1

Mixtures of Normals

In this chapter, I will review the mixture of normals model and discuss various methods for inference with special attention to Bayesian methods. The focus is entirely on the use of mixtures of normals to approximate possibly very high dimensional densities. Prior specification and prior sensitivity are important aspects of Bayesian inference and I will discuss how prior specification can be important in the mixture of normals model. Examples from univariate to high dimensional will be used to illustrate the flexibility of the mixture of normals model as well as the power of the Bayesian approach to inference for the mixture of normals model. Comparisons will be made to other density approximation methods such as kernel density smoothing which are popular in the econometrics literature.

The most general case of the mixture of normals model "mixes" or averages the normal distribution over a mixing distribution.

$$p(y|\tau) = \int \phi(y|\mu, \Sigma)\, \pi(\mu, \Sigma|\tau)\, d\mu d\Sigma \qquad (1.0.1)$$

Here $\pi(\,)$ is the mixing distribution. $\pi(\,)$ can be discrete or continuous. In the case of univariate normal mixtures, an important example of a continuous mixture is the scale mixture of normals.

$$p(y|\tau) = \int \phi(y|\mu, \sigma)\, \pi(\sigma|\tau)\, d\sigma \qquad (1.0.2)$$

A scale mixture of a normal distribution simply alters the tail behavior of the distribution while leaving the resultant distribution symmetric. Classic examples include the t distribution and

double exponential in which the mixing distributions are inverse gamma and exponential, respectively (Andrews and Mallows (1974)). For our purposes, we desire a more general form of mixing which allows the resultant mixture distribution sufficient flexibility to approximate any continuous distribution to some desired degree of accuracy. Scale mixtures do not have sufficient flexibility to capture distributions that depart from normality exhibiting multi-modality and skewness. It is also well-known that most scale mixtures that achieve thick tailed distributions such as the Cauchy or low degree of freedom t distributions also have rather "peaked" densities around the mode of the distribution. It is common to find datasets where the tail behavior is thicker than the normal but the mass of the distribution is concentrated near the mode but with rather broad shoulders (e.g., Tukey's "slash" distribution). Common scale mixtures cannot exhibit this sort of behavior. Most importantly, the scale mixture ideas do not easily translate into the multivariate setting in that there are few distributions on Σ for which analytical results are available (principally the Inverted Wishart distribution).

For these reasons, I will concentrate on finite mixtures of normals. For a finite mixture of normals, the mixing distribution is a discrete distribution which puts mass on K distinct values of μ and Σ.

$$p(y|\pi, \{\mu_k, \Sigma_k\}) = \sum_k \pi_k \phi(y|\mu_k, \Sigma_k) \qquad (1.0.3)$$

$\phi(\,)$ is the multivariate normal density.

$$\phi(y|\mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left\{-1/2 (y-\mu)' \Sigma^{-1} (y-\mu)\right\} \qquad (1.0.4)$$

d is the dimension of the data, y. The K mass points of the finite mixture of normals are often called the *components* of the mixture. The mixture of normals model is very attractive for two reasons: (1) the model applies equally well to univariate and multivariate settings; and (2) the mixture of normals model can achieve great flexibility with only a few components.

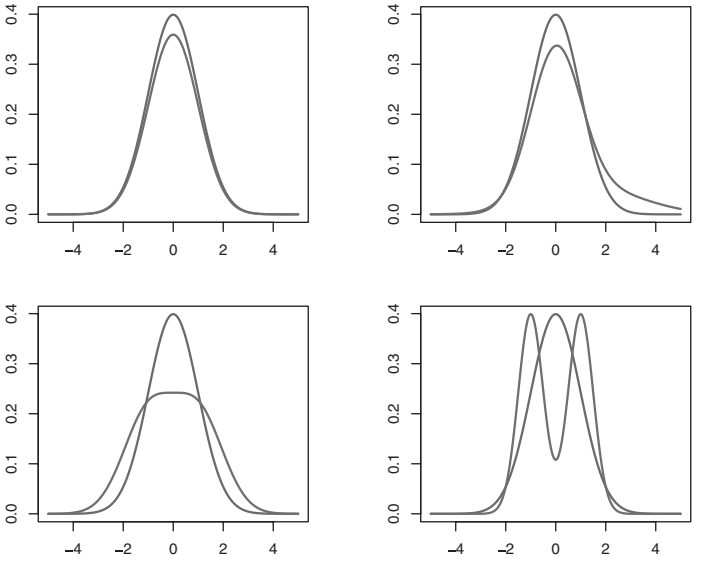

Figure 1.1. Mixtures of Univariate Normals

Figure 1.1 illustrates the flexibility of the mixture of normals model for univariate distributions. The upper left corner of the figure displays a mixture of a standard normal with a normal with the same mean but 100 times the variance (the red density curve), that is the mixture $.95\mathrm{N}(0, 1) + .05\mathrm{N}(0, 100)$. This mixture model is often used in the statistics literature as a model for outlying observations. Mixtures of normals can also be used to create a skewed distribution by using a "base" normal with another normal that is translated to the right or left depending on the direction of the desired skewness.

The upper right panel of Figure 1.1 displays the mixture, $.75\mathrm{N}(0, 1) + .25\mathrm{N}(1.5, 2^2)$. This example of constructing a skewed distribution illustrates that mixtures of normals do not have to exhibit "separation" or bimodality. If we position a number of mixture components close together and assign each component similar probabilities, then we can create a mixture distribution with a density that has broad shoulders of the type

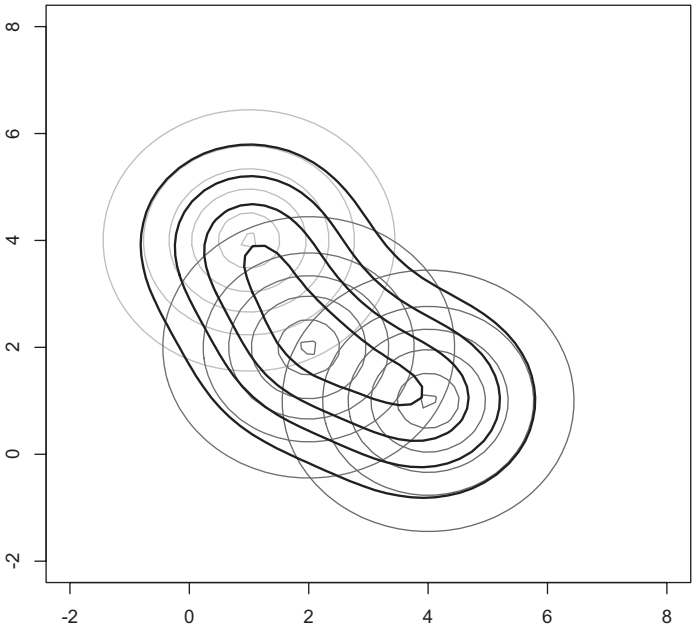

Figure 1.2. A Mixture of Bivariate Normals

displayed in many datasets. The lower left panel of Figure 1.1 shows the mixture $.5\mathrm{N}(-1, 1) + .5\mathrm{N}(1, 1)$, a distribution that is more or less uniform near the mode. Finally, it is obvious that we can produce multi-modal distributions simply by allocating one component to each desired model. The bottom right panel of the figure shows the mixture $.5\mathrm{N}(-1, .5^2) + .5\mathrm{N}(1, .5^2)$. The darker lines in Figure 1.1 show a unit normal density for comparison purposes.

In the multivariate case, the possibilities are even broader. For example, we could approximate a bivariate density whose contours are deformed ellipses by positioning two or more bivariate normal mixtures along the principal axis of symmetry. The "axis" of symmetry can be a curve allowing for the creation of a density with "banana" or any other shaped contour. Figure 1.2 shows a

mixture of three uncorrelated bivariate normals that have been positioned to obtain "bent" or "banana-shaped" contours.

There is an obvious sense in which the mixture of normals approach, given enough components, can approximate any multivariate density (see Ghosh and Ramamoorthi (2003) for infinite mixtures and Norets and Pelenis (2011) for finite mixtures). As long as the density which is approximated by the mixture of normals damps down to zero before reaching the boundary of the set on which the density is defined, then mixture of normals models can approximate the density. Distributions (such as truncated distributions) with densities that are non-zero at the boundary of the sample space will be problematic for normal mixtures. The intuition for this result is that if we were to use extremely small variance normal components and position these as needed in the support of the density then any density can be approximated to an arbitrary degree of precision with enough normal components. As long as arbitrarily large samples are allowed, then we can afford a larger and larger number of these tiny normal components. However, this is a profligate and very inefficient use of model parameters. The resulting approximations, for any given sample size, can be very non-smooth, particularly if non-Bayesian methods are used. For this reason, the really interesting question is not whether the mixture of normals can be the basis of a non-parametric density estimation procedure, but, rather, if good approximations can be achieved with relative parsimony. Of course, the success of the mixture of normals model in achieving the goal of flexible and relatively parsimonious approximations will depend on the nature of the distributions that need to be approximated. Distributions with densities that are very non-smooth and have tremendous integrated curvature (i.e., lots of wiggles) may require large numbers of normal components.

The success of normal mixture models is also tied to the methods of inference. Given that many multivariate density approximation situations will require a reasonably large number of components and each component will have a very large number of parameters, inference methods that can handle very high

dimensional spaces will be required. Moreover, the inference methods that over-fit the data will be particularly problematic for normal mixture models. If an inference procedure is not prone to over-fitting, then inference can be conducted for models with a very large number of components. This will effectively achieve the non-parametric goal of sufficient flexibility without delivering unreasonable estimates. However, an inference method that has no method of curbing over-fitting will have to be modified to penalize for over-parameterized models. This will add another burden to the user—choice and tuning of a penalty function.

1.1 Finite Mixture of Normals Likelihood Function

There are two alternative ways of expressing the likelihood function for the mixture of normals model. This first is simply obtained directly from the form of the mixture of normals density function.

$$L\left(\pi, \{\mu_k, \Sigma_k, k = 1, \ldots, K\} \,|Y\right) = \prod_i \sum_k \pi_k \phi\left(y_i | \mu_k, \Sigma_k\right) \tag{1.1.1}$$

Y is a matrix whose ith row is y_i'. A useful alternative way of expressing the likelihood function is to recall one interpretation of the finite mixture model. For each observation, an indicator variable, z_i, is drawn from a multinomial distribution with K possible outcomes each with probability π_k. y_i is drawn from the multivariate normal component corresponding to outcome of the multinomial indicator variable. That is, to simulate from the mixture of normals model is a two-step procedure:

$$\begin{aligned} z_i &\sim \text{MN}(\pi) \\ y_i &\sim \text{N}\left(\mu_{z_i}, \Sigma_{z_i}\right) \end{aligned}$$

Using this representation we can view the likelihood function as the expected value of the likelihood function given z.

$$L(\pi, \{\mu_k, \Sigma_k\} | Y) = \mathbb{E}[L(z, \mu_k, \Sigma_k)]$$
$$= \mathbb{E}\left[\prod_i \sum_k I(z_i = k)\, \phi(y_i, |\mu_k, \Sigma_k)\right] \tag{1.1.2}$$

The likelihood function for the finite mixture of normals model has been extensively studied (see, for example, McLachlan and Peel (2000)). There are several unusual features of the mixture of normals likelihood. First, the likelihood has numerous points where the function is not defined with an infinite limit (for lack of a better term, I will call these poles). In a famous example given by Quandt and Ramsey (1978), the likelihood for a mixture of two univariate normals can be driven to any arbitrary value by taking one of the means to be equal to y_i and letting σ for that mixture component go to zero.

$$\lim_{\sigma \to 0} \frac{1}{2\pi\sigma} \exp\left\{ -1/2 \left(\frac{y_i - \mu_k}{\sigma} \right)^2 \right\} |_{\mu_k = y_i} = \infty$$

This means that there are poles for every y value in the data set. Figure 1.3 plots the likelihood function for a mixture of two univariate normals and shows the log-likelihood surface around values of μ close to a particular y_i. This sort of feature may make it difficult for standard optimizers to explore the likelihood surface.

However, it is not poles that present the most difficult problem for exploring the likelihood surface using conventional optimizers that use local derivative information. The mixture of normals likelihood function has $K!$ modes, each of equal height. These modes correspond to all of the possible ways to reorder the labeling of the likelihood normal mixture component parameters. That is, there is no difference between the likelihood

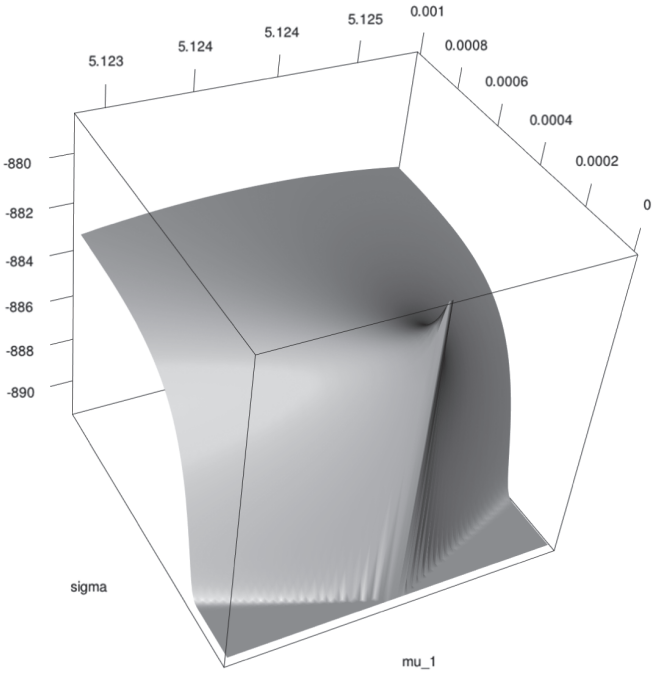

Figure 1.3. Pole in the Likelihood Function

for $\mu_1 = \mu_1^*, \mu_2 = \mu_2^*, \sigma_1 = \sigma_1^*, \sigma_2 = \sigma_2^*$ and the likelihood for $\mu_1 = \mu_2^*, \mu_2 = \mu_1^*, \sigma_1 = \sigma_2^*, \sigma_2 = \sigma_1^*$. Moreover, there are saddle points between these symmetric modes. Figure 1.4 shows what appears to be a saddle point in the likelihood of a mixture of two normals. The likelihood is only depicted in the μ_1, μ_2 space conditional on the values of the standard deviations parameters. The figure shows two local maxima near the points (2,4) and (4,2). However, if you constrain the means to be equal, there is a local maximum at the top of the saddle point near the point (1.5,1.5). This means that any standard optimizer that begins at the point of equal means (not an unreasonable starting point, for example, to start at the mean of the data for all μ parameters) will converge to local maximum that is not the global.

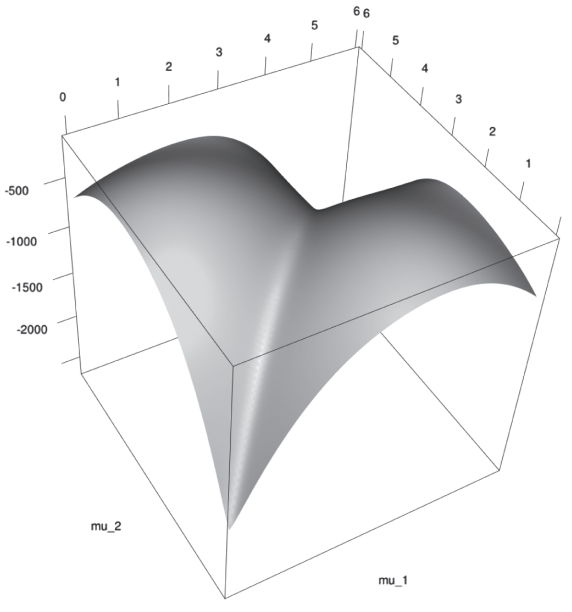

Figure 1.4. Saddle Points in the Likelihood Function

1.2 Maximum Likelihood Estimation

Given the existence of both poles and saddle points, the maximization of the mixture of normals likelihood (1.1.1) is challenging. Standard quasi-Newton optimizers will have to be started from points close to one of the $K!$ global optima in order to work well. If the mixture is designed to approximate a density in high dimensions (5 or more) and a relatively large number of components are desired for flexibility, then the number of parameters in the likelihood can be very large.

$$n_{parm} = K - 1 + K\left(\frac{d\,(d+1)}{2}\right) + Kd = o\left(d^2\right) \qquad (1.2.1)$$

d is the dimension of the data and K is the number of components in the mixture. The first term in (1.2.1) represents the number of parameters in π, the second from the set of normal component variance-covariance matrices (Σ_k), and the third from the normal component mean vectors. For example, a modest problem with $d = 5$ and $K = 10$ would feature 209 unique parameters. The sheer number of parameters does not present a computational challenge in and of itself. However, the irregular features of the mixture likelihood make high dimensional optimization difficult. Not only are the parameters large in number, but there are constraints on the parameter space imposed by the condition that all of the component covariance matrices must be positive definite symmetric. The nature of these constraints do not facilitate direct input into standard non-linear programming methods which allow for constrained optimization. Instead, it would be advisable to reparameterize to enforce symmetry and positive definiteness.

$$\begin{aligned} \Sigma_k &= U_k' U_k \\ U_k &= f(\lambda_k, \theta_k) \end{aligned} \tag{1.2.2}$$

$$f(\lambda_k, \theta_k) = \begin{bmatrix} e^{\lambda_{k,1}} & \theta_{k,1} & \cdots & \theta_{k,d-1} \\ 0 & e & \ddots & \vdots \\ \vdots & \ddots & \ddots & \theta_{k,(d-1)d/2} \\ 0 & \cdots & 0 & e^{\lambda_{k,d}} \end{bmatrix}$$

However, even with this reparameterization, the mixture of normals likelihood remains a severe challenge for standard optimization methods, particularly for cases where the dimension of the data is greater than one.

1.2.1 E-M Algorithm

Most agree that the only reliable way to compute maximum likelihood estimates for the mixture of normals model is to

employ the E-M algorithm. The E-M algorithm is particularly appropriate for those problems which can be characterized as an incomplete data problem. An incomplete data problem is one in which what we observe in our sample can be viewed as a subset of "complete" data. As we have seen, the mixture of normals model can be viewed as a sampling mechanism in which draws are made from a latent (unobserved) indicator vector, z, which indicates which of the K normal components each observation is drawn from. The complete data is then (z, y). The likelihood for the observed data is the complete data likelihood with the unobserved component integrated out. In the mixture of normals case, the integration simply consists of weighting each component by the mixing or multinomial probability and adding the components up. The E-M algorithm is an iterative procedure consisting of an "E-step" and an "M" or maximization step. Given the model parameters, the "E-step" consists of taking the expectation of the unobserved latent indicators and the "M-step" consists of maximizing the component density parameters in the expectation of the complete data likelihood function (see, for example, McLachlan and Peel (2000), section 2.8). As is well-known, the E-M algorithm provides a method by which an improvement (or at least no decrease in the likelihood) can be achieved at each step. This means that the E-M method provides a reliable, if somewhat slow, method of climbing to a local maximum in the mixture of normals likelihood.

The complete data log-likelihood can be written conveniently for application of the E-M method as follows:

$$\log\left(L_c\left(H, \Psi\right)\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} h_{ik}\left(\log\left(\pi_i\right) + \log\phi\left(y_i|\mu_k, \Sigma_k\right)\right) \tag{1.2.3}$$

h_{ik} is a matrix of indicator variables. $h_{ik} = 1$ if observation i is from component k. Ψ is the collection of all of the mixture parameters: $\pi, \{\mu_k, \Sigma_k\}$ $k = 1, \ldots, K$. The E-M method starts from an initial value, Ψ^0 , of Ψ.

E-Step: Take the expectation of the complete data log-likelihood with respect to the unobserved h_{ik} values. The expectation is taken given Ψ^0.

$$\mathbb{E}\left[\log\left(L_c\left(H,\Psi\right)\right)\right] = \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\left[h_{ik}|\Psi^0\right]\left(\log\left(\pi_i\right) + \log\phi\left(y_i|\mu_k,\Sigma_k\right)\right) \tag{1.2.4}$$

$$\mathbb{E}\left[h_{ik}|\Psi^0\right] = \tau_k\left(y_i|\Psi^0\right)$$

$$\tau_k\left(y_i|\Psi^0\right) = \pi_k^0\phi\left(y_i|\mu_k^0,\Sigma_k^0\right) / \sum_{j=1}^{K}\pi_j^0\phi\left(y_i|\mu_j^0,\Sigma_j^0\right) \tag{1.2.5}$$

M-Step: Maximize the expectation of the log-likelihood computed in 1.2.4 to form new estimates of the component mixture parameters, Ψ^1.

$$max_\Psi Q\left(\Psi|y\right) = \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_k\left(y_i|\Psi^0\right)\left(\log\left(\pi_i\right) + \log\phi\left(y_i|\mu_k,\Sigma_k\right)\right) \tag{1.2.6}$$

The solutions to the maximization problem are simply weighted averages of means and covariance matrices:

$$\pi_k^1 = \sum_{i=1}^{n}\tau_k\left(y_i|\Psi^0\right)/n \tag{1.2.7}$$

$$\mu_k^1 = \sum_{i=1}^{n}\tau_k\left(y_i|\Psi^0\right)y_i / \sum_{i=1}^{n}\tau_k\left(y_i|\Psi^0\right) \tag{1.2.8}$$

$$\Sigma_k^1 = \sum_{i=1}^{n}\tau_k\left(y_i|\Psi^0\right)\left(y_i-\mu_k^1\right)\left(y_i-\mu_k^1\right)' / \sum_{i=1}^{n}\tau_k\left(y_i|\Psi^0\right) \tag{1.2.9}$$

Thus, the E-M method is easy to program and involves only evaluating the normal density to compute the probability[1] that each observation belongs to each of the K components (1.2.5) and simple updates of the mean and covariance matrices.

The E-M method is not complete without advice regarding the choice of a starting point and a method for computing an estimate of the sampling error. Given that the E-M method can be very slow to converge, it is important to choose reasonable starting points. Some advise clustering the data using standard clustering methods and then using the cluster proportions, means, and covariance matrices as a starting point. Regarding the computation of standard errors for the parameter estimate, it appears the most practical approach would be to start a Quasi-Newton optimizer from the last E-M iterate value and use the Hessian estimate to form an approximate information matrix value which can be used for the standard asymptotic normal approximation to the sampling distribution of the MLE.

In many applications, the mixture of normals density approximation will only be one part of a larger model where the emphasis will be on inference for other model parameters. For example, suppose we are to use mixture of normals as the distribution of a regression error term. In that case, our interest is not regarding the error term density parameters but on parameters governing the regression function. Having an MLE procedure (however reliable) for the density parameters is only useful as part of a more general estimation procedure.

Even if our goal is simply to estimate the density of the data, the asymptotic distribution of the MLE for mixture of normal parameters is not directly useful. We will have to compute the asymptotic approximation to the density ordinates.

$$p\left(y|\hat{\Psi}_{MLE}\right) = f\left(\hat{\Psi}_{MLE}|y\right)$$

[1]Note that this is the posterior probability of component membership conditional on the normal mixture parameters.

Either the parametric bootstrap or the delta method would be required to obtain an asymptotic approximation to the distribution of the density ordinate and this asymptotic approximation would have to be computed for each potential value of the density ordinate.

Another major problem with a maximum likelihood approach is that the likelihood function will always increase as K increases. This illustrates the "greedy" nature of the MLE approach in which estimators are chosen via minimization of a criterion function (log-likelihood), namely that any increase in flexibility will be rewarded. At its most ridiculous extreme, a mixture of normals that allocates one component to each observation will have the highest likelihood value. In practice, this property of the MLE results in over-fitting. That is, the MLE will attempt to allocate components to tiny subsets of the data in order to fit anomalous values of the data. This propensity of m-estimators for chasing noise in the data is well-known. In order to limit this problem, procedures are used to either "test" for the number of components or to penalize models that offer slight improvements in fit at the expense of many additional parameters. In the mixture of normals literature, both the AIC and BIC criteria have been proposed to help keep the number of components small and to choose among models with differing numbers of components. The BIC criteria has been derived as an approximate Bayes Factor using asymptotic arguments.

In summary, the mixture of normals model provides a formidable challenge to standard inference methods. Even though there is a well-defined likelihood, maximization of this likelihood is far from straightforward. Even abstracting from the practical numerical difficulties in fitting high dimensional mixture of normals models, the problem of over-fitting still must be overcome. Ad hoc criteria for selecting the number of mixture components do not solve the over-fitting problem. What would be desirable is an inference procedure that is numerically reliable, not prone to over-fitting, provides accurate and easy to compute inference methods, and can be easily made a part of a more complicated model structure. The Bayesian methods

discussed in the next section of this chapter provide a solution to many of these problems.

1.3 Bayesian Inference for the Mixture of Normals Model

The likelihood function for the mixture of normals presents serious challenges for any estimator based on minimization of a criterion function (such as the MLE). Not only is it difficult to find roots of the score function, but in the normal mixtures problem the parameters, in most cases, do not have a direct meaning. Rather the model is fit with an interest in making inferences regarding an unknown joint density of the data. In a Bayesian setting, the "density estimation" problem is viewed as the problem of computing the predictive distribution of a new value of y. That is, we have an observed data set, Y (the matrix of observations, a $n \times d$ matrix). We assume that these observations are iid draws from a common but unknown distribution. Inferences are desired for the distribution of a y value drawn from the unknown population distribution given the observed data. The predictive density requires that we integrate over the posterior distribution of the unknown parameters, θ.

$$p(y|Y) = \int p(y|\theta)\, p(\theta|Y)\, d\theta \tag{1.3.1}$$

1.3.1 shows that the parameters are merely devices for implementing a solution to the density estimation-inference problem. What is desired is a method that will allow us to make inferences regarding the posterior distribution of the joint density at any particular value of y. It turns out that mixture of normals model is a model which is particularly well-suited to computationally tractable and accurate approximations to the posterior distribution of density ordinates. Simulation-based methods will be used to navigate the parameter space and avoid the problems associated with derivative-based procedures.