likelihood ratio test is available. However, we soon hit snags.

<u>Example 2</u>

H0: Single Normal.

H1: Mixture of two Normals.

We would hope to evaluate the usual "$2 \log \lambda$" test statistic and refer its value to a percentile of a χ^2 distribution with some number, ν, of degrees of freedom. What, however, should ν be? In most problems, ν is obtained as the number of constraints required to reduce H1 to H0. We may obtain this reduction here, however, by either: (i) $\pi_1 = 1$ (1 constraint), or (ii) $\mu_1 = \mu_2$, $\sigma_1 = \sigma_2$ (2 constraints).

Should we take $\nu = 1$ or $\nu = 2$ or maybe some intermediate value, as conjectured by Hartigan (1977)?

Should we even be trying to use the χ^2 table at all?

<u>Example 1</u>

H0: $\pi_1 = 1$

H1: $0 < \pi_1 < 1$.

Here asymptotically, under H0, the maximum likelihood estimator for π_1 in the H1 model is equal to 1 with probability 1/2 (Section 4.4). Thus $2 \log \lambda$ is zero, with probability 1/2, and therefore certainly not χ^2.

The problem is that the regularity conditions required for the asymptotic theory do not hold (c.f. Section 4.4). Under H0, the true value of π_1 lies on the boundary of its parameter space. In Example 2, H0 also corresponds to a region on the boundary of the parameter space for H1, a region in which identifiability fails.

So far, the treatment that has been developed for this important difficulty is far from satisfactory. Until recently, in many applications the χ^2 approximation has been used without the awkwardness about degrees of freedom being detected. For Example 2 and multivariate versions thereof, some simulations have been carried out in attempts to concoct a number of degrees of freedom to use in the χ^2 table; see Wolfe (1972), Aitkin et al. (1981) and Everitt and Hand (1981). Hardly any theoretical work has been reported. Davies (1977) mentions, but does not work through in detail, the use of a union-intersection principle for one special example.

Alternative test procedures are themselves somewhat unsatisfactory. Engelman and Hartigan (1969) use, as test statistic for Example 2, the estimated Mahalanobis distance corresponding to the optimal "maximum likelihood" clustering of the data into two components (end of Section 4.4). Also for Example 2, omnibus tests of Normality could be used.

A final possibility is to look at the degree of multimodality represented by the data. Of course, unimodality of a density is not equivalent to its corresponding to a single component density. Indeed, a symmetric mixture of two univariate Normals ($\pi_1 = 1/2$, $\sigma_1 = \sigma_2 = \sigma$) is only bimodal if $|\mu_1 - \mu_2| > 2\sigma$. The study of bimodal and multimodal densities is, however, of some interest and the two component Normal mixture is a convenient model for a bimodal density. (An alternative one with one fewer parameter is the quartic exponential density; see Mats, 1978.) Many papers, especially in fields of application, talk specifically about multimodality. What they are usually interested in, however, is the possible presence of a mixture (Murphy, 1964). What is possible, however, is to use a significance test against unimodality as a conservative test against the hypothesis of a one-component distribution. At least the asymptotic theory will not cause such problems.

Silverman (1981) has developed a technique, based on simulation and nonparametric density estimation, for assessing the modality of a data-set.

7. CLOSING REMARKS

It is hoped that we have done justice to the variety of applications and special problems that arise with distribution mixtures and that it is clear that the thorniest problems await satisfactory solution. The field is very much alive and, if anything, the publication rate on this topic is higher than ever.

We have not been able to give many details of analysis, nor even to provide anything like a full list of references. Several papers have been written which contain survey or bibliographic material; see Blischke (1963), Clark (1976), Macdonald and Pitcher (1979), Odell and Basu (1976) and Murray and Titterington (1978). Further reference may be made to sporadic sections in the quartet of books by Johnson and Kotz (1969-72), to Chapter 4 of Ord (1972) and to the recent monograph by Everitt and Hand (1981). The present contribution arose from work towards a forthcoming book by Makov et al. (1982) where, it is hoped, the missing details and references will be fully documented. In particular, a much fuller account of the sequential methods of Section 4.7 will be provided.

REFERENCES

Abraham, B. and Box, G.E.P. (1978). Linear models and spurious observations. Appl. Statist., 27, 131-138.

Aitkin, M. et al. (1981). J. R. Statist. Soc. A, 144, to appear.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory. (B. N. Petrov and F. Czaki, eds.), 267-281. Budapest: Akademiai Kiado.

Anderson, J. A. (1979). Multivariate logistic compounds. Biometrika, 66, 17-26.

Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data. New York: Wiley.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains. Ann. Math. Statist., 41, 164-171.

Bhattacharya, C. G. (1967). A simple method of resolution of a distribution into Gaussian components. Biometrics, 23, 115-135.

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

Blischke, W. R. (1962). Moment estimators for the parameters of a mixture of two binomial distributions. Ann. Math. Statist., 33, 444-454.

Blischke, W. R. (1963). Mixtures of discrete distributions. Proc. Int. Symp. on Classical and Contagious Disc. Distns. (G. P. Patil, Ed.) Pergamon.

Blischke, W. R. (1964). Estimating the parameters of mixtures of binomial distributions. J. Amer. Statist. Assoc., 59, 510-528.

Boes, D. C. (1967). Minimax unbiased estimator of mixing distribution for finite mixtures. Sankhya A, 29, 417-420.

Bowman, K. O. and Shenton, L. R. (1973). Space of solutions for a normal mixture. Biometrika, 60, 629-636.

Box, G. E. P. and Tiao, G. C. (1968). A Bayesian approach to some outlier problems. Biometrika, 55, 119-129.

Buchanan-Wollaston, H. G. and Hodgson, W. C. (1929). A new method for treating frequency curves in fishery statistics, with some results. J. Conservation, 4, 207-235.

Chandra, S. (1977). On the mixtures of probability distributions. *Scand. J. Statist., 4*, 105-112.

Charlier, C. V. L. and Wicksell, S. D. (1924). On the dissection of frequency functions. *Arkiv for Matematik, Astronomi och Fysik*, Bd. 18, No. 6.

Clark, M. W. (1976). In *Mathematical Geology*, Vol. *4*.

Cohen, A. C. (1967). Estimation in mixtures of two normal distributions. *Technometrics, 9*, 15-28.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika, 64*, 247-254.

Davis, D. J. (1952). An analysis of some failure data. *J. Amer. Statist. Assoc., 47*, 113-150

Dawid, A. P. (1976). Properties of diagnostic data distributions. *Biometrics, 32*, 647-658.

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

Deely, J. J. and Lindley, D. R. (1981). Bayes Empirical Bayes. *J. Amer. Statist. Assoc., 76*, 833-841.

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

Engelman, L. and Hartigan, J. A. (1969). Percentage points of a test for clusters. *J. Amer. Statist. Assoc., 64*, 1647-1648.

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

Fielding, A. (1977). Latent structure models. *Exploring Data Structures* (C. A. O'Muircheartaigh and C. Payne, eds.) 125-157. New York: Wiley.

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

Fryer, J. G. and Robertson, C. A. (1972). A comparison of some methods for estimating mixed normal distributions. *Biometrika, 59*, 639-648.

Gangesalingam, S. and McLachlan, G. J. (1978). The efficiency of a linear discriminant function. Biometrika, 65, 658-662.

Gregor, J. (1969). An algorithm for the decomposition of a distribution into Gaussian components. Biometrika, 25, 79-93.

Gupta, A. K. and Miyawaki, T. (1978). On a uniform mixture model. Biometrical J., 20, 631-638

Hall, P. (1981). On the nonparametric estimation of mixture proportions. J. R. Statist. Soc. B, 43, 147-156

Hartigan, J. A. (1977). Distribution problems in clustering. Classification and Clustering (J. van Ryzin, Ed.) 45-71, Academic Press.

Hartley, H. J. (1978). Comment on "Estimating mixtures of normal distributions and switching regressions." J. Amer. Statist. Assoc., 73, 738-741.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. J. Amer. Statist. Assoc., 64, 1459-1471.

Hosmer, D. W. (1973. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics, 29, 761-770.

James, I. R. (1978). Estimation of mixture proportions in a mixture of two normal distributions from simple, rapid measurements. Biometrics, 34, 265-275.

John, S. (1970). On identifying the population of origin of each observation in a mixture of observations from two Normal populations. Technometrics, 14, 553-563.

Johnson, N. L. (1973). Some simple tests of mixtures with symmetrical components. Commun. Statist., 1, 17-25.

Johnson, N. L. and Kotz, S. (1969-1972). Distributions in Statistics (4 volumes). New York: Wiley.

Kabir, A. B. M. L. (1968). Estimation of parameters of a finite mixture of distributions. J. R. Statist. Soc. B, 30, 472-482.

Kiefer, N. M. (1978). Discrete parameter variation: Efficient estimation of a switching regression model. Econometrica, 46, 427-434.

Kumar, K. D., Nickin, E. H. and Paulson, A. S. (1979). Comment on "Estimating mixtures of normal distributions and switching regressions." *J. Amer. Statist. Assoc, 74*, 52-55.

Lachenbruch, P. A. (1975). *Discriminant Analysis*. New York: Hafner.

Leonard, T. (1982). Bayes estimation of a multivariate density. *Tech. Rep.*, Math. Res. Center, U. Wis-Madison.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scand. J. Statistics, 5*, 81-91.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Statist. Soc. B, 34*, 1-41.

MacDonald, P. D. M. and Pitcher, T. J. (1979). Age-group from size-frequency data: A versatile and efficient method of analyzing distribution mixtures. *J. Fish. Res. Board of Canada, 36*, 987-1001.

Makov, U. E. (1980). Approximations of unsupervised Bayes learning procedures. *Bayesian Statistics* (J. M. Bernardo et al., Eds.) 69-81. Valencia: Univ. Press.

Makov, U. E., Smith, A. F. M. and Titterington, D. M. (1982). *Statistical Methods for Finite Mixture Distributions*. In preparation.

Mardia, K. V. and Sutton, T. W. (1975). On the modes of a mixture of two von Mises distributions. *Biometrika, 62*, 699-701.

Marsaglia, G. (1961). Expressing a random variable in terms of uniform random variables. *Ann. Math. Statist., 32*, 894-898.

Matz, A. W. (1978). Maximum likelihood parameter estimation for the quartic exponential distribution. *Technometrics, 20*, 475-484.

Medgyessy, P. (1977). *Decomposition of Functions*.

Mendenhall, W. and Hader, R. J. (1958). Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika, 45*, 504-520.

Murphy, E. A. (1969). One cause? Many causes? The argument from the bimodal distribution. *J. Chron. Dis., 17*, 301-324.

Murray, G. D. and Titterington, D. M. (1978). Estimation problems with data from a mixture. *Appl. Statist., 27*, 325-334.

Odell, P. L. and Basu, J. P. (1976). Concerning several methods for estimating crop acreages using remotely sensed data. *Commun. Statist. A, 5*, 1091-1104.

O'Neill, T. J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Assoc., 73*, 821-826.

Ord, J. K. (1972). *Families of Frequency Distributions*. Griffin.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. A*, 71-110.

Peters, C. and Coberly, W. A. (1976). The numerical evaluation of the maximum-likelihood estimate of mixture proportions. *Commun. Statist. A, 5*, 1127-1135.

Preston, E. J. (1953). A graphical method for the analysis of statistical distributions into two normal components. *Biometrika, 40*, 460-464.

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

Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Statist., 6*, 461-464.

Silverman, B. W. (1978). Density ratios, empirical likelihood and cot death. *Appl. Statist., 27*, 26-33.

Silverman, B. W. (1981). Using kernel density estimation to investigate multimodality. *J. R. Statist. Soc. B, 43*, 97-99.

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Am. Statist., 4*, 1200-1209.

Skene, A. M. (1978). Discrimination using latent structure models. *Compstat 1978*, 199-204. Vienna: Physica Verlag.

Smith, A. F. M. and Makov, U. E. (1978). A quasi-Bayes sequential procedure for mixtures. *J. R. Statist. Soc. B, 40*, 106-112.

Smith, A. F. M. and Naylor, J. (1981). *Submitted for publication*.

Sorenson, H. W. and Alspach, D. L. (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica, 7*, 465-479.

Stanat, D. F. (1968). Unsupervised learning of mixtures of probability functions. Pattern Recognition (L. Kanal, Ed.) 357-384. Washington: Thompson.

Symons, M. J. (1981). Clustering criteria and multivariate normal mixtures. Biometrics, 37, 35-43.

Tallis, G. M. and Light, R. (1968). The use of fractional moments for estimating the parameters of a mixed exponential distribution. Technometrics, 10, 161-175.

Tarter, M. and Silvers, A. (1975). Implementation and applications of bivariate Gaussian mixture decomposition. J. Amer. Statist. Assoc., 70, 47-55.

Teicher, H. (1963). Identifiability of finite mixtures. Ann. Math. Statist., 34, 1265-1269.

Tiselius, A. and Kabat, E. A. (1939). An electrophoretic study of immune sera and purified antibody preparations. J. Exper. Med., 69, 119-131.

Titterington, D. M. (1976). Updating a diagnostic system using unconfirmed cases. Appl. Statist., 25, 238-247.

Titterington, D. M. (1983). Minimum distance nonparametric estimation of mixture proportions. J. R. Statist. Soc. B, to appear.

Titterington, D. M. and Jiang, J-M. (1981). A sequential version of the EM algorithm. Submitted for publication.

Wasan, M. (1969). Stochastic Approximation. Cambridge University Press.

Wegman, E. G. (1972). Nonparametric probability density estimation: I. A summary of available methods. Technometrics, 14, 533-546.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. Multivar. Behav. Res., 5, 329-350.

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, U. S. Navy Pers. Trg. Res. Lab., San Diego.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. Ann. Math. Statist., 39, 209-214.

DMT/ed

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 2369 MRC-TSR-	2. GOVT ACCESSION NO. AD-A116175	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) SOME PROBLEMS WITH DATA FROM FINITE MIXTURE DISTRIBUTIONS		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) D. M. Titterington		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - (Statistics and Probability)
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE April 1982
		13. NUMBER OF PAGES 39
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

finite mixture distributions, incomplete data, method of moments, graphical methods, maximum likelihood, Bayesian analysis, minimum distance, curve fitting, discrimination, likelihood ratio test, multimodality

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Finite mixture distributions arise in many statistical applications. After the basic definition of mixture distributions, many of these applications are listed, sampling models are proposed and the basic statistical problems are described. More detailed study is then made of the use of the familiar statistical methodologies in mixture decomposition, of the incorporation of mixture data into discrimination procedures and of the problems that arise in hypothesis testing.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)