CHAPTER 3

Parametric Models

As an aid to understanding the role of nonparametric maximum likelihood, this chapter is a review of some of the basic features of the standard parametric models used in mixture modeling, together with basic features of maximum likelihood estimation in these models. We will consider the two different schools of modeling. In one, the latent distributions Q are assumed discrete, with a fixed number of components m. In the other, the latent distributions is assumed to come from some parametric family of continuous distributions.

3.1. Discrete versus continuous. There are certainly instances in which the latent distribution is logically and naturally modeled in either the discrete or continuous form due to the nature of the application. For example, it might be known that there are a finite and known number m of physical components in the population or, alternatively, that there is an inherent continuity expected in the latent variate Φ, such as when it represents a continuous covariate that was not measured.

More typically, however, we are on uncertain ground in specifying the number of components in a discrete latent distribution. Alternatively, in the case of a continuous variate Φ, we have no reason to believe it follows any particular parametric form. Although it is standard practice to assume normality for the latent variate, there is an element of arbitrariness in choosing the appropriate function of the latent variate to be normally distributed.

Moreover, as we have indicated already, we generally obtain very little goodness-of-fit information about the latent distribution from the data, and so there is little hope of having a large enough sample to investigate the true distributional form.

My own preference is for the use of discrete distributions in cases of doubt. The argument is that it makes maximum likelihood numerically simpler, and so is to be preferred if the form of the latent distribution specification makes little difference to the desired statistical inference. A number of investigations have found that misspecifying the latent distribution has very little effect on bias, and minimal effect on standard errors beyond the necessary correction for overdispersion when mixing is present [Butler and Louis (1992); Neuhaus, Hauk and Kalbfleisch (1992)].

One desirable feature of using a continuous family of latent distributions is that it leads to a smooth family of posterior densities $q(\phi|x)$ for empirical Bayes inference about the latent variates. However, this comes at the risk of allowing the user to imagine there is fairly precise information about the shape of the posterior, even though it is not possible to verify this shape empirically without gargantuan sample sizes.

On the behalf of using discrete mixtures, I would further argue that there is a much greater degree of flexibility—admittedly at the cost of additional parameters—without any real loss in fitting power. Moreover, as we will see, it is generally much simpler and more reliable numerically to calculate the estimates. In particular, a key limitation of the continuous approach is that, except for limited cases, the calculation of the mixture density $\int f(x;\phi)\cdot q(\phi)\,d\phi$ will require numerical integration, and unlike the discrete case, one usually cannot directly apply the reliable EM algorithm to perform the steps in the likelihood maximization.

3.1.1. *Continuous models: The conjugate family.* Although there are many instances where normality is assumed for the distribution of Φ, there is a more sophisticated approach to the construction of the latent distribution that exploits certain Bayesian structures, where the latent distribution plays the role of *prior*.

If a parametric family of prior densities $q(\phi;a,b)$ is such that the posterior densities $q(\phi|x;a,b)$ are from the same parametric family, then the family is said to be *conjugate* to the sampling density $f(x;\phi)=f(x|\phi)$. Such a conjugacy relationship holds, for example, if f and q are normal densities, because then the posterior densities are also normal.

Diaconis and Ylvisaker (1979) worked out a number of the properties for conjugate families in the case when f is from the exponential family. Suppose the density for a single observation is from the one parameter exponential family considered in the previous chapter, having density

$$f(x;\phi)=\exp(\phi x-\kappa(\phi))$$

with respect to some measure $dF_0(x)$. If we obtain a sample $X_1,\dots,X_n$, then the joint density has the form

$$f(\mathbf{x};\phi)=\exp(\phi(n\overline{x})-n\kappa(\phi))$$

with respect to the product measure, where $\bar{x}=n^{-1}\sum x_i$.

Suppose that the latent parameter Φ has a continuous prior density from the two parameter conjugate family of models:

$$dQ(\phi)=q(\phi;\mu,\tilde{n})\,d\phi=\exp(\tilde{n}\mu\phi-\tilde{n}\kappa(\phi)-\psi(\mu,\tilde{n}))\,d\phi.$$

[As an *exercise*, show that the conjugate family for the binomial distribution is the Beta(α,β) distribution, with suitable selection of parameters.] It follows by examination of the joint density $f(x;\phi) \cdot q(\phi)$ that the posterior density $q(\phi|x)$ must be from the same family, with updated parameters

$$\mu^* = \alpha\mu + (1-\alpha)\overline{x},$$

with $\alpha := \tilde{n}/(n+\tilde{n})$, and

$$\tilde{n}^* = \tilde{n} + n.$$

[*Exercise.*] The form of the updating suggests a natural interpretation of the parameter $\tilde{n}$: it is the prior "sample size". Changing it by one unit has exactly the same effect on the posterior as does changing the sample size n by one unit. Additionally, the posterior parameter μ^* is very elegantly expressed as a weighted mean of the prior parameter μ and the sample mean $\overline{x}$, with weights proportional to the "prior sample size" $\tilde{n}$ and the sample size n.

Some useful characterizations of the moments of $\overline{X}$ can be obtained by using integration-by-parts techniques. If $h(\phi)$ is a differentiable function and $q(\phi) = q(\phi;\mu,\tilde{n})$, then we can easily show

$$[h(\phi)\cdot q(\phi)]' = h'(\phi)\cdot q(\phi) + h(\phi)\cdot[\tilde{n}\mu - \tilde{n}\kappa'(\phi)]\cdot q(\phi).$$

Suppose that $h(\phi)q(\phi)$ is zero at the left and right limits of the parameter space. Then integration of the last displayed equation gives us

$$0 = E[h'(\Phi)] + \tilde{n}\cdot E[h(\Phi)\cdot\{\mu - \kappa'(\Phi)\}]. \tag{3.1}$$

Replacing $h(\phi)$ in (3.1) with various functions of interest now gives a number of useful identities regarding the marginal distribution of the observable variable X. If we let $h(\phi) = 1$, then it proves that

$$\mu = E[\kappa'(\Phi)].$$

That is, in terms of the prior, or latent, distribution, μ has a natural interpretation. This can be turned into a property for the mean of X by recalling that in a one parameter exponential family,

$$\kappa'(\phi) = E[X;\phi].$$

Hence in terms of the sampling distribution for X, the parameter μ equals $E[\ E[X|\Phi]\ ] = E[X]$, the marginal mean of X.

Applying the integration-by-parts trick to $h(\phi) = (\mu - \kappa'(\phi))$ yields a second identity

$$E[\kappa''(\Phi)] = \tilde{n}E\{\mu - \kappa'(\Phi)\}^2 = \tilde{n}\,\mathrm{Var}(\kappa'(\Phi)).$$

Since the left-hand side of the last equation is also $nE[\mathrm{Var}(\overline{X}|\Phi)]$, we have that

$$\mathrm{Var}(\overline{X}) = E[\mathrm{Var}(\overline{X}|\Phi)] + \mathrm{Var}[E(\overline{X}|\Phi)] = \mathrm{Var}(\kappa'(\Phi))\cdot\left[\frac{\tilde{n}}{n}+1\right].$$

Thus the sampling variance of the sample mean is an inflated version of the variability of the latent mean value parameter. [As an *exercise* in understanding, check that these formulas apply in the beta binomial example.]

The simple prior-to-posterior relationship makes the conjugate families very attractive in the Bayesian framework. They would also seem to have nice mathematical structures for mixture model inference. However, there are two additional points to consider. One is that the families have been constructed purely based on mathematical convenience, not modeling considerations, and so might leave the user with a goodness-of-fit problem. One solution to this is to enlarge the latent distribution model to allow finite mixtures of conjugate distributions, say mixed over the mean parameter μ, with the dispersion parameter $\tilde{n}$ fixed. Dalal and Hall (1983) show that this class is extremely rich, in that one can approximate arbitrarily closely any prior Q.

Another point to consider is that the tidy prior–posterior relationship provides no guarantees that the marginal mixture density has any nice features. Typically the resulting marginal distribution for X is no longer in the exponential family and so lacks the nice features thereof, such as reduction to sufficient statistics, guaranteed-to-be-unimodal likelihoods and uniformly most powerful test procedures. It is easily checked that the likelihood equations have the form

$$n^{-1}\sum E[\Phi|X = x_i] = E[\Phi],$$

$$n^{-1}\sum E[\kappa(\Phi)|X = x_i] = E[\kappa(\Phi)].$$

Thus in the conjugate family model, $\hat{\mu} = \bar{x}$ does not seem to be implied by the likelihood equations. This occurs in contrast to the finite mixture model, where we will show that the likelihood equations imply that the sample mean is equated to the theoretical mean.

3.2. Discrete latent distribution. We now turn to the alternative form of parametric modeling, in which one assumes that the latent distribution is discrete, with a known number of components m. We will denote a latent distribution of this type as Q_m.

3.2.1. *Known component distributions.* We return to the models of Section 1.3.1, in which the mixture density can be written as $f(x; Q_m) = \sum f_j(x)\pi_j$, with known component densities f_j, $j = 1, \ldots, m$. We now wish to set up the likelihood equations for the weight parameters. One obstacle is the need to satisfy the constraint $\sum \pi_j = 1$. One solution is to use a Lagrange multiplier technique. Another one is to eliminate one of the weights, say by setting $\pi_m = 1 - \pi_1 - \cdots - \pi_{m-1}$. If this is done, there are then $m - 1$ score functions

$$S_j(\boldsymbol{\pi}) = \frac{\partial \ln(L(Q_m))}{\partial \pi_j} = \sum \frac{f_j(x_i) - f_m(x_i)}{f(x_i; Q_m)}. \tag{3.2}$$

Unfortunately, the inequality constraints $\pi_j \geq 0$ mean that unless the maximum likelihood estimator $\hat{\pi}_j$ is in the interior of the parameter space, with all $\hat{\pi}_j$ strictly positive, the estimate does not solve the usual likelihood equations $S_j(\boldsymbol{\pi}) = 0$. Although we ignore this difficulty in this chapter, it should be pointed out that the nonparametric mixture maximum likelihood theorem of Chapter 1 does apply, and so the solution is completely described by the gradient inequality.

However, there are two pieces of good news. First, the log likelihood is *strictly concave* in the parameters π. [*Exercise*: Show that the matrix of second derivatives is negative definite.] This means that we have unique solutions for the weights, provided that they are identifiable. The second piece of good news is that the EM algorithm is easy to construct and implement.

Here is one derivation of the EM algorithm for this problem. By rearranging the likelihood equations $S_j(\boldsymbol{\pi}) = 0$, one obtains the equivalent equations

$$\pi_j = \pi_j \cdot n^{-1} \sum_{i=1}^{n} \frac{f_j(x_i)}{f(x_i; Q_m)}. \tag{3.3}$$

[*Exercise*.] This is a vector equation of the "fixed point" type, namely, $\boldsymbol{\pi} = \mathbf{F}(\boldsymbol{\pi})$, and the corresponding fixed point algorithm is simple: Given the current value $\boldsymbol{\pi}_c$, the updated value is $\boldsymbol{\pi}_u = \mathbf{F}(\boldsymbol{\pi}_c)$.

We will return to this algorithm later in order to give its missing data interpretation and some further background.

3.2.2. *Unknown component parameters.* If the m-component densities have unknown latent parameters $\xi_1, \ldots, \xi_m$, then we must also maximize the likelihood over them. Thus, in addition to the previous score functions for the weights, we will need the ξ scores:

$$U_j(\boldsymbol{\xi}, \boldsymbol{\pi}) = \frac{\partial}{\partial \xi_j} \ln(L(Q_m)) = \sum_i \left(\frac{f(x_i; \xi_j)}{f(x_i; Q_m)} \right) \left(\frac{f'(x_i; \xi_j)}{f(x_i; \xi_j)} \right). \tag{3.4}$$

We have written the scores in this fashion to emphasize that they are weighted sums of the *score functions*

$$v(x_i; \xi) := \frac{f'(x_i; \xi)}{f(x_i; \xi)}$$

from the unicomponent model.

The full m-component likelihood equations then require solving for the full set of scores (3.2) and (3.4) to equal zero. Unfortunately, the likelihood equations no longer need have unique solutions. Indeed, experience suggests that they frequently have multiple solutions. We will discuss this point further when discussing the issue of initial values for algorithms. On the other hand, despite their apparent complexity, there is the easy-to-program and reliable EM algorithm to implement for their solution.

3.3. Properties of the m-component MLE. We suppose that we have an m-component mixture estimator $\hat{Q}_m$ that satisfies the likelihood equations satisfied above. There are useful ways to paraphrase these equations that give some further insights. The following results are from Lindsay (1981).

PROPOSITION 8. *For any function $g(\phi)$ and $\hat{Q}_m$ an m-component solution determined by setting the π scores in* (3.2) *equal to zero, the following self-consistency equations hold*:

$$n^{-1}\sum E[g(\Phi)|X = x_i; \hat{Q}_m] = E[g(\Phi); \hat{Q}_m].$$

The proof is left as an *exercise*. The preceding result applies to both the known and unknown component cases. The score equations from the unknown components can also be given a prior–posterior expression as follows:

PROPOSITION 9. *For any function $h(\phi)$ and any m-component solution found by setting the ξ-score functions in* (3.4) *to zero, we have*

$$n^{-1}\sum E[\ h(\Phi)v(\Phi; x_i)\ |X = x_i; \hat{Q}_m] = 0.$$

We may put these two sets of equations together in the case of the exponential family mixture to obtain the following first moment property.

PROPOSITION 10. *If the component density $f(x;\phi)$ is a one parameter exponential family, and $\hat{Q}_m$ satisfies the m-component likelihood equations, then $E[X; \hat{Q}_m] = \bar{x}$.*

This may be proved by using $h(\phi) = 1$ and $g(\phi) = \kappa'(\phi)$ in the two preceding propositions.

3.4. EM algorithm.

The next task is to give a derivation of the EM algorithm for the m-component discrete mixture model.

3.4.1. *A description of the EM.* The EM algorithm requires a particular model structure. Suppose that we have a model with parameters η in which there is both observed data X and missing data J. We need to maximize the likelihood of the observed data X, call it $L_X(\eta)$, but the likelihood is difficult to maximize. However, we assume that if we knew the unobserved data J, then the maximization of the likelihood $L_{(X,J)}(\eta)$ of the pair (X, J) would be easy, ideally having explicit solutions. We note that the "missing data" J could be completely imaginary; the important thing is that the distribution of the variable X that is observed is exactly the same as the marginal distribution of X in some hypothetical pair (X, J) which has an "easier" likelihood.

The *E step* in the EM algorithm in this situation involves taking a current value η_c and finding the *EM log likelihood* l_{em}, which is the conditional expectation of the full data log likelihood $\ln(L_{(X,J)})$ given the observed data:

$$l_{\text{em}}(\eta; \eta_c) := E[\ln L_{(X,J)}(\eta) \mid X; \eta_c].$$

The *M step* in a cycle is to let the EM solution η_{c+1} be that value of η, that maximizes the EM likelihood $l_{\text{em}}(\eta; \eta_c)$. Ideally, both steps have explicit solutions.

It is a simple *exercise* in the use of Jensen's inequality to show that such an iterative sequence increases the true likelihood L_X at each cycle [Dempster, Laird and Rubin (1977)].

3.4.2. *The EM for finite mixtures.* We will set these equations up in the framework of the discrete component density. The advantage to this approach is that we can give a simple interpretation of the action of the algorithm in terms of the filling-in-by-expectation of the unobserved cells of a contingency table. The formulas are exactly the same in the continuous case, but this author finds the contingency table approach more insightful.

In the multinomial case we can reduce the observed data by sufficiency to the counts $n(t) = \#\{X_i = t\}$. The missing data will be the variables J_i that identify the component labels (introduced in Chapter 1, Section 1.1). Thus in a discrete setting, we can reduce the complete data by sufficiency to the counts $n_j(t)$ of the number of times the pair (t, j) appeared in the complete sample. The observed data are then the column marginal totals of an unobserved table of counts, where the *rows* correspond to the different components and the *columns* correspond to the possible values of the variable X: see Table 3.1. The corresponding table of multinomial cell probabilities is shown in Table 3.2.

The complete data log likelihood for such a multinomial model is simply

$$\sum_j \sum_t n_j(t) \ln[\pi_j f(t;\xi_j)] = \sum_j \ln(\pi_j)\Big(\sum_t n_j(t)\Big) + \sum_j \sum_t n_j(t) \ln(f(t;\xi_j)). \tag{3.5}$$

Since the entries $n_j(t)$ in the full table are unobserved, to calculate l_{em} we take the complete data log likelihood (3.5) and replace the $n_j(t)$ with their expectations, say $\tilde{n}_j(t)$, conditioned upon the observed data and the current parameter estimates. This is easily done by allocating the total observed counts in a column $n(t)$ to each row in that column proportionally to its current estimated cell probability, so that we have

$$\tilde{n}_j(t) = n(t) \cdot \frac{\pi_j f(t;\xi_j)}{f(t;Q_m)}.$$

Since the parameters π and ξ separate in the complete data likelihood (3.5), the EM algorithm updates the current estimates of the weights π by

TABLE 3.1

	$t = 0$	$t = 1$	$\cdots$	$t = T$
$J = 1$	$n_1(0)$	$n_1(1)$	$\cdots$	$n_1(T)$
$J = 2$	$n_2(0)$	$n_2(1)$	$\cdots$	$n_2(T)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$J = m$	$n_m(0)$	$n_m(1)$	$\cdots$	$n_m(T)$
Totals	$n(0)$	$n(1)$	$\cdots$	$n(T)$

TABLE 3.2

	$t = 0$	$t = 1$	$\cdots$	$t = T$
$J = 1$	$\pi_1 f(0; \xi_1)$	$\pi_1 f(1; \xi_1)$	$\cdots$	$\pi_1 f(T; \xi_1)$
$J = 2$	$\pi_2 f(0; \xi_2)$	$\pi_2 f(1; \xi_2)$	$\cdots$	$\pi_2 f(T; \xi_2)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$J = m$	$\pi_m f(0; \xi_m)$	$\pi_m f(1; \xi_m)$	$\cdots$	$\pi_m f(T; \xi_m)$
Totals	$f(0; Q)$	$f(1; Q)$	$\cdots$	$f(T; Q)$

maximizing the first summand of the likelihood above, yielding

$$\pi_{j,\mathrm{em}} = \pi_j n^{-1} \sum_t \tilde{n}_j(t),$$

the same formula derived earlier as a fixed point algorithm.

Notice also the algorithm has a natural interpretation in terms of the posterior probabilities. The term

$$\pi_{j|i} := \Pr(J_i = j | X = x_i; Q_m)$$

represents the posterior probability that the ith observation was from component j. The updated estimate of the weight parameter equals the average of the current estimated posterior probabilities:

$$\pi_{j,\mathrm{em}} = \Pr(J = j; \hat{Q}_{\mathrm{em}}) = n^{-1} \sum_i \pi_{j|i}.$$

If the component parameters ξ_j are unknown, then we must add to the previous equations a set of equations for them. In this case, we must solve $\sum_t \tilde{n}_j(t) v(\xi_{j,\mathrm{em}}; t) = 0$. This equation is related to the second proposition of the preceding section in that it specifies

$$\sum_i \pi_{j|i} v(\xi_{j,\mathrm{em}}; x_i) = 0,$$

a weighted set of score equations. If the components have an exponential family density, then these equations specify that the EM iterations in the mean value parameters $\mu(\xi)$ are

$$\mu_{j,\,\mathrm{em}} =: \frac{\sum_i \pi_{j|i} \cdot x_i}{\sum_i \pi_{j|i}};$$

that is, they are simply reweighted means, where the weights are proportional to the posterior probability of having come from group j.

3.4.3. *Algorithmic theory.* A basic overview and some algorithmic theory for the EM algorithm in the mixture model can be found in Redner and Walker (1984). In addition to the simplicity of construction in the mixture model, which is very easy to program, it must be stated that it increases the likelihood at every step and so is quite reliable at finding local maxima of the likelihood.

However, one key feature of the algorithm is that it commonly displays a *very slow* linear rate of convergence, where the rate constant is related to the amount of information in the missing portion of the data. If the components are similar in their densities, then the convergence is extremely slow. The convergence will also be slow when the maximum likelihood solution requires some of the weight parameters to be zero, because the algorithm can never reach such a boundary point. Although this is not so severe a problem for computing point estimates with modern computers, it can make simulation studies quite tedious.

An additional and related problem is that of deciding when to stop the algorithm. One risk to the naive user is the natural tendency to use a *stopping rule* for the algorithm based on the changes in the parameters or the likelihood being sufficiently small. Unfortunately, taking small steps does not mean we are close to the solution. If we were to continue, we might end up taking *many* more steps of nearly the same size and arriving at the solution a long distance away.

To combat this problem, Böhning, Dietz, Schaub, Schlattman, and Lindsay (1994) exploit the regularity of the EM algorithmic process to predict, via the device known as *Aitken acceleration*, the value of the log likelihood at the maximum likelihood solution. This method is suitable whenever one is using a linearly convergent algorithm with a slow rate of convergence. If l_{i-2}, l_{i-1} and l_i are the log likelihood values for three consecutive steps of the algorithm, then the predicted final value is

$$l_i^{\infty} = l_{i-2} + \frac{1}{1-c_i}(l_{i-1} - l_{i-2}) \quad \text{where } c_i = \frac{l_i - l_{i-1}}{l_{i-1} - l_{i-2}}.$$

If the algorithm is moving slowly, then c_i, an estimate of the rate, will be close to 1, and l_i^{∞} will be substantially larger than l_i.

Böhning, Dietz, Schaub, Schlattman and Lindsay (1994) used this device to carry out a simulation study of the likelihood ratio test for one component versus two components. It enabled them to predict the final maximized likelihood with many fewer iterations. The Aitken accelerated value can also be used to construct a stopping rule that more adequately captures the desired numerical accuracy than the usual "lack-of-progress" stopping rule criterion

$$\text{stop if } l_i - l_{i-1} < \text{tol},$$

where tol is a prespecified tolerance level. Provided that l_i^{∞} is a good estimator of the final likelihood, the rule

$$\text{stop if } l_i^{\infty} - l_i < \text{tol}$$

will cause the algorithm to stop only when the solution is near, and tol will more meaningfully represent the actual accuracy attained.

The idea of Aitken acceleration can be applied to the entire vector of parameters to speed up the algorithm itself [Louis (1982)]. However, as the number of parameters grows, this becomes more difficult to implement and less reliable. Other devices for speeding up the algorithm can be found in a variety of papers in the literature, proving that this is considered a serious problem and that no solution yet is completely satisfactory. Another point regarding the EM algorithm for mixtures is that, in fact, there are sometimes other ways to specify the "missing data," so there is not a unique EM algorithm. There are situations, such as the known component model, where there exist EM algorithms that are strictly superior to the one presented here, although the one here is the simplest.

As a final point, we relate the EM algorithm to the properties of the gradient function. First, we can express the EM steps for the weights as a form of gradient projection:

$$\pi_{j,\mathrm{em}} = \pi_j[1 + n^{-1}D_{Q_m}(\xi_j)].$$

The likelihood equations for the support point ξ_j can be written as

$$D'_{Q_m}(\xi_j) = 0.$$

The EM algorithm leaves the support point fixed if this equation is satisfied and will move it in the direction of increasing gradient value when it is violated. [*Exercise.*]

3.5. Multimodality and starting values. The presence of significant multimodality in the finite component likelihood has a number of important consequences.

For one thing, the solution of the likelihood equations can depend greatly on the initial values for the algorithmic method chosen. In an example given in Böhning, Schlattman and Lindsay (1992), it was found that in a particular problem in which the global maximum likelihood estimator for the two-component model had support points -1.6 and -6, starting the iterations of the EM algorithm with mass 0.5 at the points 0 and one of $-1, -2, -3, -4$ or -5 led to a second local maximum that had support points 0 and -1.8.

A second consequence of multimodality is that Newtonian-type algorithmic methods for obtaining solutions can be very unstable [Finch, Mendell and Thode (1989)].

Yet another important implication, one that will arise later in the consideration of the likelihood ratio test, is that the results obtained from a simulation study can be highly dependent on the stopping rules and search strategies employed. Indeed, this can make it quite difficult to compare the results of simulation studies or assess their reliability.

It is widely considered desirable to find the global maximum to the likelihood. If this is the goal, then one can adopt a number of strategies. One approach, suggested in Böhning, Schlattman and Lindsay (1992), is to calculate the nonparametric maximum likelihood estimator of Q, which can be done unambiguously. If it has more components than desired, one can often reasonably choose a way to consolidate nearby support points in a way that the likelihood stays near its maximum value. The use of this method was illustrated in the aforementioned example, and it was found that the four point NPMLE gave a good prediction of the best two-component MLE.

On the other hand, if the NPMLE has fewer components than desired, then this indicates that it will be *impossible* to find a maximum likelihood estimator with the desired number of components. (In such a situation, the EM algorithm iterations will either slowly merge the support points ξ together or force some values of π toward zero, because estimators with fewer support points have higher likelihoods than those with more.)

We should also note that this strategy for constructing initial values is not possible in some mixture models, such as the normal with unknown variance, in which a NPMLE does not exist.

Finch, Mendell and Thode (1989) suggest a strategy of multiple random starts that enables one to make a probabilistic estimate of the number of unseen modes to the likelihood. Another strategy that is possible in the two-component univariate normal is to use a normal probability plot to estimate the means and variance (or variances) from the slope and intercepts in the two tails. [See, e.g., Titterington, Smith and Makov (1985), pages 58–60.] This last approach would not be effective in a simulation study, nor would the NPMLE approach, unless they were made more systematic.

A point that is seldom raised is that it is not clear that one should insist on finding the global maximum to the likelihood. For example, if one is estimating the parameters in the normal mixture

$$\pi n(x; \mu_1, \sigma_1^2) + \bar{\pi} n(x; \mu_2, \sigma_2^2),$$

then the global maximum to the likelihood is ∞. This can be established by letting $\mu_2 = x_1$, or any other observation, and letting σ_2^2 go to zero. Then the term in the likelihood corresponding to x_1 becomes infinite while the other terms stay bounded below [Kiefer and Wolfowitz (1956)]. However, it is known that there is a consistent sequence of roots to the likelihood equation [Kiefer (1978)].

For these reasons, this author has been interested in constructing reliable and consistent estimators of the mixture parameters by *moment methods* and using these as initial values in likelihood algorithms. It is quite clear that if there is a consistent sequence of roots to the likelihood equations and we start our algorithm at the moment estimators, which will be $\sqrt{n}$ consistent, then, at least asymptotically, we should be finding a root close to the true values of the parameters. Lindsay (1989a) has an extensive development of the moment estimators, showing how they can be constructed in many important exponential family models. Among their important features is that they are unique, so that one need not hunt for the best solution, and that they are often numerically simple to compute compared to the maximum likelihood estimators.

The use of moment estimators was investigated extensively by simulation in the normal mixture model by Furman and Lindsay (1994a, b). They found that the moment estimators had high initial likelihoods, generally higher than using the true parameter values of the simulation, and that the EM iterations that were started at the true values, thereby leading to the solution of the likelihood equations closest to the truth, almost always picked the same solution as the EM iterations started at the moment estimators.

Following this successful experiment, Lindsay and Basak (1993) developed moment methods for the multivariate normal problem, which was considerably more challenging, but still resulted in a successful and numerically fast way to construct initial values.