Sankhyā : The Indian Journal of Statistics
2008, Volume 70-B, Part 1, pp. 133-155

Gibbs Sampling Based Bayesian Analysis of Mixtures with Unknown Number of Components

Sourabh Bhattacharya
Indian Statistical Institute, Kolkata, INDIA

Abstract

For mixture models with unknown number of components, Bayesian approaches, as considered by Escobar and West (1995) and Richardson and Green (1997), are reconciled here through a simple Gibbs sampling approach. Specifically, we consider exactly the same direct set up as used by Richardson and Green (1997), but put Dirichlet process prior on the mixture components; the latter has also been used by Escobar and West (1995) albeit in a different set up. The reconciliation we propose here yields a simple Gibbs sampling scheme for learning about all the unknowns, including the unknown number of components. Thus, we completely avoid complicated reversible jump Markov chain Monte Carlo (RJMCMC) methods, yet tackle variable dimensionality simply and efficiently. Moreover, we demonstrate, using both simulated and real data sets, and pseudo-Bayes factors, that our proposed model outperforms that of Escobar and West (1995), while enjoying, at the same time, computational superiority over the methods proposed by Richardson and Green (1997) and Escobar and West (1995). We also discuss issues related to clustering and argue that in principle, our approach is capable of learning about the number of clusters in the sample as well as in the population, while the approach of Escobar and West (1995) is suitable for learning about the number of clusters in the sample only.

AMS (2000) *subject classification.* Primary .
Keywords and phrases. Bayesian analysis, cross-validation, Dirichlet process, leave-one-out posterior, Markov chain Monte Carlo, pseudo-Bayes factor

1 Introduction

Mixture models are noted for their flexibility and are widely used in the statistical literature. Such models constitute a natural framework for modeling heterogeneity with strong connections with cluster analysis. The entire literature on mixture models have been almost exhaustively discussed by

Titterington et al. (1985) and McLachlan and Basford (1988). However, a particularly challenging situation arises when the number of mixture components, which we denote by k, can not be determined based on the available data set in a straightforward manner. Escobar and West (1995) (henceforth EW) were the first to propose a modeling style, involving Dirichlet processes, to introduce uncertainty within the number of components. Gibbs sampling is then used to learn about all the unknowns involved. However, the mixture model of EW does not resemble the traditional mixture models, which are often mixtures of normal distributions. For instance, conditional on the parameters, the predictive distribution in the case of EW is a mixture of non-central Student's t and normal distributions. In contrast, Richardson and Green (1997) (henceforth RG) consider a mixture of normal distributions assuming k unknown. But acknowledgment of uncertainty about k makes the dimension of the parameter space a random variable in the case of RG. They use sophisticated, but quite complicated MCMC methods, known as reversible jump MCMC (RJMCMC), introduced in the literature by Green (1995).

In this paper, we propose another alternative model for mixture analysis when the number of components is unknown. Broadly, as in RG, we consider a mixture of normal distributions, but instead of complicated (and often inefficient) RJMCMC methods, view the parameters of the components as samples from a Dirichlet process. We propose a simple Gibbs sampling algorithm, thus freeing ourselves of the burden of implementing RJMCMC, while maintaining the same variable dimensional framework as RG. This is the implementation advantage of our proposed methodology. On the conceptual side, we demonstrate with simulated and real data sets that our model is better supported by the data than that of EW. We argue tentatively that the unconventional mixture of the thick-tailed Student's t with the normal distributions of the predictive distributions of EW is the reason for its relatively poor performance compared to our conventional mixture of normal distributions. Moreover, the implementation time of our model is far less than that of EW and RG. For example, with the largest of the three real data sets that we consider, the MCMC algorithm for the implementation of EW's method does about 93 iterations per second on a Mac OS X laptop, RG's RJMCMC does 160 iterations per second (see Richardson and Green (1997), page 755), but the MCMC algorithm for implementation of our proposed model does about 715 iterations per second.

The remainder of our paper is structured as follows. In Section 2 we introduce our proposed methodology for mixture analysis with unknown number of components. A simple Gibbs sampling approach for implementing our model and methodology is proposed in Section 3. Details on the advantages of our approach as compared to the approaches of RG and EW are provided in Sections 4 and 5 respectively. Model comparison using pseudo-Bayes factor is outlined in Section 6. Illustration of our methodology and comparisons with the proposal of EW using simulated data, are carried out in Section 7. Applications of our methodology to three real data sets are discussed in Section 8. We summarize the main points and provide directions for future research in Section 9.

2 Dirichlet Process for Learning Unknown Number of Parameters

Generically denoting by $\boldsymbol{\Theta}_p$ the set of parameters $\{\boldsymbol{\theta}_j; 1 \leq j \leq p\}$, for any p, where $\boldsymbol{\theta}_j = (\mu_j, \lambda_j)$, we show that the mixture of the form

$$[y_i \mid \boldsymbol{\Theta}_p] = \sum_{j=1}^{p} \pi_j \sqrt{\frac{\lambda_j}{2\pi}} \exp\left\{-\frac{\lambda_j}{2}(y_i - \mu_j)^2\right\}, \tag{2.1}$$

although a mixture of a random number of components, may be expressed simply as an average of a fixed (but perhaps, large) number of components. To this end, we first express the distribution of y_i as the following:

$$[y_i \mid \boldsymbol{\Theta}_m] = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{\lambda_j}{2\pi}} \exp\left\{-\frac{\lambda_j}{2}(y_i - \mu_j)^2\right\} \tag{2.2}$$

In the above, $m(\geq p)$ can be interpreted as the maximum number of distinct mixture components of $[y_i \mid \boldsymbol{\Theta}_m]$, with $\pi_j = 1/m$ for each j. We now show that, under an interesting nonparametric prior assumption for $\{(\mu_j, \lambda_j); \ell = 1, \ldots, m\}$, (2.2) boils down to the form of (2.1).

In $\boldsymbol{\Theta}_m$, the parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ are samples drawn from a Dirichlet process (see Ferguson (1974), Escobar and West (1995)). In other words, we assume that, $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ are samples from some *unknown* prior distribution $G(\cdot)$ on $\Re \times \Re^+$. We further suppose that for $G \sim \mathcal{D}(\alpha G_0)$, a Dirichlet process defined by α, a positive scalar, and $G_0(\cdot)$, a specified bivariate distribution function over $\Re \times \Re^+$. Put more simply, we assume that

$$[\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m \mid G] \sim iid\ G$$

and

$$G \sim \mathcal{D}(\alpha G_0)$$

A crucial feature of our modelling style concerns the discreteness of the prior distribution G, given the assumption of Dirichlet process; that is, under these assumptions, the parameters $\boldsymbol{\theta}_\ell$ are coincident with positive probability. In fact, this is the property that we exploit to show that (2.2) reduces to (2.1) under the above modelling assumptions. The main points regarding this are sketched below.

On marginalising over G we obtain,

$$[\boldsymbol{\theta}_j \mid \boldsymbol{\Theta}_{-jm}] \sim \alpha a_{m-1} G_0(\boldsymbol{\theta}_j) + a_{m-1} \sum_{l=1, l\neq j}^{m} \delta_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}_j) \tag{2.3}$$

In the above, $\boldsymbol{\Theta}_{-jm} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}, \boldsymbol{\theta}_{j+1}, \ldots, \boldsymbol{\theta}_m)$ and $\delta_{\boldsymbol{\theta}_l}(\cdot)$ denotes a unit point mass at θ_l and $a_l = 1/(\alpha + l)$ for positive integers l.

The above expression shows that the $\boldsymbol{\theta}_j$ follow a general Polya urn scheme. In other words, it follows that the joint distribution of $\boldsymbol{\Theta}_m$ is given by the following: $\boldsymbol{\theta}_1 \sim G_0$, and, for $j = 2, \ldots, m$, $[\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}] \sim \alpha a_{j-1} G_0(\boldsymbol{\theta}_j) + a_{j-1} \sum_{l=1}^{j-1} \delta_{\boldsymbol{\theta}_l}(\boldsymbol{\theta}_j)$. Thus, given a sample $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}\}$, $\boldsymbol{\theta}_{j-1}$ is drawn from G_0 with probability αa_{j-1} and is otherwise drawn uniformly from among the sample $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{j-1}\}$. In the former case, $\boldsymbol{\theta}_j$ is a new, distinct realisation and in the latter case, it coincides with one of the realisations already obtained. Thus, there is a positive probability of coincident values. For more on the relationship between a generalized Polya urn scheme and the Dirichlet process prior, see Blackwell and McQueen (1973) and Ferguson (1974).

Now, supposing that a sample from the joint distribution of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ yields p^* distinct realisations given by $\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_{p^*}$, and if m_ℓ denotes the number of times $\boldsymbol{\theta}^*_\ell$ appears in the sample, then $\pi^*_\ell = m_\ell/m$. Certainly, it holds that $\sum_{\ell=1}^{p^*} \pi^*_\ell = 1$. Hence, (2.2) reduces to the form (2.1).

With our modelling style using Dirichlet process, the prior for m_ℓ is implicitly induced; for more details, see Antoniak (1974), Escobar and West (1995). For recent use of Dirichlet process to estimate many arbitrary functions in the context of palaeo environmental reconstruction, see Bhattacharya (2006).

2.1. Choice of G_0*.* It is now necessary to specify the prior mean $G_0(\cdot)$ of $G(\cdot)$. Following EW, we assume that under $G_0(\cdot)$, $\boldsymbol{\theta}_j = (\mu_j, \lambda_j)$ is normal-gamma. In other words, we assume that, under G_0,

$$[\lambda_j] \sim Gamma(s/2, S/2) \tag{2.4}$$

$$[\mu_j \mid \lambda_j] \sim N\left(\mu_0, \frac{\psi}{\lambda_j}\right) \tag{2.5}$$

In the above, we define $Gamma(a, b)$ to mean a *Gamma* distribution with mean a/b and variance a/b^2. The choices of the prior parameters s, S, μ_0, ψ will generally depend upon the application at hand. Hence, at this moment, we leave them unspecified. One may also specify prior distributions on μ_0 and ψ; following EW it is possible to consider that, *a priori*, $\mu_0 \sim N(a, A)$ and $\psi^{-1} \sim Gamma(w/2, W/2)$, for specified hyperparameters a, A, w, and W. In the applications, our choices will closely follow those of EW and RG. The choice of the maximum number of components, m, will also depend upon the given problem; however, following RG, who chose 30 as the maximum number of components for all their illustrations, we will generally take $m = 30$. In the applications we present in this paper, almost all posterior probability mass is concentrated on less than 30 components. We also need to either specify a value for α or put a prior distribution on it. Because of our ignorance of the actual value of α, we resort to quantify the uncertainty about α by assigning to it a $Gamma(a_\alpha, b_\alpha)$ prior distribution; the values of a_α and b_α will be chosen based on the values used by EW.

We have thus defined a semiparametric model with priors on the parameters $\boldsymbol{\Theta}_m = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ being non-parametric, viewed as samples from the Dirichlet process, but the distribution of $[y_i \mid \boldsymbol{\Theta}_m]$ has a parametric form, given by (2.2). In the next section we discuss implementation of our methodology by Gibbs sampling.

3 Markov Chain Monte Carlo Implementation of the Proposed Model

3.1. Representation of the mixture using allocation variables and associated full conditional distributions. The distribution of $[y_i \mid \boldsymbol{\Theta}_m]$ given by (2.2) can be represented by introducing the allocation variables z_i, as follows:

For $i = 1, \ldots, n$ and $j = 1, \ldots, m$,

$$\begin{aligned}
[y_i \mid z_i = j, \boldsymbol{\Theta}_m] &= \sqrt{\frac{\lambda_j}{2\pi}} \exp\left\{-\frac{\lambda_j}{2}(y_i - \mu_j)^2\right\} && (3.1)\\
[Z_i = j] &= \frac{1}{m} && (3.2)
\end{aligned}$$

It follows that the full conditional distribution of z_i $(i = 1, \ldots, n)$ given the rest is given by

$$[z_i = j \mid \mathbf{Y}, \boldsymbol{\Theta}_m, \mathbf{Z}_{-i},] \propto \sqrt{\lambda_j} \exp\left\{-\frac{\lambda_j}{2}(y_i - \mu_j)^2\right\}; \;\; j = 1, \ldots, m \qquad (3.3)$$

In the above, $\mathbf{Z}_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)'$. Note that the allocation variables play the role of clustering the observation vector $\mathbf{Y}$, into maximum m components, assuming initially that all the parameters are distinct.

Next we use the allocation variables $\mathbf{Z}$ to determine the full conditional distributions of $\boldsymbol{\theta}_j = (\mu_j, \lambda_j)$ given the rest.

3.1.1. Full conditionals of $\boldsymbol{\theta}_j$. To write down the full conditional distribution of $\boldsymbol{\theta}_j$ given the rest, we first define $n_j = \#\{i : z_i = j\}$ and $\bar{y}_j = \sum_{i:z_i=j} y_i/n_j$. Then, using the Polya urn scheme given by (2.3) and the discussion in Section 3.1 it can be shown that the full conditional distribution of $\boldsymbol{\theta}_j$ given the rest is given by

$$[\boldsymbol{\theta}_j \mid \mathbf{Y}, \mathbf{Z}, \boldsymbol{\Theta}_{-jm}] = q_{0j} G_j(\boldsymbol{\theta}_j) + \sum_{\ell=1,\ell\neq j}^{m} q_{\ell j} \delta_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_j) \qquad (3.4)$$

In the above, $G_j(\boldsymbol{\theta}_j)$ is a bivariate normal/gamma distribution such that under $G_j(\boldsymbol{\theta}_j)$:

$$\begin{aligned}
[\lambda_j] &\sim \textit{Gamma}\left(\frac{s + n_j}{2}, \frac{1}{2}\left\{S + \frac{n_j(\bar{y}_j - \mu_0)^2}{n_j\psi + 1} + \sum_{i:z_i=j}(y_i - \bar{y}_j)^2\right\}\right) && (3.5)\\
[\mu_j \mid \lambda_j] &\sim N\left(\frac{n_j\bar{y}_j\psi + \mu_0}{n_j\psi + 1}, \frac{\psi}{\lambda_j(n_j\psi + 1)}\right) && (3.6)
\end{aligned}$$

In (3.4) q_{0j} and $q_{\ell j}$ are given by the following:

$$\begin{aligned} q_{0j} &\propto \alpha\frac{\left(\frac{S}{2}\right)^{\frac{s}{2}}}{\Gamma(\frac{s}{2})}\times\frac{1}{\sqrt{n_j\psi+1}}\times\left(\frac{1}{2\pi}\right)^{\frac{n_j}{2}} \\ &\times \frac{2^{\frac{s+n_j}{2}}\Gamma(\frac{s+n_j}{2})}{\left\{S+\frac{n_j(\bar{y}_j-\mu_0)^2}{n_j\psi+1}+\sum_{i:z_i=j}(y_i-\bar{y}_j)^2\right\}^{\frac{s+n_j}{2}}} \end{aligned} \tag{3.7}$$

and,

$$q_{\ell j} \propto \left(\frac{\lambda_\ell}{2\pi}\right)^{\frac{n_j}{2}}\exp\left[-\frac{\lambda_\ell}{2}\left\{n_j(\mu_\ell-\bar{y}_j)^2+\sum_{i:z_i=j}(y_i-\bar{y}_j)^2\right\}\right] \tag{3.8}$$

The proportionality constant is chosen such that $q_{0j}+\sum_{\ell=1,\ell\neq j}q_{\ell j}=1$.

Note that the full conditional distribution of $\boldsymbol{\theta}_j$, given by (3.4) is a mixture of normal-gamma distribution and point masses $\delta_{\boldsymbol{\theta}_\ell}$, where $\delta_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_j)=1$ if $\boldsymbol{\theta}_j=\boldsymbol{\theta}_\ell$ and zero otherwise. The mixing probability of the former is q_{0j} and that of the latter are $q_{\ell j}$. Thus, to sample from (3.4) one must either choose with probability q_{0j} the normal-gamma distribution jointly given by (3.5) and (3.6), and draw a realization from it or draw an already observed value, say, $\boldsymbol{\theta}_\ell$ ($\ell\neq j$) with probability $q_{\ell j}$.

We note that the approach we have provided so far is applicable when each of the data points $y_1,\ldots,y_n$, are univariate. The reader might wonder if our approach will be applicable in case the observations are multivariate. We assure that our approach is, of course, applicable even when each of the data points $\mathbf{y}_1,\ldots,\mathbf{y}_n$ are vectors, consisting of d components. Also, configuration indicators (see, for example, MacEachern (1994), Müller et el (1996)) may be used for updating only the distinct parameters; this method has the additional advantage of being applicable even when the base measure G_0 is not conjugate with (3.1). The full conditional distributions of α, μ_0 and ψ have forms that are similar to those corresponding to the approach of EW. Due to reasons of space, we omit details on these interesting facts, but these are available from the author on request.

4 Comparison With the Approach of RG

In fixed dimensional problems, sampling from a non-standard posterior distribution is usually carried out via Metropolis-Hastings algorithms. For instance, if the posterior of a fixed-dimensional parameter $\phi \in \mathbf{\Phi}$ is proportional to prior times the likelihood, given by $\pi(\phi)L(\phi)$, then staring with an initial value $\phi^{(0)}$, a new iterate $\phi^{(1)}$ may be obtained by proposing from an arbitrary distribution $q(\cdot \mid \phi^{(0)})$, and accepting the proposed value with probability

$$\min\left\{\frac{\pi(\phi^{(1)})L(\phi^{(1)})q(\phi^{(0)} \mid \phi^{(1)})}{\pi(\phi^{(0)})L(\phi^{(0)})q(\phi^{(1)} \mid \phi^{(0)})}, 1\right\} \tag{4.1}$$

In this way, under very mild conditions, a Markov chain $\{\phi^{(0)}, \phi^{(1)}, \ldots\}$ is generated, which converges to the stationary distribution, which is also the posterior distribution of interest. However, construction of a Markov chain which converges to the posterior distribution is not straightforward in the case of varying dimensional model, since the ratio in the acceptance probability (4.1) would not make sense when the numerator and denominator are associated with parameters of different dimensionalities. Green (1995) was the first to propose "dimension-matching" transformations to circumvent the problem. We briefly illustrate the ideas below.

Defining a collection of models by $\mathcal{M}_p = \{f(\cdot \mid \phi_p); \phi_p \in \Phi_p\}$; $p = 1, \ldots, P$, the prior distribution of (p, ϕ_p) may be represented as $\pi(p, \phi_p) = \pi(p)\pi(\phi_p \mid p)$. The latter is a density with respect to the Lebesgue measure on the union of spaces $\Phi = \cup_p \{p\} \times \Phi_p$. In our set up, $\mathcal{M}_p$ is the p-component normal mixture distribution (2.1). Following Robert and Casella (2004) we motivate MCMC for general variable dimensional models, such as (2.1), from the perspective of fixed dimension MCMC. In other words, when considering a move from $\mathcal{M}_{p_1}$ to $\mathcal{M}_{p_2}$, where $d_{p_1} = dim\mathbf{\Phi}_{p_1} < dim\mathbf{\Phi}_{p_2} = d_{p_2}$, it is possible to describe an equivalent move in a fixed dimensional setting.

The crux of the RJMCMC algorithms described in Green (1995) (see also RG), lies in adding an auxiliary variable $u_{p_1p_2} \in \mathcal{U}_{p_1p_2}$ to ϕ_{p_1} so that $\mathbf{\Phi}_{p_1} \times \mathcal{U}_{p_1p_2}$ and $\mathbf{\Phi}_{p_2}$ are in bijection relation. Observe that, proposing a move from $(\phi_{p_1}, u_{p_1p_2})$ to ϕ_{p_2} is the same as the *fixed dimensional* proposal when the corresponding stationary distributions are $\pi(p_1, \phi_{p_1})\pi(u_{p_1p_2})$ and $\pi(p_2, \phi_{p_2})$ respectively. As for the proposal distribution, consider that the move from $(\phi_{p_1}, u_{p_1p_2})$ to ϕ_{p_2} proceeds by generating

$$\phi_{p_2} \sim N_{d_{p_2}}(T_{p_1p_2}(\phi_{p_1}, u_{p_1p_2}), \epsilon I);\quad \epsilon > 0 \tag{4.2}$$

and that the reverse proposal is to take $(\phi_{p_1}, u_{p_1p_2})$ as the $T_{p_1p_2}$-inverse transform of the normal distribution $N_{d_{p_2}}(T_{p_1p_2}(\phi_{p_1}, u_{p_1p_2}), \epsilon I)$. Taking into account the Jacobian of the aforementioned transformations and probabilities of the moves between $\mathcal{M}_{p_1}$ and $\mathcal{M}_{p_2}$, we obtain, after taking $\epsilon \to 0$, the

Metropolis-Hastings acceptance probability from a fixed dimensional perspective, as

$$\min\left(\frac{\pi(p_1,\phi_{p_1})\pi_{p_2p_1}}{\pi(p_1,\phi_{p_1})\pi(u_{p_1p_2})\pi_{p_1p_2}}\left|\frac{\partial T_{p_1p_2}(\phi_{p_1},u_{p_1p_2})}{\partial(\phi_{p_1},u_{p_1p_2})}\right|,1\right) \tag{4.3}$$

In the above, π_{ij} is the probability of choosing a jump to $\mathcal{M}_{p_j}$ while in $\mathcal{M}_{p_i}$.

The efficiency of the algorithm very much depends upon the dimension matching transformation selected. Indeed, this is a potential cause for inefficiency of the algorithm and requires extremely demanding tuning steps. In the words of Robert and Casella (2004), "...this is a setting where wealth is a mixed blessing, if only the total lack of direction in the choice of the jumps may result in a lengthy or even impossible calibration of the algorithm". Indeed, the analysis of (2.1) using RJMCMC as described by RG is extremely complicated even for univariate data, so much so that an error crept into the analysis of RG (the corrigendum of the error is provided in Richardson and Green (1998)). Obviously, the computational complexity increases a lot with increasing dimensionality making the method extremely inefficient and error-prone. In contrast, the Gibbs sampling algorithm we proposed in this paper retains its efficiency in all dimensions.

5 Comparison With the Approach of EW

EW proceed by assuming a hierarchical model for the data $\mathbf{Y}$: for $i=1,\ldots,n$, $y_i \sim N\left(\mu_i,\lambda_i^{-1}\right)$; $\boldsymbol{\theta}_i=(\mu_i,\lambda_i)\sim G$; $G\sim\mathcal{D}(\alpha G_0)$, where G_0 is normal/inverse Gamma. Observe that, unlike our model, y_i in the approach of EW are not *iid*. Note that the literature on modelling based on Dirichlet processes invariably refers to the set up where the observed data set is clustered into number of clusters less than or equal to the total number of observations. In other words, in the approach of EW, the number of distinct components in the mixture, denoted by k is less than or equal to n, the total number of observations. But this set up is not appropriate in cases where the true model is a mixture of many distributions, but a relatively less number of data are obtained. To clarify, suppose that, for $i=1,\ldots,n$, data $y_i\sim\sum_{j=1}^{k}\pi_j f_j$ where $\sum_{j=1}^{k}\pi_j=1$, and $f_j; j=1,\ldots$ are the mixture components. Now if $n<k$, then all components of the mixture are not represented in the data, and one must use prior information to learn about all the mixture components, and the mixing probabilities π_j. But approaches available in the literature do not allow for using such prior information. Our

approach to mixture modelling using Dirichlet processes allow for incorporation of prior knowledge by assuming the mixture model $y_i \sim \frac{1}{m}\sum_{j=1}^{m} f_j$, for typically large m, allowing for $m >> n$, and modelling the parameters of f_j as samples from an unknown distribution G, which is modelled as a Dirichlet process. Coincidences among the m parameters will then reduce the effective number of parameters, taking the resulting mixture close to the truth. In practice, m may be elicited using prior opinion of experts. For instance, to classify vegetation of a forest, the collected data may not include all possible vegetation, so classification based on the data only will be misleading. Experts can, however, provide some information, at least an upper bound, of the actual number of distinct vegetation in the forest. In our approach, we might take m to be the upper bound provided by the expert. To summarize, the traditional approaches that use Dirichlet process for clustering, are only appropriate for learning sample number of clusters, but they can never learn about the population number of clusters. In our approach, however, it is possible in principle to learn about the population number of clusters using available prior information. We will illustrate this with an example.

We also note that the upper bound provided by the expert may even be much smaller than the number of observations collected. This can again be utilised by our approach to rule out minor, unimportant clusters, that may get significant probabilities in the traditional approaches as of EW, that gives positive probabilities to all possible clusters associated with the observations.

A technical point related to the approach of EW must also be taken into account. The predictive distribution of y_{n+1}, a new observation, given $\mathbf{\Theta}_n$ (note that, $n \neq m$), is a mixture of a Student's t distribution and n normal distributions, given by (see equations (4) and (5) of EW)

$$\begin{aligned}
[y_{n+1} \mid \mathbf{\Theta}_n] &= \int [y_{n+1} \mid \boldsymbol{\theta}_{n+1}]\,[\boldsymbol{\theta}_{n+1} \mid \mathbf{\Theta}_n]\, d\boldsymbol{\theta}_{n+1} \\
&= \alpha a_n \frac{\Gamma\left(\frac{s+1}{2}\right)}{\Gamma\left(\frac{s}{2}\right)} \left(\frac{1}{Ms\pi}\right)^{\frac{1}{2}} \frac{1}{\left\{1 + \frac{(y_{n+1}-\mu_0)^2}{Ms}\right\}^{\frac{s+1}{2}}} \\
&+ a_n \sum_{i=1}^{n} \sqrt{\frac{\lambda_i}{2\pi}} \exp\left\{-\frac{\lambda_i}{2}(y_i - \mu_i)^2\right\} \qquad (5.1)
\end{aligned}$$

where $a_n = 1/(\alpha+n)$ and $M = (1+\psi)S/s$. It may be pointed out that when a normal component can fit the data, it will perform better than a thick-tailed Student's t component. As a result, it is arguable that our mixture model, which consists of normal components only, will perform better than the model of EW. We demonstrate with both simulation studies and real data examples that this is indeed the case.

6 Model Comparison Using Pseudo-Bayes Factor

The pseudo Bayes factor (PBF) owes its origin to Geisser and Eddy (1979). Using cross-validation ideas (Stone (1974), Geisser (1975)), PBF has been proposed as a surrogate for the Bayes factor by Geisser and Eddy (1979). Before discussing the advantages of PBF over Bayes factor (BF), it is useful to briefly touch upon the latter.

6.1. Bayes factor. The Bayes factor is given by

$$BF = \frac{[\mathbf{Y} \mid M_1]}{[\mathbf{Y} \mid M_2]} \tag{6.1}$$

where $[\mathbf{Y} \mid M_j]; j = 1, 2$ are the marginals of the data under the competing models M_j. In particular, denoting by $\mathbf{\Phi}$ the entire set of model parameters, and letting $[\mathbf{\Phi}]$ denote the prior for $\mathbf{\Phi}$, we note that,

$$[\mathbf{Y} \mid M_j] = \int [\mathbf{Y} \mid \mathbf{\Phi}][\mathbf{\Phi}]d\mathbf{\Phi} \tag{6.2}$$

In the above, $[\mathbf{Y} \mid \mathbf{\Phi}]$ is simply the likelihood under the observed data set $\mathbf{Y}$. Jeffreys (1961) recommends selection of model M_1 if $BF > 2$. A well-known property of the Bayes factor is that it tends to put too much weight on parsimonious models; this is also known as Lindley's paradox. Also observe that the above marginal is improper if the prior $[\mathbf{\Phi}]$ is improper. Since improper priors are used very widely in realistic problems, it is useful to seek alternatives to the traditional Bayes factor.

6.2. Pseudo Bayes factor. The pseudo Bayes factor, which is based on cross-validation, can be defined as

$$PBF(M_1/M_2) = \prod_{i=1}^{n} \frac{[y_i \mid \mathbf{Y}_{-i}, M_1]}{[y_i \mid \mathbf{Y}_{-i}, M_2]} \tag{6.3}$$

In the above, $\mathbf{Y}_{-i} = \{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n\}$. The factors $[y_i \mid \mathbf{Y}_{-i}, M_j]; j = 1, 2$ are the cross-validation predictive densities, given by

$$[y_i \mid \mathbf{Y}_{-i}, M_j] = \int [y_i \mid \mathbf{\Phi}][\mathbf{\Phi} \mid \mathbf{Y}_{-i}, M_j] d\mathbf{\Phi} \tag{6.4}$$

Note that the cross-validation predictive density $[y_i \mid \mathbf{Y}_{-i}, M_j]$ is proper whenever the posterior $[\mathbf{\Phi} \mid \mathbf{Y}_{-i}, M_j]$ is proper. Thus, (6.4) avoids the impropriety problem generally encountered by the traditional Bayes factor given by (6.1). Another point worth mentioning is, it follows from Brook's lemma (see Brook (1964)) that the set of cross-validation predictive densities $\{[y_i \mid \mathbf{Y}_{-i}, M_j]; i = 1, \ldots, n\}$ is equivalent to the marginal $[\mathbf{Y} \mid M_j]$, given by (6.2) whenever the latter exists. Gelfand and Dey (1994) show that asymptotically,

$$PBF(M_1/M_2) \approx \log(\nu_n) + \frac{p_2 - p_1}{2} \tag{6.5}$$

where

$$\nu_n = \frac{L(\hat{\mathbf{\Phi}}_1; \mathbf{Y}, M_1)}{L(\hat{\mathbf{\Phi}}_2; \mathbf{Y}, M_2)} \tag{6.6}$$

is the likelihood ratio, $\hat{\mathbf{\Phi}}_1, \hat{\mathbf{\Phi}}_2$ being maximum likelihood estimates of model parameters $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$, of dimensionalities p_1 and p_2, respectively. Very importantly, unlike BF, PBF does not suffer from Lindley's paradox (Gelfand and Dey (1994)).

6.3. Computation of pseudo Bayes factor. Computation of all n cross-validation predictive densities must preceed computation of PBF. However, this is a very challenging task, since it requires computation of n leave-one-out posteriors $[\mathbf{\Phi} \mid \mathbf{Y}_{-i}, M_j]; i = 1, \ldots, n$. In order to handle this computational challenge, Gelfand (1996) proposed an importance sampling strategy. In this paper, we arrive at the same result as Gelfand (1996), but in a more direct manner.

Note that

$$\begin{aligned}
[y_i \mid \mathbf{Y}_{-i}, M_j] &= \int [y_i \mid \mathbf{\Phi}][\mathbf{\Phi} \mid \mathbf{Y}_{-i}, M_j] d\mathbf{\Phi} \\
&= \int [y_i \mid \mathbf{\Phi}] \frac{[\mathbf{\Phi} \mid \mathbf{Y}_{-i}, M_j]}{[\mathbf{\Phi} \mid \mathbf{Y}, M_j]} [\mathbf{\Phi} \mid \mathbf{Y}, M_j] d\mathbf{\Phi} \\
&= E_{[\mathbf{\Phi} \mid \mathbf{Y}, M_j]} \left([y_i \mid \mathbf{\Phi}, M_j] \frac{[\mathbf{\Phi} \mid \mathbf{Y}_{-i}, M_j]}{[\mathbf{\Phi} \mid \mathbf{Y}, M_j]} \right)
\end{aligned} \tag{6.7}$$

Note that $[\mathbf{\Phi} \mid \mathbf{Y}_{-i}, M_j]/[\mathbf{\Phi} \mid \mathbf{Y}, M_j] \propto 1/[y_i \mid \mathbf{\Phi}, M_j]$. Now assuming that a sample $\{\mathbf{\Phi}^{(\ell)}; \ell = 1, \ldots, N\}$ is available (usually via MCMC) from the full posterior $[\mathbf{\Phi} \mid \mathbf{Y}, M_j]$, (6.4) can be approximated as

$$[y_i \mid \mathbf{Y}_{-i}, M_j] \approx \frac{N}{\sum_{\ell=1}^{N} \frac{1}{[y_i \mid \mathbf{\Phi}^{(\ell)}, M_j]}} \tag{6.8}$$

Thus, (6.8) is a harmonic mean of $\{[y_i \mid \mathbf{\Phi}^{(\ell)}, M_j]; \ell = 1, \ldots, N\}$; this is exactly the estimator obtained by Gelfand (1996) which they obtained in a somewhat indirect manner. Considering model M_1 to be our proposed model, $[y_i \mid \mathbf{\Phi}^{(\ell)}, M_1]$ is given by

$$[y_i \mid \mathbf{\Phi}^{(\ell)}, M_1] = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{\lambda_j^{(\ell)}}{2\pi}} \exp\left\{ -\frac{\lambda_j^{(\ell)}}{2} \left(y_i - \mu_j^{(\ell)} \right)^2 \right\} \tag{6.9}$$

Denoting by M_2 the model of EW, we note that $[y_i \mid \mathbf{\Phi}^{(\ell)}, M_2]$ follows from (5.1), and is given by

$$\begin{aligned} \left[y_i \mid \mathbf{\Phi}^{(\ell)}, M_2\right] &= \alpha a_{n-1} \frac{\Gamma\left(\frac{s+1}{2}\right)}{\Gamma\left(\frac{s}{2}\right)} \left(\frac{1}{Ms\pi}\right)^{\frac{1}{2}} \frac{1}{\left\{1 + \frac{(y_i - \mu_0^{(\ell)})^2}{Ms}\right\}^{\frac{s+1}{2}}} \\ &+ a_{n-1} \sum_{j=1; j \neq i}^{n} \sqrt{\frac{\lambda_j^{(\ell)}}{2\pi}} \exp\left\{ -\frac{\lambda_j^{(\ell)}}{2} (y_i - \mu_j^{(\ell)})^2 \right\} \end{aligned} \tag{6.10}$$

Hence, the sets $\{[y_i \mid \mathbf{Y}_{-i}, M_1]; i = 1, \ldots, n\}$ and $\{[y_i \mid \mathbf{Y}_{-i}, M_2]; i = 1, \ldots, n\}$ can be easily obtained using (6.8). In our experience, computation of PBF is much more stable than the more traditional BF, particularly when the data set $\mathbf{Y}$ is large. This is because PBF can be computed by taking the sums of logarithms of the univariate cross-validation densities; this computational procedure ensures stability. This is not the case with computation of BF. Moreover, the set of cross-validation predictive densities $\{[y_i \mid \mathbf{Y}_{-i}, M_j]; i = 1, \ldots, n\}$, j=1,2, can be used as unconditional density estimators at the points $\{y_1, \ldots, y_n\}$ respectively, while $\{[y_i \mid \mathbf{\Phi}^{(\ell)}, M_j]; i = 1, \ldots, n\}$ may be used as sample density estimates at the points $\{y_1, \ldots, y_n\}$ under model M_j, given a particular MCMC realization $\mathbf{\Phi}^{(\ell)}$. In other words, given k distinct $\boldsymbol{\theta}^*$'s in $\mathbf{\Theta}_l^{(\ell)}$ ($l = n$ or $l = m$, depending upon whether the model referred to is EW's model or our proposed model), the sample density estimates $\{[y_i \mid \mathbf{\Phi}^{(\ell)}, M_j]; i = 1, \ldots, n\}$ are to interpreted as conditional on k components. We now illustrate our proposed methodologies with simulation examples.

7 Simulation Study

Since our modelling ideas are similar to that of RG, the major difference being implementation issues, we confine ourselves to comparing our proposed model and methodology with that of EW.

Based upon simulated data, we chose to simulate just 5 observations from a mixture distribution with 10 mixture components. Very obviously, EW will now always put zero posterior probability on 10 components, since the model is based upon clustering of the data only, and the data size is only 5. For our model, instead of putting a prior on α, we set it equal to 5. This choice of α implies a relatively strong belief in G_0. We ran MCMC algorithms for our proposed model as well as for the model of EW for a burn-in of 300,000 iterations, and a further 15,00,000 iterations, storing one in 150 iterations, thus obtaining a total of 10,000 realizations from the posterior distribution. Convergence of our MCMC algorithms have been confirmed by Kolmogorov-Smirnov tests as described in pages 466–470 of Robert and Casella (2004).

The posterior probabilities of the possible number of clusters ranging from 1 to 30 are given by {0.0000, 0.0003, 0.0022, 0.0091, 0.0274, 0.0497, 0.0727, 0.0965, 0.1232, 0.1275, 0.1230, 0.1088, 0.0876, 0.0651, 0.0412, 0.0309, 0.0176, 0.0091, 0.0052, 0.0017, 0.0006, 0.0001, 0.0004, 0.0001, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000}. Thus, the true number of components, 10, get the highest posterior probability.

Apart from the above simulation study, we also compared our model with that of EW using a different simulated data set, where the data size, 15, is greater than the true number of components, 10. Brief results follow. With a relatively weak prior on α, taken to be $Gamma(5,1)$, the posterior probabilities of the number of clusters with respect to our model, are given by {0.0000, 0.0000, 0.0003, 0.0017, 0.0063, 0.0186, 0.0335, 0.0650, 0.0899, 0.1260, 0.1315, 0.1352, 0.1167, 0.0958, 0.0698, 0.0507, 0.0300, 0.0140, 0.0082, 0.0046, 0.0013, 0.0008, 0.0001, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000}. Thus 10 components gets very significant posterior probability, 0.1260, which is close to the highest posterior probability. For EW's model, these are given by {0.0000, 0.0011, 0.0100, 0.0388, 0.0906, 0.1550, 0.1765, 0.1913, 0.1570, 0.0987, 0.0521, 0.0214, 0.0057, 0.0017, 0.0001}; 10 components getting posterior probability 0.0987, which is close that of our model. However, $PBF(M_1/M_2) = 335.6571$ still says that our model is supported by the data much better than EW's model. This is accordance with the discussion in Section 5. Several other simulation studies supported the same conclusion.

8 Application to Real Data Problems

We now illustrate our methodologies on the three real data sets used by RG, and available at `http://www.stats.bris.ac.uk/~peter/mixdata`. The first data set concerns the distribution of enzymatic activity in the blood, for an enzyme involved in the metabolism of carcinogenic substances, among a group of 245 unrelated individuals. The second data set, the 'acidity data', concerns an acidity index measured in a sample of 155 lakes in north-central Wisconsin. The third and the last data set, the 'galaxy data', has been analysed under different mixture models by several researchers. It consists of the velocities of 82 distant galaxies, diverging from our own galaxy. For details on these data sets, see RG. We remind the reader again that since our set up is the same as in RG, it is not meaningful to compare our model with theirs, the only important difference being that our implementation is much more straightforward and simple, as compared to RJMCMC used by RG. We, however, compare our model with that of EW and demonstrate that our model performs much better. As in the case of simulation studies, in the real data situation also we ran MCMC algorithms for our proposed model as well as for the model of EW for a burn-in of 300,000 iterations, and a further 15,00,000 iterations, storing one in 150 iterations, thus obtaining a total of 10,000 realizations from the posterior distribution.

8.1 Enzyme data. Following both RG and EW, we chose the prior parameters as $s = 4.0; S = 2 \times (0.2/1.22) = 0.3278689; \mu_0 = 1.45; a_\alpha = 2; b_\alpha = 4; \psi = 33.3$. In the value of S, the factor 0.2/1.22 is actually the expectation of a *Gamma*-distributed hyperparameter considered by RG. We do not think that the extra level of hierarchy is necessary; in fact, it only adds to the computational burden. Hence, in all the applications we considered, we fixed the value of the hyperparameter as the expected value of the prior distribution of the hyperparameter. Following RG, we chose $m = 30$. Apart from fixing the values of μ_0 and ψ above, we also experimented with vaguely informative and non-informative priors on the hyperparameters as described in EW. In particular, we assigned that, *a priori*, $\mu_0 \sim N(a, A)$ and $\psi^{-1} \sim Gamma\,(w/2, W/2)$, where $A^{-1} \to 0, w \to 0$ and $W \to 0$. However, the results remained almost exactly same as in the case of fixed values of μ_0 and ψ, in all three applications we discuss below.

According to the model of EW, the posterior probability of the number of components {1,2,3,4,5,6,7,8,9,10} are given, respectively, by {0.0000, 0.0006, 0.2721, 0.3417, 0.2287, 0.1016, 0.0354, 0.0141, 0.0043, 0.0015}, the rest having zero posterior probabilities. In our contrast, our model gives the respective probabilities as {0.0000, 0.0010, 0.4483, 0.4026, 0.1228, 0.0226, 0.0023, 0.0004, 0.0000, 0.0000}, the rest having zero posterior probabilities. These are not exactly same as the results obtained by RG, but broadly the results are similar, the posterior of number of components, k, favouring 3–5 components (see page 743 of RG). That our model performs much better than EW is reflected in $PBF(M_1/M_2) = 3656.644$. Figure 1 shows the density estimates of the histogram of the enzyme data using EW's model and our model. The density estimates look similar. It is however, worth mentioning in this context, that the density estimates we present in this application and other two applications below, are not the same as presented by RG or EW (the latter consider the galaxy data only), as they did not use cross-validation to obtain the density estimates or sample density estimates. It is apparent from the papers by RG and EW that their densities were not evaluated at the observed data points, but all the observed data points were used to estimate the densities at arbitrary points on the sample space (the Y-space). As a result, our figures are different from theirs. In our opinion, for the purpose of model comparison, use of our cross-validation predictive density estimates makes more sense.

It is important to note that while implementation of our approach took 42 minutes, implementation of that of EW took 5 hours and 24 minutes on a Mac OS X laptop.

8.2 Acidity data. Based on the priors of RG and EW we take $s = 4; S = 2 \times (0.2/0.573) = 0.6980803; \mu_0 = 5.02; a_\alpha = 2; b_\alpha = 4; \psi = 33.3; m = 30$.

In this case, according to the model of EW, the posterior probabilities of the number of components {1,2,3,4,5,6,7,8,9,10,11} are {0.0000, 0.1620, 0.3185, 0.2761, 0.1470, 0.0651, 0.0212, 0.0078, 0.0019, 0.0001 0.0003}, the rest having zero posterior probabilities. With our model these are {0.0000, 0.1091, 0.3444, 0.3092, 0.1628, 0.0564, 0.0147, 0.0028, 0.0003, 0.0002, 0.0001}. In this example as well, 3–5 components are favoured by our model, a result which is once gain in agreement with the analysis of RG. The model of EW seems to favour 2–5 components. The pseudo Bayes factor again prefers our model to that of EW; $PBF(M_1/M_2) = 38.97128$. Figure 2 shows the density estimates of the histogram of the acidity data using EW's model and our model; again, the density estimates look similar.

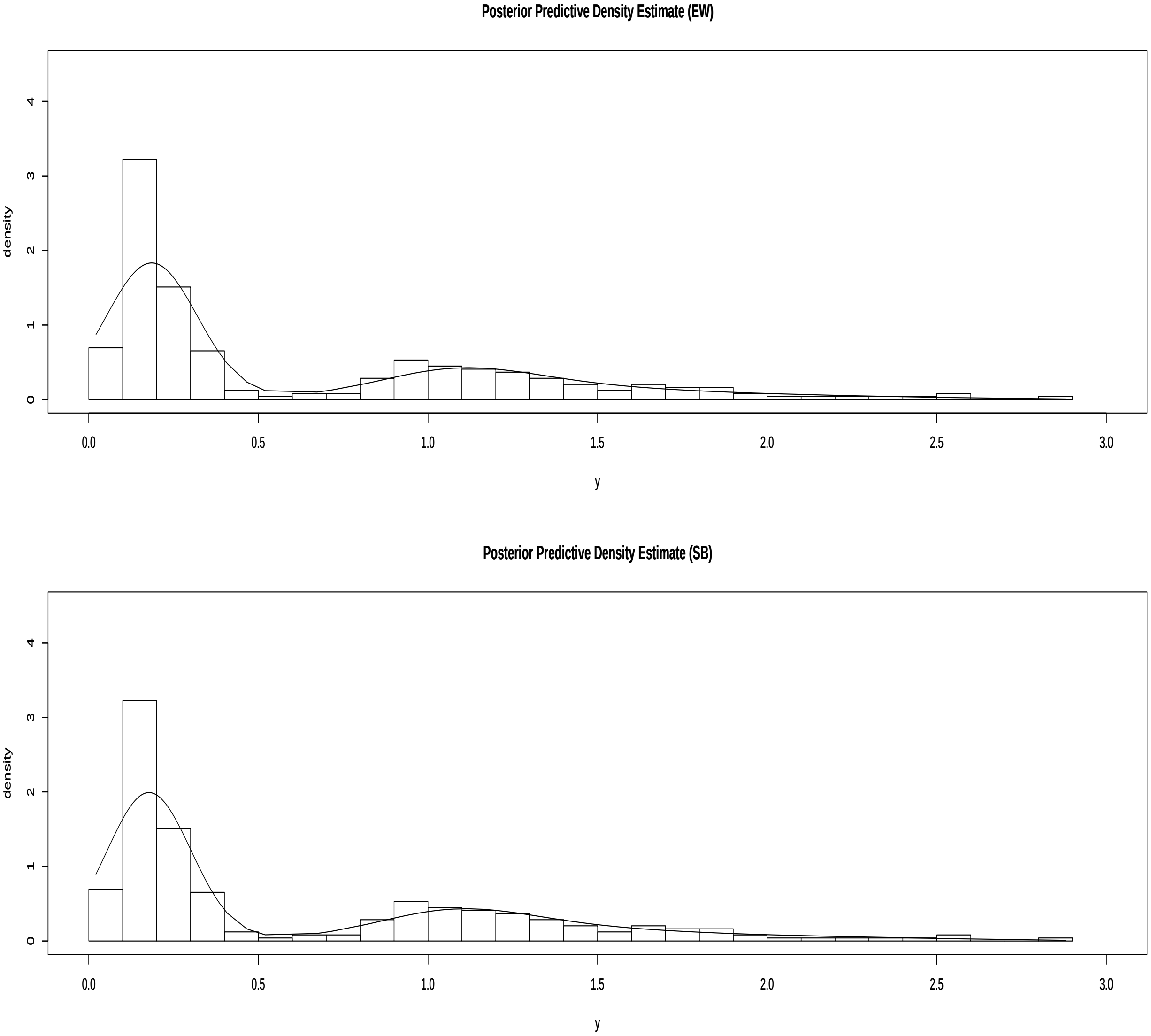
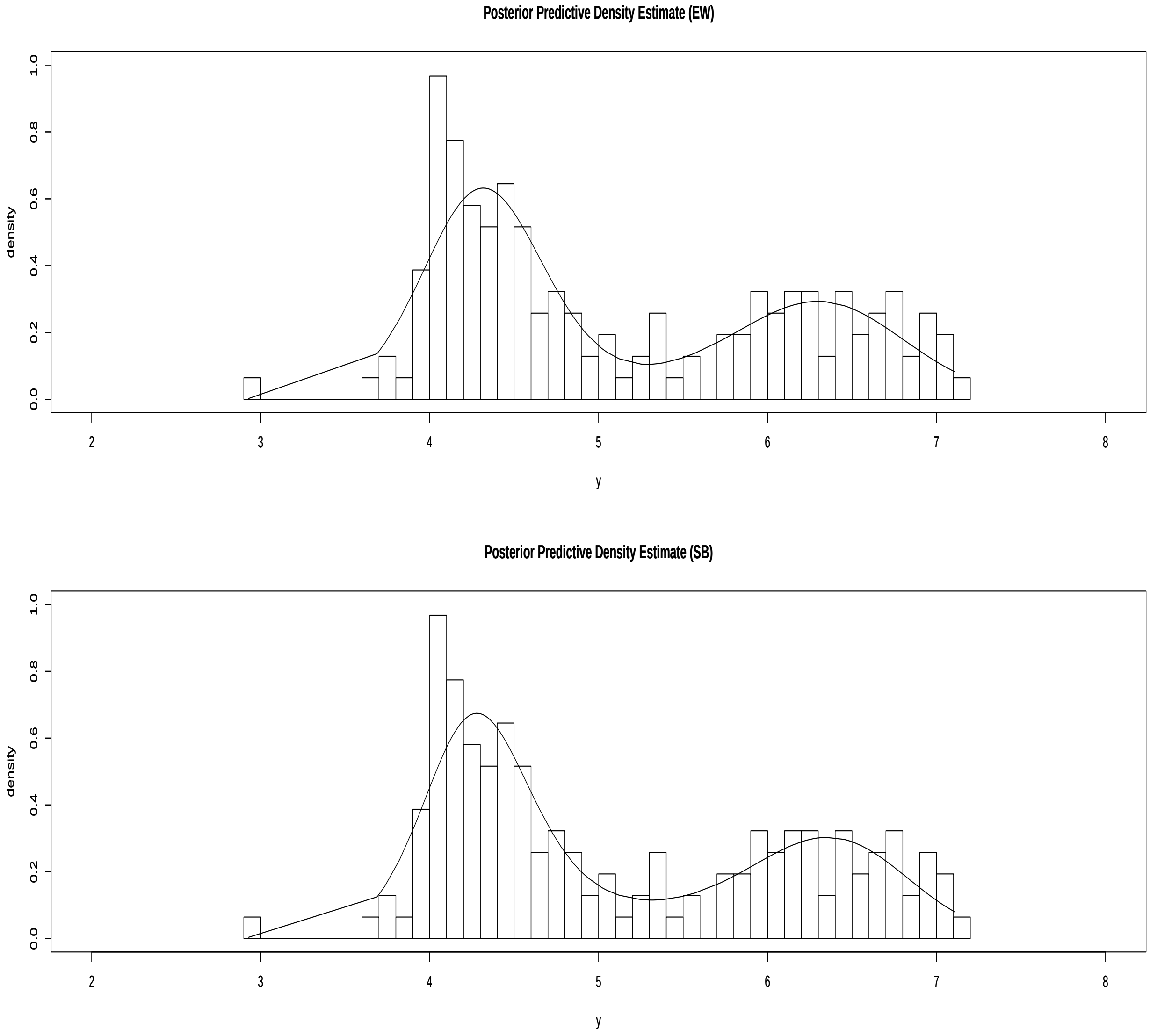


Figure 1: Density estimates of the histogram of the enzyme data using the model of EW (upper panel) and our model (lower panel).


Posterior Predictive Density Estimate (EW)

Posterior Predictive Density Estimate (SB)

Figure 2: Density estimates of the histogram of the acidity data using the model of EW (upper panel) and our model (lower panel).

In this case, implementation of our approach took 29 minutes, while that of EW took 2 hours and 4 minutes.

8.3 Galaxy data. Following EW we consider the following values of the prior parameters: $s = 4; S = 2; \mu_0 = 20; a_\alpha = 2; b_\alpha = 4; \psi = 33.3; m = 30$.

For the galaxy data, under EW's model, the posterior probabilities of the number of components {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16} are given by {0.0000, 0.0000, 0.0893, 0.1760, 0.2148, 0.1951, 0.1495, 0.0891, 0.0487, 0.0233, 0.0091, 0.0035, 0.0008, 0.0006, 0.0001, 0.0001} and the others have zero posterior probabilities. So, in this example, $k = 4, 5, 6, 7$ are well supported by the model of EW. For our model, the posterior probabilities of k are given by {0.0000, 0.0000, 0.0035, 0.0322, 0.1210, 0.2072, 0.2354, 0.1895, 0.1210, 0.0574, 0.0247, 0.0055, 0.0021, 0.0004, 0.0001}, thus putting most posterior mass on $k = 5, 6, 7$; RG too mention that 5–7 components are indicated by the galaxy data with their RJMCMC analysis. $PBF(M_1/M_2) = 495439.7$ shows that our model is much better supported by the data, as compared to the model of EW. Figure 3 displays the density estimates of the histogram of the galaxy data using the model of EW and the model we proposed. However, unlike in the cases of enzyme and acidity data, in this case, our density estimate, even in the naked eye, looks much accurate than that of EW.

Again, our approach turned out to be much faster than that of EW, the respective implementation time being 18 minutes and 36 minutes.

9 Conclusion

We have proposed a new and simple approach for Bayesian modelling and inference on mixture models, and argued that our proposal is better than that of RG implementation-wise, and is better than that of EW model-wise (in terms of pseudo Bayes factors) and in terms of the ability to take account of prior information about the number of mixture components in the population. Moreover, computationally of our approach is much less demanding than the approaches of EW and RG. Numerical experiments, conducted with simulated as well as real data sets, have demonstrated considerable power and flexibility of our approach, and confirmed the arguments we put forward in support of our proposal.

Another advantage of our proposed methodology is the ease with which regression function estimation can be done. Suppose, for example, that we have d-variate data, $\{\mathbf{y}_i; i = 1, \ldots, n\}$, where $\mathbf{y}_i = (y_{1i}, \ldots, y_{di})'$. Now

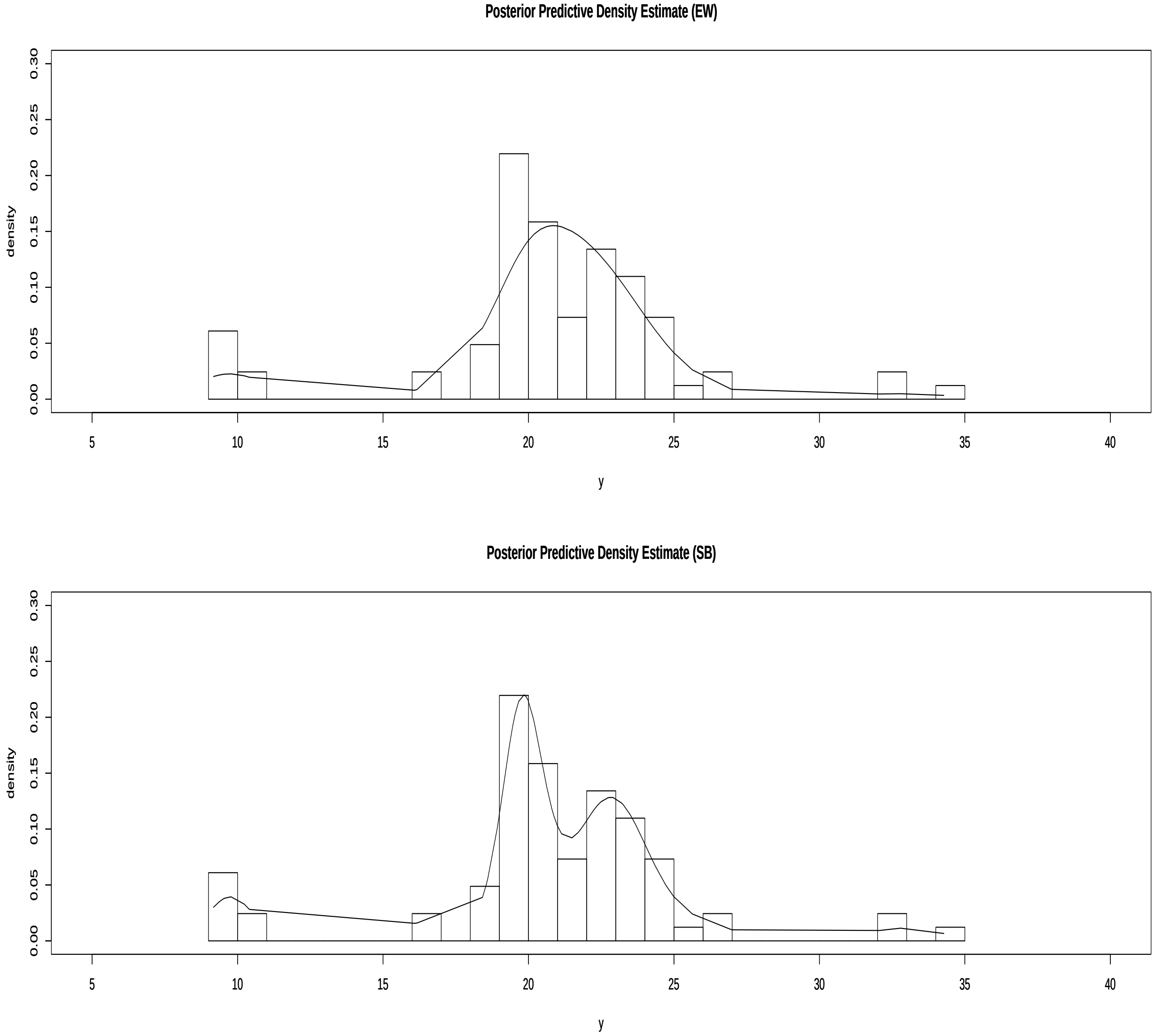


Figure 3: Density estimates of the histogram of the galaxy data using the model of EW (upper panel) and our model (lower panel).

suppose that it is of interest to obtain a weighted regression of y_1 on the other components $y_2, \ldots, y_d$. Now, according to our model, $\mathbf{y}$ is a mixture of d-variate normal distributions, given by

$$[\mathbf{y} \mid \mathbf{\Theta}_m] = \frac{1}{m} \sum_{j=1}^{m} \frac{|\mathbf{\Lambda}_j|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\left\{ -\frac{1}{2} \left(\mathbf{y} - \boldsymbol{\mu}_j\right)' \mathbf{\Lambda}_j \left(\mathbf{y} - \boldsymbol{\mu}_j\right) \right\} \tag{9.1}$$

where $\boldsymbol{\mu}_j$ and $\mathbf{\Lambda}_j$ are, respectively, the multivariate means and the inverse of the dispersion matrix (precision matrix) of $\mathbf{y}$. In simplified notation, we write (9.1) as

$$[\mathbf{y} \mid \mathbf{\Theta}_m] = \frac{1}{m} \sum_{j=1}^{m} N_d \left(\mathbf{y} : \boldsymbol{\mu}_j, \mathbf{\Lambda}_j^{-1} \right) \tag{9.2}$$

It follows that the conditional distribution of y_1 given $\mathbf{y}_{-1} = (y_2, \ldots, y_d)'$ is given by

$$[y_1 \mid \mathbf{\Theta}_m, \mathbf{y}_{-1}] \propto \frac{1}{m} \sum_{j=1}^{m} N_{d-1} \left(\mathbf{y}_{-1} : \boldsymbol{\mu}_{-1j}, \mathbf{\Lambda}_{-1j}^{-1} \right) \times N \left(y_1 : \mu_{1|2,\ldots,d}^{(j)}, \lambda_{1|2,\ldots,d}^{(j)} \right) \tag{9.3}$$

where $\mu_{1|2,\ldots,d}^{(j)}$ and $\lambda_{1|2,\ldots,d}^{(j)}$ are, respectively, the univariate conditional mean $E(y_1 \mid \mathbf{y}_{-1}, \mathbf{\Theta}_m)$ and the inverse precision $1/V(y_1 \mid \mathbf{y}_{-1}, \mathbf{\Theta}_m)$ under the assumption that $\mathbf{y} \sim N_d(\boldsymbol{\mu}_j, \mathbf{\Lambda}_j)$. The $(d-1)$ dimensional parameters $\boldsymbol{\mu}_{-1j}, \mathbf{\Lambda}_{-1j}$ stand for $\boldsymbol{\mu}_j, \mathbf{\Lambda}_j$ but without the first component.

As a result, assuming k distinct components $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*$ in $\mathbf{\Theta}_m$, and assuming further that each distinct component $\boldsymbol{\theta}_j^*$ occurs m_j times, we have

$$E[y_1 \mid \mathbf{\Theta}_m, \mathbf{y}_{-1}] = \sum_{j=1}^{k} w^{(j)}(\mathbf{y}_{-1}) \mu_{1|2,\ldots,d}^{(j)}, \tag{9.4}$$

which is a weighted sum of the component regression functions $\mu_{1|2,\ldots,d}^{(j)}$, where the associated weight $w^{(j)}(\mathbf{y}_{-1})$ is given by

$$w^{(j)}(\mathbf{y}_{-1}) \propto \frac{m_j}{m} N_{d-1} \left(\mathbf{y}_{-1} : \boldsymbol{\mu}_{-1j}^*, {\mathbf{\Lambda}^*}_{-1j}^{-1} \right) \tag{9.5}$$

and the proportionality constant is chosen such that $\sum_{j=1}^{k} w^{(j)}(\mathbf{y}_{-1}) = 1$.

Note that the regression function estimator developed above is structurally quite different from that given by Müller et al (1996). It will an interesting future work to compare our regression function estimator (9.4) with that of Müller et al (1996) after subjecting both the methodologies to various challenging applications.

Acknowledgements. We are grateful to Mr. Mriganka Chatterjee for his assistance in the preparation of this manuscript, and to an anonymous referee, whose comments have led to an improved presentation of the paper.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes With Applications to Nonparametric Problems., *Ann. Statist,*, **2**, 1152–1174.

Bhattacharya, S. (2006). A Bayesian semiparametric model for organism based environmental reconstruction, *Environmetrics*, **17** , 763–776.

Blackwell, D. and McQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes, *Ann. Statist,*, **1**, 353–355.

Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems, *Biometrika*, **51**, 481–483.

Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference using Mixtures, *J. Amer. Statist. Assoc.*, **90**, 577–588.

Ferguson, T. S. (1973). A Bayesian Analysis of some Nonparametric Problems, *Ann. Statist.*, **1**, 209–230.

Geisser, S. (1975). The predictive sample reuse method with applications, *J. Amer. Statist. Assoc.*, **70**, 320–328.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.*, **74**, 153–160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, Interdisciplinary Statistics, pages 145–162, London. Chapman and Hall.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations, *J. Roy. Statist. Soc. B*, **56**, 501–514.

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

Jeffreys, H. (1961). *Theory of Probability.* 3rd edition. Oxford University Press, Oxford.

MacEachern, S. N. (1994). Estimating normal means with a conjugate-style Dirichlet process prior, *Communications in Statistics: Simulation and Computation*, **23**, 727–741.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*, New York: Dekker.

Müller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**, 67–79.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *J. Roy. Statist. Soc. B*, **59**, 731–792.

RICHARDSON, S. and GREEN, P. J. (1998). Corrigendum: On Bayesian analysis of mixtures with an unknown number of components (with discussion), *J. Roy. Statist. Soc. B*, **560**, 661.

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods.* Springer-Verlag, New York.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion), *J. Roy. Statist. Soc. B*, **36**, 111–147.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: John Wiley & Sons.

SOURABH BHATTACHARYA
BAYESIAN AND INTERDISCIPLINARY RESEARCH UNIT
INDIAN STATISTICAL INSTITUTE
203, B. T. ROAD
KOLKATA 700108
INDIA
E-mail: sourabh@isical.ac.in

Paper received March 2008; revised November 2008.