Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information

Hemant ISHWARAN and Lancelot F. JAMES

A rich nonparametric analysis of the finite normal mixture model is obtained by working with a precise truncation approximation of the Dirichlet process. Model fitting is carried out by a simple Gibbs sampling algorithm that directly samples the nonparametric posterior. The proposed sampler mixes well, requires no tuning parameters, and involves only draws from simple distributions, including the draw for the mass parameter that controls clustering, and the draw for the variances with the use of a nonconjugate uniform prior. Working directly with the nonparametric prior is conceptually appealing and among other things leads to graphical methods for studying the posterior mixing distribution as well as penalized MLE procedures for deriving point estimates. We discuss methods for automating selection of priors for the mean and variance components to avoid over or undersmoothing the data. We also look at the effectiveness of incorporating prior information in the form of frequentist point estimates.

Key Words: Almost sure truncation; Blocked Gibbs sampler; Nonparametric hierarchical model; Penalized MLE; Pólya urn Gibbs sampling; Random probability measure.

1. INTRODUCTION

The finite normal mixture model has been the subject of much research interest from a Bayesian perspective. See, for example, Ferguson (1983), Escobar (1988, 1994), Diebolt and Robert (1994), Escobar and West (1995), Chib (1995), Richardson and Green (1997), and Roeder and Wasserman (1997). As far back as Ferguson (1983) it has been realized that the Dirichlet process (Ferguson 1973, 1974) can be used as a powerful nonparametric approach for studying this model. However, earlier attempts for Dirichlet process computing involving mixtures of normals were based on Monte Carlo simulation methods which were difficult to implement for large sample sizes and tended to produce limited posterior inference. See

Hemant Ishwaran is Associate Staff, Department of Biostatistics/Wb4, Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195 (E-mail: ishwaran@bio.ri.ccf.org). Lancelot F. James is Assistant Professor, Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218-2692 (E-mail: james@brutus.mts.jhu.edu).

Journal of Computational and Graphical Statistics, Volume 11, Number 3, Pages 1–26
DOI: 10.1198/106186002411

Ferguson (1983), Lo (1984), and Kuo (1986). It was not until the work of Escobar (1988, 1994) and Escobar and West (1995) using Gibbs sampling methods that a comprehensive approach first became available for Dirichlet process computing in the normal mixture model. Also see MacEachern (1994), West, Müller, and Escobar (1994), Müller, Erkanli, and West (1996) and MacEachern and Müller (1998). This article looks at the use of a new Gibbs sampling method described by Ishwaran and Zarepour (2000) and Ishwaran and James (2001), which differs from the Gibbs sampling approaches mentioned above, by its direct involvement of the nonparametric prior in the updating scheme. The key to this method involves exploiting a precise truncation approximation to the Dirichlet process; which as a by-product allows us to draw values directly from the nonparametric posterior, thus leading to several computational and inferential advantages.

1.1 HIERARCHICAL DESCRIPTION OF THE MODEL

In the finite normal mixture problem, we observe data $\mathbf{X} = (X_1, \ldots, X_n)$, where the X_i are iid random variables with the "true" finite normal mixture density

$$f_{Q_0}(x) = \int_{\Re\times\Re^+} \phi(x|\mu(y), \tau(y))\, dQ_0(y) = \sum_{k=1}^{d} p_{k,0}\phi(x|\mu_{k,0}, \tau_{k,0}), \tag{1.1}$$

where $\phi(\cdot|\mu, \tau)$ represents a normal density with a mean of μ and a variance of $\tau > 0$ and where we write $Y = (\mu(Y), \tau(Y))$ for the two-dimensional mean and variance, where $\mu(\cdot)$ extracts the first coordinate of Y (the mean) and $\tau(\cdot)$ extracts the second coordinate (the variance).

Based on the data $\mathbf{X}$, we would like to estimate the unknown mixture distribution Q_0, which is completely unspecified except for the assumption that it is a finite distribution. Thus, not only are the number of support points $1 \le d < \infty$ unknown, but so are the weights $\{p_{k,0}\}$ and the atoms $\{(\mu_{k,0}, \tau_{k,0})\}$, all of which are to be estimated. It is worth emphasizing at this point that the problem studied here where the number of support points d is unknown is different than the case where d is unknown but bounded by some fixed known value d_0: $1 \le d \le d_0 < \infty$. In the bounded dimension case, it was argued by Ishwaran, James, and Sun (2001) that one could use a finite dimensional Dirichlet prior as an effective method for modeling Q_0. Also see Chen (1995) for more on the inference in finite mixtures with bounded dimension. However, for the unbounded case considered here we adopt the method of modeling Q_0 through the use of a Dirichlet process to allow for mixture models of arbitrary dimension d. Although here we use a truncated Dirichlet process, we will see that in Theorem 1 and Corollary 1 of Section 2.1 that these lead to asymptotic approximations to the posterior that are exponentially accurate.

Notice that the model derived from (1.1) also contains hidden variables Y_i, since it can also be expressed as

$$\begin{aligned}(X_i|Y_i) &\overset{\text{ind}}{\sim} \mathrm{N}(\mu(Y_i), \tau(Y_i)), \quad i = 1, \ldots, n\\ (Y_i|Q_0) &\overset{\text{iid}}{\sim} Q_0(\cdot) = \sum_{k=1}^{d} p_{k,0}\, \delta_{Z_{k,0}}(\cdot),\end{aligned}$$

where $\delta_{Z_{k,0}}(\cdot)$ denotes the discrete measure concentrated at $Z_{k,0} = (\mu_{k,0}, \tau_{k,0})$. Therefore, a full analysis of the normal mixture problem should involve inference for *both* the unknown mixing distribution Q_0, as well as the unknown hidden variables Y_i. However, Dirichlet process computing based on the Pólya urn Gibbs sampling method of Escobar (1988, 1994) has traditionally focused only on the analysis for the hidden variables Y_i. This is an artifact of the Pólya urn approach which is used to circumvent the difficulty in working directly with the Dirichlet process. Although this clever method leads to a versatile Gibbs sampler for Y_i, one needs to convert these posterior values into inference for Q_0, which in the end will require some form of approximation to the Dirichlet process [see Theorem 3 from Ishwaran and James (2001) for a general method for converting posterior Y_i values into draws from the posterior random measure]. Our argument is that one might as well start with a Dirichlet process approximation, which as a by-product naturally produces draws from the posterior of Q_0, while at the same time leading to several computational/inferential advantages.

1.2 OUTLINE AND GOALS OF THE ARTICLE

The original Pólya urn Gibbs sampler developed by Escobar (1988, 1994) has evolved over time to deal with various issues. In Escobar and West (1995), a method for updating the Dirichlet mass parameter for controlling clustering was developed, while MacEachern (1994), West, Müller, and Escobar (1994), and Bush and MacEachern (1996) presented various solutions for dealing with the slow convergence seen with the original sampler. Another delicate issue is the problem associated with nonconjugacy, which has led to various approaches and modifications to the original algorithm. See MacEachern and Müller (1998), Neal (2000), and Walker and Damien (1998).

The goal of this article is to introduce a new type of Gibbs sampler, which we refer to as a *blocked Gibbs sampler*, as a competitive computational procedure for Dirichlet process computing in finite normal mixture models (the details are presented in Section 3). The proposed Gibbs sampler is conceptually easy to understand, even for novices to Bayes nonparametric MCMC methods, and is relatively straightforward to implement; requiring only the ability to draw values from simple conditional distributions, and requires no tuning parameters. It handles all the issues mentioned earlier, including (a) the ability to draw posterior values for Q_0; (b) a simple update for the Dirichlet mass parameter; (c) the ability to deal with nonconjugacy; and (d) good mixing properties. The methodology for the blocked Gibbs sampler was given by Ishwaran and Zarepour (2000) and Ishwaran and James (2001) in a general setting. The contribution of this article will be to give the many details for applying this method to the normal mixture problem, such as the details surrounding the selection of priors and hyperparameters for the mean and variance components, which can be *critical to the amount of smoothing of the data* and hence whose choice are critical for inference of Q_0. In particular, as an automated procedure for dealing with smoothing, we develop an inverse sampling method for the variance based on a nonconjugate uniform prior (see Section 3.2.1). Another contribution are the graphical methods we have developed for converting the large amount of posterior information contained in draws from the posterior

random measure into interpretable inference for Q_0. We also look at the use of a Monte Carlo penalized MLE as a method for converting posterior information into a simple point estimate for the mixing distribution. In Section 4 we study the use of informative priors using frequentist point estimates for Q_0 and study their effect on the posterior through graphical methods as well as by considering the resulting penalized estimators (see also Section 2.4). Finally, we present Theorem 1 and Corollary 1 in Section 2.1 as a tool for choosing Dirichlet process truncations which adequately approximate the posterior. The results are easy to use in practice and can be used in conjunction with the output of the blocked Gibbs sampler. The methods are illustrated by our examples of Sections 4 and 5.

2. HIERARCHICAL PRIORS FOR THE RANDOM MEASURE

The Bayesian nonparametric approach for estimating the true normal mixture model (1.1) is based on the following hierarchical model

$$\begin{aligned} (X_i|Y_i) &\overset{\text{ind}}{\sim} \mathrm{N}(\mu(Y_i),\tau(Y_i)), \quad i=1,\ldots,n \\ (Y_i|P) &\overset{\text{iid}}{\sim} P \\ P &\sim \mathcal{P}_N, \end{aligned} \tag{2.1}$$

where

$$\mathcal{P}_N(\cdot) = \sum_{k=1}^{N} p_k\, \delta_{Z_k}(\cdot)$$

is a random probability measure and $Z_k = (\mu_k, \tau_k)$ are iid variables with distribution H independent of $\mathbf{p} = (p_1,\ldots,p_N)$. Therefore, the use of the prior $\mathcal{P}_N$ is a nonparametric method for modeling the unknown mixture distribution Q_0.

The prior $\mathcal{P}_N$ is an approximate Dirichlet process, defined by choosing its random weights p_k by the stick-breaking construction

$$p_1 = V_1 \quad \text{and} \quad p_k = (1-V_1)(1-V_2)\ldots(1-V_{k-1})\,V_k \qquad k=2,\ldots,N, \tag{2.2}$$

where $V_1, V_2, \ldots V_{N-1}$ are iid $\mathrm{Beta}(1,\alpha)$ random variables and we set $V_N = 1$ to ensure that $\sum_{k=1}^{N} p_k = 1$. By the construction given by Sethuraman (1994) (see also McCloskey 1965; Sethuraman and Tiwari 1982; Donnelly and Joyce 1989; Perman, Pitman, and Yor 1992), it easily follows that $\mathcal{P}_N$ converges almost surely to a Dirichlet process with measure αH, written as $\mathrm{DP}(\alpha H)$, that is, $\mathcal{P}_N \overset{\text{a.s.}}{\to} \mathrm{DP}(\alpha H)$. We refer to H as the reference distribution and α as the Dirichlet mass parameter. See also Muliere and Tardella (1998) who discussed "ϵ-truncation" approximations to the Dirichlet process.

2.1 TRUNCATION VALUES FOR N

It is straightforward to choose a value for N that leads to a precise approximation. A useful method for selecting N is to choose a value that yields a marginal density for $\mathbf{X}$

almost indistinguishable from its limit. Let

$$m_N(\mathbf{X}) = \int \left(\prod_{i=1}^{n} \int_{\Re \times \Re^+} \phi(X_i|\mu(Y_i), \tau(Y_i))\, P(dY_i) \right) \mathcal{P}_N(dP)$$

denote the marginal density of (2.1). Similarly, let m_∞ denote the marginal density of the normal mixture hierarchical model (2.1) subject to a a DP(αH) random measure for P. See the Appendix for a proof of the following $\mathcal{L}_1$ error bound.

Theorem 1. *We have,*

$$\begin{aligned} \int_{\Re^n} |m_N(\mathbf{X}) - m_\infty(\mathbf{X})|\, d\mathbf{X} &\leq 4 \left[1 - I\!E \left\{ \left(\sum_{k=1}^{N-1} p_k \right)^n \right\} \right] \\ &\approx 4n \exp(-(N-1)/\alpha), \end{aligned} \tag{2.3}$$

where p_k are the stick-breaking random weights defined by (2.2).

Notice that the sample size has a modest effect on the bound for a reasonably large value of N. For example, if $n = 1{,}000$, and if we use a truncation value of $N = 50$, then even for the fairly large value $\alpha = 3$, we get an $\mathcal{L}_1$ bound of 3.2×10^{-4}. Therefore, even for fairly large sample sizes, a mere truncation of $N = 50$ leads to a hierarchical model that is effectively indistinguishable from one based on the DP(αH). Of course the adequacy of this truncation will also depend upon α, but even if this is an unknown parameter we can still monitor (2.3) by looking at the value for α in our Gibbs sampler. See Ishwaran and Zarepour (2000) for more details.

The bound provided by Theorem 1 also implies an error bound for the truncated Dirichlet process posterior. This may be somewhat expected, as the marginal density is a key component in the posterior. The posterior error bound is described in terms of the posterior clustering behavior of the hidden variables $Y_1, \ldots, Y_n$, or equivalently in terms of the posterior behavior of classification variables $K_1, \ldots, K_n$. Later in Section 3 we will see that one of the keys to the blocked Gibbs sampler is that it exploits the equivalent representation for Y_i in (2.1) as Z_{K_i}, where

$$(K_i|\mathbf{p}) \stackrel{\text{iid}}{\sim} \sum_{k=1}^{N} p_k\, \delta_k(\cdot), \quad i = 1, \ldots, n,$$

are classification variables identifying the Z_k corresponding to a specific Y_i. Thus, given the classification vector $\mathbf{K} = (K_1, \ldots, K_n)$ one can describe the clustering behavior of the Y_i.

Notice that $\mathbf{K} \in \{1, \ldots, N\}^n$ under $\mathcal{P}_N$, while $\mathbf{K}$ under the DP(αH) is the vector of K_i variables defined by

$$(K_i|\mathbf{p}) \stackrel{\text{iid}}{\sim} \sum_{k=1}^{\infty} p_k\, \delta_k(\cdot), \quad i = 1, \ldots, n,$$

for random weights p_k defined by the stick-breaking procedure (2.2) for $k = 1, 2, \ldots$. Thus, under the Dirichlet process, $\mathbf{K} \in \mathcal{K}_\infty$, where $\mathcal{K}_\infty = \{1, 2, \ldots\}^n$. As a consequence of Theorem 1 we can prove the following:

Corollary 1. *We have,*

$$\int_{\Re^n} \left(\sum_{\mathbf{K} \in \mathcal{K}_\infty} |\pi_N(\mathbf{K}|\mathbf{X}) - \pi_\infty(\mathbf{K}|\mathbf{X})| \right) m_\infty(\mathbf{X})\, d\mathbf{X} = O(n \exp(-(N-1)/\alpha)),$$

where $\pi_N(\mathbf{K}|\mathbf{X})$ *and* $\pi_\infty(\mathbf{K}|\mathbf{X})$ *are the posteriors for* $\mathbf{K}$ *under* $\mathcal{P}_N$ *and the Dirichlet process, DP*(αH)*, respectively.*

Thus, Corollary 1 tells us that the posterior for $\mathbf{K}$ under $\mathcal{P}_N$ is exponentially accurate when integrated with respect to the marginal density m_∞ under the Dirichlet process. Notice that the bound also shows how N could be selected to depend upon the sample size n to ensure that the posterior will be asymptotically accurate. See the Appendix for a proof of the corollary.

2.2 PRIORS FOR THE REFERENCE DISTRIBUTION AND DIRICHLET MASS PARAMETER

To complete the prior specification for $\mathcal{P}_N$ we use the following priors for $Z_k = (\mu_k, \tau_k)$ and α:

$$\begin{aligned} (\mu_k|\theta, \sigma_\mu) &\overset{\text{iid}}{\sim} \text{N}(\theta, \sigma_\mu), \quad k = 1, \ldots, N \\ (\tau_k^{-1}|\nu_1, \nu_2) &\overset{\text{iid}}{\sim} \text{Gamma}(\nu_1, \nu_2) \\ (\alpha|\eta_1, \eta_2) &\sim \text{Gamma}(\eta_1, \eta_2) \\ \theta &\sim \text{N}(0, A). \end{aligned} \tag{2.4}$$

Here we are writing Gamma(ν_1, ν_2), for example, to denote a gamma distribution with shape parameter ν_1 and scale parameter ν_2, so that the mean for the distribution in this parameterization is ν_1/ν_2.

In (2.4) we include a mean parameter θ for the μ_k values. This is very useful when the data are uncentered (as in our two examples of Sections 4 and 5). The conditional mean for θ is approximately

$$\mathbb{E}(\theta|\mu) \approx \frac{1}{N} \sum_{k=1}^{N} \mu_k$$

under a noninformative prior (i.e., for large values of A; in our examples we used $A =$ 1,000), and thus θ should be centered near the mean of $\mathbf{X}$ which allows the prior $\mathcal{P}_N$ to properly model $\mu_{k,0}$ when the data are uncentered. It is also important to select a value for σ_μ that will produce values μ_k that blanket an interval where we might anticipate the true mean values $\mu_{k,0}$ will lie in. Therefore, as we anticipate θ to be centered at the mean for the data, a good choice is to set $\sqrt{\sigma_\mu}$ equal to four times the standard deviation of the data $\mathbf{X}$.

To properly model $\tau_{k,0}$ we should be careful in selecting the choice of the shape and scale parameters ν_1 and ν_2 in the inverse gamma prior used for τ_k. The eventual choice plays a critical role in the amount of smoothing of the data, and directly effects the number of estimated clusters. One good choice is to let $\nu_1 = \nu_2 = 2$, which ensures that τ_k will

take values between 0 and 3 with high probability. This selection works well when the data have been rescaled so that there are no unreasonably large or small values, and have been rescaled so that the true variances $\tau_{k,0}$ will lie in the range of values between 0 and 3. However, trying to rescale the data to satisfy these constraints can sometimes require a fair amount of tuning, and if not properly done the inverse gamma prior for τ_k will act as an informative prior. A more automated procedure dispenses with conjugacy and instead employs a uniform prior for τ_k:

$$\tau_k \overset{\text{iid}}{\sim} \text{Uniform}[0, T], \quad k = 1, \ldots, N. \tag{2.5}$$

Selecting the upper bound T can be based on various automated rules, such as setting T to equal the variance of the data. Another nice feature of the uniform prior is that it allows for a simple inverse sampling method for updating τ_k in the Gibbs sampler. See Section 3.2.1 for details. We will investigate the use of a uniform prior in Section 5.

Finally, the values for η_1 and η_2 in the prior for α (the Dirichlet mass parameter) should be selected with care because they directly control the amount of clustering. For example, larger values for α in the approximate Dirichlet process will encourage more distinct Y_i values, and this will encourage more smoothing and an estimate for Q_0 with many components. A good choice for the hyperparameters is to use the values $\eta_1 = \eta_2 = 2$, which will encourage both small and large values for α.

2.3 EQUAL VARIANCE MODELS

In some normal mixture examples we may only be interested in modeling the mean with a mixture distribution, with the variance component modeled parametrically as a positive parameter [see the analysis of Section 4; also consult Ishwaran and Zarepour (2000) for further applications of this method via the blocked Gibbs sampler]. This is easily accommodated within the framework here by setting $\tau_1 = \tau_2 = \cdots = \tau_N$ to equal some parameter, say, τ_0 and defining

$$\mathcal{P}_N(\cdot) = \sum_{k=1}^{N} p_k \, \delta_{\mu_k}(\cdot).$$

(Note that now H is a prior only for μ_k.) A convenient prior for τ_0 is

$$(\tau_0^{-1} | a_0, b_0) \sim \text{Gamma}(a_0, b_0), \tag{2.6}$$

where we choose small values $a_0 = b_0 = 0.01$ to ensure that the prior is noninformative. A uniform prior for τ_0 can also be used.

2.4 SUBJECTIVE PRIORS: INFORMATIVE *H*

We can also incorporate prior information for Q_0 to allow the prior $\mathcal{P}_N$ to more subjectively model the mixing distribution. In this approach we replace the reference distribution

H with the mixture of distributions

$$H^*(\cdot) = wH(\cdot) + (1-w)H_M(\cdot),$$

where $0 < w < 1$ is used to quantify our belief in H and H_M is a prior based on subjective information. In our later applications we will explore the use of informative distributions for the mean in equal variance models as discussed above. We will use mixtures of normals of the form

$$H^*(\cdot) = w\phi(\cdot|\widehat{Y}_0,\widehat{\sigma}_0)+(1-w)\sum_{k=1}^{M}\widehat{\gamma}_k\phi(\cdot|\widehat{Y}_k,\widehat{\sigma}), \quad \text{where} \quad \widehat{\gamma}_k > 0, \ \sum_{k=1}^{M}\widehat{\gamma}_k = 1. \tag{2.7}$$

The first normal density on the right represents the effect of the prior H. We will set $\widehat{Y}_0$ to equal the sample average and $\sqrt{\widehat{\sigma}_0}$ to equal four times the standard deviation of the data in our examples. This follows our earlier approach of selecting the hyperparameters for H and ensures that H^* will, with probability w, produce values for μ_k that should cover an appropriate region of the sample space. The values $\widehat{\gamma}_k$ and $\widehat{Y}_k$ in $H_M(\cdot) = \sum_{k=1}^{M}\widehat{\gamma}_k\phi(\cdot|\widehat{Y}_k,\widehat{\sigma})$ will be taken to be frequentist point estimates for Q_0. For example, we could select $\widehat{\gamma}_k$ and $\widehat{Y}_k$ to be the weights and atoms obtained from the NPMLE for Q_0. Observe that the variance $\widehat{\sigma}$ in H_M can be used to further quantify our prior beliefs, with smaller values used to reflect a stronger prior belief in our point estimates. We return to these points again in in Section 4. See also Section 3.2.3 for computational details.

3. BLOCKED GIBBS SAMPLING

The trick to obtaining direct inference for $\mathcal{P}_N$, and to constructing an efficient Markov chain Monte Carlo method, is to recast the nonparametric hierarchical model completely in terms of random variables. By using the identity $Y_i = (\mu_{K_i}, \tau_{K_i})$, it follows that we can rewrite (2.1) as

$$\begin{aligned}
(X_i|\boldsymbol{\mu},\boldsymbol{\tau},\mathbf{K}) &\overset{\text{ind}}{\sim} N(\mu_{K_i},\tau_{K_i}) \quad i=1,\ldots,n \\
(K_i|\mathbf{p}) &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} p_k\,\delta_k(\cdot) \\
(\mu_k,\tau_k^{-1}|\theta) &\overset{\text{iid}}{\sim} N(\theta,\sigma_\mu)\otimes\text{Gamma}(\nu_1,\nu_2), \quad k=1,\ldots,N \\
\alpha &\sim \text{Gamma}(\eta_1,\eta_2) \\
\theta &\sim \text{N}(0,A),
\end{aligned} \tag{3.1}$$

where $\mathbf{p}$ is defined by the stick-breaking construction (2.2), and $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_N)$ and $\boldsymbol{\tau} = (\tau_1,\ldots,\tau_N)$.

By rewriting the model as (3.1), we can devise a Gibbs sampling scheme for exploring the posterior $\mathcal{P}_N|\mathbf{X}$. To implement the blocked Gibbs sampler we iteratively draw values

from the following conditional distributions:

$$
\begin{array}{l}
(\boldsymbol{\mu}|\boldsymbol{\tau}, \mathbf{K}, \theta, \mathbf{X}) \\
(\boldsymbol{\tau}|\boldsymbol{\mu}, \mathbf{K}, \mathbf{X}) \\
(\mathbf{K}|\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{X}) \\
(\mathbf{p}|\mathbf{K}, \alpha) \\
(\alpha|\mathbf{p}) \\
(\theta|\boldsymbol{\mu}).
\end{array}
$$

This method eventually produces values drawn from the distribution $(\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{K}, \mathbf{p}, \alpha, \theta|\mathbf{X})$ and in each cycle of the sampler we can keep track of $(\boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \mathbf{p}^*)$ which are sampled values for $(\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{p})$. These values produce a random probability measure

$$\mathcal{P}_N^*(\cdot) = \sum_{k=1}^{N} p_k^* \, \delta_{(\mu_k^*, \tau_k^*)}(\cdot)$$

which is a draw from the posterior $\mathcal{P}_N|\mathbf{X}$. Thus, $\mathcal{P}_N^*$ can be used to directly estimate $\mathcal{P}_N|\mathbf{X}$ and its functionals.

For example, to derive an estimate for a future observation $Y_{n+1} = (\mu_{K_{n+1}}, \tau_{K_{n+1}})$, we randomly draw a value Y_{n+1}^* from $\mathcal{P}_N^*$. We can also estimate the predictive density for a future observation X_{n+1}. If $f(X_{n+1}|\mathbf{X})$ denotes the predictive density for X_{n+1} conditioned on the data $\mathbf{X}$, then

$$
\begin{aligned}
f(X_{n+1}|\mathbf{X}) &= \int \phi(X_{n+1}|\mu(Y_{n+1}), \tau(Y_{n+1}))) \, \pi(dY_{n+1}|\mathbf{X}) \\
&= \int\int \phi(X_{n+1}|\mu(Y_{n+1}), \tau(Y_{n+1}))) \, P(Y_{n+1}) \, \mathcal{P}_N(dP|\mathbf{X}).
\end{aligned}
$$

For a probability measure P drawn from $\mathcal{P}_N|\mathbf{X}$,

$$\int \phi(X_{n+1}|\mu(Y_{n+1}), \tau(Y_{n+1})) \, P(dY_{n+1}) = \sum_{k=1}^{N} p_k^* \phi(X_{n+1}|\mu_k^*, \tau_k^*). \tag{3.2}$$

Consequently, the predictive density $f(X_{n+1}|\mathbf{X})$ can be approximated by computing the mixture of normal densities (3.2) averaged over different sampled values $(\boldsymbol{\mu}^*, \boldsymbol{\tau}^*, \mathbf{p}^*)$.

The draw $\mathcal{P}_N^*$ can also be used to derive a Monte Carlo penalized MLE. First notice that many of the random weights p_k^* in the draw may be near zero, and thus the effective dimension for $\mathcal{P}_N^*$ will typically be much smaller than N, its number of atoms. In this penalization approach, we replace $\mathcal{P}_N^*$ with $\widetilde{\mathcal{P}}_N^*$, a random measure including only those non-negligible random weights p_k^*. An effective method for selecting such weights is to use only those values whose corresponding atoms (μ_k^*, τ_k^*) have been selected for some Y_i. That is, since $Y_i = (\mu_{K_i}, \tau_{K_i})$, we use only those p_k^* for which $r_k = \#\{K_i = k\}$ is positive. We define

$$\widetilde{\mathcal{P}}_N^*(\cdot) = \sum_{k=1}^{N} \frac{I\{r_k > 0\} p_k^*}{\sum_{k=1}^{N} I\{r_k > 0\} p_k^*} \, \delta_{(\mu_k^*, \tau_k^*)}(\cdot).$$

The optimal $\widetilde{\mathcal{P}}_N^*$ is the draw over a large number of draws with the largest value

$$l_n(\widetilde{\mathcal{P}}_N^*) - a_n(\widetilde{\mathcal{P}}_N^*),$$

where $l_n(Q) = \sum_{i=1}^n \log f_Q(X_i)$ is the log-likelihood evaluated at a mixing distribution Q and $a_n(Q)$ is the penalty for Q. We will consider two different penalties: (1) Schwartz's BIC criteria (Schwartz 1978), which corresponds to the penalty

$$a_n(\widetilde{\mathcal{P}}_N^*) = \frac{1}{2}\log n \times \text{dimension}(\widetilde{\mathcal{P}}_N^*) = \log n \times \left(\sum_{k=1}^N I\{r_k > 0\} - \frac{1}{2}\right),$$

and (2) Akaike's AIC criteria (Akaike 1973),

$$a_n(\widetilde{\mathcal{P}}_N^*) = \text{dimension}(\widetilde{\mathcal{P}}_N^*) = 2\sum_{k=1}^N I\{r_k > 0\} - 1.$$

See Section 4 for an example illustrating this method.

Remark 1. Notice that the blocked Gibbs algorithm makes use of blocked updates for parameters. This allows the unobserved Y_i values to be updated simultaneously and is one of reasons for its success in producing a rapidly mixing Markov chain. In contrast, due to their use of one-coordinate-at-a-time updates, Pólya urn Gibbs samplers like those discussed by Escobar (1988, 1994) and Escobar and West (1995) tend to suppress the ability for similar Y_i values to change easily as the sampler iterates. To deal with this particular problem, one needs to apply various acceleration methods as discussed by MacEachern (1994), West, Müller, and Escobar (1994), and Bush and MacEachern (1996). An empirical comparison of the mixing performance of the blocked Gibbs sampler to various Pólya urn Gibbs samplers was given by Ishwaran and James (2001).

3.1 BLOCKED GIBBS ALGORITHM

The arguments used in Ishwaran and Zarepour (2000) can be extended to derive the required conditional distributions. Let $\{K_1^*, \ldots, K_m^*\}$ denote the current m unique values of $\mathbf{K}$. In each iteration of the Gibbs sampler we simulate:

(a) Conditional for $\boldsymbol{\mu}$: For each $j \in \{K_1^*, \ldots, K_m^*\}$, draw

$$(\mu_j|\boldsymbol{\tau}, \mathbf{K}, \theta, \mathbf{X}) \stackrel{\text{ind}}{\sim} \text{N}(\mu_j^*, \sigma_j^*), \quad \text{where} \quad \mu_j^* = \sigma_j^* \left(\sum_{\{i:K_i=j\}} X_i/\tau_j + \theta/\sigma_\mu\right),$$

$\sigma_j^* = (n_j/\tau_j + 1/\sigma_\mu)^{-1}$, and n_j is the number of times K_j^* occurs in $\mathbf{K}$. Also, for each $j \in \mathbf{K} - \{K_1^*, \ldots, K_m^*\}$, independently simulate $\mu_j \sim \text{N}(\theta, \sigma_\mu)$.

(b) Conditional for $\boldsymbol{\tau}$: For each $j \in \{K_1^*, \ldots, K_m^*\}$, draw

$$(\tau_j^{-1}|\boldsymbol{\mu}, \mathbf{K}, \mathbf{X}) \stackrel{\text{ind}}{\sim} \text{Gamma}(\nu_1 + n_j/2, \nu_{2,j}^*),$$
$$\text{where} \quad \nu_{2,j}^* = \nu_2 + \sum_{\{i:K_i=j\}} (X_i - \mu_j)^2/2.$$

Also, for each $j \in \mathbf{K} - \{K_1^*, \ldots, K_m^*\}$, independently simulate $\tau_j^{-1} \sim \text{Gamma}(\nu_1, \nu_2)$.

(c) Conditional for $\mathbf{K}$:

$$(K_i|\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{X}) \overset{\text{ind}}{\sim} \sum_{k=1}^{N} p_{k,i}\, \delta_k(\cdot), \quad i = 1, \ldots, n,$$

where

$$(p_{1,i}, \ldots, p_{N,i}) \propto \left(\frac{p_1}{\sqrt{\tau_1}} \exp\left(\frac{-1}{2\tau_1}(X_i - \mu_1)^2 \right), \right.$$
$$\left. \ldots, \frac{p_N}{\sqrt{\tau_N}} \exp\left(\frac{-1}{2\tau_N}(X_i - \mu_N)^2 \right) \right).$$

(d) Conditional for $\mathbf{p}$:

$$p_1 = V_1^* \quad \text{and} \quad p_k = (1 - V_1^*)(1 - V_2^*) \ldots (1 - V_{k-1}^*)V_k^*, \quad k = 2, \ldots, N-1,$$

where

$$V_k^* \overset{\text{ind}}{\sim} \text{Beta}\left(1 + r_k, \alpha + \sum_{l=k+1}^{N} r_l \right), \quad \text{for} \quad k = 1, \ldots, N-1$$

and (as before) r_k records the number of K_i values which equal k.

(e) Conditional for α:

$$(\alpha|\mathbf{p}) \sim \text{Gamma}\left(N + \eta_1 - 1, \eta_2 - \sum_{k=1}^{N-1} \log(1 - V_k^*) \right),$$

for the *same values* of V_k^* used in the simulation for $\mathbf{p}$.

(f) Conditional for θ:

$$(\theta|\boldsymbol{\mu}) \sim \text{N}(\theta^*, \sigma^*), \quad \text{where} \quad \theta^* = \frac{\sigma^*}{\sigma_\mu} \sum_{k=1}^{N} \mu_k$$

and $\sigma^* = (N/\sigma_\mu + 1/A)^{-1}$.

3.2 EXTENSIONS TO THE BLOCKED GIBBS ALGORITHM

3.2.1 Inverse Sampling for τ_k

As mentioned in Section 2.2, there is a simple inverse cdf method for sampling τ_k under a uniform prior (2.5). For notational ease, let $\tau = \tau_{K_j^*}$. Then τ has the conditional density

$$f(\tau) \propto \tau^{-n_j/2} \exp(-C_j/\tau)\{0 < \tau < T\}, \quad \text{where} \quad C_j = \sum_{\{i: K_i = K_j^*\}} (X_i - \mu_{K_j^*})^2/2$$

and n_j is the cardinality of $\{i : K_i = K_j^*\}$, as before. If $\sigma = C_j/\tau$, then σ has the density

$$f(\sigma) \propto \sigma^{n_j/2-2} \exp(-\sigma)\{C_j/T < \sigma < \infty\}.$$

Therefore, to sample τ we first sample σ and then set $\tau = C_j/\sigma$. Sampling σ depends upon the value of n_j. There are three distinct possibilities:

1. $n_j > 2$: In this case, σ is a truncated Gamma$(n_j/2 - 1)$ random variable. Let

$$\Gamma(a,t) = \frac{1}{\Gamma(a)} \int_0^t u^{a-1} \exp(-u)\, du, \quad a > 0$$

be the normalized incomplete gamma function. Then by inverse sampling (see Devroye 1986), it follows that

$$\sigma \stackrel{\mathcal{D}}{=} \Gamma^{-1}\left(a_j\,,\, \Gamma(a_j, C_j/T) + U\left(1 - \Gamma(a_j, C_j/T)\right)\right),$$

where $a_j = n_j/2 - 1$ and $U \sim$ Uniform$[0,1]$. The functions $\Gamma(a,\cdot)$ and $\Gamma^{-1}(a,\cdot)$ are easy to compute in various software packages. For example, in S-Plus, $\Gamma(a,\cdot)$ is called by the function pgamma$(\cdot, a)$, while $\Gamma^{-1}(a,\cdot)$ is called by the function qgamma$(\cdot, a)$.

2. $n_j = 2$: In this case, we can approximate the density for σ by

$$f(\sigma) \propto \sigma^{\epsilon-1} \exp(-\sigma)\{C_j/T < \sigma < \infty\}$$

for some small value for $\epsilon > 0$, say $\epsilon = 10^{-6}$. Then, applying the same reasoning as before,

$$\sigma \stackrel{\mathcal{D}}{=} \Gamma^{-1}\left(\epsilon\,,\, \Gamma(\epsilon, C_j/T) + U\left(1 - \Gamma(\epsilon, C_j/T)\right)\right).$$

3. $n_j = 1$: In this case, we sample σ using $\sigma \stackrel{\mathcal{D}}{=} F^{-1}(U)$, where

$$F(t) = \frac{\int_{C_j/T}^t \sigma^{-3/2} \exp(-\sigma)\, d\sigma}{\int_{C_j/T}^\infty \sigma^{-3/2} \exp(-\sigma)\, d\sigma}, \quad t > C_j/T.$$

Using integration by parts ($\sigma^{-3/2} = -2\frac{\partial}{\partial\sigma}\sigma^{-1/2}$), this can be rewritten as

$$F(t) = \frac{\Gamma(0.5,t) + (\pi t)^{-1/2}\exp(-t) - \left[\Gamma(0.5, C_j/T) + (\pi C_j/T)^{-1/2}\exp(-C_j/T)\right]}{1 - \left[\Gamma(0.5, C_j/T) + (\pi C_j/T)^{-1/2}\exp(-C_j/T)\right]}.$$

Computing the inverse, F^{-1}, is fairly straightforward using standard root finders, such as the method of bisection. Note that this part of the algorithm is *applied very infrequently*, since clusters of size $n_j = 1$ will rarely occur.

3.2.2 Equal Variances

As discussed in Section 2.3, we can fit the model containing only one variance component by setting $\tau_1 = \tau_2 = \cdots = \tau_N = \tau_0$. In this case, with an inverse-gamma prior (2.6) for τ_0, we replace Step (b) by drawing a value from the conditional distribution of τ_0:

$$(\tau_0^{-1}|\boldsymbol{\mu}, \mathbf{K}, \mathbf{X}) \sim \text{Gamma}\left(a_0 + n/2, b_0 + \sum_{i=1}^{n}(X_i - \mu_{K_i})^2/2\right).$$

Note that the algorithm described in the previous section can be employed if we use a uniform prior for τ_0.

3.2.3 Mixture Reference Distribution H^*

The blocked Gibbs algorithm is easily adjusted to incorporate a mixture reference distribution H^* as described by (2.7) for the equal variance model discussed above. In this case, we replace Step (b) as above and replace Step (a) with a draw for μ_j from H^* for each $j \in \mathbf{K} - \{K_1^*, \ldots, K_m^*\}$, and a draw for μ_j for each $j \in \{K_1^*, \ldots, K_m^*\}$ from the normal mixture density

$$w\, q_{0,j}\phi(\cdot|m_{0,j}, s_{0,j}) + (1-w)\sum_{k=1}^{M} q_{k,j}\widehat{\gamma}_k\phi(\cdot|m_{k,j}, s_{k,j}),$$

where $s_{k,j} = (n_j/\tau_0 + 1/\widehat{\sigma}_k)^{-1}$ and

$$m_{k,j} = s_{k,j}\left(\sum_{\{i:K_i=j\}} X_i/\tau_0 + \widehat{Y}_k/\widehat{\sigma}_k\right), \quad k = 0, \ldots, M,$$

where $\widehat{\sigma}_k = \widehat{\sigma}$ for $k = 1, \ldots, M$. Also,

$$q_{k,j} \propto \sqrt{\frac{s_{k,j}}{\widehat{\sigma}_k}}\exp\left(\frac{m_{k,j}^2}{2s_{k,j}} - \frac{\widehat{Y}_k^2}{2\widehat{\sigma}_k}\right), \quad k = 0, \ldots, M,$$

where these values are subject to the constraint: $wq_{0,j} + (1-w)\sum_{k=1}^{M} q_{k,j}\widehat{\gamma}_k = 1$.

Remark 2. Note that there is no longer a draw for θ given $\boldsymbol{\mu}$ in the blocked Gibbs sampler since we have replaced θ with the point estimate $\widehat{Y}_0$.

4. THE 1872 HIDALGO STAMP ISSUE OF MEXICO

The 1872–1874 Hidalgo postage stamps of Mexico were known to have been printed on different paper types, as was customary for stamps of this era. Izenman and Sommer (1988) tested this assumption extensively by reanalyzing Wilson's (1983) data consisting of the stamp thickness in millimeters of $n = 485$ unwatermarked Hidalgo stamps dating from 1872 through 1874. Applying Silverman's (1981) critical bandwidth test with a normal

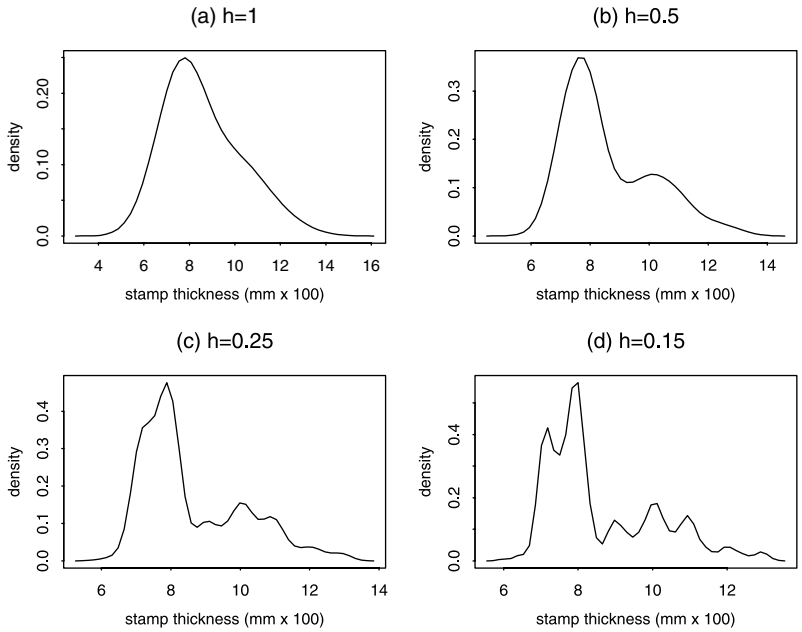

Figure 1. Normal kernel density estimate for stamp thickness using bandwidth values of (a) $h = 1$, (b) $h = 0.5$, (c) $h = 0.25$, (d) $h = 0.15$ using the "density()" function in S-Plus. The value $h = 0.15$ was the critical bandwidth value discovered in Izenman and Sommer (1988).

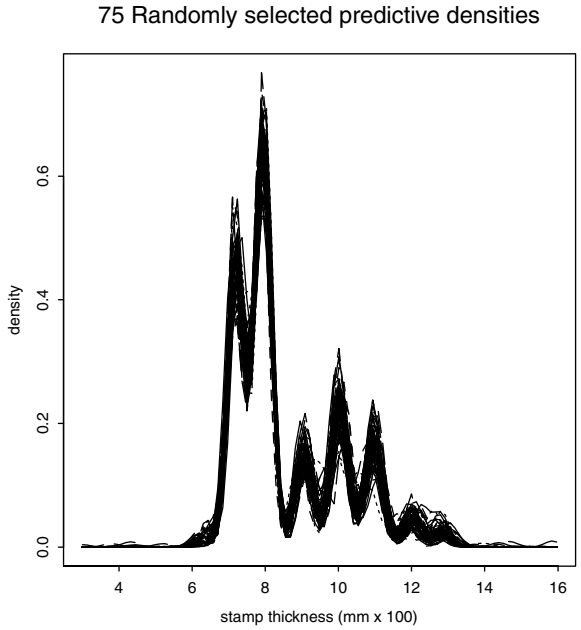

Figure 2. Bayes nonparametric density estimate (3.2) for stamp thickness data. Seventy-five density estimates randomly selected from 3,500 iterations of the blocked Gibbs sampler following a 2,000 iteration burn-in.

kernel, they concluded there were seven modes in the data located at 0.072, 0.080, 0.090, 0.100, 0.110, 0.120, and 0.130 mm, thus supporting the hypothesis that there were at least seven different types of paper used in printing the stamp. Also see Minnotte and Scott (1993) and Efron and Tibshirani (1993, chap. 16) who have also analyzed these data.

A similar analysis to Izenman and Sommer (1988) can be obtained by fitting a finite normal mixture density

$$f_{Q_0}(x) = \int_{\Re} \phi(x|\mu, \tau_0)\, dQ_0(\mu) = \sum_{k=1}^{d} p_{k,0}\, \phi(x|\mu_{k,0}, \tau_0)$$

to the data, where $\sqrt{\tau_0}$ is an unknown bandwidth value and Q_0 is an unknown mixing distribution for the mean. As discussed in Section 2.3 and 3.2.2, there is a simple adjustment to the blocked Gibbs sampler for fitting normal mixtures with fixed variance values.

The smoothed data using a normal kernel density estimate is given in Figure 1 for different bandwidth values. This can be compared to the Bayes nonparameteric density estimate in Figure 2, which seems to support the hypothesis of at least seven distinct modes in the data. The Bayes density estimate was based on 3,500 iterations from the blocked Gibbs sampler following an initial 2,000 iteration burn-in. A Dirichlet process truncation value of $N = 150$ was used for the nonparametric prior and the choice of priors and hyperparameters in the hierarchy for Z_k and the Dirichlet mass parameter α followed the guidelines suggested in Section 2.2. We also used the inverse gamma prior (2.6) for the variance τ_0.

We also ran the blocked Gibbs past the 3,500 iterations (following burn-in) until 25,000 iterations were obtained. Over these 25,000 iterations we computed a penalized MLE subject to a BIC and AIC penalty as outlined in Section 3. The large number of iterations used in determining the Monte Carlo MLE is usually necessary to ensure that the resulting point estimate is at or near the optimal penalized value. The results are presented in Table 1. Under both BIC and AIC the penalized MLE is an eight-point model with mean values not too dissimilar from those observed by Izenman and Sommer (1988).

With the MLE there is a substantial amount of information contained in the draws

Table 1. Stochastic MLE subject to BIC and AIC penalties from the blocked Gibbs sampler using a 2,000 iteration burn-in followed by 25,000 sampled iterations (data values are mm$\times$ 100). Both a noninformative and informative prior for *H* were used.

Noninformative H				*Informative H*			
BIC		*AIC*		*BIC*		AIC	
prob	*atoms*	*prob*	*atoms*	*prob*	*atoms*	*prob*	*atoms*
0.35	7.93	0.36	7.95	0.34	7.95	0.36	7.93
0.27	7.18	0.27	7.20	0.28	7.19	0.27	7.19
0.13	10.02	0.12	10.02	0.14	10.02	0.14	10.02
0.10	10.96	0.11	10.94	0.10	11.00	0.10	10.89
0.10	9.08	0.08	9.07	0.09	9.06	0.08	9.06
0.03	12.03	0.03	12.00	0.03	12.00	0.03	12.00
0.01	12.91	0.02	12.78	0.02	12.92	0.01	12.83
0.01	6.23	0.01	6.38	0.01	6.44	0.005	6.37
						0.005	11.14

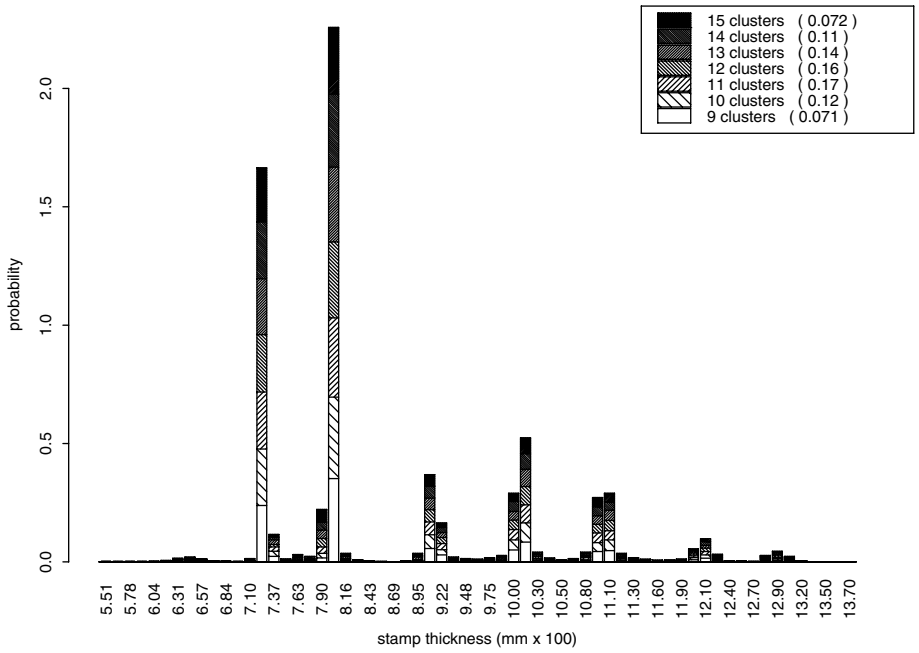

Figure 3. Averaged mixing distribution for the mean indexed by number of clusters and percentage of times clusters occurred during 3,500 iterations following a 2,000 iteration burn-in. Atoms are plotted on the x-axis and corresponding probabilities indexed by cluster on the y-axis. Barplots are stacked; thus the range of values on the y-axis is larger than one.

of $\mathcal{P}_N^*$ that goes unused. In fact, one of the key aspects to using the blocked Gibbs is to devise a way to convert the large amount of posterior information into simple interpretable inference for Q_0. One method that we find quite useful is to draw values from the posterior distribution function for the mean mixing measure,

$$F_{N,\mu}(t) = \mathcal{P}_N([-\infty, t)|\mathbf{X}).$$

Thus, from the output of the blocked Gibbs we estimate $F_{N,\mu}(t)$ with

$$\mathcal{P}_N^*([-\infty, t_l)) = \sum_{k=1}^{N} p_k^* \, \delta_{\mu_k^*}([-\infty, t_l)), \quad l = 1, \ldots, L,$$

where $t_1 < t_2 < \cdots < t_L$ are fixed values that define some refined grid over $\Re$. We then average these values, index them by the number of clusters (i.e., the number of distinct Y_i values for that draw), and then convert the averaged distribution functions into a stacked barplot with t_k values on the x-axis and probabilities on the y-axis. Thus, we convert the cumulative probabilities into probabilities

$$\mathcal{P}_N^*([-\infty, t_{l+1})) - \mathcal{P}_N^*([-\infty, t_l)), \quad l = 1, \ldots, L-1;$$

thus effectively converting the distribution function into a density histogram. When averaging over these values, many of these probabilities become near zero where the posterior has

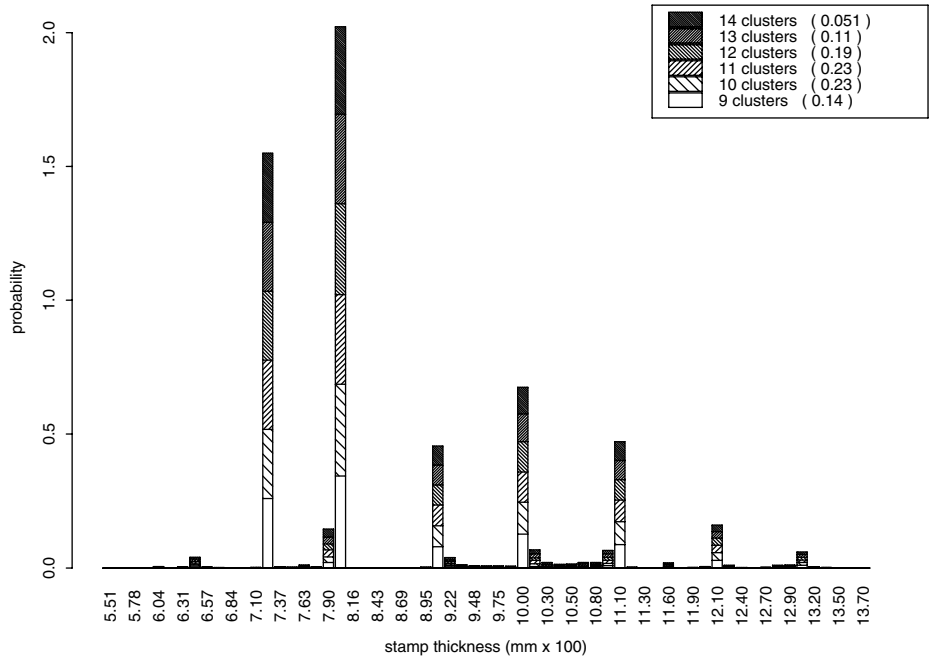

Figure 4. Averaged mixing distribution for the mean using informative reference distribution H^* with distribution H_M weighted by 50%. Analysis based on 3,500 iterations following an initial 2,000 iteration burn-in

little mass so that the corresponding barplot may often appear to have few spikes. Figure 3 presents such a plot based on the information collected from our 3,500 draws. Here we see the presence of at least seven distinct modes with a possible eighth mode appearing in the left tail. This pattern is fairly consistent across the different distributions.

To test the effect on the posterior due to the use of an informative prior, we re-ran the previous analysis using a mixture reference distribution H^* as discussed in Section 2.4 (see Equation (2.7)). In this case, we used a weight of $w = 0.50$, and for our prior guess H_M the $\widehat{Y}_k$ were values selected from the $M = 11$ points

$$\{0.060, 0.064, 0.072, 0.075, 0.080, 0.090, 0.100, 0.110, 0.115, 0.120, 0.130\},$$

which includes the orignal seven modes found in Izenman and Sommer (1988) using Silverman's critical bandwidth test, as well as two additional modes at .060 mm and .064 mm they found using a square root transformation, and also two more modes at .075 mm and .115 mm found by Minnotte and Scott (1993) using a mode tree analysis. We set $\widehat{\gamma}_k = 1/11$ so that each component of H_M is selected with equal probability. For the variance $\widehat{\sigma}$ in H_M, we set $\widehat{\sigma} = 0.001$.

Table 1 and Figure 4 present the results of our analysis, which were derived using the same configurations for the Gibbs sampler as before (to ensure a similar amount of clustering we set $\alpha = 1.7$ to equal the previous posterior mean for α). Figure 4 reveals a mean mixing distribution similar to our previous analysis. However, the use of an informative prior seems to have sharpened the posterior with the modes appearing more defined and the plot more

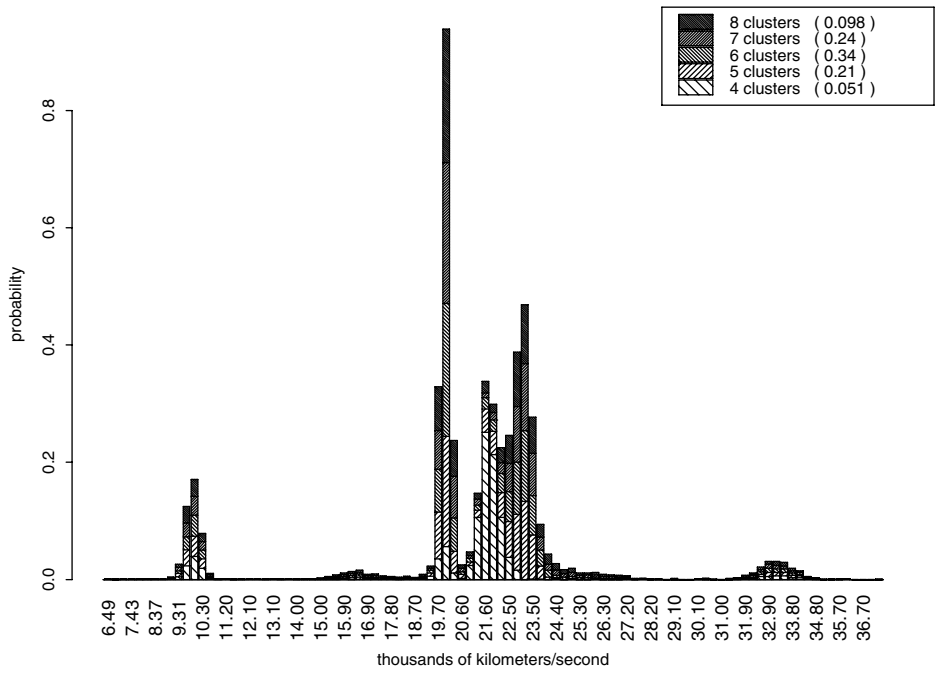

Figure 5. Averaged mixing distribution for the mean from 3,500 sampled posterior distribution functions from the galaxy data.

closely resembling an eight-point discrete distribution. The penalized MLE estimates in Table 1 are also quite similar to the values observed earlier, with BIC finding an eight-point model as before, although now AIC discovers a larger nine-point model. However, the probability for this new atom is quite small.

5. GALAXY DATA

As a second illustration of the method, we reanalyzed the galaxy data in Roeder (1990) which consists of the relative velocities in kilometers per second of $n = 82$ galaxies from six well-separated conic sections of space. As discussed by Roeder (1990), there is strong evidence to believe that modes in the data correspond to clumped galaxies, and that the observed velocities are values derived from a finite mixture of normals. Thus, estimating Q_0 in this problem corresponds to identifying different clustered galaxies.

We applied the blocked Gibbs sampler to this data, following the same strategy used in the previous example of Section 4, although we used a Gamma$(2, 4)$ prior for α in order to facilitate a more even comparison to the results in Escobar and West (1995), who also studied this data but who used a Pólya urn Gibbs sampling algorithm.

Figures 5 and 6 present plots for the averaged posterior distribution functions for the

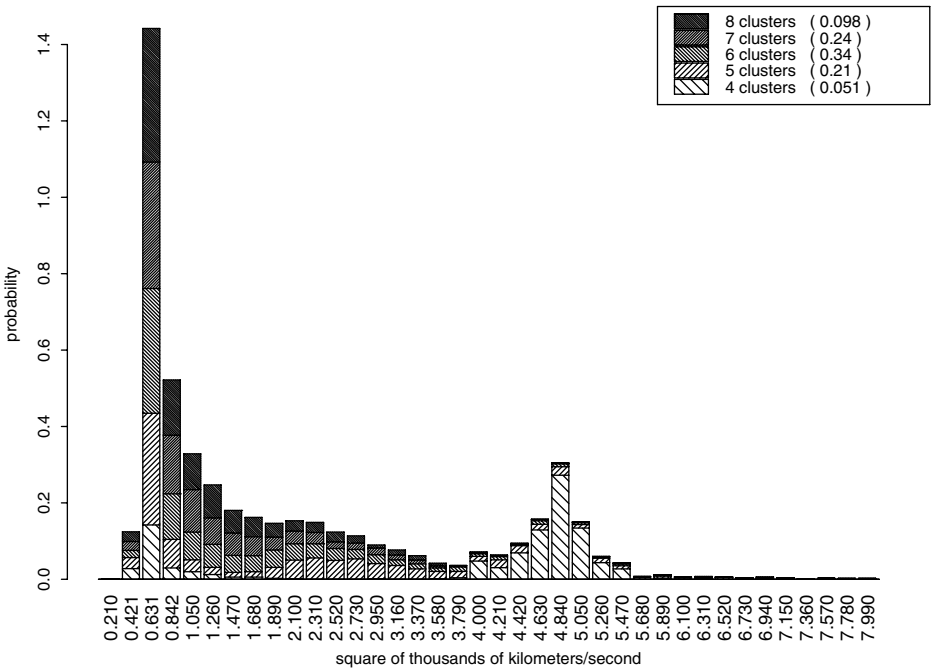

Figure 6. Averaged mixing distribution for the variance computed over the same sampled values used in Figure 5.

mean and variance (marginal) mixing distributions. That is, the marginal distributions,

$$\mathcal{P}_N^*([-\infty, t) \times \Re^+) = \sum_{k=1}^{N} p_k^* \, \delta_{\mu_k^*}([-\infty, t))$$

$$\text{and} \quad \mathcal{P}_N^*(\Re \times [-\infty, t)) = \sum_{k=1}^{N} p_k^* \, \delta_{\tau_k^*}([-\infty, t)),$$

respectively. The barplots are constructed using the same approach discussed in Section 4 (thus, as before, cumulative probabilities have been converted to probabilities). For the mean, we see that the greatest difference in the distribution functions occur in the range of values 20–22 and is mostly due to the difference between the four cluster model and models derived from five to eight clusters. The same sort of effect is also seen for the variance, with the averaged four cluster model exhibiting the greatest bimodal behavior (also see the right-hand side plot of Figure 8). Given the relative infrequency of observing a four cluster model (5.1% of the time), it seems from looking at Figure 5 and the predictive density estimate in Figure 7, that the data contains at least five or six distinct components for the mean.

However, to test how much smoothing of the data is due to our choice of the inverse gamma prior for τ_k, we re-estimated the model using a uniform prior (2.5) for τ_k, where we selected the upper bound $T = 20.83$ corresponding to the variance of the data (expressed here in thousands of kilometers per second). The results are depicted in Figures 9 and 10 and are based on the same configuration for the Gibbs sampler and choice for hyperparameters

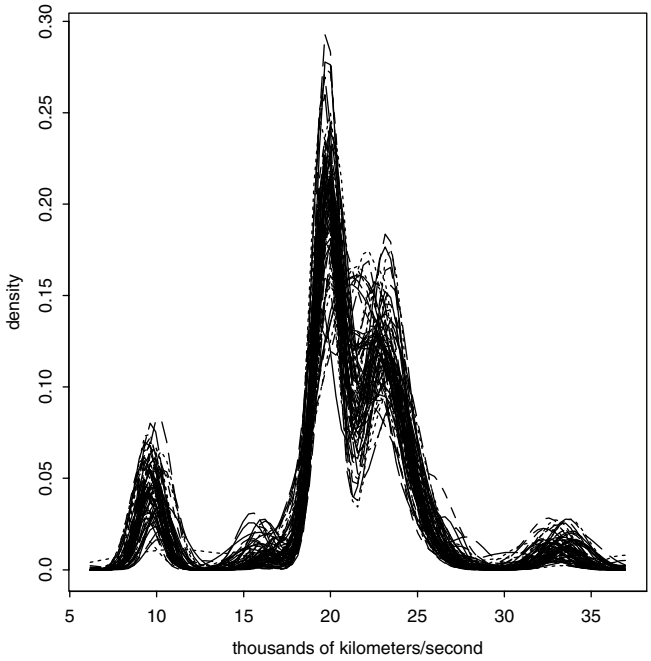

Figure 7. Seventy-five predictive densities (3.2) for the galaxy data selected randomly over 3,500 iterations.

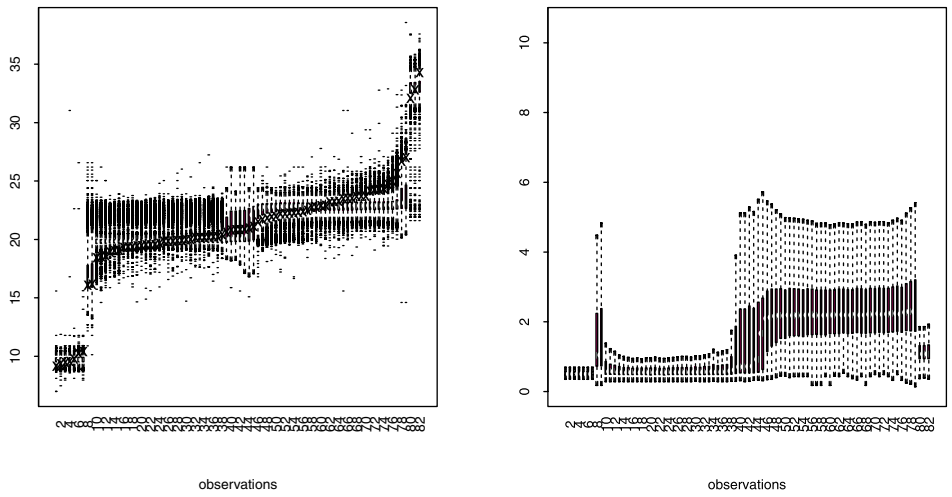

Figure 8. Boxplots for each posterior mean value μ_{K_i} (left-hand side) and variance value τ_{K_i} (right-hand side) for $i = 1,\ldots,n$ from the galaxy data. Observed values of X_i are indicated by an x (left-hand plot).

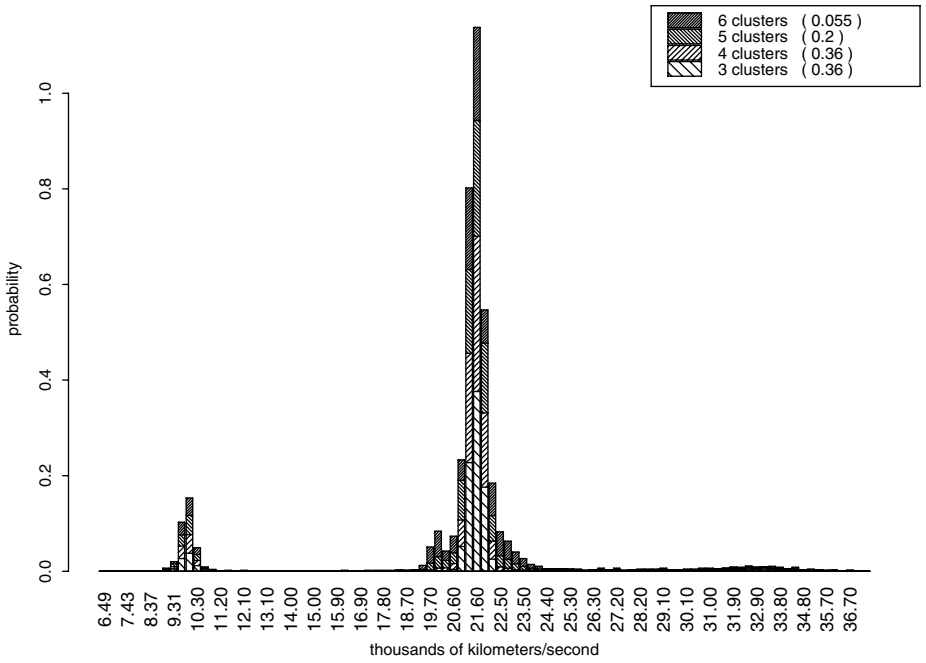

Figure 9. Averaged mixing distribution for the mean from the galaxy data obtained with a uniform prior for τ_k.

as before. With a uniform prior, we find that there is less smoothing, with the posterior concentrating predominately on three and four cluster models now (each occurring 36% of the time). With fewer clusters, we also clearly see a bimodal posterior distribution for the variance, with the values in the right-tail causing the averaged posterior distributions for the mean to have a large cluster at the values around 20–22. The same results were observed for smaller values of T, such as $T = 10$. This analysis shows us some of the hidden dangers with working with an inverse gamma prior, which can sometimes act as an informative prior, either undersmoothing, or oversmoothing the data as seen here.

6. DISCUSSION

Sections 4 and 5 have presented finite normal mixture examples that show the range of inference possible when working with a Dirichlet process approximation in a Gibbs sampling setting. In particular, we illustrated how to use sampled values from the posterior of the random measure $\mathcal{P}_N$ to graphically study the posterior mixing distribution, while at the same time we have also demonstrated how the method can be used for the more traditional analysis of the hidden Y_i variables.

The proposed Gibbs sampler mixes well due to its method for blocking the nonparametric parameters $\boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{K}$, and $\mathbf{p}$. Each of these parameters are updated simultaneously in a multivariate step, which encourages good mixing for these parameters, and thus good mixing for the unobserved Y_i values. The blocked Gibbs sampler is relatively simple to

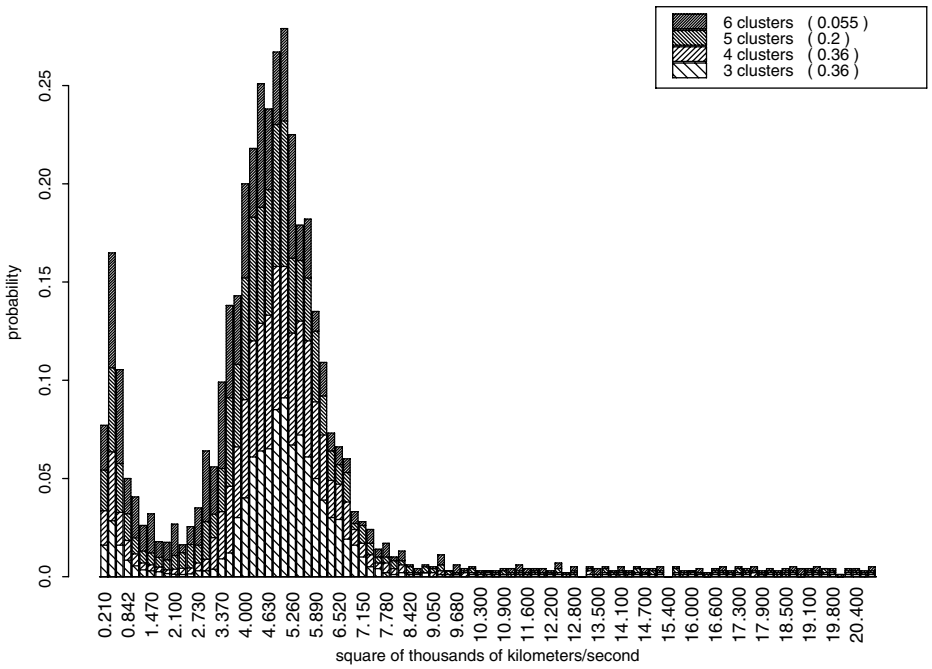

Figure 10. Averaged mixing distribution for the variance distribution based on a uniform prior for τ_k. Based on same sampled distribution functions used in deriving Figure 9.

program and requires only the ability to draw values from simple conditional distributions, including the important update for the Dirichlet mass parameter α. The algorithm is also easily extended to the nonconjugate setting involving a flat uniform prior for the variances τ_k. This extension can sometimes be important as we saw in Section 5, where the use of an inverse gamma prior for the variance appeared to oversmooth the data and to overestimate the number of mixture components. Selecting the upper bound T in the uniform prior is fairly automatic and avoids the difficulty of trying to correctly scale the data when using an inverse gamma prior.

We also explored the use of a Bayesian Monte Carlo penalized MLE in our example of Section 4. Although producing this estimate for Q_0 requires a much larger number of iterations then estimates based on averaged distribution functions, it has the advantage that it conveys an easily interpretable summary result of the analysis. We have also indicated a method for incorporating data dependent prior information that can include for example frequentist point estimates for Q_0. We saw in Section 4 that an informative prior can sometimes help to "sharpen" our estimates, but in general we believe that such priors should be used cautiously.

APPENDIX: PROOFS

Proof of Theorem 1: By integrating over P we can write m_N and m_∞ in terms of

the distributions for $\mathbf{Y} = (Y_1, \ldots, Y_n)$ under $\mathcal{P}_N$ and DP(αH) respectively. Call these two sampling distributions $\pi_N(d\mathbf{Y})$ and $\pi_\infty(d\mathbf{Y})$. Thus,

$$\begin{aligned}
&\int |m_N(\mathbf{X}) - m_\infty(\mathbf{X})|\, d\mathbf{X} \\
&\quad = \int \left| \int \prod_{i=1}^{n} \phi(X_i|\mu(Y_i), \tau(Y_i)) \left(\pi_N(d\mathbf{Y}) - \pi_\infty(d\mathbf{Y})\right) \right| d\mathbf{X} \\
&\quad \leq \int \int \prod_{i=1}^{n} \phi(X_i|\mu(Y_i), \tau(Y_i))\, d\mathbf{X}\ |\pi_N(d\mathbf{Y}) - \pi_\infty(d\mathbf{Y})| \\
&\quad = 2D(\pi_N, \pi_\infty),
\end{aligned}$$

where $D(\mathbb{P}_1, \mathbb{P}_2) = \sup_A |\mathbb{P}_1(A) - \mathbb{P}_2(A)|$ is the total variation distance between two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$.

Recall that we can write $Y_i = Z_{K_i}$. The sampled values $\mathbf{Y}$ under π_N and π_∞ are identical when K_i is sampled from a value smaller than the Nth term. Thus,

$$\begin{aligned}
D(\pi_N, \pi_\infty) &\leq 2\left(1 - \pi_N\{K_i < N, \text{ for } \quad i = 1, \ldots, n\}\right) \\
&= 2\left[1 - I\!E\left\{\left(\sum_{k=1}^{N-1} p_k\right)^n\right\}\right] \approx 2n \exp(-(N-1)/\alpha), \qquad \text{(A.1)}
\end{aligned}$$

where the right most approximation follows by observing that

$$\begin{aligned}
\sum_{k=1}^{N-1} p_k &= 1 - (1 - V_1)(1 - V_2)\ldots(1 - V_{N-1}) \\
&\stackrel{\mathcal{D}}{=} 1 - \exp(-E_1/\alpha)\exp(-E_2/\alpha)\ldots\exp(-E_{N-1}/\alpha) \\
&\approx 1 - \exp(-(N-1)/\alpha),
\end{aligned}$$

where $E_1, \ldots, E_{N-1}$ are iid exp(1) random variables. □

Proof of Corollary 1: Write $\mathcal{K}_m$ for the set $\{1, 2, \ldots, m\}^n$ for $m = 1, 2, \ldots$. We have,

$$\begin{aligned}
&\sum_{\mathbf{K} \in \mathcal{K}_\infty} |\pi_N(\mathbf{K}|\mathbf{X}) - \pi_\infty(\mathbf{K}|\mathbf{X})| \\
&\qquad = \sum_{\mathbf{K} \in \mathcal{K}_N} |\pi_N(\mathbf{K}|\mathbf{X}) - \pi_\infty(\mathbf{K}|\mathbf{X})| + \sum_{\mathbf{K} \in \mathcal{K}_\infty - \mathcal{K}_N} \pi_\infty(\mathbf{K}|\mathbf{X}).
\end{aligned} \qquad \text{(A.2)}$$

Consider the first sum on the right-hand side of (A.2). Note that

$$\pi_N(\mathbf{K}|\mathbf{X}) = \frac{\Pr_N(\mathbf{K}) m_N(\mathbf{X}|\mathbf{K})}{m_N(\mathbf{X})}, \quad \mathbf{K} \in \mathcal{K}_N,$$

where $\Pr_N(\mathbf{K})$ is the prior for $\mathbf{K}$ under $\mathcal{P}_N$ and

$$m_N(\mathbf{X}|\mathbf{K}) = \prod_{j \in \mathbf{K}^*} \int_{\Re \times \Re^+} H(dY) \prod_{\{i: K_i = j\}} \phi(X_i|\mu(Y), \tau(Y)),$$

where $\mathbf{K}^*$ denotes the set of unique K_i values. It is clear that $m_N(\mathbf{X}|\mathbf{K}) = m_\infty(\mathbf{X}|\mathbf{K})$ for each $\mathbf{K} \in \mathcal{K}_N$. Moreover, it is also clear that $\Pr_N(\mathbf{K}) = \Pr_\infty(\mathbf{K})$ for each $\mathbf{K} \in \mathcal{K}_{N-1}$ where $\Pr_\infty(\mathbf{K})$ is the prior under the Dirichlet process. However, the priors for $\mathbf{K}$ are not necessarily equal over $\mathcal{K}_N - \mathcal{K}_{N-1}$. Thus writing $\Pr_\infty(\mathbf{K})$ as $\Pr_N(\mathbf{K}) + [\Pr_\infty(\mathbf{K}) - \Pr_N(\mathbf{K})]$ we have,

$$\begin{aligned}\sum_{\mathbf{K}\in\mathcal{K}_N} |\pi_N(\mathbf{K}|\mathbf{X}) - \pi_\infty(\mathbf{K}|\mathbf{X})| \quad &\leq \quad \left|1 - \frac{m_N(\mathbf{X})}{m_\infty(\mathbf{X})}\right| \sum_{\mathbf{K}\in\mathcal{K}_N} \frac{\Pr_N(\mathbf{K}) m_N(\mathbf{X}|\mathbf{K})}{m_N(\mathbf{X})} \\ &\quad + \sum_{\mathbf{K}\in\mathcal{K}_N - \mathcal{K}_{N-1}} \frac{m_N(\mathbf{X}|\mathbf{K})}{m_\infty(\mathbf{X})} \, |\Pr_\infty(\mathbf{K}) - \Pr_N(\mathbf{K})|.\end{aligned}$$

The first sum on the right-hand side is $\pi_N(\mathcal{K}_N|\mathbf{X})$ which is bounded by one. Thus, integrating with respect to $m_\infty(\mathbf{X})$, the right-hand side is bounded by

$$\int_{\Re^n} |m_N(\mathbf{X}) - m_\infty(\mathbf{X})| \, d\mathbf{X} + \sum_{\mathbf{K}\in\mathcal{K}_N - \mathcal{K}_{N-1}} |\Pr_\infty(\mathbf{K}) - \Pr_N(\mathbf{K})|.$$

Both terms are order $O(n \exp(-(N-1)/\alpha))$. The first term by using Theorem 1, and the second term by using a similar argument as in (A.1) used in the proof of Theorem 1.

Finally, integrating the second sum on the right-hand side of (A.2) with respect to $m_\infty(\mathbf{X})$ gives $\Pr_\infty(\mathcal{K}_\infty - \mathcal{K}_N)$ which is order $O(n \exp(-(N-1)/\alpha))$ by using the same argument as in (A.1). □

ACKNOWLEDGMENTS

The authors are grateful to Yinsheng Qu for helpful discussion on earlier drafts of this article. We also thank two referees for advice that greatly improved the overall presentation. This work was supported in part by the Acheson J. Duncan Fund for the Advancement of Research in Statistics, Award #00-1, Department of Mathematical Sciences, Johns Hopkins University.

[Received June 2000. Revised May 2001.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, eds. B.N. Petrov and F. Csaki, Budapest: Akademiai Kiado, pp. 267–281.

Bush, C.A., and MacEachern, S. N (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275–285.

Chen, J. (1995), "Optimal Rate of Convergence for Finite Mixture Models," *The Annals of Statistics*, 23, 221–233.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

Devroye, L. (1986), *Non-Uniform Random Variate Generation*, New York: Springer-Verlag.

Diebolt, J., and Robert, C. P. (1994), "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society*, Series B, 56, 363–375.

Donnelly, P., and Joyce, P. (1989), "Continuity and Weak Convergence of Ranked and Size-Biased Permutations on the Infinite Simplex," *Stochastic Processes and Their Applications*, 31, 89–103.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Escobar, M. D. (1988), "Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means," Unpublished dissertation, Yale University.

——— (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statisticial Association*, 89, 268–277.

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

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

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

——— (1983), "Bayesian Density Estimation by Mixtures of Normal Distributions," in *Recent Advances in Statistics*, eds. M. H. Rizvi, J. Rustagi, and D. Siegmund, New York: Academic Press, pp. 287–302.

Ishwaran, H., and James, L. F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Ishwaran H., James, L. F., and Sun, J. (2001), "Bayesian Model Selection in Finite Mixtures by Marginal Density Decompositions," conditionally accepted by *Journal of the American Statistical Association*.

Ishwaran, H., and Zarepour, M. (2000), "Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models," *Biometrika*, 87, 371–390

Izenman, A. J., and Sommer, C. J. (1988), "Philatelic Mixtures and Multimodal Densities," *Journal of the American Statistical Association*, 83, 941–953.

Kuo, L. (1986), "Computations of Mixtures of Dirichlet Processes," *SIAM Journal of Scientific and Statistical Computing*, 7, 60–71.

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

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

MacEachern, S. N., and Müller, P. (1998), "Estimating Mixtures of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 2, 223–238.

McCloskey, J. W. (1965), "A Model for the Distribution of Individuals by Species in an Environment," Unpublished Ph.D. thesis, Michigan State University.

Minnotte, M. C., and Scott, D. W. (1993), "The Mode Tree: A Tool for Visualization of Nonparametric Density Features," *Journal of Computational and Graphical Statistics*, 2, 51–68.

Muliere, P., and Tardella, L. (1998), "Approximating Distributions of Random Functionals of Ferguson–Dirichlet Priors," *Canadian Journal of Statistics*, 26, 283–297.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian Curve Fitting Using Multivariate Normal Mixtures," *Biometrika*, 83, 67–79.

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Perman, M., Pitman, J., and Yor, M. (1992), "Size-Biased Sampling of Poisson Point Processes and Excursions," *Probability Theory and Related Fields*, 92, 21–39.

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

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.

Roeder, K., and Wasserman, L. (1997), "Practical Bayesian Density Estimation Using Mixtures of Normals," *Journal of the American Statistical Association*, 92, 894–902.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Sethuraman, J., and Tiwari, R. C. (1982), "Convergence of Dirichlet Measures and the Interpretation of Their Parameters," *Statistical Decision Theory and Related Topics III*, 2, 305–315.

Silverman, B. W. (1981), "Using kKernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society*, Series B, 43, 97–99.

Walker, S., and Damien, P. (1998), "Sampling Methods for Bayesian Nonparametric Inference Involving Stochastic Processes," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Mueller, and D. Sinha, New York: Springer Lecture Notes, pp. 243–254.

West, M., Müller, P., and Escobar, M. D. (1994), "Hierarchical Priors and Mixture Models, With Applications in Regression and Density Estimation," in *A Tribute to D. V. Lindley*, eds. A.F.M Smith and P.R. Freeman, New York: Wiley.

Wilson, I. G. (1983), "Add a New Dimension to your Philately," *The American Philatelist*, 97, 342–349.