Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models

Sanjib BASU and Siddhartha CHIB

We present a method for comparing semiparametric Bayesian models, constructed under the Dirichlet process mixture (DPM) framework, with alternative semiparameteric or parameteric Bayesian models. A distinctive feature of the method is that it can be applied to semiparametric models containing covariates and hierarchical prior structures, and is apparently the first method of its kind. Formally, the method is based on the marginal likelihood estimation approach of Chib (1995) and requires estimation of the likelihood and posterior ordinates of the DPM model at a single high-density point. An interesting computation is involved in the estimation of the likelihood ordinate, which is devised via collapsed sequential importance sampling. Extensive experiments with synthetic and real data involving semiparametric binary data regression models and hierarchical longitudinal mixed-effects models are used to illustrate the implementation, performance, and applicability of the method.

KEY WORDS: Bayesian model comparison; Bayes factor; Dirichlet process mixture; Marginal likelihood; Semiparametric binary data model; Semiparametric longitudinal data model.

1. INTRODUCTION

Advances in Markov chain Monte Carlo (MCMC) simulation methods have facilitated the study of Bayesian models under far weaker and more realistic assumptions than was previously possible. As a result of these developments, semiparametric Bayesian modeling has become a practical option, and under the Dirichlet process mixture (DPM) framework, for example, novel and appealing statistical models can be formulated and estimated. It turns out, however, that although the methodology for fitting DPM models is more or less established, there is a paucity of work on methods that can be used to compare these models with competing semiparametric or parametric models.

The general problem of comparing semiparametric models with other alternative model specifications has a relatively recent history. Florens, Richard, and Rolin (1996) and Carota and Parmigiani (1996) developed procedures to compare parametric models with nonparametric alternatives modeled with Dirichlet process (DP) and mixture of Dirichlet processes (MDP). It is important to bear in mind that, despite the similarity in nomenclature, the MDP specification is rather different from the DPM model that we consider in the sequel. In more recent work, Berger and Guglielmi (2001) modeled the nonparametric alternative by a Pólya tree process and computed the Bayes factor for a default reference prior, and Ishwaran, James, and Sun (2001) compared models with different number of unique mixture components by subsuming the models within a finite mixture model.

Significantly, earlier work on this general topic (except for that in Ishwaran et al. 2001) is concerned with nonparametric instead of semiparametric models, because the unknown distribution of the observations is modeled directly by a nonparametric prior process. Even more importantly, all available methods explicitly assume an independent and identically distributed model for the data, which rules out models that contain covariates. Therefore, none of the existing approaches can be used to compare the fit of flexible semiparameteric regression models of the type discussed by, for example, Bush and MacEachern (1996), Kleinman and Ibrahim (1998), and Basu and Mukhopadhyay (2000).

One purpose of this article is to introduce a Bayesian model comparison method for semiparametric models that can be applied even when the model contains covariates and (possibly) an involved hierarchical prior structure. The method that we devise for finding the "weight of evidence" from the marginal likelihood of the semiparametric model is apparently the first to be proposed for Bayesian semiparametric regression models. In this method, which relies on the framework of Chib (1995), the rather difficult computation of the marginal likelihood (which entails integration of the likelihood with respect to the prior density of the parameters) is reduced to the more tractable problems of finding estimates of the likelihood and of the posterior at a single point. Crucially, both of these quantities are readily available. To estimate the posterior ordinate, we need primarily the MCMC procedures that simulate the parameters from the posterior distribution, whereas, to estimate the likelihood ordinate, we develop an interesting, low-variability method based on collapsed sequential importance sampling (SIS). Collapsed SIS is a variant of the SIS method introduced by Kong, Liu, and Wong (1994), and Liu (1996, 2001). This variant was discussed by Lo, Brunner, and Chan (1996), Ishwaran and James (2001b), Ishwaran et al. (2001), and Ishwaran and Takahara (2002) in the general context of weighted Chinese restaurant processes and by Quintana (1998), MacEachern, Clyde, and Liu (1999), and Quintana and Newton (2000) in the setting of categorical DPM models.

The article is organized as follows. In Section 2 we present the DPM model and the model comparison problem of interest. In Section 3 we discuss computation of the marginal likelihood of the DPM model with a view to computing the Bayes

Sanjib Basu is Associate Professor, Division of Statistics, Northern Illinois University, Dekalb, IL 60115 (E-mail: *basu@math.niu.edu*). Siddhartha Chib is Harry C. Hartkopf Professor of Econometrics and Statistics, John M. Olin School of Business, Washington University, St. Louis MO 63130 (E-mail: *chib@olin.wustl.edu*). The authors are grateful to the editor, associate editor, and referees for constructive and valuable comments.

Journal of the American Statistical Association
March 2003, Vol. 98, No. 461, Theory and Methods
DOI 10.1198/01621450338861947

factor of alternative parametric and semiparametric models. In Sections 4 and 5 we delineate the specifics of the method. In Section 6 we provide applications of the method to three examples, each of which contains covariates. We also consider two synthetic datasets to illustrate the usefulness of Bayes factors for finding the true model. We give concluding remarks in Section 7.

2. DIRICHLET PROCESS MIXTURE MODEL

Let $\mathbf{y}_i (i \le n)$ denote a collection of scalar or vector-valued independent observations whose distribution is modeled by a general DPM model described as

$$\begin{aligned}
&\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\phi}, x_i \sim f(\cdot|\boldsymbol{\theta}_i, \boldsymbol{\phi}, x_i), \qquad i = 1, \ldots, n;\\
&\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|G \overset{iid}{\sim} G;\\
&G|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0(\cdot|\boldsymbol{\kappa}));\\
&\boldsymbol{\psi} = (\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha) \sim \pi,
\end{aligned} \tag{1}$$

where x_i are fixed covariates, $\boldsymbol{\phi}$ is a vector parameter associated with the distribution of $\mathbf{y}_i$, $\{\boldsymbol{\theta}_i\}$ are latent or subject-specific *random vectors* that are conditionally independent given the distribution G, and $\{f(\cdot|\boldsymbol{\theta}_i, \boldsymbol{\phi}, x_i)\}$ is a parametric family of densities with respect to a dominating measure μ. Given G, therefore, the density of the data $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$ with regard to μ (and suppressing dependence on the covariates) is given by the mixture

$$f(\mathbf{y}|\boldsymbol{\phi}, G) = \prod_{i=1}^{n} \int f(\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\phi}) dG(\boldsymbol{\theta}_i). \tag{2}$$

The key feature of the model is the assumption that the distribution G is unknown and is modeled by a DP prior (Ferguson 1973) with concentration parameter α and specified base probability measure $G_0(\cdot|\boldsymbol{\kappa})$ that depends on an unknown parameter vector $\boldsymbol{\kappa}$. Here G and G_0 denote probability measures, although we often refer to them as distributions. The Bayesian model is completed by assuming that the parameter vector $\boldsymbol{\phi}$, the hyperparameter vector $\boldsymbol{\kappa}$ of G_0, and the concentration parameter α follow a parametric distribution π.

The DPM model was introduced by Ferguson (1983) and Lo (1984). Kuo (1986) first described Monte Carlo techniques for fitting these models by sampling from the prior. The clever trick of exploiting the Blackwell and MacQueen (1973) Pólya urn characterization of the DP [see (4) and (5)] within a Markov chain sampling setting was elucidated by Escobar (1988, 1994), and Escobar and West (1995). The collapsed cluster sampling method of MacEachern (1994) and the "no-gaps" algorithm of MacEachern and Müller (1998) for nonconjugate DPM models also use the Pólya urn structure.

An important point is that the foregoing setup and the method developed later can be extended to any prior process that follows a generalized Pólya urn scheme. In particular, if $\mathcal{P}(\cdot|\alpha, G_0(\cdot|\boldsymbol{\kappa}))$ denotes such a prior process with hyperparameters α and $G_0(\cdot|\boldsymbol{\kappa})$ and

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|G \overset{\text{iid}}{\sim} G, \qquad G \sim \mathcal{P}(\cdot|\alpha, G_0(\cdot|\boldsymbol{\kappa})) \tag{3}$$

is a sample from this prior process, then the prequential prediction rule of $\boldsymbol{\theta}_i$ is given by

$$\begin{aligned}
P(\boldsymbol{\theta}_i \in \cdot|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}) &= q_{k_{i-1}+1,\,i}(\alpha) G_0(\cdot|\boldsymbol{\kappa})\\
&\quad + \sum_{j=1}^{k_{i-1}} q_{j,\,i}(\alpha) \delta_{\boldsymbol{\theta}^*_{j,\,i-1}}(\cdot), \qquad i \le n,
\end{aligned} \tag{4}$$

where $\delta_{\boldsymbol{\theta}}(\cdot)$ denotes the degenerate measure at $\boldsymbol{\theta}$, $\{\boldsymbol{\theta}^*_{1,\,i-1}, \ldots, \boldsymbol{\theta}^*_{k_{i-1},\,i-1}\}$ are the set of k_{i-1} unique values in $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}\}$, $\{q_{j,\,i}(\alpha)\}_{j=1}^{k_{i-1}+1}$ are the probabilities of the different components (that may functionally depend on α) adding to 1; for the case where $i = 1$, vacuous sets and sums are treated as empty. The DP prior is of course the most popular class of priors with a Pólya urn representation. If $\mathcal{P}(\cdot|\alpha, G_0(\cdot|\boldsymbol{\kappa})) = \mathrm{DP}(\cdot|\alpha, G_0(\cdot|\boldsymbol{\kappa}))$ with concentration parameter α and base measure $G_0(\cdot|\boldsymbol{\kappa})$, then (4) holds with

$$\begin{aligned}
q_{j,\,i}(\alpha) &= \frac{n_{j,\,i-1}}{\alpha + i - 1}, \qquad j = 1, \ldots, k_{i-1},\\
q_{k_{i-1}+1,\,i}(\alpha) &= \frac{\alpha}{\alpha + i - 1},
\end{aligned} \tag{5}$$

where $n_{j,\,i-1}$ denotes the frequency of the unique label $\boldsymbol{\theta}^*_{j,\,i-1}$ among $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}\}$, $j = 1, \ldots, k_{i-1}$.

Another point is that the model described in (1) can alternatively be expressed in terms of the "stick-breaking" construction of the DP as given by Sethuraman (1994),

$$\begin{aligned}
&G(\cdot) = \sum_{l=1}^{\infty} p_l \delta_{Z_l}(\cdot), \quad \text{where } Z_l \overset{\text{iid}}{\sim} G_0(\cdot|\kappa), l = 1, \ldots, \text{ and}\\
&p_1 = V_1, p_l = V_l \prod_{j=1}^{l-1} (1 - V_j), l = 2, \ldots,\\
&\qquad \text{with } V_l \overset{\text{iid}}{\sim} \mathrm{beta}(1, \alpha), l = 1, \ldots
\end{aligned} \tag{6}$$

If the sum in (6) is truncated at a large integer N, we obtain the finite-dimensional prior considered by Ishwaran and James (2001a), who developed a blocked Gibbs sampler for the model under this prior by updating blocks of parameters in multivariate steps instead of the one-at-a-time updates that appear in the Pólya urn–based samplers.

Finally, we note that the approach that we develop can also be applied to the two-parameter Poisson–Dirichlet process discussed by Ishwaran and James (2001a), which includes the DP as a special case.

3. MODEL COMPARISON PROBLEM

Suppose that we are given a collection of models $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$, where one (or more) of the models is a DPM model, and the objective is to compare the different models given the data $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$. The formal Bayesian approach for doing this comparison is via the pairwise Bayes

factors, defined for any two models $\mathcal{M}_r$ and $\mathcal{M}_s$ by the ratio of marginal likelihoods

$$B_{rs} = \frac{m(\mathbf{y}|\mathcal{M}_r)}{m(\mathbf{y}|\mathcal{M}_s)}.$$

In the semiparametric DPM context, calculation of the marginal likelihood is a largely unexplored problem. In fact, the problem in this case is somewhat deeper, because even the computation of the likelihood function of the DPM model (an input into the marginal likelihood) has not been satisfactorily tackled in the literature. Specifically, if we let $\mathcal{P}(\cdot|\alpha, G_0, \boldsymbol{\kappa})$ denote the DP measure, then the likelihood $L(\mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha, G_0)$ of the DPM model (on suppressing the model index) is given by

$$L(\mathbf{y}|\phi, \boldsymbol{\kappa}, \alpha, G_0) = \int f(\mathbf{y}|\boldsymbol{\phi}, G) d\mathcal{P}(G|\alpha, G_0, \boldsymbol{\kappa}), \quad (7)$$

which requires an integration over the space of the infinite-dimensional parameter G. Additionally, let $\pi(\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha)$ denote the prior density of the parameters. Then the marginal likelihood is obtained by integrating the likelihood function over the prior distribution of the parameters,

$$\begin{aligned} m(\mathbf{y}) &= \int L(\mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha, G_0)\pi(\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha)\, d\boldsymbol{\phi} d\boldsymbol{\kappa} d\alpha \\ &= \iint f(\mathbf{y}|\boldsymbol{\phi}, G)\, d\mathcal{P}(G|\alpha, G_0, \boldsymbol{\kappa})\pi(\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha)\, d\boldsymbol{\phi} d\boldsymbol{\kappa} d\alpha \\ &= \iint \left\{ \prod_{i=1}^{n} \int f(\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\phi})\, dG(\boldsymbol{\theta}_i) \right\} \\ &\quad \times d\mathcal{P}(G|\alpha, G_0, \boldsymbol{\kappa})\pi(\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha)\, d\boldsymbol{\phi} d\boldsymbol{\kappa} d\alpha \end{aligned} \quad (8)$$

where the last step uses (2). Clearly, direct evaluation of these integrals is impossible. Therefore, a feasible approach to this problem must tackle the problem by different means.

In this article, we focus on the approach of Chib (1995), which is based on a representation of the marginal likelihood that is amenable to calculation by MCMC methods. Because the marginal likelihood is the normalizing constant of the posterior density, one can write

$$m(\mathbf{y}) = \frac{L(\mathbf{y}|\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*, G_0)\pi(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*)}{\pi(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*|\mathbf{y})},$$

where $(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*)$ is some point in the parameter space, $\pi(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*)$ is the prior density at that point, and $\pi(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*|\mathbf{y})$ is the posterior density of the parameters also evaluated at that same point. None of the quantities in this expression is conditioned on an estimate of the unknown distribution G, because otherwise the efficiency of the estimate would be severely compromised. Now, if we let $\widehat{L}(\mathbf{y}|\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*, G_0)$ and $\hat{\pi}(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*|\mathbf{y}, G_0)$ denote estimates of the likelihood and posterior ordinates (methods for finding these estimates are given later), it follows that we can conveniently estimate the marginal likelihood as

$$\begin{aligned} \log \hat{m}(\mathbf{y}) = \log \widehat{L}(\mathbf{y}|\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*, G_0) + \log \pi(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*) \\ - \log \hat{\pi}(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*|\mathbf{y}, G_0). \end{aligned} \quad (9)$$

Han and Carlin (2001) recently reported that the marginal likelihood estimates from the Chib approach are quite accurate compared with those from other methods. Now, given the marginal likelihood estimates for any two models $\mathcal{M}_r$ and $\mathcal{M}_s$, the Bayes factor is available as

$$\widehat{B}_{rs} = \exp\{\log \hat{m}(\mathbf{y}|\mathcal{M}_r) - \log \hat{m}(\mathbf{y}|\mathcal{M}_s)\}.$$

By way of interpretation, if the two models $\mathcal{M}_r$ and $\mathcal{M}_s$ are equally probable a priori, then the Bayes factor B_{rs} is the posterior odds in favor of the model $\mathcal{M}_r$. Alternatively, the Bayes factor can also be viewed as the relative success of the two models at predicting the data $\mathbf{y}$. Good (1985) has referred to the log of the Bayes factor as the "weight of evidence." According to the famous scale of Jeffreys, a log (base e) Bayes factor values in the range of $(0, 1.15)$, $(1.15, 3.45)$, $(3.45, 4.60)$, and $(4.60, \infty)$ provide "not worth a mention," "substantial," "strong," and "very strong" evidence against the $\mathcal{M}_s$ model.

An important practical consequence of devolving the marginal likelihood computation in the foregoing manner is that the problem is reduced to one of finding estimates of the likelihood and posterior ordinates. These two problems can be tackled quite effectively by separate means. Indeed, computation of the posterior ordinate is based on the output produced by the MCMC simulation algorithms currently used to estimate DPM models. Thus this step requires almost no additional programming beyond what is needed to fit the DPM model. On the other hand, computation of the likelihood ordinate requires additional computation, but the burden is not large. The method that we have developed is based on sequential importance sampling (Kong et al. 1994; Liu 1996, 2001; Lo et al. 1996; MacEachern et al. 1999; Ishwaran and James 2001b; Ishwaran et al. 2001; Ishwaran and Takahara 2002). A variant of our method is available that can be applied to the case in which the sampling density and G_0 are nonconjugate.

We conclude this section with several remarks. If the DPM model is to be compared against a suitably embedded parametric alternative, then the marginal likelihood of the parametric model can be computed by available methods (Chib 1995; Chib and Jeliazkov 2001). The ratio of the two marginal likelihoods then provides the Bayes factor for the parametric versus semiparametric model. Of course, if the alternative model is a different DPM model, then its marginal likelihood can be computed by the method developed here. Thus, with the method at hand, we can find the Bayes factor for comparing the DPM model against both parametric and semiparametric alternatives.

Finally, for appropriate model comparisons, it is desirable, if possible, to match the prior specifications in the two models, at least for similar parameters (see Berger and Guglielmi 2001 for further discussion). If we are comparing a DPM model with another nonparametric model, then this issue needs to be taken up on a case-by-case basis, as shown in our examples. When the alternative is a parametric model, however, proper embedding should allow the DPM model to be viewed as a generalization of the parametric model. Here we follow Florens et al. (1996), who recommended that the two models be specified in such a way that the predictive (or marginal)

distribution of a single observation is identical under the two models (i.e., the two models cannot be distinguished by just one observation). In that case, the relevant parametric alternative to the DPM model is given by

$$\begin{aligned}
&\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\phi} \sim f(\cdot|\boldsymbol{\theta}_i, \boldsymbol{\phi}), \qquad i = 1, \ldots, n;\\
&\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|G_0 \overset{iid}{\sim} G_0(\cdot|\boldsymbol{\kappa});\\
&(\boldsymbol{\phi}, \boldsymbol{\kappa}) \sim \pi(\boldsymbol{\phi}, \boldsymbol{\kappa}) = \int \pi(\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha)\, d\alpha, \qquad (10)
\end{aligned}$$

where $\pi(\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha)$ is the joint prior under the DPM model in (1). The predictive distribution of a single observation $\mathbf{y}_i$ under either the DPM model in (1) or the parametric model in (10) is then identical and is given by $\int_{\boldsymbol{\phi}, \boldsymbol{\kappa}} f(\mathbf{y}_i|\theta, \boldsymbol{\phi}) dG_0(\theta|\boldsymbol{\kappa}) d\pi(\boldsymbol{\phi}, \boldsymbol{\kappa})$.

4. POSTERIOR ORDINATE ESTIMATION

4.1 Markov Chain Sampling

In our method, the posterior ordinate $\pi(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*|\mathbf{y})$ in (9) is estimated from the output of the MCMC simulation of the posterior distribution of the DPM model. There are currently two broad approaches for estimating the DPM model. These two methods differ in the way in which the lower-level parameters of the model are sampled; the parameters $\boldsymbol{\phi}, \boldsymbol{\kappa}$, and α are sampled in the same way in both methods. Practically, this means that we can produce an estimate of the posterior ordinate $\pi(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*|\mathbf{y}, G_0)$ from the MCMC output of either method. To show how this is done, we begin by presenting a brief review of the two MCMC sampling schemes.

The first class of methods are based on the Pólya urn representation in (4). First proposed by Escobar (1988, 1994) and MacEachern (1994), the sampling is conducted marginalized over the random measure G and exploits the fact that the joint distribution of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ in (3) is exchangeable. The full conditional distribution of $\boldsymbol{\theta}_i$ can be deduced, by virtue of exchageability, as

$$P(\boldsymbol{\theta}_i \in \cdot|\boldsymbol{\theta}_{-i}, \boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha, G_0, \mathbf{y}) \propto q^*_{k_{-i}+1,i} P^*(\boldsymbol{\theta}_i \in \cdot|y_i) + \sum_{j=1}^{k_{-i}} q^*_{j,i} \delta_{\boldsymbol{\theta}^*_{j,-i}}(\cdot), \qquad (11)$$

where $\{\boldsymbol{\theta}^*_{1,-i}, \ldots, \boldsymbol{\theta}^*_{k_{-i},-i}\}$ denote the set of k_{-i} unique values in $\boldsymbol{\theta}_{-i} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\} \setminus \{\boldsymbol{\theta}_i\}$, $q^*_{k_{-i}+1,i} \propto q_{k_{-i}+1,n}(\alpha) \int f(y_i|\boldsymbol{\theta}, \boldsymbol{\phi}, x_i) dG_0(\boldsymbol{\theta}|\boldsymbol{\kappa})$, $q^*_{j,i} \propto q_{j,n}(\alpha) f(y_i|\boldsymbol{\theta}^*_{j,-i}, \boldsymbol{\phi}, x_i)$, and $P^*(\boldsymbol{\theta}_i \in \cdot|y_i)$ is the conditional law of $\boldsymbol{\theta}_i$ when y_i has the density $f(y_i|\boldsymbol{\theta}_i, \boldsymbol{\phi}, x_i)$ and $\boldsymbol{\theta}_i \sim G_0(\cdot|\boldsymbol{\kappa})$. The sampling approach is completed by sampling $(\boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha)$ from their respective full conditional densities. We note here that because $\boldsymbol{\kappa}$ is at the highest level of the hierarchical specification, the functional form of its conditional posterior depends only on the prior of $\boldsymbol{\kappa}$ and the base measure $G_0(\cdot|\boldsymbol{\kappa})$ and is analytically available if the prior for $\boldsymbol{\kappa}$ is conjugate to $G_0(\cdot|\boldsymbol{\kappa})$, irrespective of the complexity of the semiparametric mixture model.

One can see that the expression for $q^*_{k_{-i}+1,i}$ involves an integral which is analytically available when $f(y_i|\boldsymbol{\theta}, \boldsymbol{\phi}, x_i)$ and $G_0(\boldsymbol{\theta}|\boldsymbol{\kappa})$ are conjugate. The nonconjugate case, which is less convenient, has been considered by MacEachern and Müller (1998). For the conjugate case, MacEachern (1994) developed an improved collapsed Gibbs sampler that provides better mixing. In this approach, $\boldsymbol{\theta}_{-i}$ is reexpressed in terms of the unique values $\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_{k_{-i}}$ and the cluster memberships $s_{-i} = (s_1, \ldots, s_n) \setminus \{s_i\}$, where each s_l records which unique $\boldsymbol{\theta}^*_j$ label corresponds to the value $\boldsymbol{\theta}_l$, that is, $s_l = j$ iff $\boldsymbol{\theta}_l = \boldsymbol{\theta}^*_j$. In the collapsed sampler, only the cluster membership, s_i, of the ith observation is sampled from the categorical distribution

$$P(s_i = j|s_{-i}, \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\kappa}, \alpha) = \begin{cases} c q_{j,n}(\alpha) \int f(y_i|\boldsymbol{\theta}, \boldsymbol{\phi}, x_i) dH_{j,-i}(\boldsymbol{\theta}), & 1 \le j \le k_{-i} \\ c q_{k_{-i}+1,n} \int f(y_i|\boldsymbol{\theta}, \boldsymbol{\phi}, x_i) dG_0(\boldsymbol{\theta}|\boldsymbol{\kappa}), & j = k_{-i}+1, \end{cases} \qquad (12)$$

where c is the normalizing constant and $H_{j,-i}(\boldsymbol{\theta})$ is the posterior distribution of $\boldsymbol{\theta}$ based on the prior G_0 and observations $\{y_l : l \ne i, \text{ and } s_l = j\}$, which are in cluster j. The unique $\{\boldsymbol{\theta}^*_j\}$ are updated next, if needed, given all cluster memberships s.

The second class of methods, known as the blocked Gibbs sampler, were developed by Ishwaran and James (2001a) and Ishwaran et al. (2001). These methods use the truncation of the stick-breaking construction given in (6) to express the random mixing measure in finite-dimensional form as $G(\cdot) = \sum_{l=1}^N p_l \delta_{Z_l}(\cdot)$, where N is some large integer. Under this restriction, the mixture density in (2) can be written in hierarchical fashion as

$$\begin{aligned}
&y_i|\mathbf{Z}, s_i, \boldsymbol{\phi}, x_i \overset{\text{ind}}{\sim} f(y_i|\mathbf{Z}_{s_i}, \boldsymbol{\phi}, x_i), \quad i = 1, \ldots, n,\\
&s_i|p \overset{\text{iid}}{\sim} \sum_{l=1}^N p_l \delta_l(\cdot), \qquad (13)
\end{aligned}$$

where s_i is the latent mixture component indicator for the ith observation, $s = (s_1, \ldots, s_n)$, $\mathbf{Z}_l \overset{\text{iid}}{\sim} G_0(\cdot|\boldsymbol{\kappa})$, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_N)$, $\mathbf{Z}$ is independent of p, and the distribution of $p = (p_1, \ldots, p_N)$ is specified by the stick-breaking construction. The blocked Gibbs sampler updates $\mathbf{Z}$, s, and p in multivariate blocks as opposed to the one-at-time updating of $\boldsymbol{\theta}_i$ or cluster membership s_i in the Pólya urn samplers. The full conditional distributions of $(\mathbf{Z}|s, \boldsymbol{\phi}, \boldsymbol{\kappa}, \mathbf{y})$, $(s|\mathbf{Z}, p, \boldsymbol{\phi}, \mathbf{y})$, and $(p|s, \alpha)$ where given by Ishwaran and James (2001a, sec. 5.2). Note that the latent $\boldsymbol{\theta}_i$ are still available in this scheme as $\boldsymbol{\theta}_i = \mathbf{Z}_{s_i}$.

4.2 Estimating the Posterior Ordinate Within the Sampler

We now detail, following the framework of Chib (1995), how the posterior ordinate of (9) can be estimated from the output of either the Pólya urn scheme sampler or the blocked Gibbs sampler. Let us start with the decomposition

$$\log \pi(\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*|\mathbf{y}) = \log \pi(\boldsymbol{\phi}^*|\mathbf{y}) + \log \pi(\alpha^*|\mathbf{y}, \boldsymbol{\phi}^*) + \log \pi(\boldsymbol{\kappa}^*|\mathbf{y}, \boldsymbol{\phi}^*, \alpha^*), \qquad (14)$$

and now consider the estimation of each ordinate. Suppose for simplicity that the full conditional distributions of $\boldsymbol{\phi}$, α, and $\boldsymbol{\kappa}$ have known normalizing constants. If the normalizing

constant(s) is (are) not known, and Metropolis–Hastings sampling (Chib and Greenberg 1995) is used for updating some of these parameters, then the ordinates can be estimated along the lines of Chib and Jeliazkov (2001).

Suppose that the MCMC sampling scheme, beyond the requisite burn-in period, has been iterated for $g = 1, \ldots, G_1$ cycles. The output from this sampling can, with an obvious justification, be capitalized to estimate $\pi(\boldsymbol{\phi}^*|\mathbf{y})$ as

$$\hat{\pi}(\boldsymbol{\phi}^*|\mathbf{y}) = \frac{1}{G_1}\sum_{g=1}^{G_1} \pi(\boldsymbol{\phi}^*|\boldsymbol{\theta}^{(g)}, \boldsymbol{\kappa}^{(g)}, \alpha^{(g)}, \mathbf{y}), \tag{15}$$

where the superscript (g) denotes the values drawn at the gth iteration and the density on the right side is the one that appears in the MCMC update. To estimate the second ordinate, we fix $\boldsymbol{\phi}$ at $\boldsymbol{\phi}^*$ and continue the Markov chain simulations for an additional G_2 iterations, where all other unobservables (except $\boldsymbol{\phi}$) are updated. These draws yield the estimate

$$\hat{\pi}(\alpha^*|\mathbf{y}, \boldsymbol{\phi}^*) = \frac{1}{G_2}\sum_{g=G_1+1}^{G_1+G_2} \pi\big(\alpha^*|\boldsymbol{\theta}^{(g)}, \boldsymbol{\phi}^*, \boldsymbol{\kappa}^{(g)}, \mathbf{y}\big), \tag{16}$$

where, with the introduction of an additional latent variable u, the full conditional posterior of α was given by Escobar and West (1995) as a mixture of two gamma distributions,

$$\frac{a_0 + k_n - 1}{n(b_0 - \log u)}\text{gamma}(a_0 + d, b_0 - \log u) + \left(1 - \frac{a_0 + k_n - 1}{n(b_0 - \log u)}\right)\text{gamma}(a_0 + k_n - 1, b_0 - \log u), \tag{17}$$

k_n denotes the number of distinct values in $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$, and the latent u is generated from its full conditional distribution given by beta$(u|\alpha + 1, n)$.

Last, we fix both $\boldsymbol{\phi}$ and α at $(\boldsymbol{\phi}^*, \alpha^*)$ and run the chain for another G_3 iterations to produce an estimate $\hat{\pi}(\boldsymbol{\kappa}^*|\mathbf{y}, \boldsymbol{\phi}^*, \alpha^*)$ analogous to the preceeding one. Finally, we substitute these three estimates into (14). We mention that the numerical standard error of the resulting estimate can be found according to the method given by Chib (1995).

5. LIKELIHOOD ORDINATE ESTIMATION

5.1 Basic Sequential Importance Sampling

In this section we describe methods for estimating the likelihood ordinate $L(\mathbf{y}|\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*, G_0)$ of (9). To set the stage for the problem, let us begin by recalling that the likelihood function of the parameters at a particular point $\boldsymbol{\psi}^* = (\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*)$ in the parameter space, given the sample data, is

$$L(\mathbf{y}|\boldsymbol{\phi}^*, \boldsymbol{\kappa}^*, \alpha^*) = \int\left\{\prod_{i=1}^{n}\int f(\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\phi}^*)dG(\boldsymbol{\theta}_i)\right\}d\mathcal{P}(G|\alpha^*, G_0, \boldsymbol{\kappa}^*). \tag{18}$$

The problem is that neither the integral in braces [even when $f(\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\phi}^*)$ and G_0 are conjugate] nor the outside integral over the infinite-dimensional parameter G is analytic. This raises an interesting question of how the integrals should be calculated. (We show in Sec. 6.1 that an exact answer can be derived by a tedious computation when n is small, but of course we need to develop a general approach that is efficient and valid for any sample size.) Earlier, Ferguson (1983) also did some exact calculations for small sample sizes. After a thorough study of this problem and extensive comparative evaluation of different techniques, we have developed a method that is both accurate and computationally efficient.

To describe our method, we first show how the likelihood ordinate can be found as a byproduct of the SIS method, where we use the subscript (i) (e.g. $\mathbf{y}_{(i)}$) to generically denote the first i elements of a vector [i.e., $\mathbf{y}_{(i)} = (\mathbf{y}_1, \ldots, \mathbf{y}_i)$]. In sequential imputation, the $\boldsymbol{\theta}_i$ $(i \le n)$ are sequentially generated from the importance sampling distribution

$$\pi^*(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|\mathbf{y}, \boldsymbol{\psi}^*) = \prod_{i=1}^{n}\pi(\boldsymbol{\theta}_i|\mathbf{y}_{(i)}, \boldsymbol{\theta}_{(i-1)}, \boldsymbol{\psi}^*), \tag{19}$$

starting with $\boldsymbol{\theta}_1$ and continuing on to $\boldsymbol{\theta}_n$. Kong et al. (1994) showed that the importance weight equals

$$\frac{\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|\mathbf{y}, \boldsymbol{\psi}^*)}{\pi^*(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|\mathbf{y}, \boldsymbol{\psi}^*)} = \frac{w}{L(\mathbf{y}|\boldsymbol{\psi}^*)}, \tag{20}$$

where

$$w = w(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = f(\mathbf{y}_1|\boldsymbol{\psi}^*)\prod_{i=2}^{n} f(\mathbf{y}_i|\mathbf{y}_{(i-1)}, \boldsymbol{\theta}_{(i-1)}, \boldsymbol{\psi}^*) \tag{21}$$

and $f(\mathbf{y}_i|\mathbf{y}_{(i-1)}, \boldsymbol{\theta}_{(i-1)}, \boldsymbol{\psi}^*)$ is the prequential predictive density of $\mathbf{y}_i$.

Because $L(\mathbf{y}|\boldsymbol{\psi}^*)$ in (20) is independent of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, the expression in (20) can be used to deliver an estimate of the likelihood function $L(\mathbf{y}|\boldsymbol{\psi}^*)$. Suppose that the sequential sampling procedure is repeated M times and at that each cycle g, we obtain the draws $\boldsymbol{\theta}_1^{(g)}, \ldots, \boldsymbol{\theta}_n^{(g)}$ from $\pi^*(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|\mathbf{y}, \boldsymbol{\psi}^*)$ and calculate $w^{(g)}$ following (21). Then the average $\bar{w} = M^{-1}\sum_{g=1}^{M} w^{(g)}$ over the M draws is a simulation-consistent Monte Carlo estimate of the likelihood ordinate, as is readily confirmed.

Interestingly, this basic idea can be applied to the DPM model when $f(\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\phi}^*)$ and G_0 are conjugate, because from (4) and (5) we know immediately that

$$f(\mathbf{y}_i|\mathbf{y}_{(i-1)}, \boldsymbol{\theta}_{(i-1)}, \boldsymbol{\psi}^*) = \frac{\alpha^*}{\alpha^* + i - 1}\int f(\mathbf{y}_i|\boldsymbol{\theta}, \boldsymbol{\phi}^*)dG_0(\boldsymbol{\theta}|\boldsymbol{\kappa}^*) + \sum_{j=1}^{k_{i-1}}\frac{n_{j,i-1}}{\alpha^* + i - 1}f(\mathbf{y}_i|\boldsymbol{\theta}_j, \boldsymbol{\phi}^*) \tag{22}$$

and

$$\pi(\boldsymbol{\theta}_i|\mathbf{y}_{(i)}, \boldsymbol{\theta}_{(i-1)}, \boldsymbol{\psi}^*) \propto \frac{\alpha^*}{\alpha^* + i - 1}f(\mathbf{y}_i|\boldsymbol{\theta}_i, \boldsymbol{\phi}^*)f_{G_0}(\boldsymbol{\theta}_i|\boldsymbol{\kappa}^*) + \sum_{j=1}^{k_{i-1}}\frac{n_{j,i-1}}{\alpha^* + i - 1}f(\mathbf{y}_i|\boldsymbol{\theta}_j, \boldsymbol{\phi}^*)\delta_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_i). \tag{23}$$

5.2 Collapsed Sequential Importance Sampling

Although the basic SIS is easy to implement, in applying this method to the DPM model we have found that the weights $w^{(g)}$ tend to be extremely variable. This problem has been observed in other settings. To overcome this problem, we turn to the collapsed SIS method developed in the context of a DPM beta-binomial model by MacEachern et al. (1999) and later extended to the multinomial and nonexchangeable beta-binomial models by Quintana (1998) and Quintana and Newton (2000). The collapsed SIS method was also discussed in the context of weighted Chinese restaurant processes by Lo et al. (1996). General weighted Chinese restaurant processes and algorithms for their posterior inference are reviewed by Ishwaran and James (2001b) and Ishwaran and Takahara (2002), and such methods were used to estimate the marginal density by Ishwaran et al. (2001). The idea behind the collapsed SIS is the elimination of $\boldsymbol{\theta}_i$ by integration, which collapses the space in which the sequential sampling operates to the set of possible cluster memberships. Because $\boldsymbol{\theta}_i$ is analytically integrated out from the computation, this method has less variability due to the Rao–Blackwellization effect (see MacEachern et al. 1999). As we illustrate in Section 6, this modified SIS method provides the correct foundation for our method.

To describe the method, we recall that each unique $\boldsymbol{\theta}_i$ forms a cluster under the Dirichlet process prior. In SIS these clusters are formed sequentially. Now in the collapsed method we do not sample $\boldsymbol{\theta}_i$ but instead sample the cluster membership, s_i, marginalized over $\boldsymbol{\theta}_i$. Recalling the notations used earlier in (4), (5), and (12), collapsed sequential sampling for the DPM model proceeds as follow. First, compute $u_1 = f(\mathbf{y}_1|\boldsymbol{\psi}^*) = \int f(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\phi}^*)\, dG_0(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)$ and set s_1 to equal 1 (because the first observation must begin with a new cluster). Then, for $i = 2, \ldots, n$, perform the following steps sequentially:

Step 1. Compute the predictive probability

$$\begin{aligned} u_i^{(g)} &= f(\mathbf{y}_i|\mathbf{y}_{(i-1)}, \mathbf{s}_{(i-1)}^g, \boldsymbol{\psi}^*, G_0) \\ &= \frac{\alpha^*}{\alpha^* + i - 1} \int f(\mathbf{y}_i|\boldsymbol{\theta}, \boldsymbol{\phi}^*)\, dG_0(\boldsymbol{\theta}|\boldsymbol{\kappa}^*) \\ &\quad + \sum_{j=1}^{k_{i-1}} \frac{n_{j,i-1}}{\alpha^* + i - 1} \int f(\mathbf{y}_i|\boldsymbol{\theta}, \boldsymbol{\phi}^*)\, dH_{j,i-1}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*), \end{aligned} \tag{24}$$

where $H_{j,i-1}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*)$ is the posterior distribution of $\boldsymbol{\theta}$ based on the prior G_0 and observations $\{\mathbf{y}_l : l \le i-1 \text{ and } s_l = j\}$. When f and G_0 are conjugate, both integrals can be obtained in closed form.

Step 2. Draw $s_i^{(g)}$ from the categorical distribution

$$\begin{aligned} &\Pr(s_i = j|\mathbf{y}_{(i)}, \mathbf{s}_{(i-1)}^{(g)}, \boldsymbol{\psi}^*) \\ &\quad = c\frac{n_{j,i-1}}{\alpha^* + i - 1} \int f(\mathbf{y}_i|\boldsymbol{\theta}, \boldsymbol{\phi}^*)\, dH_{j,i-1}(\boldsymbol{\theta}|\boldsymbol{\kappa}^*), \\ &\qquad 1 \le j \le k_{i-1} \\ &\quad = c\frac{\alpha^*}{\alpha^* + i - 1} \int f(\mathbf{y}_i|\boldsymbol{\theta}, \boldsymbol{\phi}^*)\, dG_0(\boldsymbol{\theta}|\boldsymbol{\kappa}^*), \\ &\qquad j = k_{i-1} + 1, \end{aligned} \tag{25}$$

where c is the normalizing constant. In other words, sample s_i from the set of existing unique cluster labels with probabilities given in the first line of the foregoing expression, or else assign a new cluster label with probability given in the second line.

At the end of each complete run through the observations, we calculate $w^{(g)} = u_1^{(g)} \prod_{i=2}^n u_i^{(g)}$, and then estimate the likelihood ordinate of the DPM model as $\widehat{L}(\mathbf{y}|\boldsymbol{\psi}^*, G_0) = M^{-1} \sum_{g=1}^M w^{(g)}$.

We mention that the preceeding discussion has been restricted to the conjugate DPM model, because our applications, along with most formulations of DPM models, are centered on this case. A method for the nonconjugate setting is available from the authors.

Finally, the efficiency of $\bar{w}$ as an estimate of the likelihood ordinate can be measured by its coefficient of variation, $C(\bar{w})$. As pointed out by Irwin, Cox, and Kong (1994), $C(\bar{w})$ can be shown by the delta method to be approximately the standard error of $\log \bar{w}$ in estimating the log-likelihood. The sample estimate of $C(\bar{w})$ is

$$\widehat{C}(\bar{w}) = \frac{1}{\sqrt{M}} \frac{s_w}{\bar{w}},$$

where s_w denotes the sample standard deviation of $w^{(g)}$, $g = 1, \ldots, M$.

6. EXAMPLES

6.1 An Example With an Exact Answer

We now turn to providing some empirical demonstrations and illustrations of our marginal likelihood estimation method for DPM models. Given that the method for estimating the posterior ordinate has been tested in several problems [Chib (1995), Chib and Jeliazkov (2001)] it is important to focus attention on the estimation of the likelihood ordinate proposed in Section 5. As a cross-check, we consider a special case in which the likelihood ordinate can be obtained via alternative methods, either analytically or approximately, and compare the results with those from our approach.

Consider the longitudinal study reported by Gelfand, Hills, Racine-Poon, and Smith (1990) on $n = 30$ young rats whose weights are measured weekly for five time periods. Let y_{ij} denote the weight of the ith rat measured at age x_{ij} and let

$$\begin{aligned} &\mathbf{y}_i|\boldsymbol{\theta}_i, \sigma^2, \mathbf{X}_i \sim \mathrm{N}_5(\mathbf{X}_i\boldsymbol{\theta}_i, \sigma^2\mathbf{I}); \\ &\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})'|\boldsymbol{\mu}, \mathbf{D} \sim \mathrm{N}_2(\boldsymbol{\mu}, \mathbf{D}), \qquad i = 1, \ldots, n; \\ &\mathbf{D}^{-1} \sim \mathrm{Wishart}_2(2, \mathbf{R}); \sigma^2 \sim \text{inverse gamma}(a, b), \end{aligned}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{i5})'$, $\mathbf{X}_i$ is the design matrix with units in the first column and $(x_{i1}, \ldots, x_{i5})'$ in the second column, and $\mathrm{N}_p(\mathbf{m}, \boldsymbol{\Sigma})$ is the p-variate normal distribution with mean vector $\mathbf{m}$ and covariance matrix $\boldsymbol{\Sigma}$. The hyperparameters $\mathbf{R}$, a, and b are set to equal diag(100, .1), 4.25 and 97.5, respectively.

We consider a DPM extension of this model by letting the random coefficients $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \overset{\text{iid}}{\sim} G$, where G follows a DP prior with base measure $G_0 = \mathrm{N}_2(\boldsymbol{\mu}, \mathbf{D})$, the distribution used

in the foregoing parametric model. This is a simple and effective way to relax the normality assumption in clustered data models.

Now consider the question of finding the likelihood ordinate at some high-density point $\boldsymbol{\psi}^* = (\sigma^{2*}, \boldsymbol{\mu}^*, \mathbf{D}^*)$. The n rats can cluster in a large number of possible ways. Given a set of cluster memberships $\mathbf{s} = (s_1, \ldots, s_n)$, the density $f(\mathbf{y}|\sigma^{2*}, \boldsymbol{\mu}^*, \mathbf{D}^*, \mathbf{s})$ can be obtained analytically by using the conjugate structure as

$$f(\mathbf{y}|\sigma^{2*}, \boldsymbol{\mu}^*, \mathbf{D}^*, \mathbf{s}) = \prod_{j=1}^{k_n} \left\{ \int \prod_{\{i:s_i=j\}} \mathrm{N}_5(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\theta}, \sigma^{2*})\, dG_0(\boldsymbol{\theta}|\boldsymbol{\mu}^*, \mathbf{D}^*) \right\}, \quad (26)$$

where k_n is the number of distinct clusters or the number of distinct values in $\{s_1, \ldots, s_n\}$. Then the likelihood ordinate is given as $L(\mathbf{y}|\sigma^{2*}, \boldsymbol{\mu}^*, \mathbf{D}^*) = \sum f(\mathbf{y}|\sigma^{2*}, \boldsymbol{\mu}^*, \mathbf{D}^*, \mathbf{s})\pi(\mathbf{s})$, where the sum is over all possible $\mathbf{s}$ and $\pi(\mathbf{s})$ is the prior probability of getting the partition $\mathbf{s}$. Because the sum is over all partitions, this analytic estimate is computationally feasible only when n is small. However, in these feasible cases, one can compare this analytic estimate with the likelihood value obtained by our proposed method.

Suppose that $n = 10$. For this sample size, we compare the analytic estimate of the likelihood ordinate with the estimates from the proposed basic and collapsed sequential methods. We also consider an alternative Monte Carlo estimate of the likelihood ordinate based on sampling the cluster locations $\mathbf{s} = (s_1, \ldots, s_n)$ sequentially from their prior distribution using the Pólya urn scheme in (5) and then averaging the likelihood $f(\mathbf{y}|\sigma^{2*}, \boldsymbol{\mu}^*, \mathbf{D}^*, \mathbf{s})$ of (26) over these realizations of $\{\mathbf{s}\}$. This estimate is referred to as the "prior sampling"–based estimate in Table 1. Ferguson (1983) reported a similar comparison for the case of $n = 5$ where he compared the analytic estimate of the predictive density with the Monte Carlo estimate obtained from the Kuo (1986) prior sampling method.

The estimates listed in Table 1 show that both the collapsed and basic sequential methods accurately estimate the likelihood ordinate. Whereas the inaccuracy of the likelihood estimate obtained by sampling the cluster locations from the prior may not appear significant in Table 1, note that these results are for a rather small sample size of $n = 10$. In experiments involving larger sample sizes, however, we have found that the estimate based on sampling cluster locations from the prior are smaller than the SIS-based estimates by several orders of magnitude. To get an idea of the variability of the estimates, in Figure 1 we plot the trace of the log-likelihood evaluations from the different methods. We see that the prior sampling method shows extreme fluctuations, whereas the collapsed method is the most stable (note the different vertical scales).

Table 1. Comparison of Log-Likelihood Ordinates Found by Likelihood Estimation Methods; n = 10

Analytic value	*Basic SIS*	*Collapsed SIS*	*Prior sampling*
−193.96	−193.971	−193.954	−193.75

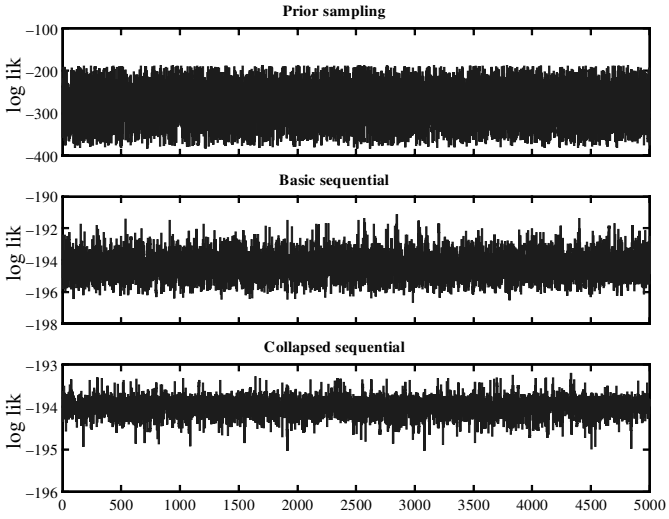

Figure 1. Rat Data: Trace of Log-Likelihood Evaluations. Note the different vertical scales.

6.2 Bayes Factors for Binary Data Models

In recent years, there has been a significant amount of Baysian work on generalizing the simple probit and logistic regression models for binary regression. Albert and Chib (1993) first showed how to fit a t-link model, and Basu and Mukhopadhyay (2000) extended this idea to DPM link models. Erkanli, Stangl, and Müller (1993) and Newton, Czado, and Chappell (1996) provided different semiparametric generalizations of the binary data model. The issue of model selection, is not addressed in these articles however. In this section we illustrate the application of our techniques to a semiparametric model for binary data that we compare with two parametric models. The data for this problem are from Brown (1980), who used the response variable $\{y_i\}$, $i \leq 53$, as an indicator of the presence of prostatic nodal involvement in patients with prostate cancer. The objective is to explain the response y_i with four covariates: log of the level of serum acid phosphate (x_2); the result of an X-ray examination, coded 0 if negative and 1 if positive (x_3); size of the tumor, coded 0 if small and 1 if large (x_4); and pathologic grade of the tumor, coded 0 if less serious and 1 if more serious (x_5). These data have been analyzed by Chib (1995) using a binary probit regression model, which yielded a log marginal likelihood value of −36.252.

Models. We start with the popular probit regression, denoted by $\mathcal{M}_1$, that models the probability of presence or "success probability" as $\Pr(y_i = 1|\mathcal{M}_1, \boldsymbol{\beta}) = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$, where $\Phi(\cdot)$ is the standard normal distribution function. This model is compared with the t-link model discussed by Albert and Chib (1993). In this case $\Pr(y_i = 1|\mathcal{M}_2, \boldsymbol{\beta}) = F_t(\mathbf{x}_i'\boldsymbol{\beta}, 1, \nu) = \int \Phi(\mathbf{x}_i'\boldsymbol{\beta}\sqrt{\lambda}) dG_0(\lambda)$, where $F_t(\cdot, \xi, \nu)$ is the cumulative distribution function of the t distribution with dispersion ξ and ν degrees of freedom and $G_0 = \text{gamma}(\nu/2, \nu/2)$. We let $\nu = 10$.

Our goal is to compare the preceeding parametric models with the semiparametric DPM model proposed by Basu and Mukhopadhyay (2000). Under the DPM model, the link function is modeled semiparametrically as a normal scale mixture

where the mixing distribution G is random,

$$\Pr(y_i = 1|\mathcal{M}_3, \boldsymbol{\beta}) = \int \Phi(\mathbf{x}_i'\boldsymbol{\beta}\sqrt{\lambda})\, dG(\lambda),$$

$$G \sim \mathrm{DP}(\alpha, G_0), \qquad G_0 = \mathrm{gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right),$$

with $\nu = 10$. Note that when G is a fixed distribution and equal to G_0, we get the t-link model $\mathcal{M}_2$.

To compare the three models on a fair basis, we assume that in each model $\boldsymbol{\beta}$ follows a $\mathrm{N}_5(\boldsymbol{\beta}_0, \mathbf{B}_0)$ prior distribution, where $\boldsymbol{\beta}_0$ is a vector with each element equal to .75 and $\mathbf{B}_0$ is a diagonal matrix with 25 on the diagonal. In addition, in the DPM model, the prior on the concentration parameter α is taken to be gamma$(5, 2)$.

Fitting of Models. We fit each of the three contending models by the Albert and Chib (1993) approach. For example, to estimate the DPM model, we express the model in terms of latent variables $\{z_i\}$ as

$$z_i|\boldsymbol{\beta}, \lambda_i \sim \mathrm{N}(\mathbf{x}_i'\boldsymbol{\beta}, \lambda_i^{-1});\ \ y_i = I(z_i > 0);$$

$$\lambda_1, \ldots, \lambda_n \overset{iid}{\sim} G, G \sim \mathrm{DP}(\alpha, G_0), G_0 = \mathrm{gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right).$$

A major benefit of this representation is that conditioned on the latent z_i, the model resembles a linear regression with all of its associated tractability. The posterior distribution of the parameters and the latent variables can now be simulated through MCMC methods by combining the procedure outlined by Albert and Chib (1993) for model $\mathcal{M}_2$ with the usual steps arising from the DPM prior. The details are suppressed to conserve space. We ran all of our MCMC simulations for 10,000 iterations after a burn-in of 500 cycles. We estimated the likelihood ordinate and the posterior ordinate at the posterior mean from these MCMC simulations.

Computing the Marginal Likelihood of the DPM Model. We now discuss marginal likelihood computation of the semiparametric DPM model $\mathcal{M}_3$. (The marginal likelihood of the parametric models is obtained from the approach outlined in Chib 1995.) As usual, we start with the decomposition in (9), where now $\boldsymbol{\phi} = \boldsymbol{\beta}$ and $\boldsymbol{\kappa}$ is nonstochastic. The posterior ordinate $\pi(\boldsymbol{\beta}^*, \alpha^*|\mathbf{y})$ can be estimated from the decomposition $\pi(\boldsymbol{\beta}^*|\mathbf{y})\pi(\alpha^*|\mathbf{y}, \boldsymbol{\beta}^*)$, where the first ordinate is calculated via (15) by averaging the conditional posterior density $\mathrm{N}_5(\boldsymbol{\beta}^*|\hat{\boldsymbol{\beta}}, \mathbf{B})$, with $\hat{\boldsymbol{\beta}} = \mathbf{B}(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{53} \lambda_i \mathbf{x}_i z_i)$ and $\mathbf{B} = (\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{53} \lambda_i \mathbf{x}_i z_i)^{-1}$, over the draws from the MCMC run. The second ordinate $\pi(\alpha^*|\mathbf{y}, \boldsymbol{\beta}^*)$ is estimated according to (16) and the mixture representation in (17).

Calculation of the likelihood ordinate $L(y|\boldsymbol{\beta}^*, \alpha^*, G_0)$ by the collapsed SIS method is rather more interesting. The presence of latent variables z_i in the model leads to some arguments that are likely to be of general interest.

One initial point is that even though the unobservables connected to the ith observation y_i include the latent variable z_i and the random precision parameters λ_i, to produce a tractable version of the collapsed SIS method, we must collapse or marginalize over only λ_i, the variable on which the DPM prior is defined. In other words, the latent variable z_i persists in the sampling in conjunction with the cluster membership variable s_i. We now describe the details.

The predictive ordinate of the first observation can be calculated easily and is given by $u_1 = F_t(\mathbf{x}_1'\boldsymbol{\beta}^*, 1, \nu)^{y_1}\{1 - F_t(\mathbf{x}_1'\boldsymbol{\beta}^*, 1, \nu)\}^{(1-y_1)}$. We next draw $\boldsymbol{\eta}_1 = (z_1, s_1)$ from its posterior distribution conditioned on y_1. This is accomplished by drawing z_1 from its conditional distribution,

$$z_1|y_1, \boldsymbol{\psi}^*, G_0 \sim \begin{cases} t(\mathbf{x}_1'\boldsymbol{\beta}^*, 1, \nu)I[0, \infty) & \text{if } y_1 = 1 \\ t(\mathbf{x}_1'\boldsymbol{\beta}^*, 1, \nu)I(-\infty, 0) & \text{if } y_1 = 0, \end{cases}$$

where $t(\mu, 1, \nu)$ is the Student t distribution with location μ, dispersion 1, and ν degrees of freedom. We next set $s_1 = 1$.

For the remaining observations $(i = 2, \ldots, n)$, we calculate the prequential predictive density of y_i in step 1 of the collapsed SIS and then draw a value of $\boldsymbol{\eta}_i = (z_i, s_i)$ from $p(\boldsymbol{\eta}_i|\mathbf{y}_{(i)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0)$ in step 2. Suppose that after completing these steps for the first $i-1$ observations, there are k_{i-1} clusters with the jth cluster with $n_{j,i-1}$ elements, $j = 1, \ldots, k_{i-1}$. Then the posterior distribution of λ based on the prior G_0 and only those latent observations $\{z_l : l < i, s_l = j\}$ in the jth cluster is

$$\begin{aligned} &H_{j,i-1}(\lambda) \\ &\quad = \mathrm{gamma}\left((\nu + n_{j,i-1})/2, (\nu + \sum_{\{l<i; s_l=j\}} (z_l - \mathbf{x}_l'\boldsymbol{\beta}^*)^2)/2\right) \\ &\quad \equiv \mathrm{gamma}\left(\frac{a_{j,i-1}}{2}, \frac{b_{j,i-1}}{2}\right). \end{aligned} \tag{27}$$

Based on this distribution, the prequential predictive density of y_i can be derived, after some algebra, as $u_i = f(y_i|\mathbf{y}_{(i-1)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0) = p_i^{y_i}\{1 - p_i\}^{(1-y_i)}$, $i = 2, \ldots, n$, where

$$\begin{aligned} p_i = &\frac{\alpha^*}{\alpha^* + i - 1} F_t(\mathbf{x}_i'\boldsymbol{\beta}^*, 1, \nu) \\ &+ \frac{1}{\alpha^* + i - 1} \sum_{j=1}^{k_{i-1}} n_{j,i-1} F_t(\mathbf{x}_i'\boldsymbol{\beta}^*, a_{j,i-1}^{-1} b_{j,i-1}, a_{j,i-1}). \end{aligned}$$

Next we move to step 2, where we apply the method of composition to draw a variate $\boldsymbol{\eta}_i = (z_i, s_i)$ from the joint distribution $\pi(z_i, s_i|\mathbf{y}_{(i)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0) = \pi(z_i|\mathbf{y}_{(i)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0) \times \pi(s_i|\mathbf{y}_{(i)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0, z_i)$. The first of these distributions can be obtained by an interesting argument, noting that

$$\begin{aligned} \lambda_i|\mathbf{y}_{(i-1)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0 \sim &\frac{\alpha^*}{\alpha^* + i - 1} G_0(\cdot|\kappa^*) \\ &+ \frac{1}{\alpha^* + i - 1} \sum_{j=1}^{k_{i-1}} n_{j,i-1} H_{j,i-1}(\cdot) \end{aligned} \tag{28}$$

and

$$z_i|\mathbf{y}_{(i-1)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0, \lambda_i \sim z_i|\boldsymbol{\psi}^*, G_0, \lambda_i \sim \mathrm{N}(\mathbf{x}_i'\boldsymbol{\beta}^*, \lambda_i^{-1}). \tag{29}$$

If we marginalize (29) with respect to the distribution of λ_i given in (28), we obtain

$$z_i|\mathbf{y}_{(i-1)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0 \sim \frac{\alpha^*}{\alpha^*+i-1} t(\mathbf{x}_i'\boldsymbol{\beta}^*, 1, \nu) + \frac{1}{\alpha^*+i-1} \sum_{j=1}^{k_{i-1}} n_{j,i-1} t(\mathbf{x}_i'\boldsymbol{\beta}^*, a_{j,i-1}^{-1} b_{j,i-1}, a_{j,i-1}). \quad (30)$$

This can be viewed as the prior distribution of z_i, and so, by the Albert and Chib (1993) approach, we immediately determine that the density $\pi(z_i|\mathbf{y}_{(i)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0)$ is the mixture distribution in (30) truncated below at 0 if y_i is 1 and truncated above at 0 if y_i is 0. Having drawn z_i from this truncated mixture distribution, we complete the composition by drawing the categorical random variable s_i according to (25) from the distribution

$$\begin{aligned} &\Pr(s_i = j|\mathbf{y}_{(i)}, \boldsymbol{\eta}_{(i-1)}, \boldsymbol{\psi}^*, G_0, z_i) \\ &\quad = c\frac{n_{j,i-1}}{\alpha^*+i-1} f_t(z_i|\mathbf{x}_i'\boldsymbol{\beta}^*, a_{j,i-1}^{-1} b_{j,i-1}, a_{j,i-1}), \qquad j \le k_{i-1} \\ &\quad = c\frac{\alpha^*}{\alpha^*+i-1} f_t(z_i|\mathbf{x}_i'\boldsymbol{\beta}^*, 1, \nu), \qquad j = k_{i-1}+1, \end{aligned} \quad (31)$$

where c is the normalizing constant and $f_t(\cdot|\mu, \xi, \nu)$ denotes the density of the t distribution with location μ, dispersion ξ, and degrees of freedom ν. A run through the observations now yields the quantity $w = u_1 \prod_{i=2}^{n} u_i$. These steps are repeated M times, and the average of the w's produces our estimate of the likelihood ordinate $L(\mathbf{y}|\boldsymbol{\beta}^*, \alpha^*, G_0)$ for the DPM model $\mathcal{M}_3$.

Finally, the marginal likelihood of the binary DPM model is estimated by inserting the posterior and likelihood ordinate estimates into (9). We note here that sampling the latent z_i within our collapsed SIS maintains the tractability of the posterior $H_{j,i-1}(\lambda)$ in (27), which in turn provides manageable expressions for all of the requisite distributions.

Results. Figure 2 illustrates the marginal likelihood estimate of the DPM model from our proposed approach. We see from the graph that the estimate stabilizes up to the first decimal place quite quickly. In Table 2 we report the marginal likelihood for the three models under contention. It is noteworthy that in this case, the Bayes factor criterion does not support the DPM model (in fact, the Bayes factors provides "substantial" evidence in favor of the Student t model against the DPM model). This shows that a model elaboration in the direction of a semiparametric model need not necessarily dominate a parametric specification. (In the next section, we reach the opposite conclusion.) The ability to evaluate such elaborations, which hitherto has not been possible, should prove useful in practical model building.

6.3 Bayes Factors for Longitudinal Data Mixed Models

Models. Carlan and Louis (2000) reported data from a clinical trial on the effectiveness of two antiretroviral drugs (didanosine and zalcitabine) in 467 persons with advanced human immunodeficiency virus infection. The response variable y_{ij} for patient i at time j is the square root (to reduce skewness) of the patient's CD4 count, recorded at study entry, and at 2, 6, 12, and 18 months after entry. Several patients have incomplete records due to dropouts, so the effective response vector for the ith patient is $\mathbf{y}_i = (y_{i1}, \ldots, y_{i,t_i})$, where $1 \le t_i \le 5$. Carlin and Louis (2000), and Chib and Jeliazkov (2001) used the following linear mixed-effects model for these data:

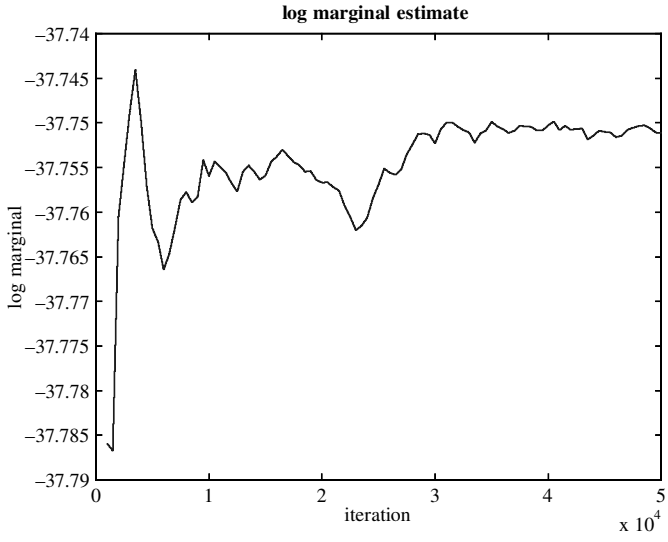

Figure 2. Binary Data: The Marginal Likelihood Estimate of the DPM Model Versus Number of Iterations.

$$\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i, \sigma^2, \mathbf{X}_i \sim \mathrm{N}_{t_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i, \sigma^2), \quad i \le n = 467; \quad \mathbf{b}_1, \ldots, \mathbf{b}_n \overset{\text{iid}}{\sim} \mathrm{N}_2(\mathbf{0}, \mathbf{D}), \quad (32)$$

where the jth row of $\mathbf{W}_i$ takes the form $\mathbf{w}_{ij} = (1, x_{ij})$, $x_{ij} \in \{0, 2, 6, 12, 18\}$, the fixed design matrix $\mathbf{X}_i$ is $\mathbf{X}_i = (\mathbf{W}_i \mid d_i\mathbf{W}_i \mid a_i\mathbf{W}_i)$, d_i is a binary variable indicating whether patient i received didanosine ($d_i = 1$) or zalecitabine ($d_i = 0$), and a_i is a binary variable indicating whether the patient was diagnosed as having AIDS at baseline ($a_i = 1$) or not ($a_i = 0$). We denote this parametric model by $\mathcal{M}_1$. The second parametric model, $\mathcal{M}_2$, provides a heavier-tailed distribution for the two-dimensional random effects, $\mathbf{b}_i$, by modeling them with a Student t distribution.

Our third model, $\mathcal{M}_3$, is a flexible semiparametric model that does not impose a parametric assumption on the random-effects distribution, but instead models it by a Dirichlet process as

$$\mathbf{b}_1, \ldots, \mathbf{b}_n|G \sim G; \quad G|G_0 \sim \mathrm{DP}(\alpha, G_0); \quad G_0(\cdot|\mathbf{D}) = \mathrm{N}_2(\mathbf{0}, \mathbf{D}).$$

Table 2. Binary Data: Estimated Log Marginal Likelihoods (on the diagonal) for Three Binary Data Models

	$\mathcal{M}_1$ (probit)	$\mathcal{M}_2$ (Student t link)	$\mathcal{M}_3$ (DPM)
$\mathcal{M}_1$	−36.252		
$\mathcal{M}_2$	(.451)	−35.801	
$\mathcal{M}_3$	(−1.488)	(−1.939)	−37.740

NOTE: The entry in brackets is the log of the Bayes factor in favor of the row model versus the column model.

If G is a fixed distribution and equal to G_0, we get the parametric model $\mathcal{M}_1$. Bush and MacEachern (1996), Kleinman and Ibrahim (1998), Tao, Palta, Yandell, and Newton (1999), and Ishwaran and Takahara (2002) also took advantage of the Dirichlet process prior in linear mixed models.

We complete the models by assuming that a priori, $\boldsymbol{\beta}: 6\times 1$ is $\mathrm{N}_6(\boldsymbol{\beta}_0, \mathbf{B}_0)$ with $\boldsymbol{\beta}_0 = (10, 0, 0, 0, -3, 0)$ and $B_0 = \mathrm{diag}(2^2, 1^2, (.1)^2, 1^2, 1^2, 1^2)$; $\mathbf{D}^{-1}$ is $\mathrm{Wishart}_2(\rho_0, \mathbf{R}_0/\rho_0)$ with $\rho_0 = 24$ and $\mathbf{R}_0 = \mathrm{diag}(.25, 16)$; and σ^2 is inverse gamma $(3, 60)$. Finally, in model $\mathcal{M}_3$, α is assumed to follow a gamma distribution with parameters 20 and 1.

Posterior Ordinate Estimation. For brevity, we only consider the marginal likelihood computation of the semiparametric model $\mathcal{M}_3$. In accordance with the general approach of (9), the notations of which are transferred here with $\boldsymbol{\phi} = (\boldsymbol{\beta}, \sigma^2)$ and $\boldsymbol{\kappa} = \mathbf{D}^{-1}$, write the posterior ordinate as

$$\pi(\mathbf{D}^{-1*}, \boldsymbol{\beta}^*, \sigma^{2*}, \alpha^*|\mathbf{y}) = \pi(\mathbf{D}^{-1*}|\mathbf{y})\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*) \times \pi(\sigma^{2*}|\mathbf{y}, \mathbf{D}^*, \boldsymbol{\beta}^*)\pi(\alpha^*|\mathbf{y}, \mathbf{D}^*, \boldsymbol{\beta}^*, \sigma^{2*}),$$

where $\boldsymbol{\psi}^* = (\boldsymbol{\beta}^*, \sigma^{2*}, \mathbf{D}^*, \alpha^*)$ is the posterior mean from the MCMC run. The first of these four ordinates is estimated from the output of the complete MCMC run by averaging the Wishart conditional density of $\mathbf{D}^{-1}$ at $\mathbf{D}^{-1^*}$. The second ordinate is estimated from a reduced run where $\mathbf{D}$ is fixed at $\mathbf{D}^*$ and the multivariate normal conditional density of $\boldsymbol{\beta}$ is evaluated at $\boldsymbol{\beta}^*$ and averaged. The third ordinate is estimated from a further reduced run with both $\mathbf{D}$ and $\boldsymbol{\beta}$ fixed, and then averaging the inverse gamma conditional density at σ^{2*}. The fourth ordinate is obtained from a final reduced run with $\mathbf{D}$, $\boldsymbol{\beta}$, and σ^2 fixed at the starred values, where the mixture gamma density of α in (17) is averaged over the sampled values at the fixed point α^*.

Likelihood Ordinate Estimation. We estimate the likelihood ordinate $L(\mathbf{y}|\boldsymbol{\psi}^*, G_0)$ at $\boldsymbol{\psi}^*$ using the collapsed SIS as described in Section 5.2. The predictive ordinate of the first observation is given by $u_1 = f(\mathbf{y}_1|\boldsymbol{\psi}^*) = \int f(\mathbf{y}_1|\mathbf{b}_1, \boldsymbol{\beta}^*, \sigma^{2*})dG_0(\mathbf{b}_1|\mathbf{D}^*) = \mathrm{N}_{t_1}(\mathbf{y}_1|\mathbf{X}_1\boldsymbol{\beta}^*, \mathbf{V}_1^*)$, where $\mathbf{V}_i^* = \sigma^{2*}\mathbf{I}_{t_i} + \mathbf{W}_i\mathbf{D}^*\mathbf{W}_i'$, $i = 1, \ldots, n$. We next set $s_1 = 1$.

For the remaining observations ($i = 2, \ldots, n$), note that the posterior distribution of $\mathbf{b}$ based on the prior G_0 and only those observations $\{y_l : l < i, s_l = j\}$ in the jth cluster is $H_{j,i-1}(\mathbf{b}) = \mathrm{N}_2(\hat{\mathbf{b}}_j, \mathbf{D}_j)$, $j = 1, \ldots, k_{i-1}$, where $\mathbf{D}_j = (\mathbf{D}^{*-1} + \sigma^{-2*}\sum_{l:s_l=j}\mathbf{W}_l'\mathbf{W}_l)^{-1}$ and $\hat{\mathbf{b}}_j = \mathbf{D}_j\sigma^{-2*}\sum_{l:s_l=j}\mathbf{W}_l'(\mathbf{y}_l - \mathbf{X}_l\boldsymbol{\beta}^*)$. The prequential predictive ordinate of $\mathbf{y}_i$ now follows from (24) as

$$u_i = \frac{\alpha^*}{\alpha^* + i - 1}\mathrm{N}_{n_i}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}^*, \mathbf{V}_i^*) + \sum_{j=1}^{k_{i-1}}\frac{n_{j,i-1}}{\alpha^* + i - 1}\mathrm{N}_{t_i}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}^* + \mathbf{W}_i\hat{\mathbf{b}}_j, \boldsymbol{\Sigma}_j),$$

where $\boldsymbol{\Sigma}_j = \sigma^{2*}\mathbf{I}_{t_i} + \mathbf{W}_i\mathbf{D}_j\mathbf{W}_i'$, which resembles $\mathbf{V}_i^*$ except that it involves $\mathbf{D}_j$, not $\mathbf{D}^*$. For step 2 of the collapsed SIS, we draw the cluster label s_i from the discrete mass distribution,

$$\begin{aligned}\Pr(s_i = j|\mathbf{y}_{(i-1)}, \mathbf{s}_{(i-1)}^{(g)}, \mathbf{y}_i, \boldsymbol{\psi}^*) &= c\frac{n_{j,i-1}}{\alpha^* + i - 1}\mathrm{N}_{t_i}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}^* + \mathbf{W}_i\hat{\mathbf{b}}_j, \boldsymbol{\Sigma}_j), \qquad 1 \le j \le k_{i-1}\\ &= c\frac{\alpha^*}{\alpha^* + i - 1}\mathrm{N}_{n_i}(\mathbf{y}_i|\mathbf{X}_i\boldsymbol{\beta}^*, \mathbf{V}_i^*), \qquad j = k_{i-1} + 1.\end{aligned}$$

These steps are repeated M times; the average of $w = \prod u_i$ from each complete sweep through the observations is our estimate of the likelihood ordinate. We obtain the marginal likelihood of the semiparametric DPM model by inserting the posterior and likelihood ordinate estimates into (9).

Two Studies With Simulated Data. We investigate the efficacy of the Bayes factor for semiparametric model comparison in two simulated datasets. Both datasets include the covariate measurements of the first 200 patients in the original CD4 count. The responses $\mathbf{y}_i$, $i = 1, \ldots, 200$ in the first dataset are simulated in two stages. In particular, the random effects are simulated from the four-component bivariate normal mixture,

$$\mathbf{b}_1, \ldots, \mathbf{b}_{200} \overset{\text{iid}}{\sim} G = \sum_{l=1}^{4}\frac{1}{4}\mathrm{N}_2(\boldsymbol{\mu}_l, \mathbf{D}^*),$$

with overall mean $(1/4)\sum\boldsymbol{\mu}_l = \mathbf{0}$. Next, $\mathbf{y}_i$ is simulated from the mixed-effects model in (32). The covariance matrix $\mathbf{D}^*$ is fixed at the posterior mean of $\mathbf{D}$ based on the complete original data.

Due to the mixture structure in the data-generation model, we expect the semiparametric DPM model $\mathcal{M}_3$ to provide a better fit than the Gaussian model $\mathcal{M}_1$, and proceed to make a model comparison via the Bayes factor. We obtain the marginal likelihood of the DPM model $\mathcal{M}_3$ using the methodology described above and estimate the marginal likelihood of the parametric Gaussian model $\mathcal{M}_1$ from the MCMC algorithm of Chib and Carlin (1999), as described by Chib and Jeliazkov (2001). These estimates and the resulting Bayes factor are listed in Table 3. The Bayes factor clearly supports the DPM model $\mathcal{M}_3$ and provides "very strong" evidence (according to the scale described in Sec. 3) against the parametric Gaussian model $\mathcal{M}_1$.

A common criticism of the Bayes factor is that it does not have an explicit penalty term for the additional dimensions of an extended model. This criticism is not valid. To provide an empirical illustration of how the Bayes factor supports parsimony, when parsimony is justified, we simulate the responses $\mathbf{y}_i$, $i = 1, \ldots, 200$ in our second dataset from the parametric Gaussian model $\mathcal{M}_1$ by first simulating the random effects

Table 3. Estimated Log Marginal Likelihood (Diagonal) and Log Bayes Factor of $\mathcal{M}_1$ Versus $\mathcal{M}_3$

	Random effects simulated from four-component mixture			*Random effects simulated from the normal model $\mathcal{M}_1$*	
	$\mathcal{M}_1$ *(normal)*	$\mathcal{M}_3$ *(DPM)*		$\mathcal{M}_1$ *(normal)*	$\mathcal{M}_3$ *(DPM)*
$\mathcal{M}_1$	−1,778.06		$\mathcal{M}_1$	−1,551.30	
$\mathcal{M}_3$	20.42	−1,757.64	$\mathcal{M}_3$	−10.74	−1,562.04

$\mathbf{b}_1, \ldots, \mathbf{b}_{200}$ from the bivariate normal distribution described in (32). Table 3 lists the estimated log-marginal likelihood of the Gaussian model $\mathcal{M}_1$ and the DPM model $\mathcal{M}_3$ for these data. We find that the Bayes factor selects the correct model and, more importantly, provides "very strong" evidence for the simple correct model $\mathcal{M}_1$ against its extended complex counterpart $\mathcal{M}_3$. Note that the marginal likelihood estimates for the two models are computed based on identical priors on the hyperparameters.

CD4 Count Data. We now turn to the original CD4 count data with $n = 467$ patients and compare the linear mixed-effects semiparametric DPM model $\mathcal{M}_3$ with the parametric Gaussian model $\mathcal{M}_1$ and the parametric Student t model $\mathcal{M}_2$ with 10 degrees of freedom. The Gaussian model has been considered by Chib and Jeliazkov (2001). Note that all three models have the same prior on the common parameters $\boldsymbol{\beta}$, σ^2, and $\mathbf{D}$.

Figure 3 illustrates the estimates of the log-marginal likelihood of the DPM model $\mathcal{M}_3$ obtained by combining the posterior ordinate estimate and the collapsed SIS-based likelihood ordinate estimate. For comparison, we also computed the marginal likelihood estimate using the basic SIS method described in Section 5.1. Significantly, our preferred collapsed SIS estimate stabilizes quickly and does not change much as the number of iterations is increased. In contrast, the estimate from the basic SIS approach tends toward the collapsed estimate, but evidently does not converge even after 100,000 iterations. We have seen similar behavior in other models that we have considered.

Finally, we evaluate the three models in terms of the marginal likelihoods and Bayes factors. We stress that comparing DPM models for clustered data in this fashion has not been possible until now. Table 4 shows that the estimated log Bayes factor in favor of the DPM model versus the Student t model is 93.424 and that versus the standard Gaussian model is 79.25, providing "very strong" support for the DPM model (on the scale of Sec. 3). These Bayes factor values can be interpreted as stating that model $\mathcal{M}_3$ is "very strongly" successful in predicting the observed data relative to model $\mathcal{M}_1$ or $\mathcal{M}_2$, or that the "weight of evidence" provided by the data in favor of $\mathcal{M}_3$ is "very strong" compared to $\mathcal{M}_1$ or $\mathcal{M}_2$.

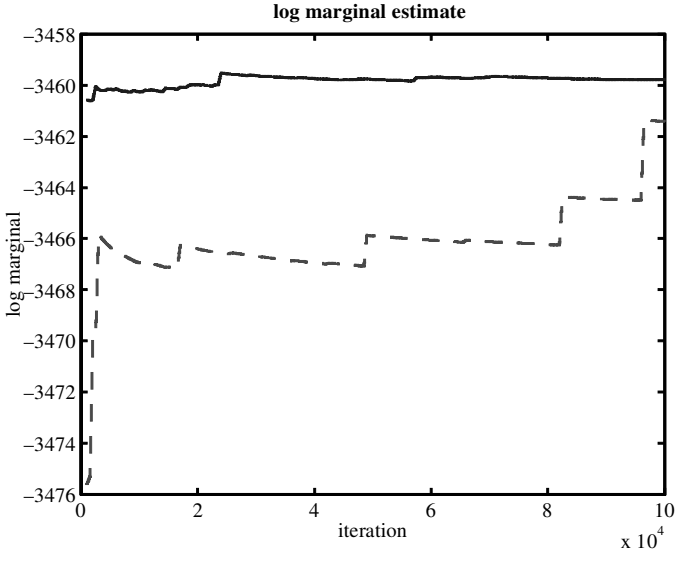

Figure 3. AIDS Data: The Marginal Likelihood Estimate of the DPM Model Versus Number of Iterations. The solid and dashed line represent estimates from the collapsed and basic sequential methods.

Table 4. Estimated Log Marginal Likelihoods (on the diagonal) for Each of Three Mixed Models for the AIDS Data

	$\mathcal{M}_1$ (DPM)	$\mathcal{M}_2$ (Student t link)	$\mathcal{M}_3$ (normal)
$\mathcal{M}_1$	−3,459.789	(93.424)	(79.25)
$\mathcal{M}_2$		−3,553.213	(−14.174)
$\mathcal{M}_3$			−3,539.039

NOTE: The entries in the upper half are the log of the Bayes factor in favor of the row model vs the column model.

7. CONCLUDING REMARKS

In this article we have developed and exemplified one of the first approaches for computing the marginal likelihood of a semiparametric DPM model. One virtue of the proposed technique, which relies on the approach of Chib (1995), is that it uses the programming done to simulate the posterior distribution of the DPM model and requires no further tuning of the MCMC algorithm. The only incidental coding needed is for the estimation of the likelihood ordinate at one fixed point, which is done by the sequential imputation method. Using a longitudinal normal regression DPM model where the value of marginal likelihood is known analytically, we have shown that our proposed estimate is accurate, stable, and efficient. The implementation and performance of the method have been further clarified in experiments involving semiparametric link models for binary response data and hierarchical mixed models for longitudinal data. Although the DPM model decisively dominates the parametric models in the longitudinal example, the binary response example leads to a different verdict. One may expect that with access to the proposed method, the practice of comparing semiparametric DPM models with parametric or other semiparametric models on the basis of marginal likelihoods and Bayes factors may become common.

[Received February 2002. Revised September 2002.]

REFERENCES

Albert, J., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of American Statistical Association*, 88, 669–679.

Basu, S., and Mukhopadhyay, S. (2000), "Binary Response Regression With Normal Scale Mixture Links," in *Generalized Linear Models: A Bayesian Perspective*, eds. D. K. Dey, S. K. Ghosh, and B. K. Mallick, New York: Marcel Dekker, pp. 231–242.

Berger, J. O., and Guglielmi, A. (2001), "Bayesian and Conditional Frequentist Testing of a Parametric Model Versus Nonparametric Alternatives," *Journal of American Statistical Association*, 96, 174–184.

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

Brown, B. W. (1980), "Prediction Analysis for Binary Data," in *Biostatistics Casebook*, eds. R. J. Miller, B. Efron, B. W. Brown, and L. E. Moses, New York: Wiley.

Bush, C. A., and MacEachern, S. N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 88, 275–285.

Carlin, B. P., and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, (2nd ed.), New York: Chapman Hall/CRC.

Carota, C., and Parmigiani, G. (1996), "On Bayes Factor for Nonparametric Alternatives," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, pp. 507–511.

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

Chib, S., and Carlin, B. (1999), "On MCMC Sampling in Hierarchical Longitudinal Models," *Statistics and Computing*, 9, 17–26.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327–335.

Chib S., and Jeliazkov, I. (2001), "Marginal Likelihood From the Metropolis–Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

Erkanli, A., Stangl, D., and Müller, P. (1993), "A Bayesian Analysis of Ordinal Data Using Mixtures," in *Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, pp. 51–56.

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

——— (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical 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.

——— (1983), "Bayesian Density Estimation By Mixtures of Normal Distributions," in *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, New York: Academic Press, pp. 287–302.

Florens, J. P., Richard, J. F. and Rolin, J. M. (1996), "Bayesian Encompassing Specification Tests of a Parametric Model Against a Nonparametric Alternative," technical report, University of Pittsburgh, Dept. of Economics.

Gelfand, A., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, 85, 972–985.

Good, I. J. (1985), "Weight of Evidence; A Brief Survey," in *Bayesian Statistics 2*, eds. J. M. Bernardo et al., New York: Elsevier, pp. 249–269.

Han, C., and Carlin, B. P. (2001), "MCMC Methods for Computing Bayes Factors: A Comparative Review," *Journal of the American Statistical Association*, 96, 1122–1132.

Irwin, M., Cox, N., and Kong, A. (1994), "Sequential Imputation for Multilocus Linkage Analysis," *Proceedings of the National Academy of Science USA*, 91, 11684–11688.

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

——— (2001b), "Generalized Weighted Chinese Restaurant Processes for Species Sampling Mixture Models," unpublished manuscript.

Ishwaran, H., James, L. F., and Sun, J. (2001), "Bayesian Model Selection in Finite Mixtures by Marginal Density Decomposition," *Journal of the American Statistical Association*, 96, 1316–1332.

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.

Kleinman, K. P., and Ibrahim, J. G. (1998), "A Semiparametric Bayesian Approach to the Random Effects Model," *Biometrics*, 54, 921–938.

Kong, A., Liu, J. S., and Wong, W. H. (1994), "Sequential Imputations and Bayesian Missing Data Problems," *Journal of the American Statistical Association*, 89, 278–288.

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

Liu, J. S. (1996), "Nonparametric Hierarchical Bayes via Sequential Imputations," *The Annals of Statistics*, 24, 911–930.

——— (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer-Verlag.

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

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

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

MacEachern, S. N., and Müller, P. (1998), "Estimating Mixtures of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

Newton, M. A., Czado, C., and Chappell, R. (1996), "Bayesian Inference for Semiparametric Binary Regression," *Journal of the American Statistical Association*, 91, 142–153.

Quintana, F. A. (1998), "Nonparametric Bayesian Analysis for Assessing Homogeneity in $k \times L$ Contingency Tables With Fixed Right Margin Totals," *Journal of the American Statistical Association*, 93, 1140–1149.

Quintana, F. A., and Newton, M. (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.

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

Tao, H., Palta, M., Yandell, B. S., and Newton, M. A. (1999), "An Estimation Method for the Semiparametric Mixed Effects Model," *Biometrics*, 55, 102–110.