Sankhyā : The Indian Journal of Statistics
2012, Volume 74-B, Part 1, pp. 77-106

Fast and efficient Bayesian semi-parametric curve-fitting and clustering in massive data

Sabyasachi Mukhopadhyay, Sisir Roy and Sourabh Bhattacharya
Indian Statistical Institute, Kolkata, India

Abstract

The problem of curve-fitting and clustering using Bayesian mixture models, treating the number of components as unknown, has received wide attention in the Bayesian statistical community. Among a number of available Bayesian methodologies specialised for the purpose, the approaches proposed in Escobar and West (1995) and Richardson and Green (1997) stand out. But in the case of massive data substantial computational challenges seem to blur the attractive theoretical advantages of such pioneering Bayesian methodologies. Based on a methodology introduced by Bhattacharya (2008), which, as we show, includes the approach of Escobar and West (1995) as a special case, we propose a very fast and efficient curve-fitting and clustering methodology. Our clustering approach is based on a new approach to analysing non-parametric posterior distributions of clusterings first proposed in Mukhopadhyay, Bhattacharya and Dihidar (2011). Significant advantages of our approach over the aforementioned established mixture modeling approaches, particularly in the case of massive data, are demonstrated theoretically and with extensive simulation studies. We also illustrate our methodologies on a real, cosmological data set consisting of 96,307 bivariate observations and demonstrate that the approach of Escobar and West (1995) is infeasible in this example and the approach of Richardson and Green (1997), although implementable, is likely to be inefficient and computationally expensive.

AMS (2000) *subject classification.* Primary 62G08; Secondary 91C20.
Keywords and phrases. Cluster analysis, Cosmology, Dirichlet process, Model validation, Markov chain Monte Carlo, Non-linear regression, Reversible jump Markov chain Monte Carlo.

1 Introduction

The theory of mixture modeling provides a very versatile framework for statistical analysis of data. In particular, such a framework provides the basis for density estimation, cluster analysis, and even semi-parametric regression. When the number of mixture components, k, is unknown, which is often the case in reality, it is either needed to estimate k, which is valid from a frequentist standpoint, or k may be treated as a unknown quantity, and uncertainty about k may be quantified by a prior distribution. The latter philosophy is exclusively Bayesian. Indeed, as the Bayesians might argue, simply providing a point estimate fails to properly account for the uncertainty about k. However, treating k as a random variable in the Bayesian paradigm introduces significant computational challenges, particularly in the case of massive data. Two prominent Bayesian methodologies that are cut out to handle statistical analyses associated with mixtures with variable number of components, are (a) The methodology based on Dirichlet process mixtures, proposed in Escobar and West (1995) (henceforth, EW) and (b) The methodology based on reversible jump Markov chain Monte Carlo (RJMCMC) proposed in Richardson and Green (1997) (henceforth, RG). Unfortunately, despite being pioneering, the above methods are also vulnerable to the problems associated with massive data.

Bhattacharya (2008) (henceforth, SB) described another methodology which may be seen as bridging the ideas of EW and RG, using the good features of both, creating, arguably, a more powerful methodology. For instance, we show in Section 1.2 that the model of EW is a special case of the model of SB. Also, the approach set out in SB permits straightforward and efficient Gibbs sampling for data sets of any size and any dimensionality; in contrast, RJMCMC is inefficient for high-dimensional data and EW's approach in infeasible computationally and inefficient for data sets with large number of observations even if the data is univariate.

1.1. Mixture model of SB extended to multivariate data. We assume that for $i = 1, \ldots, n$, the data set $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ is available, where observation $\boldsymbol{y}_i$ is $d(\geq 1)$-variate. This can be modeled as a mixture of d-variate normal distributions, having p components. Crucially, p is assumed to be unknown. Rather than assuming a prior distribution on p like RG and treating the problem as variable dimensional, we assume the following form of mixture representation of the d-variate observation $\boldsymbol{y}_i$:

$$[\boldsymbol{y}_i \mid \boldsymbol{\Theta}_M] = \frac{1}{M} \sum_{j=1}^{M} \frac{|\boldsymbol{\Lambda}_j|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_j\right)' \boldsymbol{\Lambda}_j \left(\boldsymbol{y}_i - \boldsymbol{\mu}_j\right) \right\} \tag{1.1}$$

In the above, $M(\geq p)$ is the maximum number of components the mixture can possibly have, and is known; $\boldsymbol{\Theta}_M = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, with $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j)$. We further assume that $\boldsymbol{\Theta}_M$ are samples drawn from a Dirichlet process (see, for example, Ferguson 1974):

$$\begin{aligned} \boldsymbol{\theta}_j &\overset{iid}{\sim} \boldsymbol{G} \\ \boldsymbol{G} &\sim DP(\alpha \boldsymbol{G}_0) \end{aligned}$$

In this paper, we assume that under $\boldsymbol{G}_0$,

$$\begin{aligned} [\boldsymbol{\Lambda}_j] &\sim Wishart_d\left(\frac{s}{2}, \frac{\boldsymbol{S}}{2}\right) \\ \left[\boldsymbol{\mu}_j \mid \boldsymbol{\Lambda}_j\right] &\sim N_d\left(\boldsymbol{\mu}_0, \psi \boldsymbol{\Lambda}_j^{-1}\right) \end{aligned}$$

Hence, the joint distribution of $\boldsymbol{\theta}_j$ is given by

$$\begin{aligned} [\boldsymbol{\Lambda}_j][\boldsymbol{\mu}_j \mid \boldsymbol{\Lambda}_j] &= c\,|\boldsymbol{\Lambda}_j|^{\frac{s-d-1}{2}} \exp\left\{-tr\left(\frac{\boldsymbol{S}\boldsymbol{\Lambda}_j}{2}\right)\right\} \\ &\quad \times \frac{|\boldsymbol{\Lambda}_j|^{\frac{1}{2}}}{(2\pi\psi)^{\frac{d}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_0)'\boldsymbol{\Lambda}_j(\boldsymbol{\mu}_j - \boldsymbol{\mu}_0)\right\} \end{aligned}$$

where

$$c = \frac{\pi^{-\frac{d(d-1)}{4}} \left|\frac{\boldsymbol{S}}{2}\right|^{\frac{s}{2}}}{\prod_{l=1}^{d} \Gamma\left\{\frac{1}{2}(s+1-l)\right\}}$$

Due to the discreteness of the prior distribution $\boldsymbol{G}$, the parameters $\boldsymbol{\theta}_\ell$ are coincident with positive probability. This property can be exploited to show that (1.1) reduces to the form

$$[\boldsymbol{y}_i \mid \boldsymbol{\Theta}_M] = \sum_{j=1}^{p} \pi_j \frac{\left|\boldsymbol{\Lambda}_j^*\right|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_j^*\right)' \boldsymbol{\Lambda}_j^* \left(\boldsymbol{y}_i - \boldsymbol{\mu}_j^*\right)\right\}$$

where $\left\{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_p^*\right\}$ are p distinct components in $\boldsymbol{\Theta}_M$ with $\boldsymbol{\theta}_j^*$ occuring M_j times, and $\pi_j = M_j/M$. Hence, although our model is actually variable dimensional, this is induced through the Dirichlet process prior, and does not involve complexities as in RJMCMC. Our modeling approach and the associated advantages remain valid for mixtures of any standard parametric density of the form $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ (either discrete or continuous, univariate or multivariate), not just for Gaussian mixtures. Although in this paper we confine ourselves to situations where $\boldsymbol{G}_0$ and $\boldsymbol{G}_0(\boldsymbol{\theta})$ and $f(\cdot \mid \boldsymbol{\theta})$ are conjugate, even for non-conjugate situations a Gibbs sampling-based methodology is available; see Mukhopadhyay and Bhattacharya (2012).

Table 1: Time taken to complete 20,000 MCMC iterations using SB's model and the number of iterations completed using EW's model at the same time.

Sample size	Time for 20,000 iterations with SB	#iterations in the same time with MEW
1,000	2 mins 48 secs	2371
5,000	13 mins 41 secs	29
10,000	27 mins 28 secs	6
20,000	54 mins	3

1.2. EW is a special case of SB. To show that the model of SB generalizes that of EW, we first represent (1.1) using allocation variables $\boldsymbol{Z} = (z_1, \ldots, z_n)'$, as follows:

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

$$\begin{aligned} [\boldsymbol{y}_i \mid z_i = j, \boldsymbol{\Theta}_M] &= \frac{|\boldsymbol{\Lambda}_j|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_j\right)' \boldsymbol{\Lambda}_j \left(\boldsymbol{y}_i - \boldsymbol{\mu}_j\right)\right\} \\ [z_i = j] &= \frac{1}{M} \end{aligned}$$

When $M = n$, then conditional on $z_i = i$ for $i = 1, \ldots, M(= n)$, the above reduces to EW's model, showing that the latter is a special case of SB's model.

1.3. Demonstration of computational speed in large data sets. Table 1 summarizes the computational time of SB's model to yield 20,000 Gibbs sampling simulations with relatively large sample sizes where $d = 2$ and $M = 30$ in each case. The table also shows the number of iterations of an efficient Gibbs sampling algorithm developed by Müller, Erkanli and West (1996) (henceforth, MEW), corresponding to EW's model, completed in the same computational time. All computations have been carried out in a work station of 3 GB RAM and consisting of two processors, each running at 3 GHz. The table shows that for large data sets, although SB's approach remains computationally cheap, that of EW tends to be infeasible.

Using the multivariate mixture model (1.1), we develop a semi-parametric Bayesian curve-fitting procedure and demonstrate that the approach is particularly suitable for application in massive data, where the methods based on EW and RG may be either infeasible or inefficient. Simulation studies indicate that even in the case of non-massive data our method performs satisfactorily and can outperform the EW-based curve-fitting idea of MEW.

We also adopt the ideas presented in Mukhopadhyay et al. (2011) (henceforth, MBD) for analysing the posterior distribution of clusterings associated with the Bayesian mixture model of SB. Using the model of EW, MBD showed how to obtain *a posteriori* "central clustering" and posterior credible regions of clusterings of any desired level. We demonstrate theoretical and computational advantages associated with the usage of the clustering ideas of MBD in conjunction with SB's model. We also demonstrate with simulations that for large data sets, the true clustering of the data is more likely to fall within the Markov chain Monte Carlo (MCMC)-based 95% credible regions of the posterior distribution of clusterings associated with the model of SB, compared to those associated with EW. These are all new findings and are not discussed in MBD.

The rest of the paper is structured as follows. In Section 2 we provide a brief overview of Bayesian mixture models with unknown number of components. Important features of our model are described in Section 3 and computational advantages of our model and methodology over existing ideas are considered in Section 4. Our Bayesian semi-parametric regression method is introduced in Section 5. In Section 6 we illustrate with simulation studies the advantages of our semi-parametric regression over MEW. Our clustering ideas based on MBD are introduced in Section 7; advantages of the ideas used in conjunction with SB's model are illustrated with simulation studies in Section 8. In Section 9 we consider application of our methods to a massive, real cosmological data. Conclusions and future work are enlisted in Section 10.

2 Overview of mixture models

Mixture models are noted for their flexibility. Indeed, as noted by Dalal and Hall (1983) and Diaconis and Ylvisaker (1985), mixture models composed of standard densities can, in principle, approximate any underlying distribution. For more on mixture models, see McLachlan and Basford (1988), Titterington, Smith and Makov (1985). However, a technical problem associated with classical analysis of mixture models is associated with the number of mixture components included in the model. Various methods (for a recent review, see Lee et al., 2008) may be used to obtain a point estimate of the unknown number of components. In the Bayesian paradigm, a prior distribution of k is specified, either explicitly or implicitly. The methodology of RG explicitly specifies a prior on k and uses RJMCMC to obtain samples from the resulting variable-dimensional posterior, while

that of EW implicitly induces a prior on k by assuming a Dirichlet process mixture, thus avoiding variable-dimensionality.

The RJMCMC methodology of RG is complicated and is error prone. But of more concern is its much sensitivity to the "move types", the transformations and the proposal distributions selected, often resulting in inefficiency. Diagnosis of convergence of RJMCMC is another serious problem. These problems are many times aggravated for multivariate observations; for instance, the proposal distributions suitable for univariate cases will, in general, yield poor acceptance rates in multivariate situations. An instance in which RJMCMC may be inefficient in the extreme is the so-called "large p small n" paradigm.

The model of EW is not variable dimensional and straightforward Gibbs sampling algorithms have been developed by Bush and MacEachern (1996), MacEachern (1994), MEW. However, as we argue in Section 4.3 (see also MBD for more thorough discussion), these algorithms are ineffective when the data size is massive. More recently, methods based on the Metropolis-Hastings sampler, devised by Jain and Neal (2004, 2007), although demonstrated improved mixing in certain examples, are slower than Gibbs sampling with respect to computation time per iteration (this is discussed in Section 4.3.3 of the former and in Section 7.1.3 of the latter paper). Hence these are computationally infeasible in massive data.

Moreover, the enormous clustering space of EW's model also impedes reliable MCMC-based analysis of massive data. More transparently, in the model of EW, the number of possible clusterings of the data increases exponentially with data size in accordance with the Bell number (Bell, 1934); see Section 4.2 So, in massive data, even computationally very fast MCMC algorithms with good convergence properties are unlikely to adequately explore the entire space of clusterings in finite time. Given manifold increase in computational complexity of the existing MCMC algorithms for EW's model, the problem of adequate exploration of EW's clustering space is only many times aggravated.

In an effort to reduce the computation time associated with the model of EW in massive data, Wang and Dunson (2011) have proposed the sequential updating and greedy search (SUGS) algorithm which proceeds by cycling through the data points, sequentially allocating them to the cluster that maximizes the conditional posterior allocation probability. The conditional distribution of the unknown parameter, which admits a closed form expression given the maximizing cluster, is then updated. A complete sweep of the algorithm yields the conditional posterior distribution of all the parameters, given the seuqentially optimal clusterings. The advantage

of the method of Wang and Dunson (2011) is that it is quite fast, since it does not rely upon MCMC methods. But it is not clear if the correct joint or marginal posterior distributions of the parameters or clusterings could be obtained or if the algorithm yields a global maximum *a posteriori* (MAP) estimate. Also, the algorithm of Wang and Dunson (2011) does not seem to assist in obtaining and studying the probability distribution of the clusterings.

We attempt to avoid most of the difficulties noted above by adopting the modeling idea of SB in which the increase in the number of clusterings is much slower compared to the model of EW, and the computation is much faster than all the MCMC methods associated with RG and EW. For instance, in our implementation, generation of 20,000 MCMC samples from SB's model took about 4 hours for the cosmology data. Recalling that efficient implementation of EW's model is infeasible and that the RJMCMC approach of RG suffers from the curse of dimensionality and other convergence-related problems, we recommend SB's model in massive data situations.

3 Features of SB's model

3.1. Two-stage clustering of the observations and the empty components. It is useful to provide the intuition behind the allocation variables $\boldsymbol{Z}$ and the parameters $\boldsymbol{\Theta}_M$. Given M distinct values of the parameter vector $\boldsymbol{\Theta}_M$, the allocation vector $\boldsymbol{Z}$ clusters the n-dimensional observation vector $\boldsymbol{Y}$ into M^* ($\leq M$) clusters of the form $U_j = \{i : z_i = j\}$; $j = 1, \ldots, M^*$. Empty clusters result when $M^* < M$ (which may occur when no observation is allocated to some, perhaps many, components); let these clusters be denoted by $\emptyset_j$; $j = M^* + 1, \ldots, M$. We denote by $\{U_1^*, \ldots, U_M^*\}$ the clustering $\{U_1, \ldots, U_{M^*}, \emptyset_{M^*+1}, \ldots, \emptyset_M\}$. This can be thought of as the *initial clustering*, since the Dirichlet process prior acts upon $\{U_1^*, \ldots, U_M^*\}$ to yield k ($\leq M$) distinct parameter values $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*$ out of the possible M distinct values to yield the final clustering, say, $\{V_1, \ldots, V_k\}$, of $\{U_1^*, \ldots, U_M^*\}$, with $V_\ell = \cup_{j:c_j=\ell} U_j^*$. Here $c_j = \ell$ if and only if $\boldsymbol{\theta}_j = \boldsymbol{\theta}_\ell^*$; $j, \ell = 1, \ldots, M$. Thus, the clusters V_ℓ are associated with the configuration vector $\boldsymbol{C} = (c_1, \ldots, c_M)'$. Clearly, the clustering $\{V_1, \ldots, V_k\}$ is coarser than $\{U_1^*, \ldots, U_M^*\}$ in the sense that the former consists of lesser number of blocks with more elements in each block. Note that even the final clustering $\{V_1, \ldots, V_k\}$ may consist of empty clusters, although some or all of the empty clusters in the initial clustering may get merged with the non-empty clusters. Hence, our proposal yields a two-stage clustering of the data and the empty components.

3.2. Learning about population number of clusters. The approach of EW does not allow for empty components. From this point of view, the model of SB is again more general than that of EW, and is more akin to the approach of RG in that it allows for empty components. Specifically, the main difference between the model of EW and that of RG and SB is that the former can capture only the sample number of clusters, while the latter captures both sample and population number of clusters. The empty components indicate population clusters from which data points did not, but could have arisen. Ignoring the empty clusters results in sample number of clusters while counting them along with non-empy clusters yield the population number of clusters. Although expected, it is important to note that the data seldom contains information about population number of clusters (McCullagh and Yang, 2008) and strong prior information is necessary to attempt to learn about it. This is possible in principle with SB's model by choosing $M >> n$ and choosing α large, thus probabilistically increasing the number of empty components. Such learning is also possible in RG's model by assigning large prior probabilities to large values of the number of components. But since occurrence of empty clusters is not possible in EW's model irrespective of any prior on α, it is impossible to learn about population number of clusters with EW's model, unless the sample and the population number of clusters are same. This drawback also manifests itself in the failure of EW's model to adequately learn about the true (population) regression curve associated with the mixture model, when the data size is small. We discuss these issues with simulation studies in Section 6.2.

3.3. Choice of M. At first glance it seems that rather than fixing the bound M it is more appropriate to put a prior on M, of the form $\pi_M(i) = P(M = i)$ for $i = 1, \ldots, L$, where L is either finite or tends to infinity. Let $\mathcal{M}_i$ denote the model corresponding to $M = i$. Also, let $\mathcal{C}_i$ denote the space of all clusterings supported by $\mathcal{M}_i$. Then, for finite L, $\mathcal{C}_1 \subset \mathcal{C}_2 \subset \cdots \subset \mathcal{C}_L$. Hence, $1 = P\left(\cup_{i=1}^{L}\mathcal{C}_i\right) = P(\mathcal{C}_L)$. This shows that all information is contained in $\mathcal{M}_L$, and that it is unnecessary to assign positive probabilities to $\mathcal{M}_i; i < L$. From the computational perspective, this prior would force burdensome MCMC-based exploration of $\mathcal{C}_i$ for each $i = 1, \ldots, L$ even though it is sufficient to explore only $\mathcal{C}_L$. If we let $L \to \infty$, then by the monotonicity theorem of probability, $\lim_{L\to\infty} P(\mathcal{C}_L) = P\left(\lim_{L\to\infty} \mathcal{C}_L\right) = P\left(\cup_{i=1}^{\infty}\mathcal{C}_i\right) = 1$. That is, for any $\epsilon > 0$, there exists $L_0(\epsilon) < \infty$ such that $P\left(\mathcal{C}_L\right) \geq 1-\epsilon$, for $L \geq L_0(\epsilon)$. The above arguments motivate an appropriate, finite, and deterministic choice of M. For the $L \to \infty$ case M can be likened to $L_0(\epsilon)$ for some adequate ϵ. Note that if, instead of an upper bound on the number of components, had M been the exact number of components, then

putting a prior on M would make sense. This is actually the principle used by the RJMCMC appoach of RG. Indeed, in that case, assuming non-empty components, $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ for any $i \neq j$, and the relations $P(\mathcal{C}_L) = 1$ or $\lim_{L\to\infty} P(\mathcal{C}_L) = 1$ do not hold, showing that it is essential to explore each model separately. But it is important to observe in this connection that in general the assumption of non-empty components does not hold for RG's set-up, implying $\mathcal{C}_i \subset \mathcal{C}_j$ for $i < j$. This suggests that the methodology of RG incurs inefficiency from the viewpoint of clustering the data. However, if $\mathcal{C}_i$ denotes clustering space of $\mathcal{M}_i$ with respect to the parameters instead of the data, then in RG's approach, each $\mathcal{C}_i$ is a singleton, and $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ for $i \neq j$. Hence, even prior on the exact number of components is completely sensible only from the viewpoint of clustering the parameters instead of the data.

The above arguments pertaining to our Dirichlet process-based mixture model show that fixed and finite values of M are most logical and computationally efficient, and only such values of M will be considered for the rest of this paper. Reasonable choice of M may be the prior guess of the scientific expert about the maximum number of clusters the data may possibly have. For instance, in astronomical applications the investigating astronomer may have reasons, theoretical or experimental, to believe that the number of clusters would not exceed a certain limit. Similarly, an ecologist may provide prior information about the maximum number of clusters of the different types of vegetation in a forest. However, in absence of any such information, one might proceed on a trial and error basis: if the posterior gives non-negligible mass to the fixed value of M it should be increased further until negligible posterior mass is achieved at M. One may also allow M to be a function of the data size n, but such that $M/n \to 0$ as $n \to \infty$. However, when the data size n is small and there are reasons to believe that such small data set is inadequate for inference about population characteristics, such as the population number of clusters, true regression function, true density, etc., then one must choose $M >> n$, as already discussed in Sections 5 and 3.1.

4 Computational advantages

Given a conjugate prior structure a fast and easily implementable Gibbs sampling algorithm is described in Sections 1 and 2 of the supplementary document. In Section 3 of the supplementary document we also provide

an alternative Gibbs sampling algorithm that takes advantage of the configuration vector $\boldsymbol{C}$. For reasons of efficiency, throughout we use the Gibbs sampling algorithm based upon the configuration vector.

4.1. Computational gain in updating empty components. Although the approach of SB as well as RJMCMC allow for empty components, the computation associated with empty components is more naturally and efficiently handled with SB's approach. If the j-th component is an empty component, then the fact $n_j = \#\{i : z_i = j\} = 0$, occurs naturally in SB's model, with the corresponding full conditional distribution of $\boldsymbol{\theta}_j$ boiling down to the full conditional distribution associated with the Dirichlet process prior. That is, no special care is necessary for validation of this step. But this situation requires an extra, careful, and complicated step in the method of RG.

4.2. Computational efficiency in large data sets. To compare the number of clusterings in the models of EW and SB, we first independently derive the number of ways of partitioning n items into k non-empty clusters. Using the inclusion-exclusion principle, we obtain the number of onto (surjective) mappings from the set $\{1, 2, \ldots, n\}$ to $\{1, 2, \ldots, k\}$ as $S^*(n,k) = \sum_{i=0}^{k} \binom{k}{i} (-1)^i (k-i)^n$. Since clustering is label-invariant $S^*(n,k)$ is actually $k!$ times the desired result. Hence, the required number of partitioning is $S(n,k) = S^*(n,k)/k!$, which, incidentally, is Stirling number of the second kind (see, for example, Abramowitz and Stegun, 1972).

With the above Stirling number it can be easily seen that the number of clusterings in EW's model with data size n, is the Bell number, $B(n) = \sum_{k=0}^{n} S(n,k)$. Since $B(n)$ grows exponentially with n (see Section 4 of the supplementary document), the number of clusterings in EW's model grows exponentially fast with data size. As a result, the clustering space can not be adequately explored using a finite number of MCMC simulations.

In SB's approach, a moderate value of M ensures that the number of possible clusterings of the data, given by $C(n,M) = \sum_{k=0}^{M} S(n,k)$, is very small compared to EW's case. As an illustration, let us consider the problem of obtaining the number of all possible clusterings of 10 items. Then the Bell number is given by $B(10) = \sum_{k=0}^{10} S(10,k) = 1,15,975$. In contrast, with, say, $M = 3$, the number of clusterings in SB's approach is $C(10,3) = \sum_{k=0}^{3} S(10,k) = 9,842$. Hence, an MCMC sample of size much more than $1,15,975$ is necessary to adequately explore the clustering space in EW's case, while, in SB's approach, a much smaller sample size will be adequate. Indeed, as we show in Section 4 of the supplementary document, the ratio $B(n)/C(n,M) \to \infty$ as $n \to \infty$. As a result, the Gibbs sampling algorithm can more efficiently explore the clustering space of SB's model compared to the algorithms designed for exploration of EW's model. In Section 8 we

demonstrate with simulation studies that because of the issues discussed above, Gibbs sampling in EW's model will generally fail to locate the true clustering of the data, whereas, in SB's model it will generally capture the true clustering successfully.

4.3. Discussion of computational complexity in large data sets. One might wonder whether the computational complexity of our algorithm given in Section 3 of the supplementary document is manageable. We assert that this is indeed the case, thanks to at most moderately large value of M. In our case, M is the number of possible values of each allocation variable, the number of configuration indicators and the maximum number of possible values each configuration indicator can take. In contrast, each of the n configuration indicators for EW's model can take n possible values, with n extremely large.

Since the expected number of distinct parameters is approximately $\alpha \log(1 + n/\alpha)$ (Antoniak, 1974), decreasing α will decrease the expected number of distinct parameters to be simulated, thus decreasing the computational complexity of simulation of the distinct components and hence that of each component of the configuration vector. However, in EW's approach even if α is chosen small enough so that the expected number of components is smaller than M and match that of SB, the overall computational complexity of SB's model is still negligible compared to that of EW's model. This is because, most importantly, only M-many configuration indicators need to be simulated in SB's model, while in EW's model, this is n-many, which tends to infinity as the data size n goes to infinity. Also, simulation of the entire allocation vector Z in SB's model has only negligible computational complexity; see MBD for details.

Further details provided in MBD show that marginalizing out the parameters as in MacEachern (1994) increases computational burden manifold due to increased computational complexity in each of the full conditionals, and that the non-marginalized version of SB's model utilizing the configuration vector is far more efficient computationally than any of the competing algorithms associated with EW's model.

5 Bayesian semi-parametric regression

In simplified notation, we write (1.1) as

$$[\boldsymbol{y} \mid \boldsymbol{\Theta}_M] = \frac{1}{M} \sum_{j=1}^{M} N_d \left(\boldsymbol{y} : \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j^{-1} \right)$$

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

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

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

As a result, assuming M^* distinct components $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{M^*}^*$ in $\boldsymbol{\Theta}_M$, and assuming further that each distinct component $\boldsymbol{\theta}_j^*$ occurs M_j times, we have,

$$E[y_1 \mid \boldsymbol{\Theta}_M, \boldsymbol{y}_{-1}] = \sum_{j=1}^{M^*} w^{(j)}(\boldsymbol{y}_{-1}) \mu_{1|2,\ldots,d}^{(j)}, \tag{5.1}$$

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

$$w^{(j)}(\boldsymbol{y}_{-1}) \propto \frac{M_j}{M} N_{d-1}\left(\boldsymbol{y}_{-1} : \boldsymbol{\mu}_{-1j}^*, \boldsymbol{\Lambda}^{*-1}_{-1j}\right). \tag{5.2}$$

The proportionality constant in (5.2) is chosen such that $\sum_{j=1}^{M^*} w^{(j)}(\boldsymbol{y}_{-1}) = 1$. Consistency of our mixture model and convergence of the associated Bayesian curve to the true curve are guaranteed under mild conditions; see Section 6 of the supplementary document for details.

Note that the regression function estimator developed above is structurally quite different from that given by MEW. One clear advantage of our curve over that of MEW is that for massive data the curve-fitting idea of MEW can not even be implemented due to extreme computational burden, while our curve (5.1) can be easily fitted to any data set, massive or not. Moreover, it is demonstrated in Section 6.2 that even in non-massive data, it is possible for our Bayesian curve to outperform that of MEW. The reason for better performance is connected with the difference between the sample and the population regression curve. Note that (5.1) is a weighted average of at most M linear components of the form $\mu_{1|2,\ldots,d}^{(j)}$, while the corresponding regression estimator of MEW consists of at most n linear components. Thus, for small to moderately large data set the estimator of MEW may be inadequate for approximating the true (population) regression function,

while M can be chosen large enough (in fact, one may choose $M >> n$ and a prior of α that supports large number of distinct components) in (5.1) to ensure adequate approximation with SB's model. We demonstrate this with a simulation study in Section 6.2.

Assuming that a sample $\left\{\boldsymbol{\Theta}_M^{(1)}, \ldots, \boldsymbol{\Theta}_M^{(N)}\right\}$ is available from the posterior distribution of $\boldsymbol{\Theta}_M$ (typically by MCMC), the marginalized curve $E(y_1 \mid \boldsymbol{y}_{-1})$ is estimated as

$$E(y_1 \mid \boldsymbol{y}_{-1}) = E(E(y_1 \mid \boldsymbol{\Theta}_M, \boldsymbol{y}_{-1})) \approx \frac{1}{N}\sum_{t=1}^{N} E(y_1 \mid \boldsymbol{\Theta}_M^{(t)}, \boldsymbol{y}_{-1})$$

Pointwise variability of the curve is measured by

$$Var(y_1 \mid \boldsymbol{y}_{-1}) = Var(E(y_1 \mid \boldsymbol{\Theta}_M, \boldsymbol{y}_{-1})) + E(Var(y_1 \mid \boldsymbol{\Theta}_M, \boldsymbol{y}_{-1}))$$

The first component of the above variance is estimated by the sample variance of $\{E(y_1 \mid \boldsymbol{\Theta}_M^{(t)}, \boldsymbol{y}_{-1});\ t = 1, \ldots, N\}$, and the second component is estimated by the sample mean of $\{Var(y_1 \mid \boldsymbol{\Theta}_M^{(t)}, \boldsymbol{y}_{-1});\ t = 1, \ldots, N\}$. Approximate $100(1-\tau)\%$ pointwise credible intervals of the curve are given by $E(y_1 \mid \boldsymbol{y}_{-1}) \pm z_{\tau/2}\sqrt{Var(y_1 \mid \boldsymbol{y}_{-1})}$, where z_τ is the 100τ-th quantile of a standard normal distribution.

Hence, once the MCMC realizations $\left\{\boldsymbol{\Theta}_M^{(1)}, \ldots, \boldsymbol{\Theta}_M^{(N)}\right\}$ are available, it is an easy task to obtain a Bayesian regression curve with all summaries readily available.

6 Simulation studies on semi-parametric regression

6.1. Illustration of the performance of our curve-fitting procedure. We assume a bivariate normal distribution of two random variables (y, x) (that is, $d = 2$), where the true regression function of y on x is $\mu(x) = x + \sin x$, a highly non-linear curve. We assume that $x \sim Uniform(0, 1)$ and given x, $y \sim N(\mu(x), 0.7^2)$. Pretending that the true curve is unknown, and that all we have is a sample of 1000 observations $(x_i, y_i); i = 1, \ldots, 1000$, we demonstrate that our curve-fitting idea can accurately estimate the (unknown) true curve. We obtain the data by actually simulating from the bivariate distribution of $(x, y) \sim Uniform(x : 0, 1) \times N(y : \mu(x), 0.7^2)$.

Some of the prior parameters for our model to be fitted in this example are chosen such that fast convergence to the target posterior is ensured. For example, selecting $\boldsymbol{\mu}_0$ to be the sample mean vector, and $\boldsymbol{S}$ to be the sample dispersion matrix ensured good mixing properties of our Gibbs sampler.

Table 2: Two-way table showing the deviations of the fitted curve from the true curve.

Value of α	Deviation
0.5	1.004
1.0	0.896
5.0	0.597
10.0	0.4898
15.0	0.4154
25.0	0.355

We fixed $M = 30$. Other choices (and justifications thereof) are motivated by those of EW, RG, and SB. However, it is important to select the prior parameters of α carefully, since this can significantly affect the probability distribution of the number of components, and hence the fit of the curve. We postulate $\alpha \sim Gamma(a_\alpha, b_\alpha)$ with a mode to be determined by a procedure described below. The parameters a_α and b_α will be chosen to yield the predetermined mode and a variance large enough to quantify our vagueness about the prior.

To determine the mode of the prior of α, we fit the Bayesian curve with many fixed values of α, and compute the maximum absolute difference between the true curve and the fitted curve, given by $\max_{1\le i\le 1000} |\hat{E}(y \mid x_i^*) - \mu(x_i^*)|$. Here $x_i^*, i = 1, \ldots, 1000$ are equidistant points in the interval $(0, 1)$. We choose that value of α as the prior mode for which the deviation is less than 0.4 and the fitted curve contains most of the features of the true curve. The threshold of the deviation (here 0.4) is chosen somewhat large to prevent overfitting of the model. Table 2 displays the maximum absolute deviations corresponding to a fixed value of α. To obtain each row of Table 2 we ran our Gibbs sampler for 20,000 iterations, discarding the first 5,000 iterations as burn-in. From the table we chose the value 25 as the mode of the prior distribution of α. Equating the mode of $Gamma(a_\alpha, b_\alpha)$ to 25 gives $(a_\alpha - 1)/b_\alpha = 25$, so that $a_\alpha = 25b_\alpha + 1$. Now note that the variance of $Gamma(a_\alpha, b_\alpha)$ is $a_\alpha/b_\alpha^2 = (25/b_\alpha) + 1/b_\alpha^2$. Fixing $b_\alpha = 0.1$ yields a considerably large variance of 350. Hence, we fixed $b_\alpha = 0.1$, which implies $a_\alpha = 25b_\alpha + 1 = 3.5$.

Figure 1 shows that the true regression function (black-coloured) is estimated quite accurately by the fitted Bayesian semi-parametric curve (red-coloured). Moreover, the pointwise approximate 95% credible

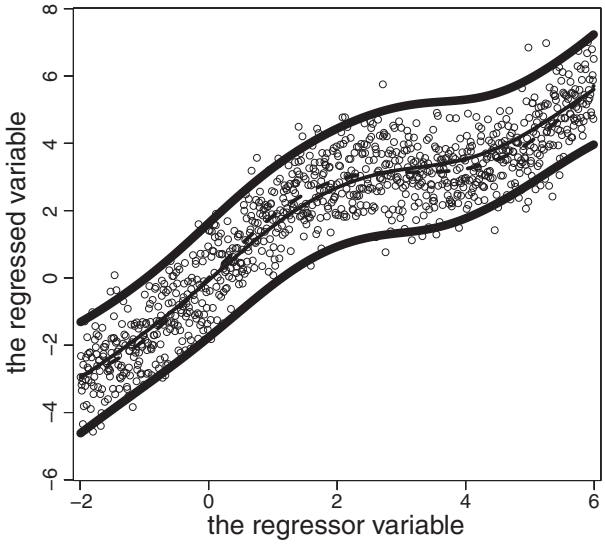

Figure 1: Bayesian curve fitting: the fitted curve (continuous line at the center) and the true curve (broken line at the center) associated with the simulation study. The thick curves denote pointwise approximately 95% credible intervals.

intervals (blue-coloured) show that the entire true curve lies within the credible limits. This is very encouraging, given that the true regression is highly non-linear.

6.2. Comparison with the curve-fitting approach of MEW. We simulated 500 data sets, each set consisting of 15 observations (x_i, y_i); $i = 1, \ldots, 15$ drawn from the distribution $Uniform(x : 0, 1) \times N(y : \mu(x), 0.7^2)$, with $\mu(x) = x + \sin(x)$, as before. We fixed $\alpha = 1000$ for both the methods. This relatively high value of α is expected to compensate for the lack of enough information in the data set. The other hyperparameters are fixed, for both the approaches. Specifically, $\boldsymbol{\mu}_0$=(7.33, 6.29), $s = 12$, the matrix $\boldsymbol{S}$ has entries $S_{11} = 5.34$, $S_{12} = S_{21} = 5.14$, $S_{22} = 6.16$, $\psi = 3.0$. We fixed $M = 50$ for our approach. For comparison of the methods, we used the quadrature versions of the L_1, L_2 metric, given by $1/15 \sum_{i=1}^{15} \left| \hat{E}(y \mid x_i^*) - \mu(x_i^*) \right|$, $\left\{ 1/15 \sum_{i=1}^{15} \left(\hat{E}(y \mid x_i^*) - \mu(x_i^*) \right) \right\}^{1/2}$, and also the maximum absolute deviation $\max_{1 \le i \le 15} \left| \hat{E}(y \mid x_i^*) - \mu(x_i^*) \right|$. As before, $\{x_i^*\}$ are equidistant points in the interval $(0, 1)$.

We found that in 80%, 84% and 82% cases the quadrature versions of L_1, L_2 and the maximum absolute deviations are smaller for our fitted curve. In 94% cases the mean length of the 95% credible intervals and in 83% of the cases the maximum length of the 95% credible intervals corresponding to our curve turned out to be less than those of the curve of MEW. For both the approaches, the true values fell within the 95% credible regions 100% times.

The above results strongly suggest that our curve significantly outperforms that of MEW. This is not unexpected, since the curve of MEW can accommodate at most 15 linear components in this case. This problem is avoided in our approach by setting $M = 50$. Larger values of M did not indicate further significant improvement, indicating that the choice is appropriate. Thus, this example demonstrates that in data sets consisting of small to moderate number of observations, our approach is expected to outperform the approach of MEW. Fundamentally, the comparatively poor performance of MEW is due to its inability in learning about the population number of clusters, as discussed in Section 3. Here the sample number of clusters is inadequate in providing information about the population regression curve associated with population number of clusters. The improved performance of our model can be attributed to learn about the population number of clusters and the population regression curve, through using available prior information of α. This example also indicates that our approach is expected to be a suitable candidate for handling the "large p small n" problem, where the data size is small and each datum is high-dimensional.

7 Bayesian posterior distribution of clustering

Scientific cluster analyses usually require summaries of clusterings (that is, how the data are partitioned into different clusters), rather than just the number of clusters in the data. It is also important to note that two different clusterings of the same data may consist of the same number of clusters. Thus, a methodology is needed which provides not only the posterior distribution of the number of clusters, but that of clustering itself, using which summaries of clusterings may be obained. This problem is much more difficult as compared to obtaining the posterior summary of a regular parameter. For instance, it is not proper to take means of clusterings produced; assuming non-empty components, although each iteration of the Gibbs sampler may yield less than M clusters, the mean clustering may

still consist of M clusters. Moreover the clusterings are permutation invariant. That is, two clusterings may be same except for a permutation of the components.

Based on product partition models, Dahl (2009) proposed an algorithm to obtain the maximum *a posteriori* (MAP) clustering (see also Jensen and Liu, 2008, Quintana and Iglesias, 2003). But we are not aware of any published work where attempt has been made to obtain appropriate credible regions of clusterings. Here we use a methodology introduced by MBD to tackle such difficulties, who rely on an appropriately defined metric to compute distances between any two clusterings of a given data set. The metric was used to compute the posterior probability distribution of clusterings, to provide a "central" clustering and the associated credible regions. In this paper, we apply their methodology to the clusterings associated with SB's Bayesian mixture model.

7.1. Definition of central clustering. Guided by the definition of mode in the case of parametric distributions, given a suitable metric d to compute the distance between any two clusterings, MBD define a clustering C^* as "central" if, for a given small $\epsilon > 0$,

$$P\left(\{C : d(C^*, C) < \epsilon\}\right) = \sup_{C'} P\left(\left\{C : d(C', C) < \epsilon\right\}\right) \tag{7.1}$$

Clearly, $\epsilon \to 0$ in (7.1) yields C^* as the mode of the distribution of clustering. Thus for a given, sufficiently small $\epsilon > 0$, the probability of an ϵ-neighbourhood of an arbitrary clustering C is highest when $C = C^*$, the central clustering. If the distribution of clustering is unimodal, then the central clustering remains unique for all $\epsilon > 0$. Otherwise, depending upon ϵ there will be different local modes of clustering, from among which the global mode is to be determined.

7.2. Empirical Definition of Central Clustering. We define that clustering $C^{(j)}$ as "approximately central" which, for a given small $\epsilon > 0$, satisfies the following equation

$$C^{(j)} = \arg \max_{1 \le i \le N} \frac{1}{N} \# \left\{ C^{(k)}; 1 \le k \le N : d(C^{(i)}, C^{(k)}) < \epsilon \right\}$$

The central clustering $C^{(j)}$ is easily computable, given $\epsilon > 0$ and a suitable metric d. Also, by the ergodic theorem, as $N \to \infty$ the empirical central clustering $C^{(j)}$ converges almost surely to the exact central clustering C^*. Given a central clustering $C^{(j)}$ one can then obtain, say, an approximate 95% credible region as the set $\left\{C^{(k)}; 1 \le k \le N : d(C^{(k)}, C^{(j)}) < \epsilon^*\right\}$, where

ϵ^* is such that

$$\frac{1}{N}\#\left\{C^{(k)}; 1 \leq k \leq N : d(C^{(k)}, C^{(j)}) < \epsilon^*\right\} \approx 0.95 \tag{7.2}$$

In (7.2) ϵ^* must be chosen by trial and error. An appropriate (say, 95%) highest posterior density credible region may be formed by considering the union of sets of clusterings which have the maximum possible probabilities, which add up to the desired level (say, 0.95).

7.3. Choice of the metric d. One way to compare two different clusterings is to find a measure of divergence between them after permuting the arbitrary indices to make the two clusterings as close to each other as possible. Ghosh, Dihidar and Samanta (2009) define the distance $d(I, II)$ between clusterings I and II as follows.

$$d(I, II) = \min[n_{00} - (n_{1j_1} + n_{2j_2} + \ldots + n_{kj_k})]/n_{00}, \tag{7.3}$$

where the minimization is over all permutations $(j_1, j_2, \ldots, j_k)$ of $(1, 2, \ldots, k)$. Here k denotes the number of clusters, n_{ij} is the number of units belonging to the i-th cluster of I and j-th cluster of II, and $n_{00} = \sum\sum n_{ij}$ is the total number of units. For justification of the above idea, and for the proof that (7.3) satisfies the properties of a metric, see Ghosh et al. (2009).

Since (7.3) requires minimization over all possible clusterings, for large number of clusters computation of (7.3) is burdensome in the extreme. To overcome this problem MBD propose an approximation to (7.3) as

$$\hat{d}(I, II) = \max\left\{\tilde{d}(I, II), \tilde{d}(II, I)\right\}$$

where

$$\begin{aligned}\tilde{d}(I, II) &= \left\{n_{00} - \sum_{i=1}^{k} \max_{1\leq j\leq k} n_{ij}\right\} \Big/ n_{00} \\ &= 1 - \frac{\sum_{i=1}^{k} \max_{1\leq j\leq k} n_{ij}}{n_{00}}\end{aligned}$$

The new quantity $\hat{d}(I, II)$ can be computed very cheaply. Quite importantly, MBD demonstrate that $\hat{d}$ provides very accurate approximations to the original metric d. It has been conjectured in MBD, for good reasons, that $\hat{d}$ is a metric. As a result, for our analysis we will always use $\hat{d}$ instead of d.

8 Simulation studies on clustering

It has been argued in Section 4.2 that a finite number of Gibbs sampling realizations from EW's model will generally not be able to adequately explore the clustering space. This problem will often manifest itself in not being able to locate the true clustering of the data, although it belongs to the clustering space. On the other hand, thanks to the much smaller clustering space, the true clustering can be easy to capture in SB's model. These we demonstrate with a simulation study, assuming that the data is of moderate size ($n = 50$ and $d = 2$).

We generate 100 data sets, each corresponding to $n = 50$ and $d = 2$, from the bi-variate mixture model $f(\boldsymbol{y}) = \sum_{j=1}^{5} \pi_j N(\boldsymbol{y} : \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. The mixing proportions are given by $\pi_1 = 0.12$, $\pi_2 = 0.32$, $\pi_3 = 0.05$, $\pi_4 = 0.25$, $\pi_5 = 0.26$, the mean vectors are $\boldsymbol{\mu}_1 = (0.2, 19.6)'$, $\boldsymbol{\mu}_2 = (0.52, 7.6)'$, $\boldsymbol{\mu}_3 = (1.2, 12.6)'$, $\boldsymbol{\mu}_4 = (0.7, 22.6)'$, $\boldsymbol{\mu}_5 = (0.4, 10.6)$, and the dispersion matrices are given by $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.43 & 0.12 \\ 0.12 & 0.25 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.23 & 0.22 \\ 0.22 & 0.15 \end{pmatrix}$, $\boldsymbol{\Sigma}_3 = \begin{pmatrix} 0.03 & 0.42 \\ 0.42 & 0.35 \end{pmatrix}$, $\boldsymbol{\Sigma}_4 = \begin{pmatrix} 0.13 & 0.02 \\ 0.02 & 0.05 \end{pmatrix}$, and $\boldsymbol{\Sigma}_5 = \begin{pmatrix} 0.33 & 0.72 \\ 0.72 & 0.45 \end{pmatrix}$. Thus, for each data set, there is a true clustering of the data, which is known in this simulation experiment.

We compute Gibbs sampling-based 95% credible regions of clusterings corresponding to the models of EW and SB, and note the number of times the true clusterings fall within their respective 95% credible regions. For specifying the prior distributions, we simulate a data set of size 50 from the mixture model $f(\boldsymbol{y})$, independently of the 100 data sets. Then we set $\boldsymbol{\mu}_0$ as the sample mean and $\boldsymbol{S}$ as the sample dispersion matrix corresponding to this independent data set. We set $M = 10$ for SB's model. For selecting an appropriate prior for α we first fit the new, independent data set using both SB and EW for different fixed values of α, and obtain the central clusterings. That value of α, which minimizes the distance between the central and the true clustering, is considered as the prior mode. In order to use the same prior distribution in both EW and SB, we take average of the prior modes of EW and SB. In this example, however, both the prior modes turned out to be 6. Using the method described in Section 6.1 we then set the prior of α as $\alpha \sim Gamma(1.6, 0.1)$ in both EW and SB. To each of the 100 data sets the two models are fitted using 20,000 iterations of the Gibbs sampler (the first 5,000 were discarded as burn-in). For SB's model the true clusterings fell within the respective 95% credible regions in all 100% cases. On the other hand, for EW's model, in only 11% cases the

Table 3: Table showing decrease of radius of 95% interval with increase of sample size.

Data Size	Radius of 95% Credible Region
50	0.780
100	0.740
500	0.730
1000	0.730
5000	0.725
10000	0.720
100000	0.680

true clusterings fell within the respective 95% credible regions, vindicating the concern expressed in Section 4.2. To explore this further, we conducted another simulation experiment with $n = 15, d = 2$. In this case, since the clustering space is much reduced for EW's model, in about 82% cases the true clusterings fell within the respective 95% credible regions. As before, however, for SB's model, in all 100% cases the true clusterings fell within the 95% credible regions. In all the above-mentioned experiments on clustering, we used $\epsilon = 0.1$ to determine the central clusterings; however, for $\epsilon > 0.1$ the results remained almost exactly the same as those with $\epsilon = 0.1$. For $\epsilon < 0.1$ too few clusterings fell in the neighborhoods of the clusterings obtained via Gibbs sampling, making the task of reliably obtaining the central clusterings very difficult.

8.1. Consistency of posterior distribution of clusterings. An interesting question that arises is whether or not consistency of the posterior distribution of clusterings is expected to be achieved at the true clustering. But the fact that even the true clustering of the data changes with the data size shows that the above question is perhaps not well-posed. Also, the clustering space increases with the data size. However, using a simulation experiment we attempted to gain some insight regarding this question, at least concerning SB's model. We remark here that since EW's model often fails to capture the true clusterings, even for moderately-sized data sets using MCMC samples, it is somewhat doubtful if practical investigation of consistency would prove to be useful. Table 3 shows the radii of the 95% credible regions of the posterior distributions of clusterings in SB's model for data sizes 50, 100, 500, 1000, 5000, 10,000 and 100,000. The radii are decreasing with increasing data

sizes, although the rate of decrease is very slow. The true clusterings fell within the 95% credible regions for each of the 7 (increasing) data sizes.

9 Application to cosmological data

We now illustrate our methodologies on a massive real data set obtained from Sloan Digital Sky Survey (SDSS) catalogue. The bivariate data consisting of 96,307 observations on logarithm of redshift (z) and apparent magnitude (m) for quasars (quasi-stellar objects) has been collected from SDSS-2007 catalogue. We are interested in studying the nature of the relationship between m and $\log(z)$ and to determine the nature of clustering of the bivariate data set. The data did not reveal any clear-cut parametric relationship between the two variables of interest. Exploratory analyses clearly ruled out the (bivariate) normality assumption for the data. Indeed, our quantile-quantile plots of each of the two variables showed that the marginal distributions of both the variables are far from univariate normal. The flexibility inherent in our approach set out in this paper and the associated computational simplicity and efficiency suggests application of our methods to this data set.

We proceed to fit our model to the SDSS data set by fixing $M = 30$. As in Section 6 we choose $\boldsymbol{\mu}_0$ and $\boldsymbol{S}$ as the sample mean and the sample dispersion matrix respectively; these choices ensured, as before, fast convergence of our Gibbs sampler. The other hyperparameters are chosen in the same way as in SB. The massive size of the data set ensures that the choices are quite robust. However, the choice of the prior distribution of α is important, and requires discussion.

9.1. Determination of the prior of α. In this real data situation, unfortunately, the true curve is unknown. Hence, we can not use exactly the same procedure as in the simulation study to determine the prior of α. Here the concept of mixture models offers an interesting method for determination of the prior of α, as detailed below. It is well-known that as the number of components in a finite mixture increases, closer is the approximation to the true curve. The price paid is the loss of parsimony of the model. We can forsake parsimony only for determining the prior of α, not for model-fitting. So, for our purpose, we first fit a finite mixture model to the cosmological data with a fixed (large) number of components. Since $M = 30$ was fixed as the maximum number of components in our Dirichlet process-based model, we use a mixture model with 30 fixed number of components. This is equivalent to letting $\alpha \to \infty$ in our model. The resulting Gibbs sampler is straightforwardly implemented.

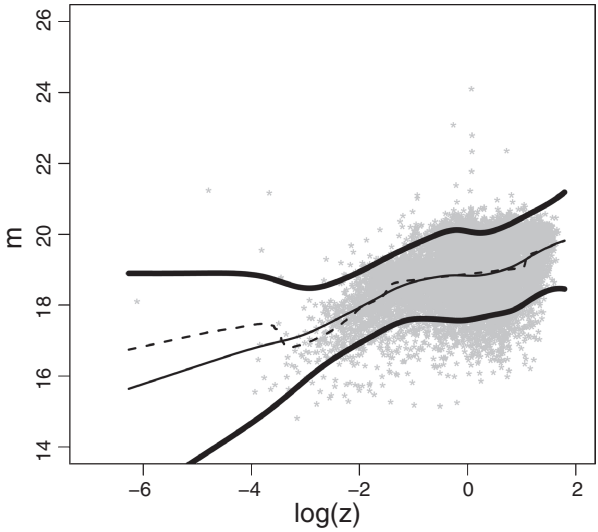

Figure 2: Real data analysis: the fitted curve is shown as the continuous line at the center, the broken line denotes the change point curve, and the pointwise 95% credible intervals are shown in thick lines. For plotting purpose we did a thinning of the original data set, plotting one in every 10 data points.

The curve thus obtained can be taken as a close approximation to the "true" curve. We further increased the value of M to 50 but noted no significant deviation of the resuting curve from that corresponding to $M = 30$. We then applied the prior determining procedure in the case of α, as described in Section 6, given the "approximately true" curve as obtained by the above method. In other words, successively fixing α and noting the maximum absolute deviations of the fitted curves from the "approximately true" curve, we chose that α for which the maximum absolute deviation fell below 0.4. This yielded 50 as the prior expectation of α in this real cosmological data case. Using ideas contained in Section 6 we obtain $Gamma(26, 0.5)$ as the prior of α.

9.2. Implementation time. Our Gibbs sampling algorithm completed 20,000 iterations in 4 hours 22 secs only in a work station with 3 GB RAM, and having two processors each with speed 3 GHz. Considering the enormity of the number of observations, this is an achievement in terms of computational speed. In contrast, the algorithm of MEW completed only 3 iterations in an 8 hours long run in the same machine, indicating that performing 20,000 iterations with MEW's algorithm is infeasible.

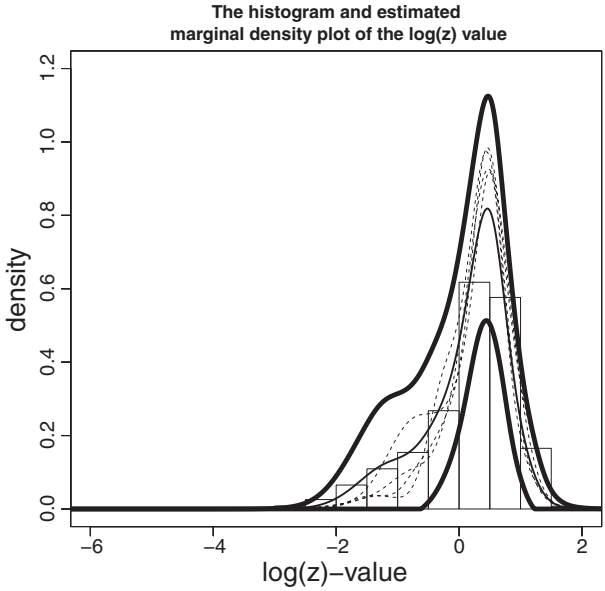

Figure 3: Marginal density estimation of $\log(z)$: the thick lines represent the 95% limits of the density, the continuous line at the center stands for the fitted density and the broken lines represent sample densities.

9.3. Fitted Bayesian cosmological curve and change point analysis. We discarded the first 5,000 MCMC realizations as burn-in and stored the remaining 15,000 for inference. Informal convergence diagnostics indicated excellent mixing properties of our algorithm. A more formal convergence diagnostic method suited for semi-parametric mixture models has been prescribed in MBD; the method confirms excellent convergence in this example. Details are provided in Section 5 of the supplementary document. The fitted curve and the associated pointwise 95% credible intervals are shown in Figure 2; the green line represents the estimated Bayesian cosmological curve, and the pointwise 95% credible intervals are shown in black colour. The difference in the nature of the lines in the same curve occurs due to variaton in the nature of red shift of the quasars of different ages. The number of different such quasars is reflected in the number of distinct components of the mixture model. The distinct components of the mixture correspond to distributions of absolute magnitude for different ages of the quasars.

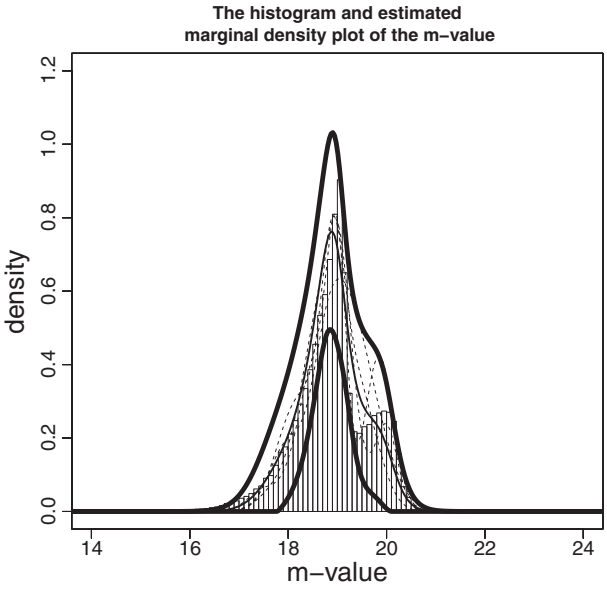

Figure 4: Marginal density estimation of m: the thick lines represent the 95% limits of the density, the continuous line represents the fitted density and broken lines stand for sample densities.

The obtained non-linear curve is linear for the first half ($z \leq -2.0$) with intercept 18.7840 and slope 0.5136 (1.182608 with respect to logarithm with base 10, which is of interest to astro-physicists). After that, however, non-linearity is exhibited. But we also note that the form of non-linearity can be approximated by line segments, indicating presence of change points. A close look will reveal presence of four change points, but to get a solid basis of belief, we performed a detailed change point analysis, assuming four change points. Although a Gibbs sampling algorithm is available on the similar lines of Carlin, Gelfand and Smith (1992), the algorithm is computationally expensive because of the massive number of observations. Instead, we resort to the Metropolis-Hastings algorithm for simulating from the posterior. We omit details to save space, but remark that we achieved excellent convergence with our Metropolis-Hastings algorithm. Figure 2 also shows that the curve obtained by the change point analysis, which is shown in red colour, nicely approximates our fitted semi-parametric Bayesian curve (the green curve) at all places except at the extreme lower end of the x-space, where there are hardly any information about the curve. Moreover, the entire change point

Table 4: Table showing the variation in the number clusters with change in the prefixed limit.

Value of prefixed limit	Number of clusters after merger
0.05	23
0.1	21
0.3	10
0.5	9
0.65	5
0.7	4
0.9	2

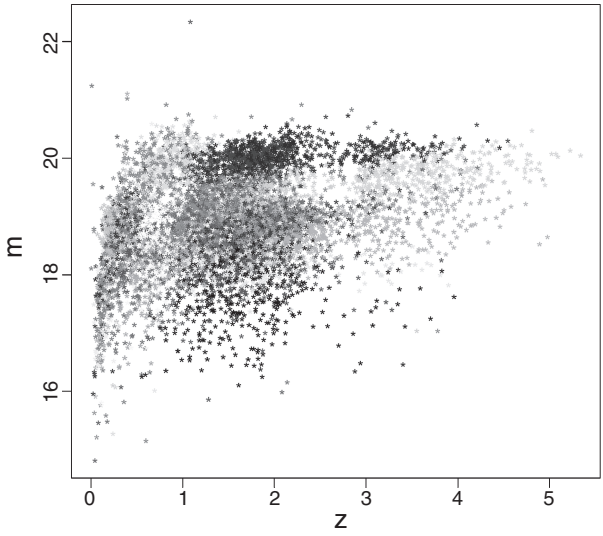

Figure 5: Central clustering with 29 clusters: different degrees of brightness in the gray scale indicate different clusters. For plotting purpose we did a thinning of the original data set, plotting one in every 10 data points.

curve falls within the (pointwise) 95% credible intervals associated with the semi-parametric Bayesian curve. This is not surprising, since the change point curve may be looked upon as a special case of our approach, with one of the linear functions having weight unity and all others having zero weight.

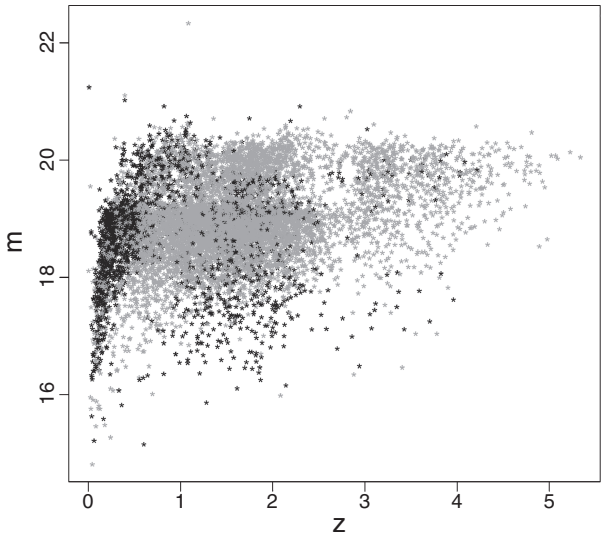

Figure 6: Merged central clustering with 2 clusters: two different degress of brightness in the gray scale indicate two different clusters. For plotting purpose we did a thinning of the original data set, plotting one in every 10 data points.

Apart from the semi-parametric Bayesian curve and the change point curve, we also fitted the least squares regression line, obtained by assuming a simple linear regression of $\log(z)$ on m. The least squares regression line falls well within the pointwise 95% credible intervals of our curve. This shows that the linear regression, although not optimal (in the sense that normality assumption does not hold for this data set, for example), is not ruled out by our semi-parametric method.

9.4. Estimation of the densities of the observed data and goodness of fit check. Note that the marginal densities of $y = m$ and $x = \log(z)$ can be estimated from our mixture model, given the MCMC-based posterior realizations $\left\{\mathbf{\Theta}_M^{(t)}; t = 1, \ldots, N\right\}$, for any $X = x$ and $Y = y$, as

$$\hat{f}_X(x) = \frac{1}{N}\sum_{t=1}^{N}\left[x \mid \mathbf{\Theta}_M^{(t)}\right]$$

$$= \frac{1}{M}\frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{M} N\left(x : \mu_{1j}^{(t)}, 1/\lambda_{1j}^{(t)}\right)$$

and

$$\hat{f}_Y(y) = \frac{1}{N}\sum_{t=1}^{N}\left[y \mid \mathbf{\Theta}_M^{(t)}\right]$$

$$= \frac{1}{M}\frac{1}{N}\sum_{t=1}^{N}\sum_{j=1}^{M} N\left(y : \mu_{2j}^{(t)}, 1/\lambda_{2j}^{(t)}\right)$$

Pointwise 95% credible intervals can be obtained for each of the marginal densities as in the case of Bayesian curve estimation. These Bayesian density estimates are useful for model validation purpose. In fact, these density estimates can be compared with the observed histograms of the individual variables of the observed data. A high degree of discrepancy between the observed histogram and the corresponding density estimate will indicate lack of model fit. Figures 3 and 4 show the observed marginal histograms, the marginal density estimates, and the associated 95% credible intervals of the true density. A few sample densities are also shown. The marginal density estimates fit the histogram very satisfactorily, leaving no reason to doubt the validity of our mixture model. In fact, the histograms (if smoothed by any means), the density estimates, and also the sample densities, all lie within their respective 95% credible intervals, which is very encouraging.

9.5. Application of the clustering ideas to the cosmology data set. On application of the central clustering ideas, we observe that for different range of values of $\epsilon > 0$ we have different central clusterings, indicating multimodality of the posterior distribution of clusterings. For $0 < \epsilon < 0.05$ the central clustering is the 351-st clustering after burn-in. For $0.05 < \epsilon < 0.1$ it is 5836-th; for $0.1 < \epsilon < 0.3$ it is 1077-th; for $0.3 < \epsilon < 0.5$ the number is 47-th clustering after the burn-in period. Since the global mode is approached by letting $\epsilon \to 0$, we identify the clustering corresponding to iteration number 351 as the global central clustering. The radius of the 95% credible region of the global mode is 0.35, which is reasonably low.

The central clustering in our case consists of 29 clusters. This is quite reasonable, given that there are more than 96,000 observations. Moreover, we note that although there are 29 clusters, many are effectively the same cluster, thanks to the small Euclidean distances between them. Driven by the above observations and discussions, we merge those clusters with Euclidean distances less than a prefixed limit. Table 4 shows how the number of clusters change if the prefixed limit is changed.

The original central clustering consisting of 29 clusters and the merged central clustering consisting of two components only (which corresponds to the prefixed limit being 0.9) are shown in Figures 5 and 6 respectively.

10 Conclusions and future work

In this article we have developed a semi-parametric curve-fitting method based on SB. We have demonstrated theoretically as well as with many simulation studies and application to a real, massive, cosmological data set, that this methodology is easily and efficiently implementable, even when the data set is massively large. In such cases the well-known methods based on EW become infeasible. The RJMCMC method, although implementable in massive data situations, can be inefficient, particularly when the dimensionality of the data is high.

Even for small to moderate data sets, we have demonstrated that our methods based on SB can outperform the other methods. In particular, our semi-parametric regression curve can more adequately approximate the underlying true curve by better utilizing the available prior information about the number of distinct components.

The issues related to clusterings are also interesting. Our clustering method, based on MBD, shows that even if the data size is moderate, the true clustering may not be learned well using approaches based on EW, while our model and methodology based on SB is well-suited for the same purpose.

We are currently investigating the usefuless of our methodologies in handling the "large p small n" problem, and have obtained encouraging preliminary results. In such problems, since the data dimensionality is extremely large, RJMCMC will be inefficient in the extreme. Also, since the data size is very small, EW's approach will be inadequate. We anticipate that the methodologies we proposed in this paper will be far more efficient than the methods associated with EW, RG, and the other existing methodologies specialized to deal with this problem. Our findings will be communicated elsewhere.

Acknowledgement. We are extremely grateful to an Associate Editor and a referee whose comments have led to improved presentation of the paper. Our sincere gratitude also goes to Prof. Jayanta K. Ghosh, Prof. Sarat Dass, Prof. Arni S. Rao and Mr. Arunabha Majumdar for providing useful feedback on this work.

References

ABRAMOWITZ, M. and STEGUN, I.A. (1972). Stirling Numbers of the Second Kind. In *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (M. Abramowitz and I.A. Stegun, eds.). Dover, New York, pp 824–825.

ANTONIAK, C.E. (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *Ann. Statist.*, **2**, 1152–1174.

BELL, E.T. (1934). Exponential numbers. *Amer. Math. Monthly*, **41**, 411–419.

BHATTACHARYA, S. (2008). Gibbs sampling based Bayesian analysis of mixtures with unknown number of components. *Sankhya B*, **70**, 133–155.

BUSH, C.A. and MACEACHERN, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika*, **83**, 275–285.

CARLIN, B.P., GELFAND, A.E. and SMITH, A.F.M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Appl. Stat.*, **41**, 389–405.

DAHL, D. (2009). Modal clustering in a class of product partition models. *Bayesian Anal.*, **4**, 243–264.

DALAL, S.R. and HALL, W.J. (1983). Approximating priors by mixtures of natural conjugate priors. *J. R. Stat. Soc. Ser. B.*, **45**, 278–286.

DIACONIS, P. and YLVISAKER, D. (1985). Quantifying Prior Opinion (with discussion). In *Bayesian Statistics 2* (J.-M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.). North-Holland, Amsterdam, pp. 133–156.

ESCOBAR, M.D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, **90**, 577–588.

FERGUSON, T.S. (1974). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.

GHOSH, J.K., DIHIDAR, K. and SAMANTA, T. (2009). On Different Clusterings of the Same Data Set. In *Felicitation Volume in Honour of Prof. B. K. Kale* (B. Arnold, U. Gather and S.M. Bendre, eds.). MacMillan, New Delhi.

JAIN, S. and NEAL, R.M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J. Comput. Graph. Statist.*, **13**, 158–182.

JAIN, S. and NEAL, R.M. (2007). Splitting and merging components of a nonconjugate Dirichlet process mixture model (with discussion). *Bayesian Anal.*, **2**, 445–472.

JENSEN, S.T. and LIU, J.S. (2008). Bayesian clustering of transcription factor binding Motiffs. *J. Amer. Statist. Assoc.*, **103**, 188–200.

LEE, K., MARIN, J.-M., MENGERSEN, K. and ROBERT, C.P. (2008). Bayesian inference on mixtures of distributions.

MACEACHERN, S.N. (1994). Estimating normal means with a conjugate-style Dirichlet process prior. *Comm. Statist. Simulation Comput.*, **23**, 727–741.

MCCULLAGH, P. and YANG, J. (2008). How many clusters? *Bayesian Anal.*, **3**, 101–120.

MCLACHLAN, G.J. and BASFORD, K.E. (1988). *Mixture models: inference and applications to clustering*. Dekker, New York.

MUKHOPADHYAY, S. and BHATTACHARYA, S. (2012). Perfect simulation for mixtures with known and unknown number of components. *Bayesian Anal.*, **7**, 675–714.

MUKHOPADHYAY, S., BHATTACHARYA, S. and DIHIDAR, K. (2011). On Bayesian central clustering: application to landscape classification of Western Ghats. *Ann. Appl. Stat.*, **5**, 1948–1977.

MÜLLER, P., ERKANLI, A. and WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**, 67–79.

QUINTANA, F.A. and IGLESIAS, P.L. (2003). Bayesian clustering and product partition models. *J. R. Stat. Soc. Ser. B.*, **65**, 557–574.

RICHARDSON, S. and GREEN, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Stat. Soc. Ser. B.*, **59**, 731–792.

TITTERINGTON, D.M., SMITH, A.F.M. and MAKOV, U.E. (1985). *Statistical analysis of finite mixture distributions*. John Wiley & Sons, New York.

WANG, L. and DUNSON, D.B. (2011). Fast Bayesian inference in Dirichlet process mixture models. *J. Comput. Graph. Statist.*, **20**, 196–216.

Sabyasachi Mukhopadhyay
Bayesian and Interdisciplinary
Research Unit
Indian Statistical Institute
203, B. T. Road
Kolkata 700 108, India

Sisir Roy
Physics and Applied
Mathematics Unit
Indian Statistical Institute
203, B. T. Road
Kolkata 700 108, India

Sourabh Bhattacharya
Bayesian and Interdisciplinary
Research Unit
Indian Statistical Institute
203, B. T. Road
Kolkata 700 108, India
E-mail: bhsourabh@gmail.com

Paper received: 23 July 2010; revised: 7 February 2011.