



Statistical Computing in Mixture Models

By MARY L. LESPERANCE and BRUCE G. LINDSAY

Technical Report #01-07-13

2001

Center for Likelihood Studies
DEPARTMENT OF STATISTICS
THE PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PA 16802

Contents

1	Introduction	2
1.1	The mixture model	2
1.2	Example	4
1.3	Resources	5
1.4	Other Sampling Schemes	6
1.5	Number of Components	7
1.6	Identifiability	8
1.7	Chapter Overview	9
2	Fitting mixture models, m fixed	11
2.1	The Method of Moments	11
2.1.1	The method and asymptotic properties	11
2.1.2	Examples	13
2.2	Maximum Likelihood Estimation	14
2.2.1	The likelihood	14
2.2.2	Asymptotic properties of MLE's	15
2.2.3	Numerical properties of MLE's	15
2.2.4	The main algorithmic choices	17

2.2.5	Computing the MLE's: Newton-Raphson	18
2.2.6	Computing the MLE's: EM algorithm	20
2.2.7	Stopping criteria	25
2.2.8	Starting values	27
2.2.9	Information Matrix	30
2.3	Bayesian methods	31
3	Fitting Mixture Models, m unknown	32
3.1	The nonparametric maximum likelihood estimator	32
3.1.1	Geometric characterization	32
3.1.2	Example	36
3.1.3	A gradient characterization	36
3.1.4	Computational Strategies.	40
3.1.5	Stopping criteria	41
3.1.6	Algorithms - EM	42
3.1.7	Algorithms - gradient based	43
3.1.8	A dual problem, semi-infinite programming problem . .	43
4	References	44

Statistical Computing in Mixture Models

Mary L. Lesperance and Bruce G. Lindsay

1 Introduction

1.1 The mixture model

The term *mixture model* encompasses many types of statistical structures including random and mixed effects models, empirical Bayes, latent class and trait models, models for overdispersion, measurement error and nuisance parameters, clustering and deconvolution. Mathematically, we denote a *finite* mixture model as follows. Let X be a random variable or vector taking on values in a sample space, \mathcal{X} , with probability density function (or probability mass function)

$$g(x; \beta, \theta, \pi) = \pi_1 f(x; \beta, \theta_1) + \pi_2 f_2(x; \beta, \theta_2) + \dots + \pi_m f(x; \beta, \theta_k). \quad (1)$$

Such a model can arise if one is sampling from a heterogeneous population that can be decomposed into m distinct homogeneous subpopulations,

called *component populations*. If these components have been “mixed” together, and we measure only the variable X without determining the component identities, then this model holds. The *component weight* $\pi_j \geq 0$ represents the proportion of the total population in the j -th component. These parameters are therefore constrained so that $\sum \pi_j = 1$. The *component densities*, $f(x; \beta, \theta_j)$ represent the densities or probability mass functions that one would use for sampling from the homogeneous components, and so are often presumed to have a nice parametric form. They depend on a parameter θ_j that describes the component population characteristics, but may also depend on a parameter β that is common to all components.

We can generalize (1) by rewriting it as an integral where G is a discrete probability distribution that puts mass π_j at the support point θ_j , so that

$$f(x; \beta, G) = \int f(x; \beta, \theta) dG(\theta). \quad (2)$$

A natural extension is to allow G to be a continuous distribution, as for example, in the study of latent class models or random effects models. We can think of this model as a missing data model by letting the complete data be (Θ_i, X_i) , where the unobserved variable Θ_i is sometimes called a *latent*

variable. The term G is sometimes called the *mixing distribution*; we will call it the *latent distribution*. The densities defined in (1) and (2) are called the *mixture densities*.

1.2 Example

As an example, consider a random sample of black widow spiders and let X represent their length. We assume that given their gender, the lengths are normally distributed with mean and variance (θ_F, σ^2) for the females and (θ_M, σ^2) for the males, with $\theta_F > \theta_M$. The component weights, $\pi_F, \pi_M = 1 - \pi_F$, represent the population proportion of spiders that are female and male respectively. If we sample their lengths without identifying their gender, the density for X is a mixture of the two normal distributions. The latent distribution is discrete, with masses (π_F, π_M) at (θ_F, θ_M) .

To illustrate problems that can arise in estimating the parameters in the mixture model, suppose that the two components have equal weights of $1/2$, and $\theta_F = \theta_M + 4\sigma$. *Figure 1* shows the curves of the component densities as dashed lines and the mixed density as a solid line. There is a clear separation between the two components, and one would expect to obtain a great deal of information about gender given length. Suppose now that $\theta_F = \theta_M + 2\sigma$.

Figure 2 shows the component and mixed densities. The population mixed density is unimodal, and there is considerably less information about gender given length. In this case, algorithms used to fit the two component mixture model will have difficulty converging or will be extremely slow to converge.

1.3 Resources

The subject of mixture models is particularly challenging, and therefore interesting, because much of the theory lies outside the traditional domain of parametric statistical modelling. One must be very cautious about applying asymptotic distributional results because the model has a high degree of irregularity. In addition, one must be cautious about applying standard numerical routines for methods such as maximum likelihood.

Although we lay out a number of the important features of the mixture model here, one should also consult some of the book-length treatments of the model found in Titterington (1985), McLachlan & Basford (1988), Lindsay (1995, pp. 108-136), Böhning (1999) and McLachlan & Peel (2000).

The world of software is constantly evolving, but hardly perfected yet.

At the time of writing, there is a website maintained by David Dowe

[http : //www.csse.monash.edu.au/~dld/mixture.modelling.page.html](http://www.csse.monash.edu.au/~dld/mixture.modelling.page.html)

that has an extensive listing of many resources for mixture modelling.

1.4 Other Sampling Schemes

In the spider example we assumed that we had data only on the spider lengths. The information in the experiment could be augmented by additional sampling of the following two types:

- (a) We may take separate samples from the male and female component populations and measure length in these groups. (This provides no new information about the relative proportion of males and females.)
- (b) The gender categorized lengths in (a) may be obtained from a random sample from the population, in which case we gain additional information about the relative frequencies of the two components.

Clearly we have successively more information about the parameters $\theta_F, \theta_M, \sigma$ and then π_F , as we move from (a) to (b) above. The additional

information can be easily incorporated into a maximum likelihood analysis as will be shown in the next section. (Lindsay 1995; Titterington et al. 1985)

1.5 Number of Components

In many examples, the number of components, m , is unknown, in which case a mixture model with the fewest components that describes the data well is sought. In effect, one is asking, “What is the minimal number of components we can clearly identify?” The resulting statistical problem is related to cluster analysis.

Inference concerning the number of components is not a standard parametric hypothesis testing problem. To see this, consider testing the null hypothesis of *one* normal component versus *two* where ϕ is the standard normal density function:

$$H_0 : g(x) = \phi(x; \mu, \sigma)$$

$$H_1 : g(x) = \pi\phi(x; \mu_1, \sigma_1) + (1 - \pi)\phi(x; \mu_2, \sigma_2).$$

The null model can be obtained from the alternative model in two ways: by setting $\pi = 0$ or 1 or letting $(\mu_1 = \mu_2, \sigma_1 = \sigma_2)$, with π any value. Thus,

although it appears that the null model is simply nested within the alternate, it is not, and the standard chi-squared asymptotic distribution theory is not appropriate when using the likelihood ratio test. This has ramifications for interval estimation of the number of components m as well. It has been shown (Donoho, 1988) that one cannot set an upper confidence limit on the true value of m , although lower confidence bounds are possible. Curiously, it does appear possible to consistently estimate m nonetheless (Chen & Kalbfleisch, 1996). Although consistency is clearly desirable for an m estimator, the lack of a confidence interval implies that we should be very cautious in its interpretation. Intuitively, although one might have a good estimator of the number of components that were in the actual sample, there could be many “unseen” components with very small values of π that we were unable to detect because of a small sample size.

1.6 Identifiability

Identifiability questions, which can affect estimation procedures, are also an important aspect of mixture models. For example, consider the case of two

component mixtures of *binomial* $(2, \theta)$ distributions:

$$g(0) = \pi(1 - \theta_1)^2 + (1 - \pi)(1 - \theta_2)^2$$

$$g(1) = 2\pi\theta_1(1 - \theta_1) + 2(1 - \pi)\theta_2(1 - \theta_2).$$

The third equation, that for $g(2)$, is superfluous because $g(0) + g(1) + g(2) = 1$.

The representation of $g(x)$ as a mixture of the two binomials in terms of $(\pi, \theta_1, \theta_2)$ is *not unique*; given the probabilities $g(0)$ and $g(1)$, we have two equations and three unknowns. Estimation procedures for the parameters in a mixture model will not be well defined without identifiability. This is especially problematic with discrete data. Fortunately many mixtures of continuous densities are identifiable. (Maritz & Lwin, 1989; McLachlan & Peel, 2000; Prakasa Rao, 1992, pp. 183-228; Titterton et al. 1985).

1.7 Chapter Overview

In the remaining sections of this chapter, we review techniques for estimating parameters and standard errors for those parameters in the mixture model (2) where we have a random sample of independent observations. The standard analyses of the mixture model involve making one of three assumptions:

- There are a fixed and known number of components m , so that the unknown parameters are the π 's, θ 's and β .
- The distribution G is continuous, from a known parametric family.
- The distribution G is unknown. In this “nonparametric” case, we can consider estimating G either with a continuous density function or a discrete distribution.

We consider the first and last cases here.

There is a vast literature on mixture models and their variants, and many methods have been used for estimation. Much of this literature has been reviewed in the aforementioned books. In *Section 2*, we review some methods for finite mixture models where the number of components m is fixed, including the method of moments, maximum likelihood and Bayesian methods. Methods for the case when m is not fixed are discussed in *Section 3*.

2 Fitting mixture models, m fixed

In this section, we assume that the number of components is known and fixed equal to m . The latent distribution G therefore has support size m .

2.1 The Method of Moments

2.1.1 The method and asymptotic properties

Karl Pearson (1894) first used the method of moments to fit a mixture of two univariate normal components to crab measurements. The method does not always yield numerically simple estimators for the parameters. However, simple solutions have been obtained for several models.

Let ψ represent the vector of all parameters, π 's, θ 's, and β 's. Let x_1, x_2, \dots, x_n represent a random sample from $f(x; \beta, G)$. In the *method of moments*, we equate a set of theoretical moments, which are functions of the parameters ψ , to the sample moments and solve for the unknown parameters. That is, if we let $\mu(\psi)$ be the vector of theoretical moments and m be the vector of corresponding sample moments, we solve

$$\mu(\hat{\psi}) = m \tag{3}$$

for $\hat{\psi}$. If we have κ functionally independent unknown parameters, we then

need κ moment equations. Explicit solution of (3) for ψ may be difficult, and the solution need not be unique. Lindsay (1989, 1995) describes a class of exponential family models with moment estimators that are well-defined and straightforward to calculate. Lindsay & Basak (1991, 1993) demonstrate the use of moment methods for multivariate normal mixtures.

Consistency of $\hat{\psi}$ typically follows from the laws of large numbers, however the estimator may not be efficient. For large samples, the covariance matrix of $\hat{\psi}$ is approximately $D(\hat{\psi})^{-1}\text{cov}(m; \hat{\psi})[D^T(\hat{\psi})]^{-1}$ where $D = \partial\mu/\partial\psi$, and $\text{cov}(m; \hat{\psi})$ is the covariance matrix of m under the assumption that $\psi = \hat{\psi}$. (Titterington et al. 1985, p. 71)

Moment methods have a long history of application from the days before there were computers fast enough to perform maximum likelihood computations. Even so, the method of moments can still be useful for determining good initial values for algorithmic methods such as those used to compute maximum likelihood estimates, for simulations, bootstrapping and computing diagnostics. We return to their use as starting values later.

2.1.2 Examples

Titterton et al. (1985) give examples of using the method of moments to fit mixtures of two known densities, two known exponentials, two univariate normals and two multivariate normals, and a general class of k component mixtures. They also provide references for other examples. See also McLachlan & Peel (2000).

As an example, consider a two component mixture of binomial random variables with index k and known success probabilities θ_j , $j = 1, 2$, so that

$$g(x; \theta, \pi) = \pi \binom{k}{x} \theta_1^x (1 - \theta_1)^{k-x} + (1 - \pi) \binom{k}{x} \theta_2^x (1 - \theta_2)^{k-x}.$$

To estimate π , we equate the first population moment, $E(X) = \pi k \theta_1 + (1 - \pi) k \theta_2$ to the first sample moment, \bar{x} , and solve for π :

$$\bar{x} = \tilde{\pi} k (\theta_1 - \theta_2) + k \theta_2.$$

Solving for $\tilde{\pi}$ yields $\tilde{\pi} = \frac{\bar{x} - k \theta_2}{k(\theta_1 - \theta_2)}$. Note that this estimator of π is not guaranteed to lie between zero and one!

2.2 Maximum Likelihood Estimation

2.2.1 The likelihood

We consider the three sampling schemes mentioned in Section 1 and the likelihoods arising from each. If none of the observations have been categorized, then the likelihood has the form:

$$L_o(\psi) = \prod_{i=1}^n f(x_i; \beta, G) = \prod_{i=1}^n \left[\sum_{j=1}^m \pi_j f(x_i; \beta, \theta_j) \right]$$

In case (a) of Section 1, we may have observations known to arise from a given category. Let observations categorized to the j 'th component be $\{x_{jh}, j = 1, \dots, m, h = 1, \dots, n_j\}$, then the contribution to the likelihood for these observations is:

$$L_*(\beta, \theta) = \prod_{j=1}^m \prod_{h=1}^{n_j} f(x_{jh}; \beta, \theta_j).$$

If the categorized observations can be assumed to be a random sample from the populations with incidence rates π_1, \dots, π_m the likelihood is augmented by:

$$L_{**}(\pi) = \prod_{j=1}^m \pi_j^{n_j}.$$

The likelihood, $L(\psi) = L_0(\psi)L_*(\beta, \theta)L_{**}(\pi)$ where only the applicable terms are included. With Maximum Likelihood Estimation, we usually obtain a root, $\hat{\psi}$, of the *Likelihood equations*, $\partial L(\psi)/\partial\psi = 0$, which corresponds to a local maxima of the $L(\psi)$ in the interior of the parameter space.

2.2.2 Asymptotic properties of MLE's

Redner & Walker (1984) show that under identifiability and regularity conditions, and given that the Fisher information, $\mathcal{I}(\psi_0) = E\{[\partial \log L(\psi)/\partial\psi][\partial \log L(\psi)/\partial\psi]^T\}$, evaluated at the true ψ_0 , is well defined and positive definite, asymptotically the probability goes to one that there is a unique solution $\hat{\psi}_n$ of the likelihood equations in a certain small neighbourhood of ψ_0 . They also show that the usual asymptotic distribution of the maximum likelihood estimator (MLE) holds as $n \rightarrow \infty$, that is

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{\mathcal{L}} N[0, \mathcal{I}(\psi_0)^{-1}].$$

2.2.3 Numerical properties of MLE's

In computing the MLE's, the parameters π are constrained so that $\sum_{j=1}^m \pi_j = 1$, $\pi_j \geq 0$. One can use Lagrange multipliers to incorporate the constraint

or eliminate one parameter by setting $\pi_m = 1 - \pi_1 - \dots - \pi_{m-1}$.

It is possible there does not exist a maximum likelihood solution with the desired number of components. That is, there exists an m^* component mixture, with $m^* < m$, that has greater likelihood than any m component mixture. Any attempt to fit m will cause the algorithmic methods to move from the initial values to the parameter space boundaries in one of two ways. First, the maximum likelihood estimates for some of the π'_j s may become zero. Secondly, some of the neighboring component parameters θ_j can become equal.

Care must be taken in computing ML estimates because the likelihood can be high dimensional and is known to be multimodal. In some cases, it is not only multimodal, but unbounded, and so there is no maximum likelihood estimator. For example, consider a two component mixture of normal densities where we have a random sample of unclassified observations. The log-likelihood is:

$$\mathcal{L}(\psi) = \mathcal{L}(\pi, \mu_1, \mu_2, \sigma_1, \sigma_2) = \sum_{i=1}^n \ln[\pi \phi(x_i; \mu_1, \sigma_1) + (1 - \pi) \phi(x_i; \mu_2, \sigma_2)].$$

If we set $\mu_1 = x_1$, and let $\sigma_1 \rightarrow 0$, then the log-likelihood tends to infinity.

The likelihood is riddled with global maxima, one at each observation x_i .

In many applications, the likelihood equations may have multiple roots corresponding to local maxima. If the likelihood is bounded, one might wish to choose the root which yields the largest local maxima. However, this would require finding all the roots with certainty, or at least with high probability. An alternative that leads to consistent efficient likelihood estimators is to use consistent but inefficient estimators for starting values; see Lehmann (1983, Chapter 6.3) for the theory. One possibility, explored by Furman & Lindsay (1994), is to use moment estimators for this purpose.

2.2.4 The main algorithmic choices

There are three main issues that must be addressed in constructing an algorithm for maximum likelihood estimation. We will deal with them over the following subsections.

- **The selection of initial values.** In a multimodal likelihood, an algorithm tends to go to a root nearest the initial value. Thus it is wise to either search over the space of initial values or to use starting values known to have good properties.
- **The algorithm.** In a multimodal likelihood, a fast algorithm also

tends to be an unreliable one. Here we present a fast algorithm, Newton Raphson, and a slow one, EM. The literature has many papers trying to find algorithms with an ideal blend of reliability and speed. Although the speed of computers is increasing, there are still limitations in using the EM algorithm with bootstrapping and Monte Carlo simulation

- **The stopping rule.** In a fast algorithm, there is no difficulty with ad hoc stopping criterion based on the changes in the parameters between iterations. In slow algorithms, however, these changes can be small even if the algorithm has a long ways to go.

2.2.5 Computing the MLE's: Newton-Raphson

Newton-Raphson or other quasi-Newton methods may be used to compute the maximum likelihood estimates. The generalized Newton-Raphson iterations are at the $(c+1)$ 'st step:

$$\psi^{(c+1)} = \psi^{(c)} - \alpha_c [D^2 \mathcal{L}(\psi^{(c)})]^{-1} D \mathcal{L}(\psi^{(c)})$$

where D and D^2 represent the first and second differentiation operators with respect to ψ , and α_c is a tuning parameter that is set to 1 in the conventional

Newton-Raphson. Note that matrix inversion is required at each iteration and there is no guarantee that the log-likelihood will increase at each iteration.

Finch et al. (1989) investigated the behaviour of a quasi-Newton method for computing the maximum likelihood estimates of the parameters in a two-component mixture of unknown normals. They found that more than 8 percent of the starting values resulted in a failure to converge in 750 iterations, and 25 percent of the starting points resulted in a solution that was not the global maximum.

In the mixture model we recommend that, at least in the initial algorithm, one check for increases in the likelihood at every step. If the algorithm gives $\psi^{(c+1)}$ that leaves the parameter space or does not increase the likelihood over $\psi^{(c)}$, a subroutine could be used to adjust the tuning parameter, say by sequentially halving its value, until a likelihood increase is found without leaving the parameter space.

If the Newton-Raphson iterations converge, they do so quickly, since the algorithm is of second order (or “quadratically convergent”, Fletcher 1987 p.20). If the standard asymptotic theory holds for the problem, then the Newton-Raphson iterations yield an estimator of the covariance matrix of $\hat{\psi}$,

$$-[D^2\mathcal{L}(\psi^{(c)})]^{-1}.$$

2.2.6 Computing the MLE's: EM algorithm

The EM algorithm is a method that is commonly used to compute maximum likelihood estimates in the mixture model with fixed number of components (Dempster et al., 1977; McLachlan & Krishnan, 1997). It has the advantages over the Newton-Raphson algorithm that it is often very easy to program and very reliable.

Suppose that we have a random sample of fully categorized data. This is the idealized *complete data*. The i 'th observation of a fully categorized random sample, can be represented as $y_i = (x_i, z_i)$, where z_i is a “multinomial indicator vector” of length m . That is, if the i -th observation came from the j -th component, then z_i has a 1 in the j -th position and zeroes elsewhere. The indicator vectors are the ideal ways to represent the component information because the complete data likelihood can then be written in product form as:

$$L_c(\psi; y) = \prod_{i=1}^n \prod_{j=1}^m \pi^{z_{ij}} f(x_i; \beta, \theta_j)^{z_{ij}}$$

with logarithm

$$\mathcal{L}(\psi; y) = \sum_{i=1}^n z_i^T V(\pi) + \sum_{i=1}^n z_i^T U_i(\beta, \theta), \quad (4)$$

where $V(\pi)^T = (\ln \pi_1, \dots, \ln \pi_m)$ and $U_i(\beta, \theta)^T = [\ln f(x_i; \beta, \theta_1), \dots, \ln f(x_i; \beta, \theta_m)]$.

In the standard mixture setting, the indicator vectors z are not observed, and so the data are ‘incomplete.’ The EM algorithm is a clever device that exploits the simplicity of the optimization problem for the complete data to obtain an algorithm for incomplete data.

In the E-step of the EM algorithm, we compute the expected value of the complete log-likelihood, $\mathcal{L}(\psi; y)$, given the *observed* data x and the current value of the parameter, $\psi^{(c)}$:

$$\mathcal{L}_{em}(\psi; \psi^{(c)}, x) = E[\mathcal{L}(\psi; y); x, \psi^{(c)}].$$

At the M-step in a cycle, we maximize $\mathcal{L}_{em}(\psi; \psi^{(c)}, x)$ with respect to ψ to find the value $\psi^{(c+1)}$. The E and M steps are iterated until convergence is obtained. The incomplete likelihood is increased at each step, that is,

$$L(\psi^{(c+1)}; x) \geq L(\psi^{(c)}; x).$$

The form of $\mathcal{L}_{em}(\psi; \psi^{(c)}, x)$ is particularly interesting in the case of mixture models. Given x_i , the indicator vectors z are the only random quantities, and they enter the likelihood linearly. It is straightforward to calculate

$$E[z_{ij}; x, \psi^{(c)}] = E[z_{ij}; x_i, \psi^{(c)}] = \pi_j^{(c)} f(x_i; \theta_j^{(c)}) / f(x_i; G^{(c)}) := w_{ij}(\psi^{(c)}).$$

This term can be interpreted as the posterior probability that the component is j given that the observation was x_i , assuming that $G^{(c)}$ is the “prior” distribution. The EM log-likelihood is then

$$\mathcal{L}_{em}(\psi; \psi^{(c)}, x) = \sum_{i=1}^n w_i^T(\psi^{(c)}) V(\pi) + \sum_{i=1}^n w_i^T(\psi^{(c)}) U_i(\beta, \theta). \quad (5)$$

Maximizing (5) over π in the M-step yields the explicit solution for the π parameters:

$$\pi_j^{(c+1)} = \sum_{i=1}^n w_{ij}(\psi^{(c)}) / n \quad j = 1, \dots, m. \quad (6)$$

If β, θ are unknown, we maximize the second term in (5) with respect to β, θ .

This involves solving a weighted set of score equations $\sum_{i=1}^n w_i^T(\psi^{(c)}) \partial U_i(\beta, \theta) / \partial \theta_j = 0$ and $\sum_{i=1}^n w_i^T(\psi^{(c)}) \partial U_i(\beta, \theta) / \partial \beta = 0$. In some problems, especially ex-

ponential family models such as the binomial, these equations have explicit solutions. Otherwise, one might need to incorporate a subalgorithm such as a Newton-Raphson-type procedure within each EM step. Fortunately, the EM likelihoods tend to be much more stable than the full likelihood and so the use of Newton Raphson can be straightforward.

As an example, consider an m -component mixture of binomial random variables each with index k and with unknown probabilities of success, θ_j , $j = 1, \dots, m$. At the c -th E-step, the weights w_{ij} are:

$$w_{ij}(\psi^{(c)}) = \frac{\pi_j^{(c)} \binom{k}{x_i} \theta_j^{(c)x_i} (1 - \theta_j^{(c)})^{k-x_i}}{\sum_{j=1}^m \pi_j^{(c)} \binom{k}{x_i} \theta_j^{(c)x_i} (1 - \theta_j^{(c)})^{k-x_i}}.$$

These are the posterior probabilities that X is from the j -th component, given that x_i is observed and the ‘prior distribution’ is defined by $(\pi^{(c)}, \theta^{(c)})$. At the M-step, we derive the terms $\partial U_{ij}(\theta) / \partial \theta_j = \frac{x_i}{\theta_j} - \frac{k-x_i}{1-\theta_j}$, and solve the $j = 1, \dots, m$ equations

$$\sum_{i=1}^n w_{ij}(\psi^{(c)}) \left[\frac{x_i}{\theta_j} - \frac{k-x_i}{1-\theta_j} \right] = 0.$$

The estimates of θ_j at the $(c + 1)$ -th step are

$$\hat{\theta}_j^{(c+1)} = \sum_{i=1}^n \left(\frac{w_{ij}(\psi^{(c)})}{\sum_{\ell=1}^n w_{\ell j}(\psi^{(c)})} \right) \frac{x_i}{k}.$$

So the updated component probabilities of success are weighted averages of the sample proportions of success, x_i/k . The updated estimates of π are given in equation (6). The next step is an E-step where the w_{ij} 's are recomputed using the updated parameter values, and the iteration proceeds.

In practice, the EM algorithm is relatively easy to program, however, it displays a very slow linear rate of convergence, especially if the components are similar in their densities. This would be the case, for example, if the two binomial components had values of θ close to each other. Convergence is also slow when the maximum likelihood solution for some of the π 's is zero, a point on the boundary. This is not particularly troublesome when one is finding maximum likelihood estimates for one data set, but it does pose problems when one is simulating a number of data sets and computing estimates.

2.2.7 Stopping criteria

One must always stop an iterative algorithm at some point; ideally that decision would be made because the estimators at the stopping time were sufficiently close to the answers one would get in the limit of the iterations. In the case of fast algorithms, like Newton Raphson, the answer might be determined simply by the limitations of the machine accuracy. However, the slow convergence rate of the EM algorithm can make it difficult to discern when one has converged to a local maximum with a sufficient accuracy. Unfortunately, stopping rules for the EM have often been based on naive ideas about the relationship between the size of the steps of the algorithm and the distance to the solution.

In a quadratically convergent algorithm, the size of the steps is very closely related to the distance remaining to the solution, and a small step implies that the remaining distance after that step is yet an order of magnitude smaller. However, in a linearly convergent algorithm like the EM, the step sizes can be very small even though there is a large distance yet to be traversed.

An additional issue is that the problems at hand are multidimensional, so one must decide how one is going to measure distance. One could use,

for example, the sums of squared differences between the current and final point estimators. This criterion ignores the fact that certain parameters will have higher statistical accuracy than others, and we would like the numerical errors in the point estimators to be small relative to the standard errors of the estimators. We need this if our statistical tests are to be meaningful.

Fortunately, we can make the errors small in this sense if we insist that the current log likelihood l_c be close to the final log likelihood \hat{l} . This is discussed in Lindsay (1995); in essence, if the log likelihood is close to \hat{l} , then likelihood based confidence intervals must be close. Thus we recommend that the likelihood be the function on which we base our stopping rules, and that an *ideal stopping rule* is to stop when $\hat{l} - l_c < \text{target}$. Of course, the ideal cannot be used because we do not know the final likelihood. Consider a naive stopping rule based on the successive differences in the log-likelihood:

$$\text{stop when } l_c - l_{c-1} < \text{tolerance.}$$

As indicated, for a slow linear algorithm, the tolerance level need not predict well the targeted difference between the current and final likelihoods, and so this is not a good surrogate for the ideal rule.

However, we can use a clever method, called *Aitkin acceleration*, to predict the final value \hat{l} based on the values of the likelihood over several iterations, and then plug this prediction into the ideal rule. (Lindsay, 1995). This is how it goes: Let l_{c-2}, l_{c-1} , and l_c be consecutive values of the log-likelihood at iterations $c-2, c-1$, and c respectively. The Aitkin predicted final value is

$$l_c^\infty = l_{c-2} + \frac{1}{1 - k_c} (l_{c-1} - l_{c-2}) \quad \text{where} \quad k_c = \frac{l_c - l_{c-1}}{l_{c-1} - l_{c-2}}.$$

If the algorithm progresses slowly, then k_c will be close to 1, and l_c^∞ will be larger than l_c . If l_c^∞ is a good estimator of the final likelihood, the rule

$$\text{stop if } l_c^\infty - l_c < \text{target}$$

will result in an actual error much closer to the targeted error.

Böhning et al. (1994) used this stopping rule in a simulation study of the likelihood ratio test for one component versus two components.

2.2.8 Starting values

The presence of multimodality in the finite component mixture likelihood has implications for how one chooses starting values. For example, Seidel,

Mosler & Alker (2000) showed that different starting and stopping strategies with the EM algorithm can lead to quite different estimates for exponential mixture models. The choice of starting value is especially important given the EM algorithm's slow convergence. Also, the sequence of estimates may diverge if the starting value is close to the boundary and the likelihood is unbounded there.

The magnitude of the problem has lead to a diverse set of approaches to deal with the problem. We mention some of them here; which one to use depends somewhat on the needs of the problem. For example, nonsystematic methods are ill-suited for simulation studies and it is therefore hard to determine their operating characteristics.

Informal. It is generally recommended that one perform several runs of the chosen optimization routine using different starting values. If the same answer reappears, one feels reassured.

Random. Finch et al. (1989) suggest using multiple random starting values. If the random starts come from a distribution, then it is possible to estimate the probability that further searching will turn up more solutions. In this, as in all cases, one would like the distribution to cover the plausible values without giving nonsense starts. One possibility, used by Markatou

(2000) is to bootstrap small samples of the data and use the moment estimators that arise therefrom as starts.

Method of moments. If one can get good enough initial values, then maybe searching is not necessary. To investigate this, Furman & Lindsay (1994) showed by simulation that the moment estimates provided good starting values for the EM algorithm. They worked as well as the true values (which generated the data) in the sense that the likelihoods were larger at the start with moment estimators, and the moment and true-value starting values proceeded almost universally to the same mode of the likelihood.

Data digging. McLachlan & Peel (2000) suggest starting the E-step with $w_{ij}(\psi^{(0)}) = z_{ij}^{(0)}$ where $z_{ij}^{(0)}$ is an initial partition of the data x_i into one of the m components. For some examples, the initial partition can be obtained from a plot of the data or from a clustering algorithm such as k-means. The initial partition could also be obtained from a random partitioning of the x_i into the m components.

NPMLE. Another approach suggested in Böhning et al. (1992) is to calculate the nonparametric maximum likelihood estimator of G , that is an estimator which does not fix the number of components, m . If the resulting estimator has more components than desired, one can often choose a way to

combine nearby support points while keeping the likelihood near its maximum value. One might be able to do this systematically by computing a penalized minimum distance estimator of G , which typically has fewer points of support than the nonparametric maximum likelihood estimator. (Chen & Kalbfleisch, 1996; Leroux, 1992; Leroux & Putterman, 1992)

We note that if the nonparametric maximum likelihood estimator has fewer than the desired number of components, then as we mentioned earlier it will not be possible to compute a maximum likelihood estimator with the desired number of components. In this case, the EM algorithm will slowly merge points of support, θ , together or force some values of π toward zero.

2.2.9 Information Matrix

The inverse of the observed information matrix, $I^{-1}(\psi; y) = [-\partial \log L(\psi) / \partial \psi \partial \psi^T]^{-1}$ is often used to estimate the covariance matrix of the MLE $\hat{\psi}$. Chapter #??? provides methods for computing and approximating $I^{-1}(\psi; y)$ at the end of an EM iteration.

Estimates of standard errors may also be obtained using resampling methods such as the bootstrap (see Chapter #???). Basford et al. (1997) and Peel (1998) found that unless sample sizes were very large, that standard

errors from the information matrix were too unstable and recommended using the bootstrap. For mixture models, a parametric bootstrap approach is suggested by McLachlan & Peel (2000).

2.3 Bayesian methods

The development of Bayesian approaches for mixture models has mushroomed with the popularization of Markov Chain Monte Carlo methods (MCMC).

In the Bayesian approach, $L(\psi)$ defined in Section 2.2.1 is combined with a prior density $p(\psi)$ for the parameter vector ψ to obtain a posterior density for ψ :

$$p(\psi; x) = c^{-1} L(\psi) p(\psi)$$

The computational effort required to compute the posterior is very large even for moderate sample sizes. Approximate posterior quantities of interest can be obtained through the use of MCMC methods (see Chapter #???). McLachlan & Peel (2000, Chapter 4) provide a nice introduction and references to the literature on using MCMC methods to Bayesian quantities of interest in the mixture model.

3 Fitting Mixture Models, m unknown

In this section, we assume that the number of components in the mixture model, m , is unknown. The problems at hand are to estimate G , the unknown mixing distribution, and any other structural parameters β . We will focus on one particular method here called nonparametric maximum likelihood.

3.1 The nonparametric maximum likelihood estimator

Kiefer & Wolfowitz (1956) developed the theoretical background for maximum likelihood estimation of the mixing distribution G , showing that the estimator is consistent. However, they did not address issues of computation nor did they characterize the solution in any way. Papers by Simar (1976), Laird (1978), Jewell (1982), Lindsay (1981, 1983a, b) and Böhning (1982) have since given us a number of tools to compute the nonparametric maximum likelihood estimator, \hat{G} .

3.1.1 Geometric characterization

One of the key tools to understanding the nature of the nonparametric maximum likelihood estimator is to characterize the maximization problem in

geometric terms. We start with the likelihood we wish to maximize. For simplicity, we consider models that do not contain structural parameters β , and so the only unknown is the latent distribution G . If we have a random sample x_1, x_2, \dots, x_n , then the log likelihood we wish to maximize has the form

$$\sum \log \int f(x_i; \theta) dG(\theta).$$

Now it is important that we write this likelihood as compactly as possible. In particular, if the data is discrete (categorical, for example), then many of the x_i will be the same. We therefore let $L_{\Delta_\theta} = [L_1(\theta), L_2(\theta), \dots, L_D(\theta)]^T$ represent the D distinct likelihoods $L_s(\theta) = f(x_i; \theta)$ arising from the data x_1, x_2, \dots, x_n , and let n_s be the multiplicity of $L_s(\theta)$. The log-likelihood of G can then be written in the form

$$\mathcal{L}(G) = \sum_{s=1}^D n_s \ln \int L_s(\theta) dG(\theta) = \sum_{s=1}^D n_s \ln L_s(G)$$

where $L_s(G) := \int L_s(\theta) dG(\theta)$.

We can geometrize the problem by noticing that the likelihood problem

depends on G only through the values taken by the D -dimensional vector

$$\mathbf{L}_G = [L_1(G), L_2(G), \dots, L_D(G)]^T.$$

That is, we can re-formulate the problem of maximizing $\mathcal{L}(G)$ over all distributions G , to the problem of maximizing the objective function

$$l(\mathbf{p}) := \sum_{s=1}^D n_s \ln p_s$$

over the elements $\mathbf{p} = (p_1, p_2, \dots, p_D)^T$ in the set $B = \{\mathbf{L}_G : G \text{ is a distribution}\}$ that correspond to \mathbf{L}_G for some G .

Now the beautiful result is that B is a convex set, and the objective function is a concave function, so we can appeal to convex optimization theory to describe our maximum likelihood solution. Let us describe the set B . First, let $\Gamma = \{\mathbf{L}_{\Delta_\theta} : \theta \in \Theta\}$, so that Γ represents the set of all possible likelihood kernel vectors that could arise from a one component mixing distribution, that is where G is Δ_θ , which is degenerate at $\theta \in \Theta$. The convex hull of Γ , the set of all convex combinations of Γ , is $\text{conv}(\Gamma)$, which is also the set B . In conclusion, the maximization problem can now be

written as

$$\sup_{\mathbf{p} \in \text{conv}(\Gamma)} \sum_{s=1}^D n_s \ln p_s = \sup_{\mathbf{p} \in \text{conv}(\Gamma)} l(\mathbf{p}) \quad \text{where } \mathbf{p} = (p_1, p_2, \dots, p_D)^T. \quad (7)$$

We next observe that if the set Γ is closed and bounded, then $\text{conv}(\Gamma)$ is a compact subset of \mathcal{R}^d . Under this assumption, Theorem 18 of Lindsay (1995) states the following:

- that if $\text{conv}(\Gamma)$ contains at least one point with positive likelihood, then there exists a unique $\hat{\mathbf{L}} \in \partial \text{conv}(\Gamma)$, the boundary of $\text{conv}(\Gamma)$, such that $\hat{\mathbf{L}}$ maximizes $l(\mathbf{p})$ over $\text{conv}(\Gamma)$
- that the solution $\hat{\mathbf{L}}$ is expressible as $\mathbf{L}_{\hat{G}}$ where \hat{G} is a discrete distribution with no more than D points of positive support. Recall that D is the number of distinct components in the likelihood and so is never larger than n , the sample size.

This characterization has now moved us closer to a solution. In fact, one strategy to find the solution would be to maximize the likelihood over m components, for each m less than or equal to D , and to choose the number of components \hat{m} with the largest likelihood. Of course, this strategy would leave us with the search problem we described in Section 2; at each stage m ,

how can we be sure we found the global maximum? Fortunately, as we will see shortly, there are some other tools we can use to identify the solution.

3.1.2 Example

As an example, consider a random sample of size two from a mixture of normal component densities with unknown means, θ_j , and variances equal to one, $\phi(x - \theta_j)$. Suppose that we observe $x_1 = 1.5$ and $x_2 = 3$. Figure 3 shows the set Γ as a solid line, and $\text{conv}(\Gamma)$ is the convex region inside the boundary of $\text{conv}(\Gamma)$. The dashed lines are contours of the objective function $l(\mathbf{p})$, and $\hat{\mathbf{L}}$ is the point (marked with an asterisk) on the boundary of $\text{conv}(\Gamma)$ that maximizes $l(\mathbf{p})$. Here $\hat{\mathbf{L}}$ is expressible as $\mathbf{L}_{\hat{G}}$ where \hat{G} places mass one on the point $\theta = 2$.

3.1.3 A gradient characterization

Returning to our set of tools, we next show that the optimal vector $\hat{\mathbf{L}} \in \text{conv}(\Gamma)$ and a corresponding mixing distribution \hat{G} , where $\hat{\mathbf{L}} = \mathbf{L}_{\hat{G}}$, can be characterized in terms of directional derivatives. The result will be that we can verify, in a straightforward way, whether we have attained a maximum of the likelihood. Recall that this is very different from the case where m

was fixed.

Consider the log-likelihood as a function of the D -dimensional vector $L_G = [L_1(G), L_2(G), \dots, L_D(G)]^T$. If we use the geometric optimization problem, the *directional derivative* of the objective function $l(\mathbf{p})$ from the point \mathbf{L}_{G_0} towards \mathbf{L}_{G_1} is defined to be:

$$d_{\mathbf{L}_{G_0}}(\mathbf{L}_{G_1}) := \lim_{\alpha \downarrow 0} \frac{l[(1 - \alpha)\mathbf{L}_{G_0} + \alpha\mathbf{L}_{G_1}] - l(\mathbf{L}_{G_0})}{\alpha}.$$

It can be easily calculated in our problem as

$$d_{\mathbf{L}_{G_0}}(\mathbf{L}_{G_1}) = \sum_{s=1}^D n_s \left(\frac{L_s(G_1)}{L_s(G_0)} - 1 \right).$$

This describes the optimization problem on the set B . For example, it must be true of the maximum point $\hat{\mathbf{L}}$ that $d_{\hat{\mathbf{L}}}(\mathbf{L}_{G_1}) \leq 0$ for every other \mathbf{L}_{G_1} , as otherwise we could increase the likelihood by moving in the direction of \mathbf{L}_{G_1} . We can think of characterizing the solution in this manner as being parallel to the way we use the likelihood equations to characterize the maximum likelihood estimator—the derivatives must be zero at the maximum.

However, as a practical matter we are more interested in how the likelihood changes with the choice of latent distribution, and so we want deriva-

tives in the space of distribution functions. However, it turns out the problem is not very different in this space. For example, it is easily shown that $d_{L_{G_0}}(L_{G_1})$ is also equal to the directional derivative of the log-likelihood, \mathcal{L} , from the point G_0 towards G_1 , which we define to be:

$$D_{G_0}(G_1) := \lim_{\alpha \downarrow 0} \frac{\mathcal{L}[(1 - \alpha)G_0 + \alpha G_1] - \mathcal{L}(G_0)}{\alpha} = d_{\mathbf{L}_{G_0}}(\mathbf{L}_{G_1}).$$

Although these directional derivatives also characterize the maximum, in the sense that $D_{\hat{G}}(G_1) \leq 0$, it would be very onerous to use this criterion, as we would have to check the inequality for every possible latent distribution G_1 . It is our good fortune that the values of $D_{G_0}(G_1)$ can be generated from a simpler function.

For the special case when $G_1 = \Delta_\theta$, a point mass distribution with mass at a single point θ , we define the *gradient function*

$$D_{G_0}(\theta) := D_{G_0}(\Delta_\theta) = \sum n_s \left(\frac{L_s(\theta)}{L_s(G_0)} - 1 \right).$$

Geometrically, this is also the directional derivative of $l(\mathbf{p})$ from a point $\mathbf{L}_{G_0} \in \text{conv}(\Gamma)$ towards a point $\mathbf{L}_{\Delta_\theta}$ on the curve Γ . The relationship $D_{G_0}(G_1) = \int D_{G_0}(\theta) dG(\theta)$ shows that the gradient function can be used to calculate all

the other directional derivatives as well, and so it is no surprise that it can be used to characterize the NPMLE.

The following fundamental theorem for nonparametric mixture maximum likelihood provides much of the basis for algorithms to compute the nonparametric MLE. It is extracted from Lindsay (1995, pp. 115-6)

Theorem 1: *The following three statements are equivalent:*

1. \hat{G} maximizes $\mathcal{L}(G)$.
2. \hat{G} minimizes $\sup_{\theta} D_G(\theta)$.
3. $\sup_{\theta} \{D_{\hat{G}}(\theta)\} = 0$.

The third criterion is the key to identifying in an algorithm when one has found the solution. One checks the gradient function $D_{G_c}(\theta)$ at any current estimator G_c to see if for any θ it violates (at some level of tolerance) the nonnegativity constraint.

The following theorem can sometimes be useful in the construction of an algorithm as well.

Theorem 2: *The support of any maximum likelihood estimator \hat{G} lies in the set $\{\theta : D_{\hat{G}}(\theta) = 0\}$.*

A result from Lindsay (1981) can be useful because it enables us to restrict our search for support points of \hat{G} to a finite interval.

Proposition 3: *Suppose that the parameter θ is real-valued and that for every i , the likelihood kernel $L_i(\theta)$ is unimodal in θ , with unique mode at $\tilde{\theta}_i$. Then all the support points of \hat{G} lie in the interval*

$$[\min_i \tilde{\theta}_i, \max_i \tilde{\theta}_i].$$

3.1.4 Computational Strategies.

As in the case of the mixture model with fixed m , there are a number of choices to be made in creating an algorithm.

- **Starting values.** Fortunately, the choice of a starting distribution G is no longer an issue, as the likelihood has a unique solution that is characterized by the gradient function. There should be no dependence on where we start, provided we use the gradient for our stopping rule.
- **Stopping Rules.** We believe that the gold standard stopping rule would be to stop when the current likelihood is sufficiently close to the maximum likelihood. Again, we are in fortune because the gradient function can help us achieve this, as discussed in the next subsection.

- **Algorithms.** Here we have a wide variety of choices, ranging from EM based to convex optimization based approaches. We will discuss a few of the possibilities in the next few subsections.

3.1.5 Stopping criteria

In theory, we should continue an iterative algorithm until we obtain the maximum, $\hat{\mathbf{L}} = \mathbf{L}_{\hat{G}}$ where $\delta = \sup_{\theta} \{D_{\hat{G}}(\theta)\} = 0$. In practice, we may not reach the exact value $\mathbf{L}_{\hat{G}}$, and one might think that a reasonable approximation to the maximum may be obtained by stopping when $D_{\hat{G}^{(e)}}(\theta) \leq \gamma$, for all $\theta \in \Theta$, where γ is a small positive number.

It turns out that this simple rule is also a gold standard stopping rule. The following theorem from Lindsay (1995, p. 118) provides upper and lower bounds on the maximum change in the log likelihood in going to the maximum from a given a candidate estimator, $\hat{G}^{(e)}$.

Theorem 4: *Let $\hat{G}^{(e)}$ be the current mixing distribution estimate in an iterative algorithm designed to compute the maximum likelihood estimator.*

Define $\delta = \sup_{\theta} \{D_{\hat{G}^{(e)}}(\theta)\}$. Then

$$A(\delta) \leq l(\hat{L}) - l(L_{\hat{G}^{(e)}}) \leq B(\delta) \leq \delta,$$

where $B(\delta) := n \ln(1 + \delta/n)$, $A(\delta) := B(\delta) - n^* \ln(1 + \delta/n^*)$ and $n^* = n - \min_s \{n_s\}$.

Lindsay (1995, p. 132) provides some guidance as to the choice of γ . In practice, it is also difficult to ensure the inequality $D_{\hat{G}^{(c)}}(\theta) \leq \gamma$ for all $\theta \in \Theta$, and Lindsay (1995, p. 127-8, 133-5) suggests testing the inequality for a carefully chosen grid of θ values, where the choices can be made so as to preserve the gold standard of stopping within a fixed tolerance of the likelihood maximum.

3.1.6 Algorithms - EM

The EM algorithm can be used to compute the nonparametric MLE. One could start the algorithm assuming D points of support and iterate. If the MLE has fewer than D points of support, the mass associated with some of the points will tend to zero. The gradient bound should be used to ensure adequate convergence, and one must be prepared to allow the algorithm to run a long time. (DerSimonian, 1986; Laird, 1978)

3.1.7 Algorithms - gradient based

Böhning (1999, 1995) provides an overview of algorithms available to compute the nonparametric MLE. These algorithms generally use the gradient function evaluated at the current location, $D_{\hat{G}^{(c)}}(\theta)$, to define new directions that will increase the likelihood.

3.1.8 A dual problem, semi-infinite programming problem

A dual form of the maximum likelihood problem yields a problem in the form of a semi-infinite programming problem. (Lindsay 1995, pp. 117-8, Coope & Watson, 1985)

Dual: Minimize $l(\mathbf{p})$ subject to the constraints $p \geq 0$, and $d_p(L_{\Delta_\theta}) \leq 0$, for all $\theta \in \Theta$.

If the nonparametric MLE solution is \hat{L} , then $p = \hat{L}$ solves the Dual problem. Lesperance & Kalbfleisch (1992) solve the dual formulation of the problem using a Lagrangian algorithm (Coope & Watson, 1985) to estimate the population distribution in cure rates for data from Laird (1978).

4 References

References

- [1] Basford, K.E. Greenway, D.R., McLachlan, G.J. and Peel, D. (1997) Standard errors of fitted means under normal mixture models. *Computational Statistics*, 12, 1-17.
- [2] Böhning, D. (1982) Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Annals of Statistics*, 10, 1006-1008.
- [3] Böhning, D., Schlattmann, P., and Lindsay, B.G. (1992) Computer assisted analysis of mixtures (C.A.MAN): Statistical algorithms. *Biometrics* 48, 283-303.
- [4] Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. and Lindsay, B.G. (1994) The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46, 373-388.

- [5] Böhning, D. (1995) A review of reliable maximum likelihood algorithms for the semi-parametric mixture maximum likelihood estimator. *Journal of Statistical Planning and Inference*, 47, 5-28.
- [6] Böhning, D. (1999) *Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others*. Chapman & Hall, New York.
- [7] Chen, J. and Kalbfleisch, J. (1996). Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, 24, 167-175.
- [8] Coope, I.D. and Watson, G.A. (1985) A projected Lagrangian algorithm for semi-infinite programming. *Mathematical Programming*. 32, 337-356.
- [9] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum likelihood estimation from incomplete data via the Em algorithm (with discussion). *Journal of the Royal Statistical Society, B.*, 39, 1-38.
- [10] DerSimonian, R. (1986). Maximum likelihood estimation of a mixing distribution. *Journal of the Royal Statistical Society, Ser. C*, 35, 302-309.

- [11] Donoho, David L. (1988) One-sided inference about functionals of a density. *The Annals of Statistics*, 16, 1390-1420.
- [12] Finch, Stephen J., Mendell, Nancy R., and Thode, Henry C. (1989) Probabilistic measures of adequacy of a numerical search for a global maximum. *Journal of the American Statistician*, 84, 1020-1023.
- [13] Fletcher, R. (1987) *Practical methods of optimization, second edition*. John Wiley & Sons.
- [14] Furman, D. and Lindsay, B.G. (1994) Measuring the relative effectiveness of moment estimators as starting values in maximizing mixture likelihoods. *Comput. Statist. Data Anal*, 17, 493-507.
- [15] Kiefer, J. and Wolfowitz, J. (1956) consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27, 886-906.
- [16] Jewell, N.P. (1982) Mixtures of exponential distributions. *Annals of Statistics*, 10, 479-484.

- [17] Laird, N. (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73, 805-811.
- [18] Lehmann, E. (1983) *Theory of Point Estimation*. John Wiley, New York.
- [19] Leroux, B. (1992) Consistent estimation of a mixing distribution. *Annals of Statistics*, 20, 1350-1360.
- [20] Leroux, B. and Puterman, M.L. (1992) Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, 48, 545-558.
- [21] Lesperance, M.L. and Kalbfleisch, J.D. (1992) An algorithm for computing the nonparametric MLE of a mixing distribution. *Journal of the American Statistical Association*, 87, 120-126.
- [22] Lindsay, B.G. (1981) Properties of the maximum likelihood estimator of a mixing distribution. In *statistical Distributions in Scientific Work* (G.P. Patil, ed.), 5, 95-109, Reidel, Boston.

- [23] Lindsay, B.G. (1983a) The geometry of mixture likelihoods: A general theory. *Annals of Statistics*, 11, 86-94.
- [24] Lindsay, B.G. (1983b) The geometry of mixture likelihoods, Part II: The exponential family. *Annals of Statistics*, 11, 783-792.
- [25] Lindsay, B.G. (1989) Moment Matrices, applications in mixtures. *Annals of Statistics*, 17, 722-740.
- [26] Lindsay, Bruce G. (1995) *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5. IMS/ASA.
- [27] Lindsay, B.G. and Basak, P. (1991) On using bivariate moment equations in mixed normal problems, in *Estimating Functions*, edited by V.P. Godambe. Oxford Science Publications, New York.
- [28] Lindsay, B.G. and Basak, P. (1993) Multivariate normal mixtures: A fast consistent method of moments. *Journal of the American Statistical Association*, 88, 468-476.
- [29] Maritz, J.S. and Lwin, T. (1989) *Empirical Bayes Methods, second edition*. Chapman and Hall, New York.

- [30] Markatou, M. (2000) A closer look at the weighted likelihood in the context of mixtures. In *Probability and Statistical Models with Applications: A Volume in honor of T. Cacoullos*, Chapman and Hall/CRC, 467-487.
- [31] McLachlan, G.J. and Basford, K.E. (1988) *Mixture Models: Inference and applications to clustering*. Marcel Dekker, New York.
- [32] McLachlan, G.J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*. Wiley, New York.
- [33] McLachlan, G.J. and Peel, D. (2000) *Finite Mixture Models*. John Wiley, New York.
- [34] Peel, D. (1998) *Mixture model clustering and related topics*. Unpublished Ph.D. thesis, University of Queensland, Brisbane.
- [35] Prakasa Rao, B.L.S. (1992) *Identifiability in Stochastic Models, Characterization of Probability Distributions*. Academic, Boston.
- [36] Seidel, W., Mosler, K., and Alker, M. (2000) A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics* (to appear).

[37] Simar, L. (1976) Maximum likelihood estimation of a compound Poisson process. *Annals of Statistics*, 4, 1200-1209.

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

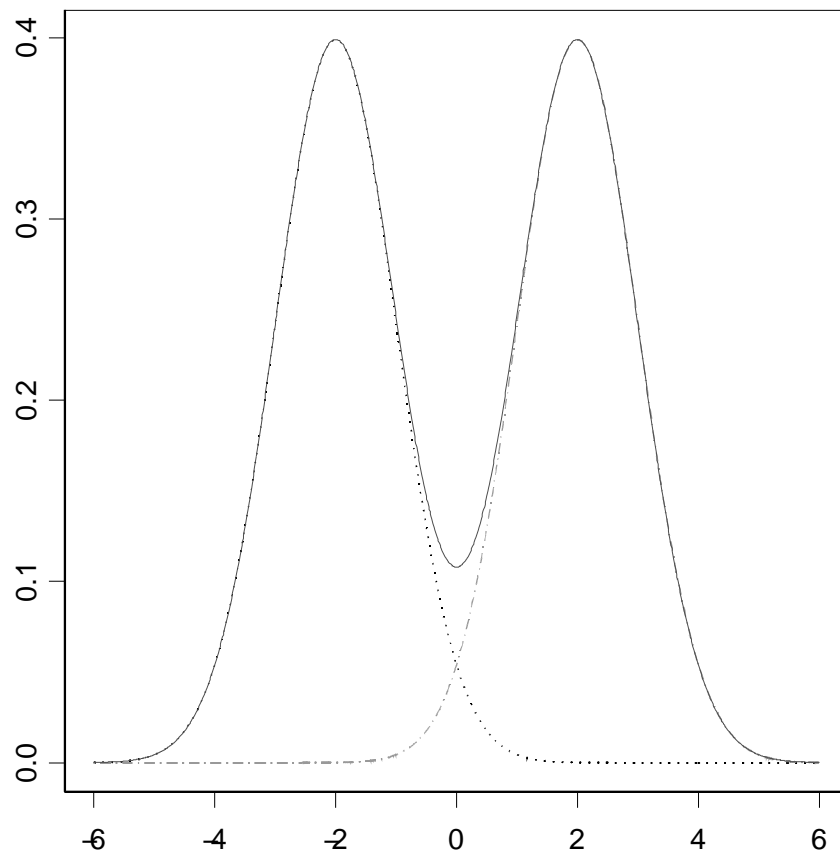


Figure 1: Mixed normal densities with means four standard deviations apart, and equal weights. The dotted lines show the component densities and the solid line is the mixture density. (Note that the mixed density is scaled to have integral equal to two for comparison purposes.)

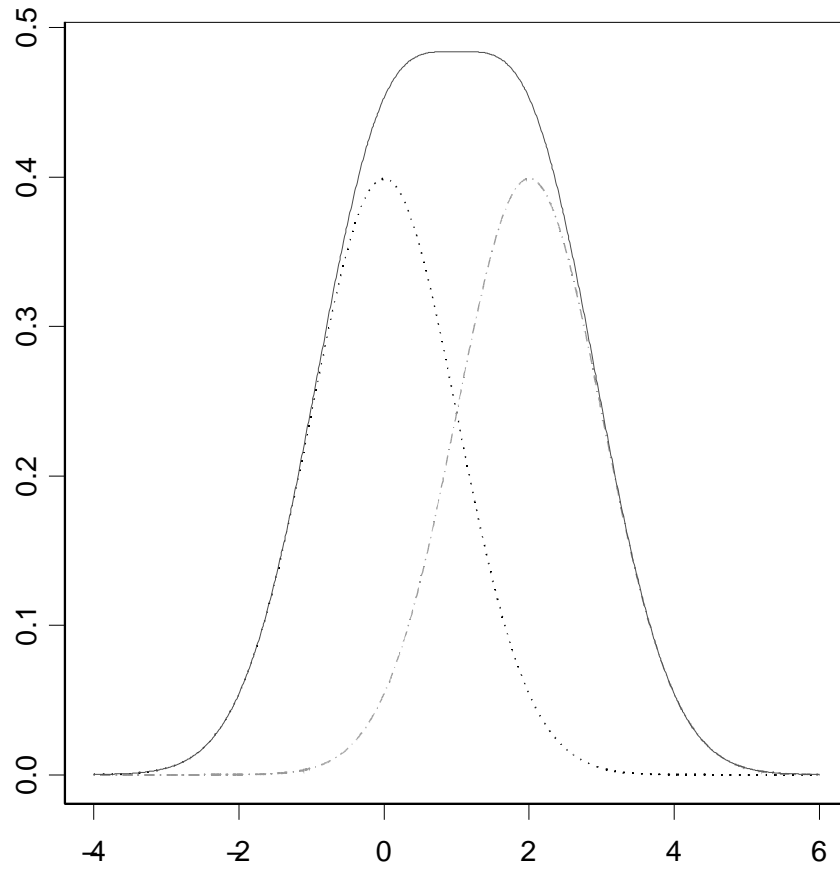


Figure 2: Mixed normal densities with means two standard deviations apart and equal weights. The dotted lines show the component densities and the solid line is the mixture density. (Note that the mixed density is scaled to have integral equal to two for comparison purposes.)

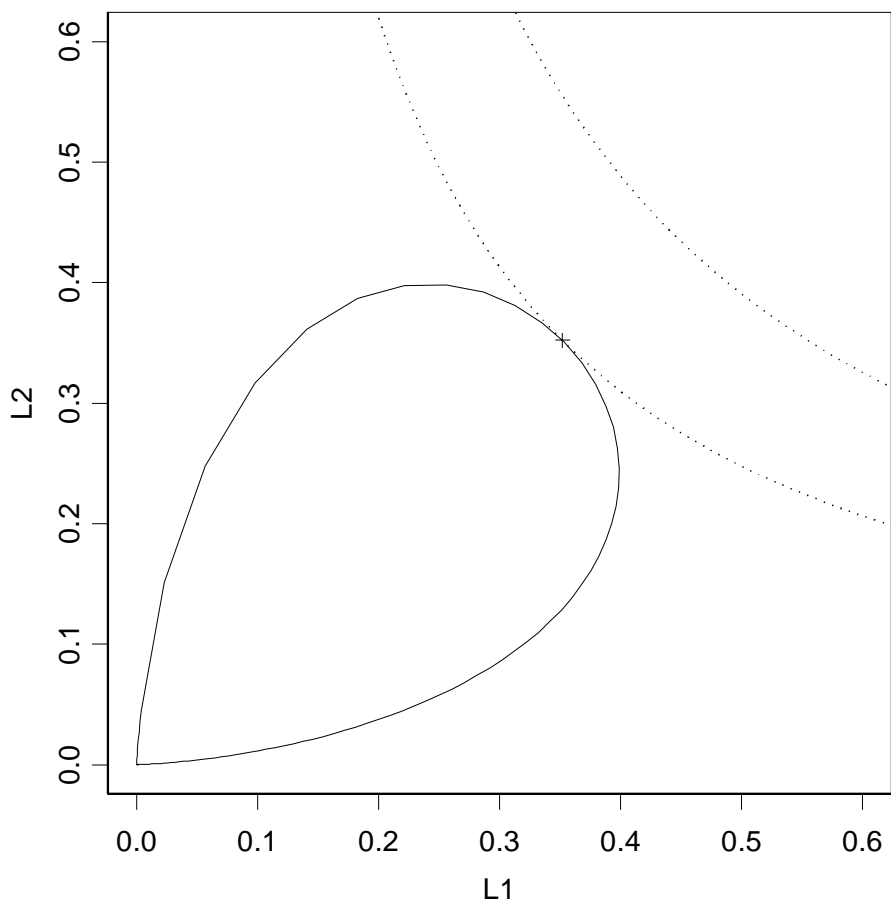


Figure 3: The solid line is the curve \clubsuit for two observations $(x_1=1.5, x_2=2.5)$ from normal component densities with variances equal to one. The region within the solid curve is $\text{conv}(\clubsuit)$. The dotted lines show the contours of the log likelihood. The point marked with an X is the point (p_1, p_2) at which the likelihood is maximized, subject to the constraint that (p_1, p_2) is in $\text{conv}(\clubsuit)$. The corresponding maximum likelihood estimator G , places mass one on