Analysis of Mixture Models using Expected Posterior Priors, with Application to Classification of Gamma Ray Bursts

José M. Pérez and James O. Berger *

Abstract

Consider observations distributed according to the mixture of component densities with different parameters. In the Bayesian framework, it is not possible to perform a *default* statistical analysis of the mixture using an improper prior for the component parameters, since the posterior distribution does not exist. To overcome this difficulty, we propose use of the expected posterior prior approach of Pérez and Berger (1999). Besides providing suitable default priors for general mixture models, a key advantage of the use of expected posterior priors here is that they can be used in conjunction with Markov Chain Monte Carlo methods, even when the number of components is unknown. An application is considered involving Gamma Ray Bursts, modeled as arising from a bivariate normal mixture model with measurement errors on the observations.

Keywords: MIXTURE MODELS; DEFAULT PRIORS.

1 Introduction

Mixture models have been used in situations including pseudo-parametric density estimation, clustering, change point problems and image analysis (McLachlan and Basford, 1985; Escobar and West, 1995; Roeder and Wasserman,1997). In Bayesian analysis, mixture models are typically analyzed using Markov Chain Monte Carlo simulation (MCMC). See, for example, Diebolt and Robert (1994), Escobar and West (1995) and Richardson and Green (1998).

Unfortunately, with mixture models it is typically not possible to perform default analysis with standard noninformative priors. In this work, we demonstrate that the expected posterior prior approach (Pérez and Berger, 1999) can be used to provide a default Bayesian analysis for this problem. In section 2, we define the expected posterior prior for mixture models. In section 3, a reversible jump MCMC scheme

*James O. Berger, Institute of Statistics and Decision Sciences (ISDS), Duke University, NC, U.S.A. José M. Pérez, Centro de Estadística y Software Matemático (CESMa), Universidad Simón Bolívar, Aptdo. 89000, Caracas 1080A, Venezuela. This work formed part of the second authors Ph.D. thesis at the Department of Statistics, Purdue University. This work was supported by the National Science Foundation (USA), Grants DMS-9303556 and DMS-9802261.

(Richardson and Green, 1998) is presented for computation in mixture models with these priors. Application involving a bivariate normal mixture is considered in section 4 in the determination of the number of classes of Gamma Ray Bursts (Meegan *et al.* , 1995).

1.1 Notation

Consider the finite mixture model, where observations are distributed independently as

$$Y_i \sim \sum_{j=1}^{k} w_j f(\cdot \,|\, \theta_j), \tag{1}$$

for $i = 1, \ldots, n$. Here, k represents the unknown number of components. Conditional on k, the weights of the mixture are given by $\boldsymbol{w}_{(k)} = (w_1, \ldots, w_k)$, where w_j is the probability of an observation coming from component j and $\sum_j w_j = 1$. As indicated by the notation, we assume that each component of the mixture is a density of a given parametric form, but with different values of the parameters θ_j (typically multivariate). Let $\boldsymbol{\theta}_{(k)} = (\theta_1, \ldots, \theta_k)$. (Strictly, we should write the components of $\boldsymbol{w}_{(k)}$ and $\boldsymbol{\theta}_{(k)}$ as w_{jk} and θ_{jk}, but we will suppress the k index to prevent notational overload.)

For convenience, it is often useful to define latent variables z_i, $i = 1, \ldots, n$, representing the unknown allocation of the observations into each component. In this way, $z_i = j$ indicates that observation i comes (uniquely) from component j.

1.2 BATSE gamma ray bursts

The proposed method will be applied in 4 to analysis of the third BATSE catalog of gamma ray bursts. The third BATSE catalog contains 1122 measurements performed by the Compton Gamma Ray Observatory between 1991 and 1994 (Meegan *et al.* , 1995). We utilize only the 745 observations that have no missing data and which have not been flagged as being possibly erroneous in some way. The catalog also includes the known standard deviations of the measurement errors for the duration and hardness ratio of each burst, and these will be taking into account in the analysis in section 4 (with errors adjusted to the log-scale).

Of interest here is classification of the gamma ray bursts in terms of their duration and hardness ratio. Hurley (1991) observed a bimodal distribution for the duration, in the log scale. A correlation between the hardness ratio and the duration had also been observed. It is thus natural to attempt to classify the gamma ray bursts based on the bivariate data.

We assume that $(\log(\mathrm{T90})_i, \log(\mathrm{HR})_i) = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}_i$, where the $\boldsymbol{\epsilon}_i$ are normal with mean zero and the standard deviations given in the BATSE catalog, and the $\boldsymbol{\mu}_i$ are a mixture of bivariate normal distributions with completely unknown parameters. The number of components of the mixture, k, is also considered unknown. This combination of a completely unknown mixture of bivariate normal distributions, with observations carrying measurement errors, has not, to our knowledge, been previously considered.

1.3 Mixture model and prior structure

The priors for the parameters of the mixture are typically defined in a hierarchical setup (Richardson and Green, 1998). The joint density of all variables can be written in a hierarchical form as

$$p(k, \boldsymbol{w}_{(k)}, \boldsymbol{z}, \boldsymbol{\theta}_{(k)}, \boldsymbol{y}) = p(k) p(\boldsymbol{w}_{(k)} \mid k) p(\boldsymbol{z} \mid \boldsymbol{w}_{(k)}, k) \pi(\boldsymbol{\theta}_{(k)}) f(\boldsymbol{y} \mid \boldsymbol{\theta}_{(k)}, \boldsymbol{z}),$$

where $\boldsymbol{z} = (z_1, \ldots, z_n)$. The following structure for the various probability distributions is assumed:

- $p(k)$ is the probability that there are k components; this is chosen to be uniform on $\{1, \ldots, K\}$ in our default analysis.
- $p(\boldsymbol{w}_{(k)} \mid k)$ is a Dirichlet distribution with known parameter $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_0)$; in our default analysis, $\alpha_0 = 1$, resulting in a uniform distribution for $\boldsymbol{w}_{(k)}$.
- The allocation variables z_i, $i = 1, \ldots, n$, are i.i.d. with
$$p(z_i = j \mid \boldsymbol{w}_{(k)}, k) = w_j. \tag{2}$$
- The likelihood $f(\boldsymbol{y} \mid \boldsymbol{\theta}_{(k)}, \boldsymbol{z}) = \prod_1^n f(\boldsymbol{y}_i \mid \theta_{z_i})$; in our examples, f will be a normal density (of the appropriate dimension), and θ_j will denote the unknown mean and covariance parameters corresponding to component j of the mixture.
- Choice of the prior for the component parameters, $\pi(\boldsymbol{\theta}_{(k)})$, is the major focus of this paper.

1.4 Background

The component parameter priors, as specified above, must be proper or the posterior distributions will themselves be improper (Diebolt and Robert, 1994; Shui, 1996; Roeder and Wasserman, 1997). The large number of parameters that typically arise in mixture models makes it very difficult to assess subjective proper priors. (Also, in our major example, the gamma ray bursts are completely unknown phenomena and so no subjective information concerning them is available.) It is thus desirable to have default or automatic priors for analysis of mixture models. Unfortunately, standard default priors are typically improper, and so cannot be used directly in the analysis.

In order to avoid this problem, Diebolt and Robert (1994) changed the probability distribution of $\boldsymbol{z}$ in (2) so that each component contains sufficient observations so that the marginal is finite. The method has seemed to be successful in practice, but might be viewed by some as a rather adhoc fix.

The use of arbitrary "vague proper priors" does not solve the problem, as has been pointed out by many (Diebolt and Robert, 1994; Shui, 1996). Thus, for the normal univariate case, Richardson and Green (1998) suggest that the degree of vagueness be chosen in a data-dependent way. Although also a bit adhoc, the results produced by this approach seem to be sensible. It is unclear, however, how to make this a general approach, especially when the mixture components can be higher dimensional distributions.

Several other approaches have been proposed in the literature (Robert, 1996; Mengersen and Robert, 1996; Roeder and Wasserman, 1997; Shui, 1996). Some of these approaches utilize a prior structure in which the components are linked to a common parameter with a flat prior, so that all observations contribute to its estimation. Shui (1996) proposes use of resampling from the observations in order to produce versions of the Intrinsic Bayes Factor (Berger and Pericchi, 1996) and the Fractional Bayes Factor (O'Hagan, 1995) which can then be used for model selection.

2 Expected posterior priors

Expected posterior priors (or EP priors) were introduced in Pérez and Berger (1999) as a means to develop priors suitable for comparison among models M_k, $k = 1, \ldots, K$, having unknown parameters $\boldsymbol{\theta}_{(k)}$, when the standard noninformative priors, $\pi^N(\boldsymbol{\theta}_{(k)})$, are unsuitable because of impropriety. The EP prior approach considers (possibly imaginary) *training samples*, y^*, such that $\pi^N(\boldsymbol{\theta}_{(k)} \mid y^*)$ is a proper density for all k. It is recommended that the size of the training samples be as small as possible while guaranteeing the propriety of these posteriors. The space of possible training samples, $\mathcal{Y}^*$, is assumed to have a density $m^*(y^*)$, common to all models under consideration, and the EP prior for $\boldsymbol{\theta}_{(k)}$ is defined as

$$\pi^*(\boldsymbol{\theta}_{(k)}) = \int \pi^N(\boldsymbol{\theta}_{(k)} \mid y^*) m^*(y^*) dy^*.$$

Even if $m^*(y^*)$ is improper, the EP priors will be "well calibrated" for model comparison, because they are all based on the same $m^*(y^*)$ (convolved with proper distributions). For further discussion of this, see Pérez and Berger (1999).

Since choice of k in mixture models can be viewed as model selection, EP priors can be directly applied to mixture models with improper priors. Thus we consider the mixture model in (1), and it is natural to assume that the initial prior for $\boldsymbol{\theta}_{(k)}$, given k, is a noninformative improper prior of the form $\pi^N(\boldsymbol{\theta}_{(k)}) = \prod_{j=1}^k \pi^N(\theta_j)$. Let K be the maximum number of components and consider imaginary training samples $\boldsymbol{y}^* = (y_1^*, \ldots, y_K^*)$, where y_j^* is viewed as being an imaginary sample arising from component j of the mixture and is of minimal size such that $\pi^N(\theta_j \mid y_j^*)$ exists. $\mathcal{Y}^*$ will denote the space of such imaginary minimal training samples.

Definition 2.1 *Let m^* be a density on $\mathcal{Y}^*$. The expected posterior prior for $\boldsymbol{\theta}_{(k)}$, conditional on k, is given by*

$$\pi^*(\boldsymbol{\theta}_{(k)}) = \int (\prod_{j=1}^{k} \pi^N(\theta_j \mid y_j^*)) m^*(\boldsymbol{y}^*) d\boldsymbol{y}^*, \tag{3}$$

provided the integral exists for all $\boldsymbol{\theta}_{(k)}$ and all $k = 1, \ldots, K$.

Pérez and Berger (1999) consider a variety of possible choices of the marginal m^*. The two main suggestions from that paper are the *empirical marginal* and the *base model marginal.* The *empirical marginal* is defined by randomly drawing training samples, with replacement, from the actual observations. (One could also do this without replacement, but this is somewhat more difficult computationally.) The empirical marginal is a proper discrete distribution, so that $p^*(\boldsymbol{\theta}_{(k)})$ is clearly proper. (The integral in (3) is, of course, then to be interpreted as a sum.)

The *base model marginal* is defined as the marginal distribution arising from the simplest model under consideration. The simplest model here corresponds to $k=1$ (i.e., the mixture consisting of a single component), and the resulting predictive for the training samples is

$$m^*(y_1^*,\ldots,y_K^*) = \int f(y_1^*,\ldots,y_K^* \,|\, \theta_1)\pi^N(\theta_1)d\theta_1. \tag{4}$$

Example 2.1 *Consider the mixture model in which the component distributions are* $N(\cdot\,|\,\mu_i,1)$ *and* $K=2$. *The usual noninformative priors are* $\pi^N(\mu_j)=1$, $j=1,2$. *In this case,* y_j^*, *a minimal imaginary training sample for component* j, *consists of a single observation. The marginal in (4) thus becomes*

$$m^*(y_1^*,y_2^*) = \int f(y_1^*,y_2^* \,|\, \mu_1,1)d\mu_1 = \frac{1}{2\sqrt{\pi}}\exp\{-(y_1^*-y_2^*)^2/4\}.$$

Given $k=1$, *the EP prior for* μ_1 *under this base model marginal can easily be seen to be the original nonininformative prior* $\pi^N(\mu_1)=1$. *Given* $k=2$, *the EP prior for* (μ_1,μ_2) *under this base model marginal is*

$$\int \pi^N(\mu_1\,|\,y_1^*)\pi^N(\mu_2\,|\,y_2^*)m^*(y_1^*,y_2^*)dy_1^*dy_2^* = \frac{1}{2\sqrt{2\pi}}\exp\{-(\mu_1-\mu_2)^2/8\}.$$

Note that this can be written as $\pi^N(\mu_1)\pi^*(\mu_2\,|\,\mu_1)$, *where* $\pi^*(\mu_2\,|\,\mu_1)$ *is a normal density with mean* μ_1 *and variance* 4. *As indicated earlier, the impropriety of EP priors in examples such as this is not a cause for concern, in that they are appropriately calibrated between different* k.

This type of prior was considered in Roeder and Wasserman (1997), therein named a "partially proper prior". For more than two components, however, the EP priors differ from the partially proper priors of Roeder and Wasserman (1997). The EP prior for more than two components was also considered in Moreno and Liseo (1998), although they developed the prior from a different perspective.

We do not review all of the attractive properties of EP priors discussed in Pérez and Berger (1999), but do highlight two of the properties that are particularly relevant to mixture models. First, the proposed EP priors are marginally coherent, in the sense that $\int \pi^*(\theta_1,\ldots,\theta_k)d\theta_k = \pi^*(\theta_1,\ldots,\theta_{k-1})$. This can easily be shown to hold for either the empirical or base model choices of m^*.

A second interesting property is that, for a single observation y, the marginal densities corresponding to the EP priors, for different k, are equal. This follows directly from the facts that we assume that the components of the mixture are of the same parametric form, and that the marginal EP priors for each θ_j are the same. The interest in this property is that it suggests that the EP priors are indeed *predictively matched* across k, as discussed in, say, Berger *et al.* (1998).

3 MCMC estimation of the mixture model

MCMC techniques have been widely used for estimation of mixture models when the number of components, k, is fixed (Diebolt and Robert, 1994; Escobar and

West, 1995; among others). In this setting, the MCMC sampler would generate a chain $\{X_1, \ldots, X_n\}$, where the state variable X consists of the model parameters, $X = (\boldsymbol{w}_{(k)}, \boldsymbol{\theta}_{(k)}, \boldsymbol{z})$. The samples can then be used to estimate posterior probabilities and expectations. However, when k is unknown, the state variable, $X = (k, \boldsymbol{w}_{(k)}, \boldsymbol{\theta}_{(k)}, \boldsymbol{z})$, assumes values among spaces of different dimensions, and so, there is not an underlying measure for the Markov chain.

One solution to this problem, the reversible jump MCMC, is given in Green (1995). Here, the author proposes a random-sweep, based on the Metropolis-Hastings method, that will allow the chain to "jump" between different dimensional spaces. Let x denote the current state of the Markov chain and consider a countable family of move types, indexed by r. Within the reversible jump MCMC framework, a new state is proposed by choosing a move type r and a destination x', with joint density $q_r(x, dx')$, where $q_r(x, \cdot)$ is arbitrary for the moment. The chain will move to x' with acceptance probability

$$\alpha_r(x, x') = \min\left\{1, \frac{\pi(dx')q_r(x', dx)}{\pi(dx)q_r(x, dx')}\right\}. \tag{5}$$

The ratio in (5) is the Radon-Nykodym derivative with respect to a "dimension-balancing measure". A sufficient condition on q_r for the existence of this measure is given in Green (1995).

For moves where x and x' have the same dimension, the algorithm described above reduces to usual Metropolis-Hastings algorithm. To jump from x to a higher dimension point x', one will typically generate continuous random variables u and set $x' = x'(x, u)$ using an invertible (deterministic) function. The new state x' is accepted with probability (5), given by

$$\min\left\{1, \frac{\pi(x')p_r(x')}{\pi(x)p_r(x)q(u)} J_{x'}(x, u)\right\},$$

where $p_r(x)$ is the probability of choosing move type r when in state x, $q(u)$ is the density of u and $J_{x'}(x, u)$ is the Jacobian of the transformation $x'(x, u)$.

3.1 MCMC for mixture models under EP priors

Given observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ and an initial state $X = (k(t), \boldsymbol{w}_{(k)}(t), \boldsymbol{\theta}_{(k)}(t), \boldsymbol{z}(t))$, a reversible jump MCMC algorithm for the mixture model can be implemented to produce inferences from the posteriors arising from the EP priors for mixtures. The key idea is to use $m^*(y_1^*, \ldots, y_K^*)$ as a prior density for the imaginary training samples, $\boldsymbol{y}^* = (y_1^*, \ldots, y_K^*)$, and, so, incorporate the generation of the training samples alongside the other parameters in the simulation. Hence, the states of the Markov chain are given by $X = (k, \boldsymbol{w}_{(k)}, \boldsymbol{\theta}_{(k)}, \boldsymbol{z}, \boldsymbol{y}^*)$

With some abuse of notation, the joint density for the observations $\boldsymbol{y}$, the parameters $(\boldsymbol{w}_{(k)}, \boldsymbol{\theta}_{(k)}, \boldsymbol{z})$ and the training samples $\boldsymbol{y}^*$ can be written as

$$\begin{aligned} p(k, \boldsymbol{w}_{(k)}, \boldsymbol{z}, \boldsymbol{\theta}_{(k)}, \boldsymbol{y}, \boldsymbol{y}^*) = \quad & p(k)p(\boldsymbol{w}_{(k)}|k)p(\boldsymbol{z}|\boldsymbol{w}_{(k)}, k)m^*(\boldsymbol{y}^*) \\ & \times \textstyle\prod_{j=1}^k \pi^N(\theta_j|y_j^*)f(\boldsymbol{y}_j|\theta_j). \end{aligned}$$

To improve the performance of the algorithm, an allocation variable for the training samples is introduced. Let $l = (l_1, \ldots, l_k)$, such that $l_j = i$ indicates that training sample y_i^* is (uniquely) assigned to component j. Denote by $\boldsymbol{y}^*(l) = (y_{l_1}^*, \ldots, y_{l_k}^*)$ and by $p(l|k)$ the uniform distribution over all possible assignments of $y_1^*, \ldots, y_K^*$ into k components.

For m^* chosen as in (4), the reversible jump MCMC algorithm is described as follows

1. Generate weights from $\boldsymbol{w}_{(k)}(t+1)|k = k(t)$, $\boldsymbol{z}(t)$, $\boldsymbol{y} \sim$ Dirichlet$(\alpha_0 + n_1, \ldots, \alpha_0 + n_k)$, where n_j is the number of observations in component j.

2. Generate component parameters from $\boldsymbol{\theta}_{(k)}(t+1)|k = k(t), \boldsymbol{z}(t), \boldsymbol{y}^*(l), \boldsymbol{y}$, with density $\prod_{j=1}^{k} \pi^N(\theta_j|\boldsymbol{y}_j, y_{l_j}^*)$.

3. Generate allocations from $\boldsymbol{z}(t+1)|k = k(t), \boldsymbol{w}_{(k)}(t+1), \boldsymbol{\theta}_{(k)}(t+1), \boldsymbol{y}$, with probability $p(z_i = j|k, \theta_j, w_j, \boldsymbol{z}, \boldsymbol{y}) \propto w_j f(\boldsymbol{y}_j|\theta_j)$.

4. Generate $\boldsymbol{y}^*(-l) = \{y_i^* : i \neq l_j, j = 1, \ldots, k(t)\}$ at time $t+1$, conditional on $y^*(l)$, with density $m^*(\boldsymbol{y}^*(-l)|\boldsymbol{y}^*(l)) = m^*(\boldsymbol{y}^*)/m^*(\boldsymbol{y}(l))$.

5. Generate $\boldsymbol{y}^*(l)|\boldsymbol{\theta}_{(k)}(t+1), l$ at time $t+1$.

6. Generate new assignments $l_j|\theta_j, l_1, \ldots, l_{j-1}, l_{j+1}, \ldots, l_k$ at time $t+1$, for $j = 1, \ldots, k$, with probabilities $p(l_j = i|\cdots) \propto \pi^N(\theta_j|y_i^*)p(l|k)$.

7. Generate $k(t+1)|\boldsymbol{w}_{(k)}(t+1), \boldsymbol{\theta}_{(k)}(t+1), \boldsymbol{z}(t+1), \boldsymbol{y}$.

Steps 1 and 3 can be carried using the Gibbs kernels described above.

In step 5, a new set of training samples, $y_{l_1}^*, \ldots, y_{l_{k(t)}}^*$, is generated at simulation time $t+1$, using a Metropolis-Hastings move, with proposal distributions $f(y_{l_j}^*|\theta_j(t))$, $j = 1, \ldots, k(t)$. The remaining training samples, $\boldsymbol{y}^*(-l)$, are generated in step 4 from their "posterior" $m^*(\cdot|\boldsymbol{y}^*(l))$. Step 6 uniquely assigns the training samples at time $t+1$ to the components.

Step 7 involves changing k by adding or deleting one component, and updating $\boldsymbol{w}_{(k)}$, $\boldsymbol{\theta}_{(k)}$, $\boldsymbol{z}$ and the assignment of the training samples, l, accordingly. At this point the reversible jump mechanism is needed. As in Richardson and Green (1998), we consider splitting or combining with probabilities c_k and $s_k = 1 - c_k$ respectively. Clearly, $s_K = c_1 = 0$.

A split move will produce components j_1 and j_2 by splitting a single component j_0 in two. The weight and component parameters for these two components are obtained by independently generating a random variable u, of the same dimension as (w_{j_0}, θ_{j_0}), and deterministically assigning $(w_{j_1}, \theta_{j_1}, w_{j_2}, \theta_{j_2}) = T(w_{j_0}, \theta_{j_0}, u)$, where T is an invertible, everywhere differentiable function. The allocation variables of those y_i such that $z_i = j_0$ are randomly set to either j_1 or j_0, with probabilities determined by the Gibbs kernel given $(w_{j_1}, \theta_{j_1}, w_{j_2}, \theta_{j_2})$.

A combine move entails merging two randomly chosen components j_1 and j_2 into a new component j_0. The allocation variables are simply updated by setting $z_i = j_0$ for all y_i such that z_i is either j_1 or j_2. The rest of the parameters, (w_0, θ_0), are obtained deterministically from the inverse of T at $(w_{j_1}, \theta_{j_1}, w_{j_2}, \theta_{j_2})$.

The jump from k to $k+1$ components in step 7, is done conditional on all the training samples $y_1^*(t), \ldots, y_K^*(t)$. In this way, splitting a component j_0 into j_1 and j_2 only involves choosing training samples $y_{l_{j_1}}^*$ and $y_{l_{j_2}}^*$ from the available pool of training samples $(\boldsymbol{y}^*(-l), y^*_{l_{j_0}})$. Similarly, the combine move for the training samples only entails choosing $y^*_{l_{j_0}}$ from the pool of unassigned training samples $(\boldsymbol{y}^*(-l), y^*_{l_{j_1}}, y^*_{l_{j_2}})$. This effectively avoids introducing unknown constants in the acceptance/rejection probabilities for the move.

In section 4 we consider mixtures of normal distributions, where the observations in the jth component have a bivariate normal density with parameters $\theta_j = (\boldsymbol{\mu}_j, \Sigma_j)$. The parameters $\boldsymbol{\mu}_j$ and Σ_j correspond to the location and covariance matrix for the first and second coordinates of the bivariate observations, respectively. The implementation of the reversible jump MCMC in this section is straight forward. The jump function T is similar to the jump function in the univariate case given in Richardson and Green (1998), but here the location parameters are split in a random direction from the center $\boldsymbol{\mu}_{j_0} = (\mu_{1j_0}, \mu_{2j_0})$. Similarly, auxiliary variables are introduced to split the component variances and correlation in an invertible way. The actual jump function T is specified in full detail in Pérez (1998) in full detail.

The mixture model components are permutation invariant, and hence their parameters are not identifiable. Although not completely satisfactory, for inference concerning individual components, one usually imposes constraints on the parameter space (see, for example, Robert, 1996; Mengersen and Robert, 1996 and Richardson and Green, 1998.) When the constraints are not appropiate, the posterior estimates of the parameters resulting from the MCMC simulation may produce multimodal marginal posteriors. This multimodality in the estimated posteriors is especially noticeable when the posterior variances are not kept small. This is, unfortunately, the case when EP priors are used, as the posterior variance for empty components is not controlled. However, if the number of observations, n, is large, the probability of observing empty components is reduced and use of a set of constraints is reasonable.

4 BATSE gamma ray burst data set

We now turn to analysis if the gamma ray burst data set discussed in the introduction, utilizing a bivariate normal mixture modes (with all parameters unknown) for log(T90) and log(HR). Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}) = (\log(\mathrm{T90})_i, \log(\mathrm{HR})_i)$ denote the ith observation and $\boldsymbol{\sigma}_i = (\sigma_{i1}, \sigma_{i2}) = (\sigma_{T90_i}, \sigma_{HR_i})$ the standard deviation of the associated observational errors. Then x_{i1} and x_{i2} are assumed to have independent normal distributions, with means y_{i1} and y_{i2} and variances σ^2_{i1} and σ^2_{i2}, respectively, where the y_{ij} represent the true values of log(T90) and log(HR) for observation i. The true values $\boldsymbol{y}_i = (y_{i1}, y_{i2})$ are assumed to arise from a mixture of k bivariate normal distributions, having means $\boldsymbol{\mu}_j$ and covariance matrices Σ_j, $j = 1, \ldots, k$. Thus,

$$\boldsymbol{x}_i \sim N(\boldsymbol{x}_i | \boldsymbol{y}_i, \boldsymbol{\sigma}_i) \quad \text{and} \quad \boldsymbol{y}_i \sim \sum_{j=1}^{k} w_j N(\boldsymbol{y}_i | \boldsymbol{\mu}_j, \Sigma_j),$$

where $N(\cdot|\boldsymbol{\mu}, \Sigma)$ denotes the normal density with location $\boldsymbol{\mu}$ and covariance matrix Σ. The initial prior distributions for the component parameters are chosen as in section 1.3, with $\alpha_0 = 1$ and $p^N(\boldsymbol{\mu}_j, \Sigma_j) = |\,\Sigma_j\,|^{-3/2}$.

Base model EP priors		Empirical EP priors	
Component 1		Group 1	
$\hat{w}_1 = 0.24$		$\hat{w}_1 = 0.24$	
$\hat{\mu}_{T90}$= -0.85	$\hat{\mu}_{HR}$= 1.61	$\hat{\mu}_{T90}$= -0.92	$\hat{\mu}_{HR}$= 1.62
$\hat{\sigma}_{T90}$= 1.04	$\hat{\sigma}_{HR}$= 0.50	$\hat{\sigma}_{T90}$= 0.98	$\hat{\sigma}_{HR}$= 0.50
$\hat{\rho}$ =-0.03		$\hat{\rho}$=-0.02	
Component 2		Group 2	
$\hat{w}_2 = 0.76$		$\hat{w}_2 = 0.76$	
$\hat{\mu}_{T90}$ = 3.31	$\hat{\mu}_{HR}$ = 0.95	$\hat{\mu}_{T90}$= 3.31	$\hat{\mu}_{HR}$ = 0.95
$\hat{\sigma}_{T90}$ =1.10	$\hat{\sigma}_{HR}$ = 0.49	$\hat{\sigma}_{T90}$=1.10	$\hat{\sigma}_{HR}$=0.49
$\hat{\rho}$ =0.01		$\hat{\rho}$ = 0.01	

Table 1: BATSE: Estimates for log(T90) and log(HR).

Analysis under the base model EP priors and under the empirical version of the EP priors was considered. The reversible jump MCMC algorithm was used in both cases. An additional step was added to the MCMC algorithm to deal with the generation of $\boldsymbol{y}_i$ conditional on all other parameters and on $\boldsymbol{x}_i$. Indeed, conditional on $\boldsymbol{\theta}_{(k)}$, $\boldsymbol{z}$ and $\boldsymbol{x}_i$, the posterior distribution for $\boldsymbol{y}_i$ is given by

$$p(\boldsymbol{y}_i|\cdots) \propto N(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\sigma}_i) \times N(\boldsymbol{y}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}). \tag{6}$$

Thus, in addition to to the other steps of the MCMC sweep, the mean location of each burst is generated from a normal density with mean and variance obtained straight forward from (6) on each sweep.

The length of the run for the MCMC simulation was 100,000 sweeps, and only the last 60,000 were used for inference. Identifiability was obtained by ordering the first coordinate of the means $\boldsymbol{\mu}_j$. The posterior probabilities for k strongly indicate that only two components are present ($P(k = 2\,|\,\boldsymbol{y}) = .99$.) Table 1 gives the corresponding estimates of the location and covariance matrices for two components under the "base model" and the empirical versions of the EP prior. Both analyses essentially agree. Notice that no evidence of correlation between y_{i1} and y_{i2} is found.

Figure 1 shows the allocation distribution $p(\boldsymbol{z}\,|\,\boldsymbol{y})$ for the gamma ray bursts, together with predictive confidence sets of levels 90%, 95% and 99% for each of the two components.

References

Berger, J. and Pericchi, L. (1996). The Intrinsic Bayes Factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109–122.

Berger, J., Pericchi, L., and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics. Series A*, **60**, 307–321.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 363–375.

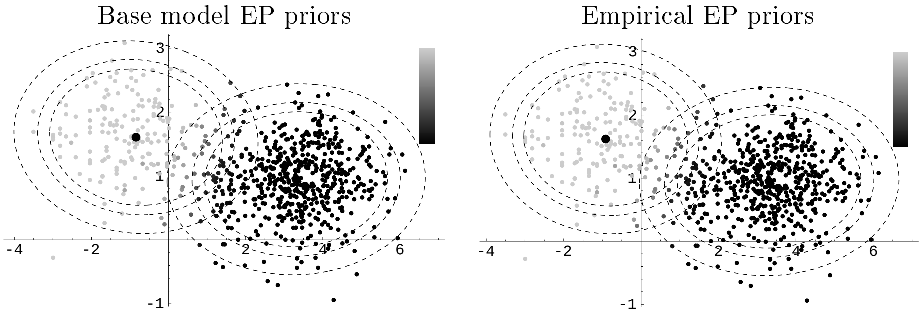

Figure 1: BATSE: Classification probabilities. Color bar indicates value of $p(z_i = 2 \mid \boldsymbol{y})$.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

Green, P. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, **57**, 97–109.

Hurley, K. (1991). Gamma-ray bursts. In Paciesas, W. S. and Fishman, G. J., editors, *AIP Conference Proceedings*, **265**, New York. AIP.

McLachlan, G. and Basford, K. E. (1985). Likelihood estimation with normal mixture models. *Journal of the Royal Statistical Society*, **34**, 282–289.

Meegan, C. A. *et al.* (1995). The third BATSE Gamma-Ray Burst catalog. *Astrophysical Journal Supplement* **106**.

Mengersen, K. and Robert, C. (1996). Testing for mixtures: a Bayesian entropy approach. In Bernardo, J. *et al.* , editors, *Bayesian Statistics 5*, London. Oxford University Press.

Moreno, E. and Liseo, B. (1998). Default Analysis of Normal Mixture Models. In Bernardo, J. *et al.* , editors, *Bayesian Statistics 6*. Oxford university Press.

O'Hagan, A. (1995). Fractional Bayes Factors for model comparison. *Journal of the Royal Statistical Society, Series B*, **57**, 99–138.

Pérez, J. M. (1998). *Development of Expected Posterior Prior Distributions for Model Comparisons*. PhD thesis, Purdue University.

Pérez, J. M. and Berger, J. O. (1999). Expected posterior prior distributions for model selection. Technical Report 99-09, CESMa. Universidad Simón Bolívar.

Richardson, S. and Green, P. J. (1998). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 60.

Robert, C. (1996). *Practical Markov Chain Monte Carlo*, chapter Mixtures of Distributions: Inference and estimation. Chapter 24, pages 441–464. Chapman and Hall, London.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

Shui, C. (1996). *Default Bayesian Analysis of Mixture Models.* PhD thesis, Purdue University.