Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Fast Bayesian Inference in Dirichlet Process Mixture Models

Lianming WANG and David B. DUNSON

There has been increasing interest in applying Bayesian nonparametric methods in large samples and high dimensions. As Markov chain Monte Carlo (MCMC) algorithms are often infeasible, there is a pressing need for much faster algorithms. This article proposes a fast approach for inference in Dirichlet process mixture (DPM) models. Viewing the partitioning of subjects into clusters as a model selection problem, we propose a sequential greedy search algorithm for selecting the partition. Then, when conjugate priors are chosen, the resulting posterior conditionally on the selected partition is available in closed form. This approach allows testing of parametric models versus nonparametric alternatives based on Bayes factors. We evaluate the approach using simulation studies and compare it with four other fast nonparametric methods in the literature. We apply the proposed approach to three datasets including one from a large epidemiologic study. Matlab codes for the simulation and data analyses using the proposed approach are available online in the supplemental materials.

Key Words: Clustering; Density estimation; Efficient computation; Large samples; Nonparametric Bayes; Pólya urn scheme; Sequential analysis.

1. INTRODUCTION

In recent years, there has been an explosion of interest in Bayesian nonparametric methods due to their flexibility and to the availability of efficient and easy to use algorithms for posterior computation. Most of the focus has been on Dirichlet process mixture (DPM) models (Lo 1984; Escobar 1994; Escobar and West 1995), which place a Dirichlet process (DP) prior (Ferguson 1973, 1974) on parameters in a hierarchical model. For DPMs, there is a rich literature on Markov chain Monte Carlo (MCMC) algorithms for posterior computation, proposing marginal Gibbs sampling (MacEachern 1994; West, Müller, and Escobar 1994; Bush and MacEachern 1996), conditional Gibbs sampling (Ishwaran and James

Lianming Wang is Assistant Professor, Department of Statistics, University of South Carolina, Columbia, SC 29208 (E-mail: *wang99@mailbox.sc.edu*). David B. Dunson is Professor, Department of Statistical Science, Duke University, Durham, NC 27708.

Journal of Computational and Graphical Statistics, Volume 20, Number 1, Pages 196–216
DOI: 10.1198/jcgs.2010.07081

2001), and split-merge (Jain and Neal 2004) algorithms. These approaches are very useful in small to moderate sized datasets when one can devote several hours (or days) for computation.

However, there is clearly a pressing need for dramatically faster alternatives to MCMC, which can be executed within seconds (or at most minutes) even for very large datasets. Such algorithms are absolutely required in large-scale data analyses, in which computational speed is paramount. In the pregnancy outcome application considered in Section 6, data were available for 34,178 pregnancies and it was infeasible to implement MCMC. Even in smaller applications, it is very desirable to obtain results quickly. Speed also has the advantage of allowing detailed simulation studies of operating characteristics and sensitivity analyses for different prior specifications. In addition to obtaining results quickly for one DPM, it is typically of interest to compare DPMs to simpler parametric models. Typical MCMC algorithms do not allow such comparisons, as marginal likelihoods are not estimated, though there has been some recent work to address this gap (Basu and Chib 2003).

The focus of this article is on extremely fast alternatives to MCMC, which allow accurate approximate Bayes inferences under one DPM, while also producing marginal likelihood estimates to be used in model comparison. For example, one may be interested in comparing a DPM to a simpler parametric model. For simplicity in exposition, we focus throughout the article on Gaussian DPMs, though the methods can be trivially modified to other cases in which a conjugate prior is chosen.

For DPM models, a number of alternatives to MCMC have been proposed, including predictive recursion (PR) (Newton and Zhang 1999; Newton 2002; Ghosh and Tokdar 2006; Tokdar, Martin, and Ghosh 2009), weighted Chinese restaurant (WCR) sampling (Lo, Brunner, and Chan 1996; Ishwaran and Takahara 2002; Ishwaran and James 2003), sequential importance sampling (SIS) (MacEachern, Clyde, and Liu 1999; Quintana and Newton 2000), and variational Bayes (VB) (Blei and Jordan 2006; Kurihara, Welling, and Vlassis 2006; Kurihara, Welling, and Teh 2007). The WCR and SIS approaches are computationally intensive because they are based on a large number of particles. For a Gaussian mixture model with unknown mean and variance, the recursive algorithm (Newton 2002; Ghosh and Tokdar 2006; Tokdar, Martin, and Ghosh 2009) needs to estimate a bivariate mixing density and involves approximating a normalizing constant in each sequential updating step. VB relies on maximization of a lower bound on the marginal likelihood using a factorization approximation to the posterior. Wang and Titterington (2005) showed a tendency of VB to underestimate uncertainty in mixture models. Also, VB is sensitive to the starting values, motiving the use of a short SIS run to choose initial values.

We propose an alternative *sequential updating and greedy search* (SUGS) algorithm. This algorithm relies on factorizing the DP prior as a product of a prior on the partition of subjects into clusters and independent priors on the parameters within each cluster. Adding subjects one at a time, we allocate subjects to the cluster that maximizes the conditional posterior probability given their data and the allocation of previous subjects, while also updating the posterior distribution of the cluster-specific parameters. Hence, viewing selection of the partition as a model selection problem, we implement a sequential greedy

search for a good partition, with the exact posterior given this partition then available in closed form. The algorithm is very fast involving only a single cycle of simple calculations for each subject. In addition, a marginal likelihood is produced that can be used for model selection and for eliminating sensitivity to the order in which subjects are added through model averaging or selection over random orders. Existing methods related to SUGS include those of Daumé III (2007), Fearnhead (2004), Minka and Ghahramani (2003), and Zhang, Ghahramani, and Yang (2005).

Section 2 describes the prior structure. Section 3 proposes the fast SUGS posterior updating algorithm, with Section 4 providing details for normal DPMs. Section 5 evaluates the approach and compares it with four other fast nonparametric methods through simulation studies. Section 6 contains three real data applications and Section 7 concludes with some remarks.

2. DIRICHLET PROCESS MIXTURES AND PARTITION MODELS

DPM models have a well-known relationship to partition models (Quintana and Iglesias 2003; Park and Dunson 2009). For example, consider a DP mixture of normals (Lo 1984):

$$y_i \sim N(\widetilde{\mu}_i, \widetilde{\tau}_i^{-1}), \qquad (\widetilde{\mu}_i, \widetilde{\tau}_i) \overset{\text{iid}}{\sim} P, \qquad i = 1, \ldots, n, \qquad P \sim \text{DP}(\alpha P_0), \tag{2.1}$$

where $\widetilde{\boldsymbol{\theta}}_i = (\widetilde{\mu}_i, \widetilde{\tau}_i)$ are parameters specific to subject i, α is the DP precision parameter, and P_0 is a base probability measure. Then, upon marginalizing out the random mixing measure P, one obtains the DP prediction rule (Blackwell and MacQueen 1973):

$$(\widetilde{\boldsymbol{\theta}}_i | \widetilde{\boldsymbol{\theta}}_1, \ldots, \widetilde{\boldsymbol{\theta}}_{i-1}) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0 + \left(\frac{1}{\alpha + i - 1}\right) \sum_{j=1}^{i-1} \delta_{\widetilde{\boldsymbol{\theta}}_j}, \qquad i = 1, \ldots, n, \tag{2.2}$$

where δ_θ is a probability measure concentrated at θ. Sequential application of the DP prediction rule for subjects $1, \ldots, n$ creates a random partition of the integers $\{1, \ldots, n\}$. Commonly used algorithms for posterior computation in DPM models rely on marginalizing out P to obtain a random partition, so that one bypasses computation for the infinitely-many parameters characterizing P (Bush and MacEachern 1996).

Taking advantage of a characterization of Lo (1984), one can express the posterior distribution in DPMs after marginalizing out P as a product of the posterior for the partition multiplied by independent posteriors for each cluster, obtained by updating the prior P_0 with the data for the subjects allocated to that cluster. Instead of obtaining this structure indirectly through marginalization of P, one could directly specify a model for the random partition, while assuming conditional independence given the allocation to clusters. This possibility was suggested by Quintana and Iglesias (2003), who focused on product partition models (PPMs) (Barry and Hartigan 1992).

We assume that there is an infinite sequence of clusters, with $\boldsymbol{\theta}_h$ representing the parameters specific to cluster h, for $h = 1, \ldots, \infty$. We use the DP prediction rule in (2.2) for sequentially allocating subjects to a sparse subset of these clusters. The first subject

will be automatically allocated to cluster $h = 1$, with additional clusters occupied as subjects are added as needed to improve predictive performance, obtaining an online updating approach. Sensitivity to ordering will be discussed later in the article.

Let γ_i be a cluster index for subject i, with $\gamma_i = h$ denoting that subject i is allocated to cluster h. Relying on the DP prediction rule, the conditional prior distribution of γ_i given $\boldsymbol{\gamma}^{(i-1)} = (\gamma_1, \ldots, \gamma_{i-1})$ is assumed to be multinomial with

$$\Pr\bigl(\gamma_i = h|\boldsymbol{\gamma}^{(i-1)}\bigr) = \begin{cases} \frac{\sum_{j=1}^{i-1} 1_{(\gamma_j=h)}}{\alpha+i-1}, & h = 1, \ldots, k_{i-1} \\ \frac{\alpha}{\alpha+i-1}, & h = k_{i-1}+1, \end{cases} \tag{2.3}$$

where $\alpha > 0$ is a DP precision parameter controlling sparseness and $k_{i-1} = \max\{\gamma_h\}_{h=1}^{i-1}$, the number of clusters after $i-1$ subjects have been sequentially added. As α increases, there is an increasing tendency to allocate subjects to new clusters instead of clusters occupied by previous subjects. The prior probabilities in (2.3) favor allocation of subject i to clusters having large numbers of subjects.

To complete a Bayesian specification, it is necessary to choose priors for the parameters within each of the clusters:

$$\pi(\boldsymbol{\theta}) = \prod_{h=1}^{\infty} p_0(\boldsymbol{\theta}_h), \tag{2.4}$$

where p_0 is the prior distribution on the cluster-specific coefficients $\boldsymbol{\theta}_h$ and independence across the clusters is implied by the result of Lo (1984).

3. SEQUENTIAL UPDATING AND GREEDY SEARCH

3.1 PROPOSED ALGORITHM

Suppose that a measurement y_i is obtained for subjects $i = 1, \ldots, n$. Updating (2.3) one can obtain the conditional posterior probability of allocating subject i to cluster h given the data for subjects $1, \ldots, i$ $[\mathbf{y}^{(i)} = (y_1, \ldots, y_i)']$ and the cluster assignment for subjects $1, \ldots, i-1$ $[\boldsymbol{\gamma}^{(i-1)} = (\gamma_1, \ldots, \gamma_{i-1})']$:

$$\Pr\bigl(\gamma_i = h|\mathbf{y}^{(i)}, \boldsymbol{\gamma}^{(i-1)}\bigr) = \frac{\pi_{ih} L_{ih}(y_i)}{\sum_{l=1}^{k_{i-1}+1} \pi_{il} L_{il}(y_i)}, \qquad h = 1, \ldots, k_{i-1}+1, \tag{3.1}$$

where $\pi_{ih} = \Pr(\gamma_i = h|\boldsymbol{\gamma}^{(i-1)})$ is the conditional prior probability in expression (2.3), and $L_{ih}(y_i) = \int f(y_i|\theta_h)\pi(\theta_h|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})\, d\theta_h$ is the conditional likelihood of y_i given allocation to cluster h and the cluster allocation for subjects $1, \ldots, i-1$, with $f(y_i|\theta_h)$ denoting the likelihood of y_i given parameters θ_h and $\pi(\theta_h|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}) \propto p_0(\theta_h) \prod_{\{j:\gamma_j=h, 1\le j\le i-1\}} f(y_j|\theta_h)$, the posterior distribution of θ_h given $\mathbf{y}^{(i-1)}$ and $\boldsymbol{\gamma}^{(i-1)}$. For a new cluster $h = k_{i-1}+1$, $\pi(\theta_h|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)}) = p_0(\theta_h)$, as none of the first $i-1$ subjects have been allocated to cluster $k_{i-1}+1$.

For conjugate p_0, the posterior $\pi(\theta_h|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})$ and likelihood $L_{ih}(y_i)$ are available in closed form. Hence, the joint posterior distribution for the cluster-specific coefficients

$\boldsymbol{\theta} = \{\theta_h\}_{h=1}^{\infty}$ given the data and cluster allocation for all n subjects,

$$\pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma}) = \prod_{h=1}^{\infty} \pi(\theta_h|\mathbf{y}, \boldsymbol{\gamma}) = \left\{\prod_{h=1}^{k_n} \pi(\theta_h|\{y_i : \gamma_i = h\})\right\}\left\{\prod_{h=k_n+1}^{\infty} p_0(\theta_h)\right\},$$

is similarly available in closed form. Note that the first k_n clusters are *occupied* in that they have at least one member from the sample.

The real challenge is addressing uncertainty in the partition of subjects to clusters, $\boldsymbol{\gamma}$. MCMC algorithms attempt to address this uncertainty by generating samples from the joint posterior distribution of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. As highlighted in Section 1, such MCMC algorithms are quite expensive computationally. This is particularly true if sufficient numbers of samples are collected to adequately explore the posterior distribution of $\boldsymbol{\gamma}$. The multimodal nature of the posterior and the tendency to remain for long intervals in local modes make this exploration quite challenging. In addition, $\boldsymbol{\gamma} \in \Gamma$ can be viewed as a model index belonging to the high-dimensional space Γ. As for other high-dimensional stochastic search procedures, for sufficiently large n, it is for all practical purposes infeasible to fully explore Γ or to draw enough samples to accurately represent the posterior of $\boldsymbol{\gamma}$.

An additional issue is that, even if one could obtain iid draws from $\boldsymbol{\gamma}$, problems in interpretation often arise due to the *label switching* issue. Viewing $\boldsymbol{\gamma}$ as a model index, samples from the joint posterior of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ can be used to obtain model-averaged predictions and inferences, allowing for uncertainty in selection of $\boldsymbol{\gamma}$. Although it is well known that model averaging is most useful for prediction, the ability to obtain interpretable inferences may be lost in averaging across models. This is certainly true in mixture models, because the meaning of the cluster labels changes across the samples, making it difficult to summarize cluster-specific results. There has been some work on postprocessing algorithms to align the clusters (Stephens 2000), though this can add considerably to the computational burden.

Motivated by these issues, there has been some recent work on obtaining an optimal estimate of $\boldsymbol{\gamma}$ (Lau and Green 2007; Dahl 2009). These approaches are quite expensive computationally, so will not be considered further. We instead propose a very fast sequential updating and greedy search (SUGS) algorithm, which cycles through subjects, $i = 1, \ldots, n$, sequentially allocating them to the cluster that maximizes the conditional posterior allocation probability. This proceeds as follows:

1. Let $\gamma_1 = 1$ and calculate $\pi(\theta_1|y_1, \gamma_1)$.
2. For $i = 2, \ldots, n$,
   (a) Choose γ_i to maximize the conditional probability of $\gamma_i = h$ given $\mathbf{y}^{(i)}$ and $\boldsymbol{\gamma}^{(i-1)}$ using (3.1).
   (b) Update $\pi(\theta_{\gamma_i}|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})$ using the data for subject i.

This algorithm only requires a single cycle of simple deterministic calculations for each subject under study, and can be implemented within a few seconds even for very large datasets. In addition, the algorithm is online so that additional subjects can be added as they become available without additional computations for the past subjects. Hence, the

algorithm is particularly suited for large-scale real-time prediction. The proposed method is similar to the hard decision method by Zhang, Ghahramani, and Yang (2005) in the field of online document clustering and the "trivial" algorithm by Daumé III (2007). However, we also propose methods to remove order dependence in sequential updating, allow unknown DP precision parameter α, use empirical Bayes for estimation of key hyperparameters, and conduct model comparison associated with SUGS.

3.2 REMOVING ORDER DEPENDENCE

The SUGS approach for selecting $\boldsymbol{\gamma} \in \Gamma$ is sequentially optimal, but will not in general produce a global maximum a posteriori (MAP) estimate of $\boldsymbol{\gamma}$. Producing the global MAP is in general quite challenging computationally given the multimodality and size of Γ. In addition, as noted by Stephens (2000), there are in general very many choices of $\boldsymbol{\gamma}$ having identical or close to identical marginal likelihoods. Hence, SUGS seems to provide a reasonable strategy for rapidly identifying a good partition without spending an enormous amount of additional time searching for alternative partitions that may provide only minimal improvement.

One aspect that is unappealing is dependence of the selection of $\boldsymbol{\gamma}$ on the order in which subjects are added. As this order is typically arbitrary, one would prefer to eliminate this order dependence. To address this issue, we recommend repeating the SUGS algorithm of Section 3.1 for multiple permutations of the ordering $\{1, \ldots, n\}$. The marginal likelihood given an ordering $\boldsymbol{\gamma}$ is calculated as

$$L\big(\mathbf{y}^{(n)}|\boldsymbol{\gamma}\big) = \prod_{h=1}^{k_n} \int \Big\{ \prod_{i:\gamma_i=h} f(y_i|\theta_h) \Big\} p_0(\theta_h)\, d\theta_h. \tag{3.2}$$

Selecting an ordering with the largest marginal likelihood works fine in eliminating the ordering effect in general. However, this marginal likelihood criterion is not perfectly reliable and sometimes leads to poor predictive density estimation. This is because the ordering with the largest marginal likelihood occasionally overfits the data in assigning subjects to more clusters than is necessary. As an alternative, we propose to use pseudo-marginal likelihood (PML) and base inferences on the ordering having the largest PML.

The pseudo-marginal likelihood is defined as the product of conditional predictive ordinates (Geisser 1980; Pettiti 1990; Gelfand and Dey 1994) as follows:

$$\begin{aligned} \mathrm{PML}_{\boldsymbol{\gamma}}(\mathbf{y}) &= \prod_{i=1}^{n} \pi\big(y_i|\mathbf{y}^{(-i)}, \boldsymbol{\gamma}^{(-i)}\big) = \prod_{i=1}^{n} \int \pi(y_i|\boldsymbol{\theta}) \pi\big(\boldsymbol{\theta}|\mathbf{y}^{(-i)}, \boldsymbol{\gamma}^{(-i)}\big)\, d\boldsymbol{\theta} \\ &= \prod_{i=1}^{n} \sum_{h=1}^{k_n+1} \Pr\big(\gamma_i = h|\mathbf{y}^{(-i)}, \boldsymbol{\gamma}^{(-i)}\big) \int f(y_i|\theta_h) \pi\big(\theta_h|\mathbf{y}^{(-i)}, \boldsymbol{\gamma}^{(-i)}\big)\, d\theta_h, \end{aligned} \tag{3.3}$$

where $\mathbf{y}^{(-i)}$ is the set of all the data but y_i for $i = 1, \ldots, n$. The $\mathrm{PML}_{\boldsymbol{\gamma}}(\mathbf{y})$ criterion is appealing in favoring a partition resulting in good predictive performance and has been used for assessing goodness of fit and Bayesian model selection by Geisser and Eddy (1979), Gelfand and Dey (1994), Sinha, Chen, and Ghosh (1999), and Mukhopadhyay,

Ghosh, and Berger (2005), among others. To speed up computation, we use $\pi(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\gamma})$ instead of $\pi(\boldsymbol{\theta}|\mathbf{y}^{(-i)}, \boldsymbol{\gamma}^{(-i)})$ in practice, approximating PML with a product of predictive densities defined in (3.8) over all subjects. This approximation is accurate, particularly for large samples. We use PML criteria in all the implementation of SUGS in the simulation studies and real data analyses unless mentioned otherwise. Because SUGS is extremely fast, repeating it for a modest number of random orderings and selecting a good ordering does not take much time.

A variety of strategies have been suggested to limit order dependence in other nonparametric sequential algorithms. The recursive algorithm (Newton 2002; Tokdar, Martin, and Ghosh 2009) and the expectation propagation method (Minka and Ghahramani 2003) proposed to take an unweighted average over a number of permutations. Daumé III (2007) proposed to present the data in ascending order of individual marginal likelihood, $\int f(y_i|\theta)\, p_0(\theta)\, d\theta$, prior to the online updating.

3.3 Allowing the DP Precision Parameter α to be Unknown

In the above development, we have assumed that the DP precision parameter α is fixed, which is not recommended because the value of α plays a strong role in the allocation of subjects to clusters. To allow unknown α, we choose the prior:

$$\pi(\alpha) = \sum_{t=1}^{T} \eta_t \delta_{\alpha_t^*}(\alpha), \tag{3.4}$$

with $\boldsymbol{\alpha}^* = (\alpha_1^*, \ldots, \alpha_T^*)'$ a prespecified grid of possible values with a large range and $\eta_t = \Pr(\alpha = \alpha_t^*)$.

We can easily modify the SUGS algorithm to allow simultaneous updating of α. Letting $\phi_t^{(i-1)} = \Pr(\alpha = \alpha_t^*|\mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})$ and $\pi_{iht} = \Pr(\gamma_i = h|\alpha = \alpha_t^*, \mathbf{y}^{(i-1)}, \boldsymbol{\gamma}^{(i-1)})$, we obtain the following modification to (3.1):

$$\Pr\big(\gamma_i = h|\mathbf{y}^{(i)}, \boldsymbol{\gamma}^{(i-1)}\big) = \frac{\sum_{t=1}^{T} \phi_t^{(i-1)} \pi_{iht} L_{ih}(y_i)}{\sum_{t=1}^{T} \phi_t^{(i-1)} \sum_{l=1}^{k_{i-1}+1} \pi_{ilt} L_{il}(y_i)},$$
$$h = 1, \ldots, k_{i-1} + 1, \tag{3.5}$$

which is obtained marginalizing over the posterior for α given the data and allocation for subjects $1, \ldots, i-1$. Then we obtain the following updated probabilities:

$$\phi_t^{(i)} = \Pr\big(\alpha = \alpha_t^*|\mathbf{y}^{(i)}, \boldsymbol{\gamma}^{(i)}\big) = \frac{\phi_t^{(i-1)} \pi_{i\gamma_i t}}{\sum_{s=1}^{T} \phi_s^{(i-1)} \pi_{i\gamma_i s}}, \qquad t = 1, \ldots, T. \tag{3.6}$$

Note that we obtain a closed-form joint posterior distribution for the cluster-specific parameters $\boldsymbol{\theta}$ and DP precision α given $\boldsymbol{\gamma}$.

In our proposed approach, we handle the DP precision parameter α from a fully Bayesian perspective and we can obtain the posterior distribution. This is more appealing than most of the fast DP mixture algorithms, which use fixed value of α, for example, the particle filter by Fearnhead (2004), the fast DPM model by Daumé III (2007), and the accelerated and collapsed variational DP mixture models by Kurihara, Welling, and

Vlassis (2006) and Kurihara, Welling, and Teh (2007), respectively. The online document clustering method by Zhang, Ghahramani, and Yang (2005) employed an empirical Bayes method to estimate α, whereas Blei and Jordan (2006) adopted a gamma prior for α in their variational DP mixture approach. We found the Blei and Jordan (2006) approach to have better performance than the newer VB variants in simulations (results not shown).

3.4 ESTIMATING PREDICTIVE DISTRIBUTIONS

From applying SUGS, we obtain a selected partition $\boldsymbol{\gamma}$ and posterior distributions in closed form for the parameters within each cluster, $\pi(\boldsymbol{\theta}_h|\mathbf{y}, \boldsymbol{\gamma})$, and for DP precision parameter α as in (3.6). From these posterior distributions, we can conduct inferences on the cluster-specific coefficients, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_n}$.

In addition, we can conduct fast online predictions for new subjects. The predicted probability of allocation of subject $i = n + 1$ to cluster h is

$$\pi_{n+1,h} = \begin{cases} \sum_{t=1}^{T} \phi_t^{(n)} \frac{\sum_{i=1}^{n} 1(\gamma_i = h)}{\alpha_t^* + n}, & h = 1, \ldots, k_n \\ \sum_{t=1}^{T} \phi_t^{(n)} \frac{\alpha_t^*}{\alpha_t^* + n}, & h = k_n + 1. \end{cases} \tag{3.7}$$

The predictive density is then

$$\begin{aligned} \widehat{f}(y_{n+1}) &= \sum_{h=1}^{k_n+1} \pi_{n+1,h} \int f(y_{n+1}|\gamma_{n+1} = h, \boldsymbol{\theta}_h)\, d\pi\left(\boldsymbol{\theta}_h|\mathbf{y}^{(n)}, \boldsymbol{\gamma}^{(n)}\right) \\ &= \sum_{h=1}^{k_n+1} \pi_{n+1,h} f\left(y_{n+1}|\gamma_{n+1} = h, \mathbf{y}^{(n)}, \boldsymbol{\gamma}^{(n)}\right), \end{aligned} \tag{3.8}$$

which is available in closed form.

To obtain pointwise credible intervals for the conditional density estimate, $\widehat{f}(y_{n+1})$, apply the following Monte Carlo procedure:

1. Draw S samples $\{\boldsymbol{\theta}_1^{(s)}, \ldots, \boldsymbol{\theta}_{k_n}^{(s)}, \alpha^{(s)}\}_{s=1}^{S}$ from the joint posterior distribution of

$$\left(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k_n}, \alpha|\mathbf{y}^{(n)}, \boldsymbol{\gamma}^{(n)}\right).$$

2. Calculate the conditional density for each of these draws:

$$f^{(s)}\left(y_{n+1}|\boldsymbol{\theta}_1^{(s)}, \ldots, \boldsymbol{\theta}_{k_n}^{(s)}, \alpha^{(s)}\right) = \sum_{h=1}^{k_n} \pi_{n+1,h}^{(s)} f\left(y_{n+1}|\gamma_{n+1} = h, \boldsymbol{\theta}_h = \boldsymbol{\theta}_h^{(s)}\right),$$

   where $\pi_{n+1,h}^{(s)}$ is calculated using formula (3.1) with $\alpha = \alpha^{(s)}$ and $i = n + 1$.

3. Calculate empirical percentiles of $\{f^{(s)}(y_{n+1})\}_{s=1}^{S}$.

Because the proposed SUGS algorithm is deterministic for a selected ordering, the resulting credible intervals tend to underestimate the uncertainty. This is observed in our simulation study and occurs in many other competing approaches, such as VB.

3.5 MODEL COMPARISON

One very appealing aspect of the SUGS approach is that we obtain a closed-form expression for the exact marginal likelihood for the selected model $\boldsymbol{\gamma}$, because each integral term in (3.2) has a simple closed form due to the conjugacy. Hence, we can obtain Bayes factors and posterior probabilities for competing models. For example, the Bayes factor for comparing the selected semiparametric model to the parametric base model is

$$BF = \frac{L(\mathbf{y}^{(n)}|\boldsymbol{\gamma})}{L_1(\mathbf{y}^{(n)})},$$

where the denominator is the marginal likelihood obtained in allocating all subjects to the first cluster. The performance of tests based on these Bayes factors is assessed through simulations in Section 5.

4. DP MIXTURES OF NORMALS AND SUGS DETAILS

4.1 SUGS DETAILS

We focus on normal mixture models as an important special case, letting $\boldsymbol{\theta}_h = (\mu_h, \tau_h)'$ represent the mean parameter μ_h and residual precision τ_h for cluster h, $h = 1, \ldots, \infty$. To specify p_0, we choose conjugate normal inverse-gamma priors as follows:

$$\pi(\mu_h, \tau_h) = N_p(\mu_h; m, \psi\tau_h^{-1})G(\tau_h; a, b), \tag{4.1}$$

with m, ψ, a, b hyperparameters that are assumed known.

After updating prior (4.1) with the data for subjects $1, \ldots, i$, we have

$$\pi\big(\mu_h, \tau_h|\mathbf{y}^{(i)}, \boldsymbol{\gamma}^{(i)}\big) \sim N_p\big(\mu_h; m_h^{(i)}, \psi_h^{(i)}\tau_h^{-1}\big)G\big(\tau_h; a_h^{(i)}, b_h^{(i)}\big), \tag{4.2}$$

where the values $m_h^{(i)}, \psi_h^{(i)}, a_h^{(i)}, b_h^{(i)}$ are obtained through sequential application of the updating equations:

$$\begin{aligned}
\psi_h^{(i)} &= \big\{\big(\psi_h^{(i-1)}\big)^{-1} + 1(\gamma_i = h)\big\}^{-1},\\
m_h^{(i)} &= \psi_h^{(i)}\big\{\big(\psi_h^{(i-1)}\big)^{-1}m_h^{(i-1)} + 1(\gamma_i = h)y_i\big\},\\
a_h^{(i)} &= a_h^{(i-1)} + 1(\gamma_i = h)/2,\\
b_h^{(i)} &= b_h^{(i-1)} + \frac{1(\gamma_i = h)}{2}\\
&\qquad \times \big[y_i^2 + \big(m_h^{(i-1)}\big)'\big(\psi_h^{(i-1)}\big)^{-1}m_h^{(i-1)} - \big(m_h^{(i)}\big)'\big(\psi_h^{(i)}\big)^{-1}m_h^{(i)}\big],
\end{aligned}$$

with $m_h^{(0)} = m, \psi_h^{(0)} = \psi, a_h^{(0)} = a, b_h^{(0)} = b$ corresponding to the initial prior in (4.1).

Letting $\pi_{ih} = \Pr(\gamma_i = h|\boldsymbol{\gamma}^{(i-1)})$ as shorthand for the conditional prior probabilities in (2.3) and updating with the data for subject i, we obtain

$$\widehat{\pi}_{ih} = \Pr\big(\gamma_i = h|\boldsymbol{\gamma}^{(i-1)}, \mathbf{y}^{(i)}\big) = \frac{\pi_{ih}f(y_i|\gamma_i = h, \boldsymbol{\gamma}^{(i-1)})}{\sum_{l=1}^{k_{i-1}+1}\pi_{il}f(y_i|\gamma_i = l, \boldsymbol{\gamma}^{(i-1)}, \mathbf{y}^{(i-1)})} \tag{4.3}$$

for $l = 1, \ldots, k_{i-1} + 1$, where $f(y_i|\gamma_i = h, \boldsymbol{\gamma}^{(i-1)}, \mathbf{y}^{(i-1)})$ corresponds to a noncentral t-distribution, a special case in (A.1) in the Appendix for $x = 1$ and one-dimensional β, with $m_h^{(i-1)}, \psi_h^{(i-1)}, a_h^{(i-1)}, b_h^{(i-1)}$ used in place of ξ, Ψ, a, b in (A.1).

4.2 EMPIRICAL SUGS

In implementing SUGS, we have found some sensitivity to the prior specification, which is an expected feature of analyses based on DP mixture models. To reduce this sensitivity, we recommend routinely normalizing the data prior to analysis. Then, one can let $m = 0$, $\psi = 1$, and $a = 1$ in prior (4.1) as a default. However, there is still some sensitivity to the choice of b, which we propose to address through the following procedure. We first choose a prior for b, $\pi(b) = G(c, d)$, with $c = 1, d = 10$ providing a good default choice. We then propose to update this prior sequentially within a preliminary SUGS run to obtain an estimate of b. This estimate will then be plugged in for b in a subsequent SUGS analysis. We find this modification leads to good performance in terms of estimation and model selection in a very wide variety of cases.

Let $\hat{b}^{(i)}$ be an estimate of b after the first $i - 1$ subjects have been incorporated, with

$$\hat{b}^{(i)} = \frac{c + ak_{i-1}}{d + \sum_{h=1}^{k_{i-1}} a_h^{(i-1)}/b_h^{(i-1)}}.$$

The updating equation for $b_h^{(i)}$ is then modified to be

$$b_h^{(i)} = b_h^{(i-1)} + \frac{1(\gamma_i = h)}{2}\left[y_i^2 + \left(m_h^{(i-1)}\right)'\left(\psi_h^{(i-1)}\right)^{-1} m_h^{(i-1)} - \left(m_h^{(i)}\right)'\left(\psi_h^{(i)}\right)^{-1} m_h^{(i)}\right] - \hat{b}^{(i-1)} + \hat{b}^{(i)}.$$

The final estimate $\hat{b}^{(n+1)}$ is used as the value for b in the subsequent SUGS analyses. Although this increases the computational cost, the result is a more robust estimate.

5. SIMULATION STUDY

5.1 PERFORMANCE OF SUGS

Simulation studies were conducted to evaluate the performance of the proposed algorithm. We focused on the normal DPM model of Section 4 and considered two cases for the true density: (1) mixture of three normals:

$$g(y) = 0.3N(y; -2, 0.4) + 0.5N(y; 0, 0.3) + 0.2N(y; 2.5, 0.3),$$

and (2) a single normal with mean 0 and variance 0.4. In each case, we considered 100 simulated datasets each with sample size $n = 500$. For all the analyses reported in this article, we used the default priors recommended in Section 4.2, and took the prior for α to be a discretized Gamma(1, 1) distribution with support on the points $\{0.01,\ 0.05\} \cup \{0.1 + 0.2k, k = 0, 1, \ldots, 20\}$. In addition, SUGS was repeated for 10 random orderings.

For each sampled dataset, we calculated the predictive density using SUGS, the typical frequentist kernel density estimate, and the Bayes factor of the selected model against

the parametric null model (a single normal distribution). The kernel density estimate was obtained using the "ksdensity" function in Matlab, with default settings including using a normal kernel (Bowman and Azzalini 1997) and an optimal default value for kernel width. To measure the closeness of the proposed density estimate and the true density, we calculated the Kullback–Leibler divergence (KLD) between densities f and g defined as follows:

$$K(f, g) = \int f(x) \log\left\{ \frac{f(x)}{g(x)} \right\} dx,$$

with f being the true density and g being an estimate.

Figures 1 and 2 plot the true density (solid) and the 100 predictive densities (dotted) given by the SUGS algorithm in case 1 and case 2, respectively. Clearly, the predictive densities are very close to the true density. The averages of 100 KLDs of the proposed density estimates and the kernel density estimates relative to the true density are 0.0111 and 0.041 in case 1 and 0.0027 and 0.0080 in case 2, respectively. The results suggest that the proposed density estimates are closer to the true density than the kernel density estimates.

Table 1 summarizes the estimated Bayes factors across the simulations. To obtain the Bayes factor, we calculate the marginal likelihood using the formula in (3.2), with the hyperparameter $b = 1$ for the normal baseline model. When data are generated from a mixture of normals in case 1, the Bayes factors provide decisive support in favor of the true model as shown in Table 1. When data are generated from the null model as in simulation 2, the Bayes factors pick up the base normal model over 90% of the datasets. These results show that SUGS has good performance in selecting between a single normal and a mixture of normals.

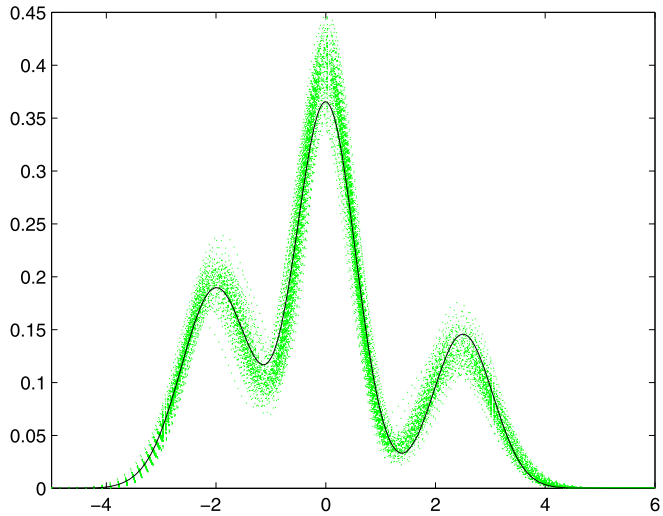

Figure 1. SUGS density estimates in simulation case 1 for $n = 500$. The estimated densities (dotted) from 100 datasets and the true density (solid). A color version of this figure is available in the electronic version of this article.

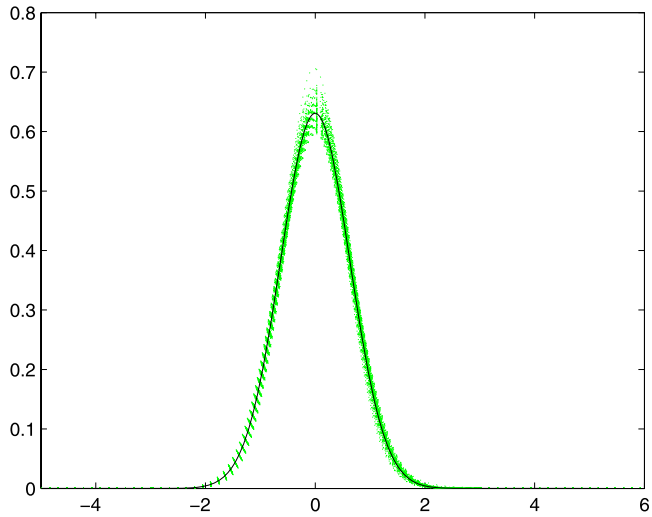

Figure 2. SUGS density estimates in simulation case 2 for $n = 500$. The estimated densities (dotted) from 100 datasets and the true density (solid). A color version of this figure is available in the electronic version of this article.

In the above implementation of SUGS, we treat α as unknown and assign a discretized Gamma prior for it. To see the benefit of this, we run SUGS with several fixed values of α separately. The comparison results are presented in Table 2 in terms of the average of 100 KLDs and the computational time per dataset. From Table 2, SUGS using random α performs better than using fixed value of α in estimating the predictive density because it produces the smallest average of KLDs. This is more apparent in case 1 of the simulation, for which all the averages of KLDs from the SUGS using fixed α values are relatively large. It is observed that the computation time increases as the value of α becomes large. The reason is that large values of α tend to induce more clusters and thus need more computation cost. It is appealing that SUGS using random α has better performance than using fixed value of α whereas it does not necessarily take more time.

In the above simulations, we adopt the PML criterion to eliminate the effect of sequential ordering of subjects in SUGS. For comparison, we also run SUGS using marginal likelihood (ML) criterion mentioned in Section 3.2 together with SUGS using simply averaging (SAV) over 10 random orderings for each dataset. The first part of Table 3 lists the sample standard deviation (SSD) of 100 log marginal likelihood estimates obtained from 100 datasets using these three criteria in SUGS. From Table 3, the PML criterion clearly performs best with smallest SSD compared to the ML and SAV criteria in both cases 1 and

Table 1. Performance of Bayes factor under null and alternative models.

	$BF \leq 1$	$1 < BF \leq 100$	$BF > 100$
Case 1	0	0	100
Case 2	92	4	4

Table 2. Effect of using random and fixed value of α in SUGS.

		Random α	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
KLD	Case 1	0.0111	0.1182	0.0643	0.0488	0.0430
	Case 2	0.0027	0.0037	0.0064	0.0069	0.0067
Time	Case 1	2.98	2.27	3.88	5.55	8.28
	Case 2	2.93	2.71	3.75	5.36	8.03

2 of simulation. The large value of SSD from using the SAV criterion indicates that SUGS is very sensitive to the ordering of subjects being added to the model. The proposed PML criterion does a great job in eliminating the effect of such ordering as clearly indicated from Table 3.

The SUGS algorithm is very fast. In both case 1 and case 2, the analyses for all 100 simulated datasets were completed in approximately 3 minutes for sample size 500. We also ran simulations with sample size $n = 5000$ and obtained excellent results (not shown). All programs including the simulations in Section 5.2 and data analysis in Section 6 were executed in Matlab version 7.3 running on Dell desktop with Intel(R) Xeon(R) CPU and 3.00 GB of RAM.

5.2 COMPARISON WITH FOUR OTHER FAST NONPARAMETRIC DPM ALGORITHMS

To compare SUGS with competing fast nonparametric methods, we reanalyzed the same simulated data in Section 5.1 using the VB method of Blei and Jordan (2006), the PR approach of Newton (2002), the SIS of MacEachern, Clyde, and Liu (1999), and the inadmissible (Inad) approach of Daumé III (2007). We evaluated their performance in terms of predictive density estimation and running time.

The idea of variational inference is to formulate the computation of the posterior distribution as an optimization problem (Blei and Jordan 2006; Wainwright and Jordan 2008). To implement the variational Dirichlet process Gaussian mixture model, we applied the code created by Dr. Kenichi Kurihara; the code is available at *http://sato-www.cs.titech.ac.jp/kurihara/vdpmog.html*. It is known that VB is very sensitive to the initial values and poorly chosen initial values usually result in poor estimates. To overcome this problem, the code adopts sequential importance sampling to find good initial values.

Table 3. Sample standard deviation of log marginal likelihood estimates for running SUGS and the inadmissible approach using different criteria regarding the online clustering of subjects based on 100 datasets: PML, ML, and SAV criteria for SUGS and AS and ML criteria for the inadmissible approach.

	SUGS			Inad	
	PML	ML	SAV	AS	ML
Case 1	17.4	62.4	209.9	99.2	21.6
Case 2	4.1	88.3	143.9	62.9	19.0

The recursive algorithm of Newton (2002) sequentially updates the mixing density $\pi(\theta)$ via the following equation:

$$\pi_i(\theta) = (1 - w_i)\pi_{i-1}(\theta) + w_i \frac{f(y_i|\theta)\pi_{i-1}(\theta)}{c(y_i, \pi_{i-1})},$$

where $\mathbf{w} = (w_1, \ldots, w_n)$ is a sequence of weights satisfying some conditions, $c(y, \pi) = \int f(y|\theta)\pi(\theta)\, d\theta$, and π_n is used as the estimated mixing density. The predictive density estimate is then $f_n(y) = \int f(y|\theta)\pi_n(\theta)\, d\theta$, which is strongly consistent (Tokdar, Martin, and Ghosh 2009). Because π_n depends on the ordering of $(y_1, \ldots, y_n)$, following the recommendation of Newton (2002) and Tokdar, Martin, and Ghosh (2009), we use the average of $f_n(y)$ over 10 permutations.

We generalize the SIS of MacEachern, Clyde, and Liu (1999) for DP mixture of normals. The SIS is similar to SUGS in the sense of sequential updating but differs in that it adopts a random assignment of allocation instead of finding the allocation that maximizes the posterior probabilities in step 2(a) of Section 3.1 in SUGS. Here we implement SIS with 10 particles (permutations) and calculate the predictive density by taking the average for each dataset.

The admissible approach is the one that performs fastest and best among the three algorithms proposed by Daumé III (2007). It aims to find the maximum posteriori assignment of data points to clusters. To achieve this goal, it sequentially updates multiple clusterings in a queue to obtain clusterings that have highest scores. The clustering with the highest score is chosen to be the allocation of all subjects. To be fair to the admissible approach in comparison with SUGS, we adopt only one clustering in the queue. We fix the DP precision parameter α to be 1 in the simulation and calculate marginal likelihood and predictive density based on the final clustering that has the highest score. Because the ordering of subjects can affect the clustering result, Daumé III (2007) recommended to present the data in the ascending ordering (AS) of individual marginal likelihood. We evaluate this criterion and also implement a marginal likelihood (ML) criterion, in which one runs the inadmissible approach over 10 random orderings and chooses the one that leads to the highest marginal likelihood among the 10 final clusterings. The results in the second part of Table 3 suggest that AS is not a good criterion as it produces a large uncertainty. In contrast, the ML criterion seems to be more reliable because the SSD is much smaller as seen in Table 3. Also, Inad ML gives larger marginal likelihoods than Inad AS from our simulation (results not shown).

Table 4 shows the comparison of SUGS with VB, PR, SIS, Inad AS, and Inad ML in terms of the average of 100 KLDs and the running time per dataset. The values of the average of KLDs obtained from SUGS and VB are comparable and are smaller than the corresponding value obtained from PR and the Inad methods in case 1 of simulation. In case 2, the average of KLDs is small for all the approaches except the inadmissible approach although the latter gives acceptable predictive density estimates using the ML criterion.

The second part of Table 4 shows the runtime (in seconds) of all the methods including SUGS per dataset. From Table 4, SUGS is only slower than the SIS and Inad AS, both of which perform poorly in estimating the predictive density, however. Also, we observed

Table 4. Comparison of SUGS, VB, PR, SIS, and the inadmissible algorithm proposed by Daumé III (2007) in terms of the average of KLD and running time.

		SUGS	VB	PR	SIS	Inad AS	Inad ML
KLD	Case 1	0.0111	0.0101	0.0265	0.1395	0.0559	0.0205
	Case 2	0.0027	0.0040	0.0051	0.0032	0.0427	0.0298
Time	Case 1	2.98	33.40	47.18	2.60	0.37	6.94
	Case 2	2.93	33.70	49.01	2.74	0.53	9.05

from our simulations that SUGS works over 10 times faster than VB and and over 15 times faster than the recursive algorithm. The reason that VB is slow here is that the applied code of VB adopts a sequential importance sampling step to obtain feasible initialization for VB, which is quite time-consuming. However, the SIS step is necessary because VB gives poor results if it is used alone. The recursive algorithm is slow because it involves estimating a two-dimensional mixing density in the sequential updating and also averages on 10 permutations to eliminate the ordering effect for each dataset.

6. APPLICATIONS

We applied the SUGS algorithm to three data examples. The first two are galaxy data and enzyme data, which have been studied thoroughly by many people in the literature. The third is gestational age at delivery data from the Collaborative Perinatal Project (CPP), which was a very large epidemiologic study conducted in the 1960s and 1970s. The official CPP data and documentation are available at *ftp://sph-ftp.jhsph.edu/cpp/* provided by Johns Hopkins University School of Public Health. Here we focus on 34,178 pregnancies that had their gestational ages at delivery (GADs) recorded in the CPP data, which provide a large sample size example. The Matlab codes of the SUGS algorithms for the simulation in Section 5.1 and data analyses as well as all the datasets are provided in the Supplemental Materials.

The galaxy data are a commonly used example in assessing methods for Bayesian density estimation and clustering; see, for example, the works of Roeder (1990), Escobar and West (1995), and Richardson and Green (1997), among others. The data contain measured velocities of 82 galaxies from six well-separated conic sections of space. The SUGS algorithm gives five clusters and the corresponding predictive density is shown in Figure 3, which is similar to kernel estimate and the results of Escobar and West (1995), Richardson and Green (1997), and Fearnhead (2004).

The enzyme data record enzyme activities in blood for 245 individuals. One interest in analyzing this dataset is the identification of subgroups of slow or fast metabolizers as a marker of genetic polymorphisms (Richardson and Green 1997). Bechtel et al. (1993) concluded that the distribution is a mixture of two skewed distributions based on a maximum likelihood analysis. Richardson and Green (1997) analyzed the data using Bayesian normal mixtures with an unknown number of components. The application of the SUGS algorithm using the default priors on the enzyme data gives a partition of three clusters.

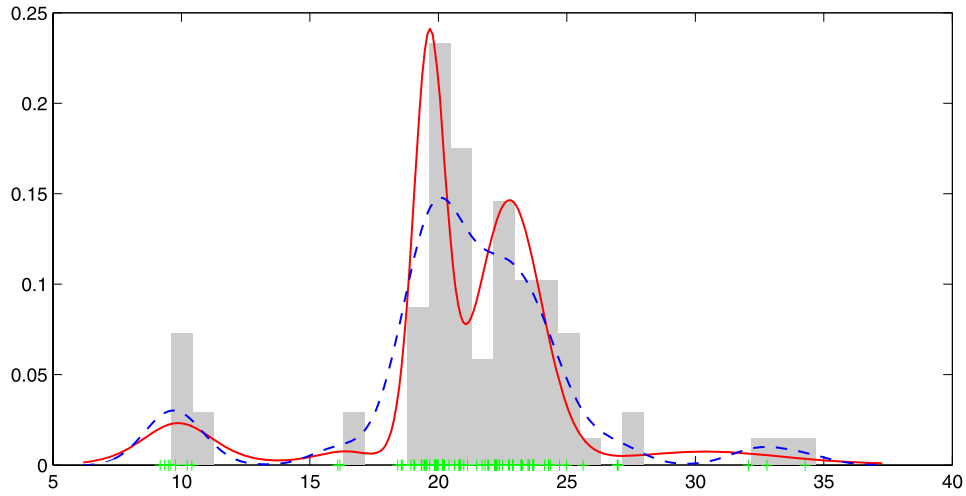

Figure 3. Predictive density estimate (solid), kernel density estimate (dashed), histogram, and plots of galaxy data (+). A color version of this figure is available in the electronic version of this article.

The predictive density shown in Figure 4 agrees closely with the findings in the above mentioned papers.

For the third example, we consider the GADs in weeks for 34,178 births in the CPP. We are interested in the relationship of GAD and the covariates race, sex, maternal smoking status during pregnancy, and maternal age. We use indicators X_1, X_2, X_3, and X_4 to denote these four variables, with 1 indicating black, female, smoker, and maternal age less than 35, respectively, and with 0 indicating nonblack, male, nonsmoker, and maternal age no less than 35, respectively. Table 5 gives the observed frequencies for these covariates.

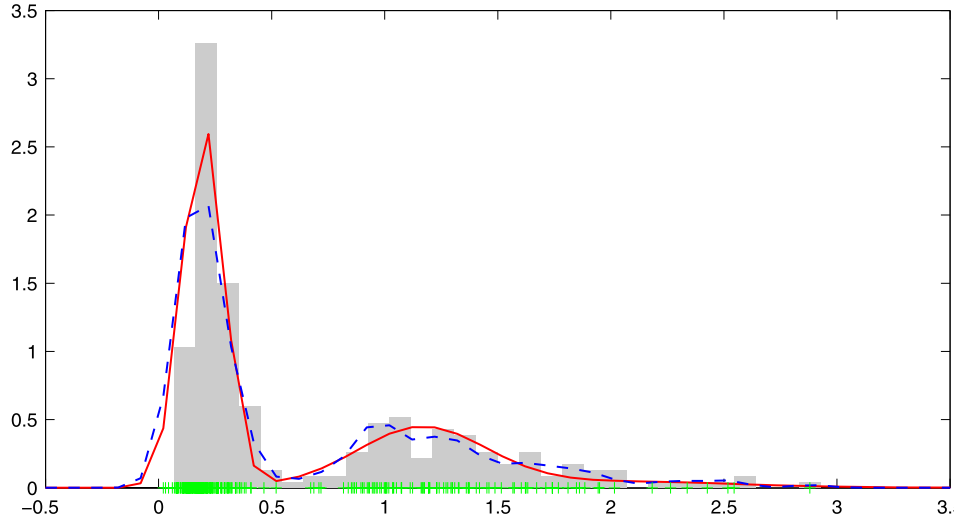

Figure 4. Predictive density estimate (solid), kernel density estimate (dashed), histogram, and plots of enzyme data (+). A color version of this figure is available in the electronic version of this article.

Table 5. The frequency of observations in categories of each covariate.

Value	x_1 (race)	x_2 (sex)	x_3 (smoke)	x_4 (age)
0	51.23%	49.72%	48.51%	7.72%
1	48.77%	50.28%	51.49%	92.28%

The distribution of GAD is known to be nonnormal and have heavy left tails by previous research done by, for example, Smith (2001), among others. In the following, we apply the proposed SUGS algorithm on this dataset. The left tail behavior of the distribution of GAD corresponding to premature deliveries is particularly of interest.

Let $\mathbf{z}_i = (1\ x_{i1}\ x_{i2}\ x_{i3}\ x_{i4})'$ and y_i denote the GAD for subject i. We consider the following model:

$$y_i \sim N(\mathbf{z}_i'\boldsymbol{\beta}_i, \tau_i^{-1}), \qquad (\boldsymbol{\beta}_i, \tau_i)|P \sim P, \qquad P \sim \mathrm{DP}(\alpha P_0)$$
$$P_0(\beta, \tau) = N(\beta; \xi, \Psi\tau^{-1})\,\mathrm{Ga}(\tau; a, b),$$

where $\boldsymbol{\beta}_i = (\beta_{i0}\ \beta_{i1}\ \beta_{i2}\ \beta_{i3}\ \beta_{i4})'$ denotes the random effects of intercept and covariates for subject i. To apply SUGS, we set $\xi = 0$, $\Psi = n(\mathbf{z}\mathbf{z}')^{-1}$, $a = 1$, and estimated b as described in Section 4.2, where $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$. We run 20 permutations for CPP data to eliminate the ordering effect. In terms of computational speed, the analysis was completed within a few seconds for the galaxy and enzyme data, whereas a single permutation took approximately 2 minutes for the CPP data.

Figure 5 shows the estimated predictive densities and cumulative distribution functions of GAD for nonblack babies and black babies controlling other covariates equal to zero. As seen in Figure 5, the predictive density of GAD for black babies is shifted left by around

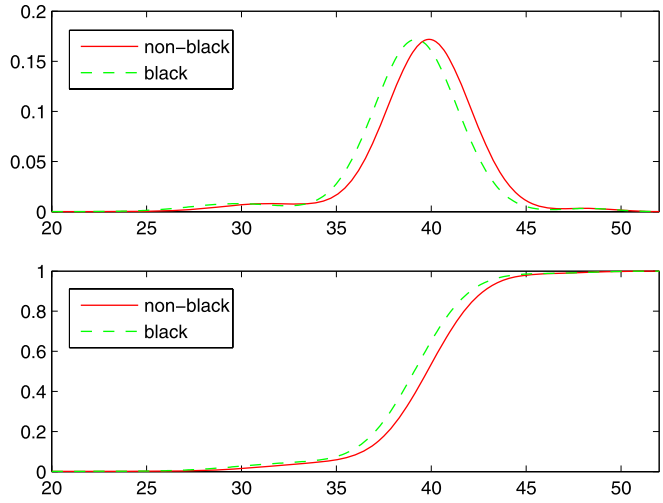

Figure 5. Top: Densities of GAD for black race (dashed) and other race (solid). Bottom: Cumulative distribution functions of GAD for black race (dashed) and other race (solid). A color version of this figure is available in the electronic version of this article.

Table 6. Cluster-specific coefficients obtained by SUGS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
β_0	48.26 (48.19, 48.32)	39.88 (39.88, 39.88)	31.40 (31.36, 31.44)	19.98 (16.19, 23.77)
β_1	−0.01 (−0.02, 0.01)	−0.70 (−0.70, −0.70)	−1.62 (−1.63, −1.61)	−3.28 (−4.43, −2.12)
β_2	−0.11 (−0.12, −0.09)	0.12 (0.12, 0.12)	0.05 (0.04, 0.06)	−0.56 (−1.48, 0.37)
β_3	0.03 (0.02, 0.05)	0.02 (0.01, 0.02)	0.04 (0.03, 0.05)	0.16 (−0.83, 1.14)
β_4	0.04 (−0.02, 0.10)	0.15 (0.15, 0.15)	1.06 (1.03, 1.10)	1.52 (−1.01, 4.04)
$\bar{y}_j$	48.28	39.75	31.39	18.42
n_j	385	32,024	1702	67

1 week compared to the density of GAD for nonblacks. This result suggests that black babies are more likely to be born premature than nonblack babies. Here, we only show the comparison of densities and CDFs of GAD for different race groups, and the corresponding densities and CDFs of GAD for different other covariate groups are very close (not shown).

Note that SUGS will produce clusters of subjects having identical coefficients for the different predictors. Table 6 summarizes the results of the cluster-specific coefficients in the original data scale given by SUGS, including the posterior means and the corresponding 95% credible intervals. Table 6 also presents the sample mean of GAD and the number of subjects for each cluster in the last two rows. Clearly, the four clusters represent different groups of babies with the first cluster at the right tail of the distribution of GAD and the third and fourth clusters at the left tail. Cluster 2 is the dominant cluster containing about 94% babies and the covariate effects are all significant in this cluster due to the extremely large sample size. Seen across the clusters, the effect of black race on GAD tends to increase in the clusters corresponding to lower GAD babies. This implies an interaction in which black race has a significant impact in shifting GAD slightly for full-term deliveries, with a larger impact on timing of premature deliveries.

7. DISCUSSION

We have proposed a fast algorithm for posterior computation and model selection in Dirichlet process mixture models. The proposed SUGS approach is very fast and can be implemented easily in very large datasets when priors are chosen to be conjugate. In the simulations and real data examples we considered, we obtained promising results. Extensions to nonconjugate cases are conceptually straightforward. In such cases, instead of obtaining the exact marginal likelihoods conditional on the allocation to clusters, one can utilize an approximation. A promising strategy for many models would be to use a Laplace approximation. The performance of such an approach remains to be evaluated.

Although our focus was on DPMs, the same type of approach can conceptually be applied in a much broader class of models, including species sampling models and general product partition models.

APPENDIX: DESCRIPTION OF THE PREDICTIVE DISTRIBUTION

Suppose that $(y|\mathbf{x}, \boldsymbol{\beta}, \tau) \sim N(\mathbf{x}'\boldsymbol{\beta}, \tau^{-1})$ with $\pi(\boldsymbol{\beta}, \tau) = N_p(\boldsymbol{\beta}; \boldsymbol{\xi}, \Psi\tau^{-1})G(\tau; a, b)$ the prior. Then, the marginal density of y given $\mathbf{x}$ follows the noncentral t-distribution:

$$f(y|\mathbf{x}) = \frac{\Gamma((\nu+1)/2)}{(\pi\nu)^{1/2}\Gamma(\nu/2)\sigma}\left(1 + \frac{1}{\sigma^2\nu}(y-\mu_y)^2\right)^{-(\nu+1)/2} = t_\nu(y; \mu_y, \sigma^2), \quad \text{(A.1)}$$

where $\nu = 2a$ is the degrees of freedom, $\widehat{\Psi} = (\Psi^{-1} + \mathbf{x}\mathbf{x}')^{-1}$,

$$\mu_y = \frac{\mathbf{x}'\widehat{\Psi}\Psi^{-1}\boldsymbol{\xi}}{1 - \mathbf{x}'\widehat{\Psi}\mathbf{x}} \qquad \text{and} \qquad \sigma^2 = \frac{1}{\nu}\left(\frac{2b + \boldsymbol{\xi}'(\Psi^{-1} - \Psi^{-1}\widehat{\Psi}\Psi^{-1})\boldsymbol{\xi}}{1 - \mathbf{x}'\widehat{\Psi}\mathbf{x}} - \mu_y^2\right),$$

with μ_y the mean and $\sigma^2\nu/(\nu-2)$ the variance for $\nu > 2$.

SUPPLEMENTAL MATERIALS

Data Files: The galaxy data, enzyme data, and CPP data used in Section 6. (Data_application.zip, WinZip archived file)

Matlab Codes for Data Analysis: The SUGS codes used for analyzing the galaxy data, enzyme data, and CPP data. (SUGS_application.zip, WinZip archived file)

Matlab Code for Simulation: The SUGS code used in case 1 of the simulation study in Section 5.1. (SUGSsimu.m, Matlab m file)

ACKNOWLEDGMENTS

The authors thank the editor, the associate editor, and two anonymous reviewers for their critical and constructive comments that greatly improved the presentation of this article.

[Received July 2007. Revised September 2009.]

REFERENCES

Barry, D., and Hartigan, J. A. (1992), "Product Partition Models for Change Point Problems," *The Annals of Statistics*, 20, 260–279. [198]

Basu, S., and Chib, S. (2003), "Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models," *Journal of the American Statistical Association*, 98, 224–235. [197]

Bechtel, Y. C., Bonaïti-Pellié, C., Poisson, N., Magnette, J., and Bechtel, P. R. (1993), "A Population and Family Study of N-Acetyltransferase Using Caffeine Urinary Metabolites," *Clinical Pharmacology & Therapeutics*, 54, 134–141. [210]

Blackwell, D., and MacQueen, J. (1973), "Ferguson Distributions via Polya Urn Schemes," *The Annals of Statistics*, 1, 353–355. [198]

Blei, D. M., and Jordan, M. I. (2006), "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, 1, 121–144. [197,203,208]

Bowman, A. W., and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis*, New York: Oxford Univeristy Press. [206]

Bush, C. A., and MacEachern, S. N. (1996), "A Semiparametric Bayesian Model for Randomized Block Designs," *Biometrika*, 83, 175–185. [196,198]

Dahl, D. B. (2009), "Modal Clustering in a Class of Product Partition Models," *Bayesian Analysis*, 4, 243–264. [200]

Daumé, III, H. (2007), "Fast Search for Dirichlet Process Mixture Models," *Conference on Artificial Intelligence and Statistics*. [198,201,202,208-210]

Escobar, M. D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268–277. [196]

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588. [196,210]

Fearnhead, P. (2004), "Particle Filters for Mixture Models With an Unknown Number of Components," *Statistics and Computing*, 14, 11–21. [198,202,210]

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [196]

——— (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629. [196]

Geisser, S. (1980), Discussion on "Sampling and Bayes Inference in Scientific Modeling and Robustness," by G. E. P. Box, *Journal of the Royal Statistical Society, Ser. A*, 143, 416–417. [201]

Geisser, S., and Eddy, W. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153–160. [201]

Gelfand, A. E., and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 56, 501–514. [201]

Ghosh, J., and Tokdar, S. (2006), "Convergence and Consistency of Newton's Algorithm for Estimating a Mixing Distribution," in *The Frontiers of Statistics*, eds. J. Fan and H. Koul, London: Imperial College Press, pp. 429–443. [197]

Ishwaran, H., and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 101, 179–194. [197]

——— (2003), "Generalized Weighted Chinese Restaurant Processes for Species Sampling Mixture Models," *Statistica Sinica*, 13, 1211–1235. [197]

Ishwaran, H., and Takahara, G. (2002), "Independent and Identically Distributed Monte Carlo Algorithms for Semiparametric Linear Mixed Models," *Journal of the American Statistical Association*, 97, 1154–1166. [197]

Jain, S., and Neal, R. M. (2004), "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," *Journal of Computational and Graphical Statistics*, 13, 158–182. [197]

Kurihara, K., Welling, M., and Teh, Y. W. (2007), "Collapsed Variational Dirichlet Process Mixture Models," in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI07)*, San Francisco, CA: Kaufmann, pp. 2796–2801. [197,203]

Kurihara, K., Welling, M., and Vlassis, N. (2006), "Accelerated Variational Dirichlet Mixture Models," in *Advances in Neural Information Processing Systems, 19 (NIPS 2006)*, Vancouver, British Columbia, Canada. [197,203]

Lau, J. W., and Green, P. J. (2007), "Bayesian Model Based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 16, 526–558. [200]

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I, Density Estimates," *The Annals of Statistics*, 12, 351–357. [196,198]

Lo, A. Y., Brunner, L. J., and Chan, A. T. (1996), "Weighted Chinese Restaurant Processses and Bayesian Mixture Models," Research Report 1, Hong Kong University of Science and Technology. [197]

MacEachern, S. N. (1994), "Estimating Normal Means With a Conjugate Style Dirichlet Process Prior," *Communications in Statistics: Simulation and Computation*, 23, 727–741. [196]

MacEachern, S. N., Clyde, M., and Liu, J. S. (1999), "Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation," *Canadian Journal of Statistics*, 27, 251–267. [197,208,209]

Minka, T., and Ghahramani, Z. (2003), "Expectation and Propagation for Infinite Mixtures," in *NIPS'03 Workshop on Nonparametric Bayesian Methods and Infinite Models*, Vancouver, British Columbia, Canada. [198,202]

Mukhopadhyay, N., Ghosh, J. K., and Berger, J. O. (2005), "Some Bayesian Predictive Approaches to Model Selection," *Statistics & Probability Letters*, 73, 369–379. [202]

Newton, M. A. (2002), "On a Nonparametric Recursive Estimator of the Mixing Distribution," *Sankhyā, Ser. A*, 64, 306–322. [197,202,208,209]

Newton, M. A., and Zhang, Y. (1999), "A Recursive Algorithm for Nonparametric Analysis With Missing Data," *Biometrika*, 86, 15–26. [197]

Park, J.-H., and Dunson, D. B. (2009), "Bayesian Generalized Product Partition Models," *Statistica Sinica*, to appear. [198]

Pettit, L. I. (1990), "The Conditional Predictive Ordinate for the Normal Distribution," *Journal of the Royal Statistical Society, Ser. B*, 52, 175–184. [201]

Quintana, F. A., and Iglesias, P. L. (2003), "Bayesian Clustering and Product Partition Models," *Journal of the Royal Statistical Society, Ser. B*, 65, 557–574. [198]

Quintana, F. A., and Newton, M. A. (2000), "Computational Aspects of Nonparametric Bayesian Analysis With Applications to the Modeling of Multiple Binary Sequences," *Journal of Computational and Graphical Statistics*, 9, 711–737. [197]

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components," *Journal of Royal Statistical Society, Ser. B*, 59, 731–792. [210]

Roeder, K. (1990), "Density Estimation With Confidence Sets Exemplified by Superclusters and Voids in the Galaxies," *Journal of the American Statistical Association*, 85, 617–624. [210]

Sinha, D., Chen, M. H., and Ghosh, S. K. (1999), "Bayesian Analysis and Model Selection for Interval-Censored Survival Data," *Biometrics*, 55, 585–590. [201]

Smith, G. C. S. (2001), "Use of Time to Event Analysis to Estimate the Normal Duration of Human Pregnancy," *Human Reproduction*, 16, 1497–1500. [212]

Stephens, M. (2000), "Dealing With Label Switching in Mixture Models," *Journal of the Royal Statistical Society, Ser. B*, 62, 795–809. [200,201]

Tokdar, S. T., Martin, R., and Ghosh, J. K. (2009), "Consistency of a Recursive Estimate of Mixing Distributions," *The Annals of Statistics*, 37, 2502–2522. [197,202,209]

Wainwright, M. J., and Jordan, M. I. (2008), "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends in Machine Learning*, 1, 1–305. [208]

Wang, B., and Titterington, M. (2005), "Inadequacy of Interval Estimates Corresponding to Variational Bayesian Approximations," in *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, eds. R. G. Cowell and Z. Ghahramani, London: Society for Artificial Intelligence and Statistics, pp. 373–380. [197]

West, M., Müller, P., and Escobar, M. D. (1994), "Hierarchical Priors and Mixture Models, With Application in Regression and Density Estimation," in *Aspects of Uncertainty: A Tribute to D. V. Lindley-Smith AFM*, ed. P. R. Freedman, London: Wiley, pp. 363–386. [196]

Zhang, J., Ghahramani, Z., and Yang, Y. (2005), "A Probabilistic Model for Online Document Clustering With Application to Novelty Detection," in *Advances in Neural Information Processing Systems 17 (NIPS-2004)*, Vancouver, British Columbia, Canada. [198,201,203]