Some Issues in Nonparametric Bayesian Modeling Using Species Sampling Models

Carlos Navarrete[1], Fernando A. Quintana[1], and Peter Müller[2],

[1]Departamento de Estadística, Facultad de Matemáticas,
Pontificia Universidad Católica de Chile, Santiago, CHILE

[2]Department of Biostatistics & Applied Mathematics, The University of Texas
M. D. Anderson Cancer Center, Houston, Texas, U.S.A.

Abstract

We review some aspects of nonparametric Bayesian data analysis with discrete random probability measures. We focus on the class of species sampling models (SSM). We critically investigate the common use of the Dirichlet process (DP) prior as a default SSM choice. We discuss alternative prior specifications from a theoretical, computational and data analysis perspective. We conclude with a recommendation to consider SSM priors beyond the special case of the DP prior and make specific recommendations on how different choices can be used to reflect prior information and how they impact the desired inference. We show the required changes in the posterior simulation schemes, and argue that the additional generality can be achieved without additional computational effort.

1 Introduction

Bayesian methods have become increasingly popular over the past few decades. This has been largely due to work by Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990), which revolutionized the way Bayesian inference is carried out. The availability of Markov Chain Monte Carlo (MCMC) simulation-based

methods allows researchers to implement Bayesian inference in problems that were otherwise intractable. As an example, the technically convenient restriction to conjugate posterior analysis (Bernardo and Smith 1994), has been mostly replaced by the use of richer families of prior distributions that reflect subjective beliefs and/or prior information more accurately. In recent years a similar revolution in Bayesian data analysis has been related to the increasing use of non-parametric models.

When the probability model is assumed to belong to a family of candidate distributions that can be indexed by a d-dimensional parameter vector, a *parametric* Bayesian model is obtained. In many cases, however, this could be judged to be too restrictive, particularly in the case where some of the motivating questions in the study have answers that critically depend on the specific parametric form chosen. A richer and potentially more realistic class of models is obtained by letting the family of prior distributions be indexed by an infinite-dimensional hyperparameter. Such constructions are typically used to express uncertainty on a distribution function. Probability models for infinite-dimensional random parameters are known as non-parametric Bayesian models. Data analysis based on such models is known as *non-parametric Bayesian inference*. When used to describe uncertainty about a distribution function, the models are called *random probability measures* (RPMs) and can be thought of as probability distributions defined on the space of distribution functions.

It is customary to make a distinction between *continuous* and *discrete* nonparametric models, depending on whether the RPMs generate distributions almost surely supported on the set of continuous or discrete distributions. The distinction, however, is not a clear-cut separation. Some of the continuous models admit discrete distributions as (extreme) special cases. Such is the case, for example, for the Polya trees discussed in Lavine (1992, 1994). See also Hanson and Johnson (2002), Paddock et al. (2003) and Hanson (2006). Additional RPMs of the continuous type include beta processes (Hjort 1990), beta-Stacy processes (Walker and Muliere 1997), extended and weighted Gamma processes (Dykstra and Laud 1981, Nieto-Barajas and Walker 2002, Ishwaran and James 2004), Gaussian processes (O'Hagan 1992, Angers and Delampady 1992), random Bernstein polynomials (Petrone, 1999a, 1999b), and logistic normal processes (Lenk 1988). For a review of these and other nonparametric

priors and additional references, see Dey, Müller and Sinha (1998), Walker et al. (1999), Ghosh and Ramamoorthi (2003), and Müller and Quintana (2004).

The focus of this article is on discrete RPMs and their applications. Specifically, we consider the class of RPMs that can be represented as

$$F(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\mu_h}(\cdot), \tag{1.1}$$

where $w_1, w_2, \ldots$ are weights that verify $0 \leq w_h \leq 1$ for all h, $P(\sum_h w_h = 1) = 1$ and $\mu_1, \mu_2, \ldots$ are random locations of the point masses, independent of the $\{w_h\}$ collection.

The rest of this article is organized as follows. Section 2 reviews the class of species sampling models (SSMs), the main focus of this article, discussing some special and well-known cases. We emphasize the clustering structure induced by the discreteness of SSMs. Section 3 discusses some issues underlying nonparametric Bayesian models using SSMs. Aspects of MCMC implementation are described in Section 4. Section 5 illustrates the ideas in an example. In particular, we explore sensitivity to various specifications of the prior SSM. Some final remarks are given in Section 6.

2 Species Sampling Models and Some Special Cases

A very flexible collection of models based on discrete RPMs is given by the class of *species sampling models* (SSMs), discussed by Pitman (1996), Ishwaran and James (2003a) and Quintana (2006). Suppose a random sample $\theta_1, \theta_2, \ldots$ is collected from a large population and that θ_n is the tag assigned to the species of the n-th selected individual. We use the term species to generically identify the variable θ_i, keeping in mind that θ_i need not literally identify a biological species. Having observed $\theta_1, \ldots, \theta_n$, denote by $\theta_1^*, \ldots, \theta_k^*$ the unique values that have been recorded, in order of appearance, where $k \equiv k(n)$ is the number of such items. Thus, θ_j^* represents the j-th sampled species. Let the indicators $\boldsymbol{s} = (s_1, \ldots, s_n)$ be defined as $\theta_i = \theta_{s_i}^*$ for $i = 1, \ldots, n$. The set of observations sharing a common tag value is referred to as a *cluster*. The clusters define a partition of $\{1, \ldots, n\}$ as $\rho = (S_1, \ldots, S_k)$, where $S_j = \{i \in \{1, \ldots, n\} : s_i = j\}$. We explicitly note the implied sequential order, i.e. clusters are numbered

consecutively, with $1 \in S_1$ and there is a "no-gaps" restriction in the sense that $s_i = \ell > 1$ for some i implies that the set of unique values of $s_1, \ldots, s_i$ is exactly $1, \ldots, \ell$. The cluster sizes are denoted by $m_j(n) = \sum_{i=1}^{n} I\{\theta_i = \theta_j^*\}$, where $I\{A\} = 1$ if A occurs and 0 otherwise. Let F_0 be a fixed probability model (baseline). Consider a prediction rule of the form $P(\theta_1 \in B) = F_0(B)$ and

$$P(\theta_{n+1} \in B \mid \theta_1, \ldots, \theta_n) = \sum_{j=1}^{k(n)} \rho_j(\boldsymbol{m}(n))\, \delta_{\theta_j^*}(B) + \rho_{k(n)+1}(\boldsymbol{m}(n))\, F_0(B), \tag{2.1}$$

where $\boldsymbol{m}(n) = (m_1(n), \ldots, m_{k(n)}(n))$ is the vector of cluster sizes, and the collection of functions $\rho_j(\cdot)$ are weights that add up to 1, i.e., $0 \leq \rho_j(\boldsymbol{m}(n)) \leq 1$ and $\sum_{j=1}^{n+1} \rho_j(\boldsymbol{m}(n)) = 1$ for all n and $\boldsymbol{m}(n)$. To avoid practical complications (see Section 3 below) we will assume F_0 to be continuous. The predictive distribution (2.1) is a mixture of point-masses at the already observed tags and the baseline distribution F_0. The weights of the mixture are given by the $\rho_j(\boldsymbol{m}(n))$ functions. Assuming that the sequence $\{\theta_n\}$ is exchangeable, it follows (Pitman 1996) that $F_{n+1}(B) = P(\theta_{n+1} \in B \mid \theta_1, \ldots, \theta_n)$ as defined in (2.1) converges in variation norm to a RPM F of the form

$$F(B) = \sum_{h=1}^{\infty} w_h \delta_{\theta_h^*}(B) + \left(1 - \sum_{h=1}^{\infty} w_h\right) F_0(B), \tag{2.2}$$

where the $\{\theta_h^*\}$ are i.i.d draws from F_0, independent of $\{w_h\}$, and w_h represents the weight of the h-th species to appear. Also, $P(\lim_{n\to\infty} m_j(n)/n = w_j) = 1$ for all $j \geq 1$ and given F, $\{\theta_n\}$ is a random sample from F. A RPM of the form (2.2) together with a random sample $\{\theta_n\}$ from F is referred to as SSM. The continuity of F_0 implies that (2.2) defines a discrete distribution provided that $P(\sum_h w_h = 1) = 1$. Furthermore, if the predictive probabilities $\{\rho_j\}$ are such that $P(\lim_{n\to\infty} k(n)/n = 0) = 1$ then $P(\sum_h w_h = 1) = 1$. In this case the model is said to be *proper*, and the RPM is of the form (1.1). This property depends primarily on the specification of $\{\rho_j\}$. Thus, SSMs can be defined by means of the predictive probabilities $\rho_j(\boldsymbol{m}(n))$ and the baseline distribution F_0. The choice of ρ_j is not arbitrary. See below. The almost sure convergence result $P(\lim_{n\to\infty} m_j(n)/n = w_j) = 1$ allows us to give an interpretation of the weights w_h in (1.1) as limit proportions of the recorded tags. Letting ρ generically denote the collection of predictive probabilities $\{\rho_j\}$, we use the

notation $F \sim \text{SSM}(\rho, F_0)$ to designate the limit RPM in (2.2). It follows (Pitman 1996) that for any SSM, marginalizing over the RPM F leads to a joint distribution $p(\boldsymbol{\theta})$ that can be expressed as the product of conditional distributions as in (2.1). This observation will turn out to be quite important later in Section 4. See additional general properties of SSMs in Pitman (1996).

A pragmatic use of SSMs might involve specifying the predictive probabilities and carrying out the analysis without ever worrying about the specific form of the limit RPM (2.2). However, one must make sure that the resulting sequence $\{\theta_n\}$ is indeed exchangeable. To do so, it is useful to consider the *exchangeable partition probability function* (EPPF), defined for a partition $\rho = (S_1, \ldots, S_k)$ as

$$P\left(\bigcap_{j=1}^{k}\{\theta_\ell = \theta_j^* \text{ for all } \ell \in S_j\}\right) = p(\boldsymbol{m}(n)). \tag{2.3}$$

If $\{\theta_n\}$ is exchangeable then $p(\cdot)$ is a symmetric function of all possible k-tuples of positive integers summing up to $n \geq 1$, constrained to satisfy a *coherence* condition

$$p(1) = 1 \qquad \text{and} \qquad p(\boldsymbol{m}(n)) = \sum_{j=1}^{k(n)+1} p(\boldsymbol{m}(n)^{+j}).$$

Here, $\boldsymbol{m}(n)^{+j}$ represents $\boldsymbol{m}(n)$ with the j-th component increased by 1 (Pitman 1996). EPPFs are important because SSMs can be alternatively defined by an EPPF plus the baseline measure F_0. In this case, the predictive probability functions are easily shown to be given by

$$\rho_j(\boldsymbol{m}(n)) = \frac{p(\boldsymbol{m}(n)^{j+})}{p(\boldsymbol{m}(n))}, \qquad 1 \leq j \leq k(n)+1.$$

An interesting example of a EPPF is that corresponding to the Pitman-Yor (PY) process (Pitman and Yor 1987). Let $[x]_\ell = \prod_{j=1}^{\ell}(x+j-1)$. The EPPF for the PY process is

$$p_{\alpha,M}(m_1(n), \ldots, m_{k(n)}(n)) = \frac{\left(\prod_{j=1}^{k-1}(M+j\alpha)\right)\left(\prod_{j=1}^{k(n)}[1-\alpha]_{m_j(n)-1}\right)}{[1+M]_{n-1}},$$

for $\alpha = -\kappa$ and $M = \ell\kappa$ for some $\kappa > 0$ and $\ell = 2, 3, \ldots$, *or* $0 \leq \alpha < 1$ and $M > -\alpha$. The corresponding predictive probability functions are given by

$$\rho_j(\boldsymbol{m}(n)) = \frac{1}{M+n}\begin{cases} m_j(n) - \alpha & \text{if } 1 \leq j \leq k(n) \\ M + k(n)\alpha & \text{if } j = k(n)+1. \end{cases} \tag{2.4}$$

The special case $\alpha = 0$ and $M > 0$ corresponds to the Dirichlet process (DP) introduced by Ferguson (1973). Sethuraman (1994) showed that the DP corresponds to (1.1) with $\mu_1, \mu_2, \ldots$ being i.i.d. from a baseline distribution F_0 and with weights defined as $w_1 = V_1$ and $w_h = \prod_{j=1}^{h-1}(1 - V_j)V_h$ for $h > 1$, where $V_1, V_2, \ldots \overset{iid}{\sim} Beta(1, M)$, for some $M > 0$ called *total mass parameter*. We denote it as $F \sim DP(M, F_0)$ and note that the corresponding predictive probabilities reduce to

$$\rho_j(\boldsymbol{m}(n)) = \frac{1}{M+n} \begin{cases} m_j(n) & \text{if } 1 \leq j \leq k(n) \\ M & \text{if } j = k(n) + 1. \end{cases}$$

These probabilities have been interpreted as the *Chinese restaurant process* by Arratia, Barbour and Tavaré (1992).

Due to its simplicity and the availability of efficient posterior simulation schemes (Bush and MacEachern 1996, MacEachern and Müller 1998, Neal 2000, Jain and Neal 2004), the DP is the most popular non-parametric Bayesian model. DP-based models have been considered in an ever growing range of applications and settings. See details on computation in Section 4 below. A survey of applications involving the DP can be found in MacEachern and Müller (2000). Some recent applications include accelerated failure time (AFT) models (Hanson and Johnson 2004), analysis of developmental toxicology data (Dominici and Parmigiani 2001), animal breeding (van der Merwe and Pretorius 2003), competing risks models (Tiwari et al. 1997), homogeneity assessments in contingency tables (Kuo and Yang 2006), linear mixed models (Ishwaran and Takahara 2002), median regression (Kottas and Gelfand 2001), meta-analysis (Müller et al. 2004, Doss and Burr 2005), modeling differential gene expression (Dahl 2003, Do et al. 2005), modeling matched case-control studies (Sinha et al. 2004), multivariate ordinal data analysis (Kottas et al. 2005), regression for count data (Carota and Parmigiani 2002), spatial modeling (Gelfand et al. 2004), sports data (Quintana and Müller 2004), and many others. Discussions of properties and applications of DPs can be found in Ferguson (1973), Korwar and Hollander (1973), Antoniak (1974), Diaconis and Freedman (1986), Cifarelli and Regazzini (1990), Rolin (1992), Diaconis and Kemperman (1996), Florens et al. (1999), Cifarelli and Melilli (2000), Regazzini, Guglielmi and Di Nunno (2002), James (2005), Hanson et al. (2005), Hjort

and Ongaro (2005) and in other references therein.

The class of SSM models admits several other important special cases. These include the Dirichlet-multinomial process (Muliere and Secchi 1995), the beta-two process (Ishwaran and Zarepour 2000) and the stick-breaking priors (Ishwaran and James 2001, 2003b). Additional properties of SSMs can be found in Pitman (1996). For a related class of RPMs see Lijoi, Mena and Prünster (2006).

3 Nonparametric Bayesian Modelling

A primary motivation for the use of nonparametric Bayesian models is their inherent flexibility. The SSM allows extensive flexibility in the specification of the predictive rule (2.1). However, there are limitations to the use of the SSM in data analysis. One important limitation arises from the fact that (2.1) implies positive probability for ties. This makes it inappropriate, for example, to use a SSM prior for density estimation with continuous outcomes. This limitation motivated Ferguson (1983) to consider mixtures of DPs, i.e., a convolution of a DP RPM F with a continuous kernel. This approach became very popular for non-parametric Bayesian data analysis. (see, e.g. Lo 1984, Escobar 1988, MacEachern 1994, Escobar and West 1995). Extending this idea to SSMs, a typical non-parametric model based on SSMs is expressed as

$$X_i \sim F(x) = \int p(x \mid \theta)\, dG(\theta), \tag{3.1}$$

where $G \sim \text{SSM}(\rho, G_0)$, for a suitable continuous kernel $p(x \mid \theta)$ and a continuous distribution G_0. The continuity of G_0 guarantees that clusters are defined only in terms of the different values sampled from G_0 (i.e. the tags); otherwise different clusters may have a common tag value, complicating the interpretation. Model (3.1) can be thought of as the mixture of a continuous distribution with respect to a discrete RPM. The mixture generates a continuous distribution $F(x)$. Introducing latent parameters θ_i, the mixture can be written as a hierarchical model:

$$X_i \mid \theta_i \overset{ind}{\sim} p(X_i \mid \theta_i), \qquad \theta_i \overset{iid}{\sim} G, \qquad G \sim \text{SSM}(\rho, G_0). \tag{3.2}$$

Optionally, G_0 and/or the likelihood may include additional hyperparameters. The flexibility of a model like (3.1) is reflected in the fact that many distributions can be

well approximated by such a construction. For instance, Lo (1984) points out that the closure of the family of distributions defined as $F(x) = \int p(\tau(x-\mu))\, dG(\mu,\tau)$ with $G \sim DP(M, G_0)$ contains all continuous distributions on the real line for reasonable choices of p and G_0. A similar result can be shown to hold for the more general (proper) SSMs with RPMs as in (1.1).

The latent variables θ_i in (3.2) are interpreted as subject-specific random effects. The use of these random effects allows for great flexibility but also entails complexity. The SSM prior assumption includes the possibility of ties among the θ_i values. In practice this means that *a priori* some individuals share a common parameter value θ_j^*, their differences being explained by sampling variability in the likelihood.

From a data analysis perspective it is interesting to note that model (3.2) allows for two extreme cases: all the θ_i parameters are equal, and all of them are distinct, reducing inference to parametric models. But more generally, a discrete RPM prior for the unknown distribution represents an intermediate choice between models with all parameters equal or different. By adequately choosing the predictive probabilities $\{\rho_j\}$ the analyst can favor different partition structures. In the DP case, for instance, a large value of M implies many clusters, while small values of M favor a reduced number of clusters. The prior expectation and variance of the number of clusters are given by (Liu 1996)

$$\sum_{i=1}^{n} \frac{M}{M+i-1} \qquad \text{and} \qquad \sum_{i=1}^{n} \frac{M(i-1)}{(M+i-1)^2}.$$

The extreme cases mentioned earlier follow by letting $M \to 0$ and $M \to \infty$, respectively. Some authors (e.g. Escobar and West 1995) treat M as an unknown parameter itself, choosing a prior distribution (usually Gamma) to reflect uncertainty. The above expressions for prior mean and variance can be used for prior elicitation purposes (Kottas et al. 2005). In contrast, the PY process has a more flexible partition structure, where an increased number of clusters may be attained by increasing within the valid ranges either M or α.

In a typical application of RPM priors based on SSMs one would consider a model like (3.2) regarding the θ_i parameters as random effects. The posterior process of G would seldom be of interest itself. In that sense, the modeling flexibility would

be seen as a useful device to account for variability among subjects and to model the heterogeneity of the underlying population of experimental units. An important consequence of this is that other population-level parameters or hyperparameters can be better understood, because much of the underlying noise has thus been considered.

4 Posterior Computation

A crucial aspect of any modeling effort is the availability of efficient posterior simulation schemes that facilitate the implementation. This is particularly true for nonparametric Bayesian models. To fix ideas, consider the following variation of model (3.2)

$$X_i \mid \beta, \theta_i \overset{ind}{\sim} p(X_i \mid \beta, \theta_i), \qquad \theta_i \overset{iid}{\sim} G, \qquad G \mid \phi \sim \mathrm{SSM}(\rho, G_0(\cdot \mid \phi)), \tag{4.1}$$

where $\beta \sim p(\beta)$ and $\phi \sim p(\phi)$ represent additional parameters and hyperparameters in the likelihood and baseline distribution, respectively. The need for such additional (hyper-)parameters is a very common feature in applied data analysis. Assume for now that the likelihood $p(x \mid \beta, \theta)$ is conjugate with respect to $G_0(\cdot \mid \phi)$. Let $g_0(\cdot \mid \phi)$ denote the density of G_0 for any fixed value of ϕ.

We describe an implementation of posterior inference by MCMC simulation. The MCMC scheme is described by specifying the Markov chain transition probabilities. Conditional on currently imputed values for all other parameters and the data we describe how each of the parameters in the model is updated.

The conditional posterior distribution of θ_i given all other model parameters can be explicitly derived to find a mixture between point masses at the already imputed θ_j, $j \neq i$ and a distribution $g_i(\theta_i \mid \beta, \phi) \propto p(X_i \mid \beta, \theta_i) g_0(\theta_i \mid \phi)$. The assumed conjugacy of the sampling model and the base measure g_0 imply that $g_i(\cdot)$ can be worked out analytically. By the exchangeability of $\theta_1, \ldots, \theta_n$ built in the SSMs, we find that

$$\begin{aligned} \theta_i \mid \boldsymbol{\theta}_{-i}, \beta, \phi, \boldsymbol{X} \propto & \sum_{j=1}^{k^-} \rho_j(\boldsymbol{m}(n-1)) p(X_i \mid \theta_j^{*-}, \beta) \delta_{\theta_j^{*-}}(\theta_i) \\ & + \rho_{k^-+1}(\boldsymbol{m}(n-1)) \left[\int p(X_i \mid \theta, \beta) g_0(\theta \mid \phi)\, d\theta \right] g_i(\theta_i \mid \beta, \phi), \end{aligned} \tag{4.2}$$

where $\boldsymbol{X} = (X_1, \ldots, X_n)$, $\boldsymbol{\theta}_{-i} = (\theta_j)_{j \neq i}$, and $k^- \equiv k(n)^-$ and $\theta_1^{*-}, \ldots, \theta_{k^-}^{*-}$ represent the number of clusters and the locations left after removing the i-th observation.

A basic Gibbs sampler can then be defined as follows:

Step 1. *Updating locations* $\boldsymbol{\theta}$: For $i = 1, \ldots, n$ resample θ_i from (4.2). After sampling each θ_i, update $k(n)$, $\rho_j(\boldsymbol{m}(n))$, $j = 1, \ldots, k(n) + 1$ and $\theta_1^*, \ldots, \theta_k^*$.

Note that updating $\boldsymbol{\theta}$ implicitly changes $\boldsymbol{s}$ and $\boldsymbol{\theta}^\star$.

Step 2. *Resampling locations*: As noted by Bush and MacEachern (1996) it is advisable to include a step to resample $\boldsymbol{\theta}^\star$ conditional on the imputed configuration $\boldsymbol{s}$. This greatly increases the mixing of the Markov chain. Therefore, conditional on $\boldsymbol{s}$, we consider draws

$$\theta_j^* \sim p(\theta_j^* \mid \ldots) \propto \prod_{i \in S_j} p(X_i \mid \beta, \theta_j^*) g_0(\theta_j^* \mid \phi), \qquad j = 1, \ldots, k,$$

which can be obtained exactly by the conjugacy assumptions. Then define $\theta_i = \theta_{s_i}^*$, $i = 1, \ldots, n$.

Step 3. *Updating the remaining parameters*: The remaining parameters are resampled according to their complete conditional posterior distributions:

$$p(\beta \mid \ldots) \propto \prod_{i=1}^{n} p(X_i \mid \beta, \theta_i) p(\beta) \qquad \text{and} \qquad p(\phi \mid \ldots) \propto \prod_{j=1}^{k} g_0(\theta_j^* \mid \phi) p(\phi).$$

A Metropolis within Gibbs step might be required in case any of these is not available in closed form.

However, it is usually even more efficient, in the sense of a faster mixing Markov chain, to consider updating the configurations vector $\boldsymbol{s}$ with respect to a reduced model resulting after analytically marginalizing with respect to the θ_i parameters (MacEachern and Müller 2000). Representing the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ as $(\boldsymbol{\theta}^*, \boldsymbol{s})$, and marginalizing with respect to the RPM G we find that the joint distribution of $(\boldsymbol{X}, \boldsymbol{\theta}^*, \boldsymbol{s}, \beta, \phi)$ is

$$p(\boldsymbol{X}, \boldsymbol{\theta}^*, \boldsymbol{s}, \beta, \phi) = \prod_{j=1}^{k(n)} \left\{ \prod_{i \in S_j} p(X_i \mid \beta, \theta_j^*) \right\} \cdot \prod_{j=1}^{k(n)} g_0(\theta_j^* \mid \phi) \cdot p(\boldsymbol{s}) p(\beta) p(\phi). \tag{4.3}$$

The joint prior $p(\boldsymbol{s})$ is evaluated by writing $p(\boldsymbol{s}) = p(s_1)\prod_{i=2}^{n} p(s_i \mid s_1, \ldots, s_{i-1})$ and noting that the definition of SSM implies $P(s_1 = 1) = 1$ and

$$P(s_i = \ell \mid s_1, \ldots, s_{i-1}) = \rho_\ell(\boldsymbol{m}(i-1)).$$

Exploiting exchangeability we note that $P(s_i = \ell \mid \boldsymbol{s}_{-i} = \tilde{\boldsymbol{s}}) = P(s_n = \ell \mid \boldsymbol{s}_{-n} = \tilde{\boldsymbol{s}})$, i.e., the complete conditional prior for the i-th configuration indicator is the same as for the last, n-th configuration indicator. For instance, in the case of the PY process, the latter is just (2.4) substituting $n-1$ for n. The assumed conjugacy of the sampling model and the base measure allows us to analytically marginalize with respect to $\boldsymbol{\theta}^*$ in (4.3), leading to

$$p(\boldsymbol{X}, \boldsymbol{s}, \beta, \phi) = \prod_{j=1}^{k(n)} p(\boldsymbol{X}_{S_j} \mid \beta) p(\boldsymbol{s}) p(\beta) p(\phi), \tag{4.4}$$

where $\boldsymbol{X}_{S_j} = (X_i : i \in S_j)$ and $p(\boldsymbol{X}_{S_j} \mid \beta) = \int \prod_{i \in S_j} p(X_i \mid \beta, \theta_j^*) g_0(\theta_j^*)\, d\theta_j^*$. In summary, posterior updating proceeds by replacing Step 1 in the basic algorithm by Step 1′ below:

Step 1′ *Updating cluster indicators* $\boldsymbol{s}$: For $i = 1, \ldots, n$ update s_i from the conditional distribution

$$P(s_i = \ell \mid \boldsymbol{s}_{-i}, \beta, \phi) = \begin{cases} p(X_i \mid \boldsymbol{X}_{-i}, s_i = \ell, \beta)\rho_\ell(\boldsymbol{m}(n-1)) & \text{if } \ell = 1, \ldots, k^- \\ p(X_i \mid \beta)\rho_{k^-+1}(\boldsymbol{m}(n-1)) & \text{if } \ell = k^- + 1, \end{cases} \tag{4.5}$$

where $k^- \equiv k(n)^-$ is the number of clusters left after removing the i-th observation. When recording the newly imputed value s_i by generating from (4.5) note the following convention about labelling:

(a) If $m_{s_i}(n) > 1$ then we resample s_i directly from (4.5).

(b) If $m_{s_i}(n) = 1$ then removing the i-th observation also eliminates a cluster. Consequently, we first set k^- as $k^- - 1$ and relabel clusters so as to avoid gaps and keep the ordering, and then resample according to (4.5).

Also, we note $p(X_i \mid \boldsymbol{X}_{-i}, s_i = \ell, \beta)$ depends on $\boldsymbol{X}_{-i}$ only through those observations with index in $S_\ell^- = (t : t \in S_\ell - \{i\})$, that is

$$p(X_i \mid \boldsymbol{X}_{-i}, s_i = \ell, \beta) = p(X_i \mid \boldsymbol{X}_{S_\ell^-}, s_i = \ell, \beta) = \frac{p(\boldsymbol{X}_{S_\ell^- \cup \{i\}} \mid \beta)}{p(\boldsymbol{X}_{S_\ell^-} \mid \beta)},$$

which by hypothesis can be evaluated analytically.

The algorithms just described can be modified and/or adapted in a number of ways. For instance, in the special case of assuming the RPM to be a finite-dimensional version of stick-breaking priors, the blocked Gibbs sampling described in Ishwaran and James (2001) may be used. Other alternative approaches to various cases are sequential importance sampling (Liu 1996, MacEachern et al. 1999, Quintana and Newton 2000, Ishwaran and James 2003a) and Metropolis-Hastings moves (Neal 2000, Dahl 2003, Jain and Neal 2004, Dahl 2005).

The algorithm relies on the conjugacy of the sampling model and the base measure g_0. In the absence of this conjugacy, the algorithm by MacEachern and Müller (2000) can be used, for example. Alternatively, Dahl (2005) has described a nonconjugate version of his Metropolis-Hastings algorithm.

5 Illustrations

We illustrate some of the concepts discussed earlier using the galaxy dataset presented in Roeder (1990). The same example has been analyzed by a number of authors, including Escobar and West (1995), Richardson and Green (1997), Stephens (2000), Ishwaran and James (2003a) and Quintana (2006). The dataset includes $n = 82$ measured velocities (in 10^3 km/s), relative to our own galaxy, of galaxies from six well-separated conic sections in space. We consider a model based on SSMs using various choices of predictive probabilities $\{\rho\}$ defining special cases of the PY process. We assess the impact of such choices on the density estimate, given by the posterior predictive density $p(X_{n+1} \mid X_1, \ldots, X_n)$.

We assume the model

$$\begin{aligned} X_i \mid (\mu_i, V_i) &\sim N(\mu_i, V_i) \\ (\mu_1, V_1), \ldots, (\mu_n, V_n) \mid F &\sim F \\ F &\sim PY(\alpha, M, F_0(\phi)), \end{aligned}$$

where $F_0(\phi)$ is a normal-inverse Gamma distribution with hyperparameters ϕ. That is, $(\mu, V) \sim F_0(\phi)$ means $\mu \mid V \sim N(\eta, \tau V)$ and $V^{-1} \sim \Gamma(r/2, R/2)$ with $\phi = (\eta, \tau)$ and assuming r and R to be fixed. Finally, we assume the components of ϕ to be a priori independent, with $\eta \sim N(m_0, V_0)$ and $\tau^{-1} \sim \Gamma(w/2, W/2)$ for known values of m_0, V_0, w and W.

The PY predictive probabilities are given in (2.4). In what follows we assign the following values to the hyperparameters: $m_0 = 20$, $V_0 = 1000$, $w = 1$, $W = 100$, $r = 4$ and $R = 2$, matching the choices in Escobar and West (1995), except for the prior assumptions for M. We start by considering the case $\alpha = 0$ and $M > 0$, i.e. the specification that leads to the DP. We alternatively treat M as a fixed and known constant or as an additional parameter, with prior density given by $M \sim \Gamma(a_0, b_0)$, so that the prior mean and variance are a_0/b_0 and a_0/b_0^2. The posterior predictive density is presented in Figure 1 for each of four cases: (i) M=1 fixed; (ii) $a_0 = 2$ and $b_0 = 4$, matching the choices in Escobar and West (1995); (iii) $a_0 = 10$ and $b_0 = 2$; and (iv) $a_0 = 20$ and $b_0 = 2$. In (ii) through (iv) the prior mean increases from 1/2 to 10, thus favoring the creation of more clusters. In all cases the resulting distributions can be interpreted as mixtures, concentrated on essentially the same region, but with weights that vary slightly according to the prior. In that sense, model (iv) appears to make a more marked distinction between the two central mixture components. In fact, the corresponding posterior density of M for models (ii) through (iv), presented in Figure 2, simply reflect a kind of stochastic monotonicity of the posterior with respect to the prior. Nevertheless, the impact of such prior definitions, including the extreme case (i) for which $P(M = 1) = 1$, seems minor.

Next we consider specifications of PY predictive probabilities with the following parameter choices: (i) $\alpha = 0$ and $M = 1$ (the usual DP); (ii) $\alpha = 0.3$ and $M = 1$; (iii) $\alpha = 0.9$ and $M = 1$; and (iv) $\alpha = 0.9$ and $M = 5$. By increasing α we decrease the

predictive probability of joining an already formed cluster (uniformly across cluster), while increasing the probability of creating new clusters. The posterior predictive densities are shown in Figure 3, left panel, together with the corresponding histograms for the posterior distribution of number of clusters (right panel). The effect of increasing α is clearly reflected in the posterior distribution of $k(n)$. However, it is also found that increasing α tends to slightly oversmooth the posterior predictive density.

Finally, we consider the effect of moving both parameters as follows: (i) $\alpha = -1$ and $M = 2$; (ii) $\alpha = -1$ and $M = 4$; and (iii) $\alpha = -5$ and $M = 20$. These represent some widely varying specifications. The corresponding posterior predictive and posterior distribution for the number of clusters are presented in Figure 4. We observe some substantial variability in the posterior predictive densities. Unlike all other cases, choice (i) produces unimodal shapes, and case (iii) nearly misses the rightmost mode, which is best captured by case (ii). The effect on the posterior distribution of the number of clusters is also dramatic. All these choices lead to a distribution that is strongly concentrated on a few small values.

6 Discussion

We have reviewed some aspects of nonparametric Bayesian modeling, specifically in the context of discrete random probability measures as contained in the class of species sampling models. We discussed some critical considerations in the modeling process, together with appropriate posterior simulation schemes for the conjugate case.

We focused our comparison on the special case of the PY process, which contains the DP as a particular case, exploring the effect of various specifications on density estimation. Our results suggest that the posterior predictive density is strongly influenced by such specifications, as reflected by the number of implied mixture components. The problem of formally selecting or estimating the number of such components is beyond the scope of this article. A possible solution can be implemented via the clustering algorithms discussed in Quintana (2006). For the purpose of data analysis with SSM priors in general and with PY priors in particular, we recommend to select the predictive probabilities on the basis of the desired number of clusters and the relative sizes of

clusters. For example, if the goal is to identify a large subpopulation of normal cases and smaller subsets of outliers, then the choice of a prior that favors a skewed distribution of cluster sizes is reasonable. If, however, the intention is to identify subpopulations that are a priori expected to be likely of comparable size, then we recommend to use a prior that favors approximately equal cluster sizes. The choice of the hyperparameters α and M in the PY process control the a priori expected number of clusters. In summary, we recommend against using the DP as an unreflected default choice. Instead we suggest to use the flexibility of SSM priors to reflect prior information about the number and nature of clusters. The additional flexibility allows the analyst to improve the data analysis by including more relevant prior information. This can be done with essentially no additional computational cost.

Acknowledgments

This work was partially funded by grant FONDECYT 1060729.

References

Angers, J.-F. and Delampady, M. (1992). Hierarchical Bayesian curve fitting and smoothing, *The Canadian Journal of Statistics* **20**: 35–49.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *The Annals of Statistics* **2**: 1152–1174.

Arratia, R., Barbour, A. D. and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula, *Annals of Applied Probability* **2**: 519–535.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*, New York: Wiley.

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

Carota, C. and Parmigiani, G. (2002). Semiparametric regression for count data, *Biometrika* **89**: 265–281.

Cifarelli, D. M. and Melilli, E. (2000). Some New Results for Dirichlet Priors, *The Annals of Statistics* **28**: 1390–1413.

Cifarelli, D. M. and Regazzini, E. (1990). Distribution functions of means of a Dirichlet process, *Annals of Statistics* **18**: 429–442.

Dahl, D. (2005). Sequentially-Allocated Merge-Split Sampler for Conjugate and Non-conjugate Dirichlet Process Mixture Models, *Technical report*, Texas A&M University, Department of Statistics.

Dahl, D. B. (2003). Modeling differential gene expression using a Dirichlet Process mixture model, *Proceedings of the American Statistical Association, Bayesian Statistical Sciences Section [CD-ROM]*, Alexandria, VA: American Statistical Association.

Dey, D., Müller, P. and Sinha, D. (eds) (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*, New York: Springer-Verlag.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion), *The Annals of Statistics* **14**: 1–67.

Diaconis, P. and Kemperman, J. (1996). Some new tools for Dirichlet priors, *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 5. Proceedings of the Fourth Valencia International Meeting*, Oxford University Press, pp. 97–106.

Do, K.-A., Müller, P. and Tang, F. (2005). A Bayesian mixture model for differential gene expression, *Journal of the Royal Statistical Society Series C* **54**(3): 1–18.

Dominici, F. and Parmigiani, G. (2001). Bayesian Semiparametric Analysis of Developmental Toxicology Data, *Biometrics* **57**: 150–157.

Doss, H. and Burr, D. (2005). A Bayesian semi-parametric model for random effects meta-analysis, *Journal of the American Statistical Association* **100**: 242–251.

Dykstra, R. L. and Laud, P. (1981). A Bayesian nonparametric approach to reliability, *The Annals of Statistics* **9**: 356–367.

Escobar, M. (1988). *Estimating the means of several normal populations by estimating the distribution of the means*, PhD thesis, Yale University.

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.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions, *in* M. H. Rizvi, J. S. Rustagi and D. Siegmund (eds), *Recent Advances in Statistics*, Academic Press, New York, pp. 287–302.

Florens, J.-P., Mouchart, M. and Rolin, J.-M. (1999). Semi- and Non-Parametric Bayesian Analysis of Duration Models with Dirichlet Priors: A Survey, *International Statistical Review / Revue Internationale de Statistique* **67**(2): 187–210.

Gelfand, A. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**: 398–409.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2004). Bayesian Nonparametric Spatial Modeling With Dirichlet Processes Mixing, *Technical Report ams2004-05*, UCSC Department of Applied Mathematics and Statistics.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*, New York: Springer.

Hanson, T. (2006). Inference for mixtures of finite Polya trees models, *Journal of the American Statistical Association*. To appear.

Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of polya trees, *Journal of the American Statistical Association* **97**: 1020–1033.

Hanson, T. and Johnson, W. O. (2004). A Bayesian semiparametric AFT model for interval-censored data, *Journal of Computational and Graphical Statistics* **13**: 341–361.

Hanson, T., Sethuraman, J. and Xu, L. (2005). On choosing the centering distribution in Dirichlet process mixture models, *Statistics and Probability Letters* **72**: 153–162.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data, *The Annals of Statistics* **18**: 1259–1294.

Hjort, N. L. and Ongaro, A. (2005). Exact inference for random Dirichlet means, *Statistical Inference for Stochastic Processes.* **8**(3): 227–254.

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. and James, L. F. (2003a). Generalized weighted Chinese restaurant processes for species sampling mixture models, *Statistica Sinica* **13**: 1211–1235.

Ishwaran, H. and James, L. F. (2003b). Some further developments for stick-breaking priors: finite and infinite clustering and classification, *Sankhya Series A* **65**: 577–592.

Ishwaran, H. and James, L. F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes and panel count data, *Journal of the American Statistical Association* **99**: 175–190.

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.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models, *Biometrika* **87**: 371–390.

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

James, L. F. (2005). Functionals of Dirichlet processes, the Cifarelli-Regazzini identity and beta-gamma processes, *The Annals of Statistics* **33**(2): 647–660.

Korwar, R. M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes, *The Annals of Probability* **1**: 705–711.

Kottas, A. and Gelfand, A. (2001). Bayesian Semiparametric Median Regression Modeling, *Journal of the American Statistical Association* **96**: 1458–1468.

Kottas, A., Müller, P. and Quintana, F. (2005). Nonparametric Bayesian modeling for multivariate ordinal data, *Journal of Computational and Graphical Statistics* **14**(3): 610–625.

Kuo, L. and Yang, T. Y. (2006). An improved collapsed Gibbs sampler for Dirichlet process mixing models, *Computational Statistics & Data Analysis* **50**(3): 659–674.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling, *The Annals of Statistics* **20**: 1222–1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling, *The Annals of Statistics* **22**: 1161–1176.

Lenk, P. (1988). The logistic normal distribution for bayesian,nonparametric predictive densities, *Journal of the American Statistical Association* **83**: 509–516.

Lijoi, A., Mena, R. H. and Prünster, I. (2006). Bayesian nonparametric estimation of the probability of discovering new species, *Technical report*, U. degli Studi di Pavia.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations, *The Annals of Statistics* **24**: 911–930.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates, *The Annals of Statistics* **12**: 351–357.

MacEachern, S. (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 Mixture of Dirichlet Process Models, *Journal of Computational and Graphical Statistics* **7**(2): 223–338.

MacEachern, S. N. and Müller, P. (2000). Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichlet Process Mixture Models, *in* F. Ruggeri and D. Ríos-Insua (eds), *Robust Bayesian Analysis*, New York:Springer-Verlag, pp. 295–316.

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.

Muliere, P. and Secchi, P. (1995). A note on a proper Bayesian Bootstrap, *Technical Report 18*, Università degli Sudi di Pavia, Dipartamento di Economia Politica e Metodi Quantitativi.

Müller, P. and Quintana, F. A. (2004). Nonparametric Bayesian Data Analysis, *Statistical Science* **19**: 95–110.

Müller, P., Quintana, F. and Rosner, G. (2004). Hierarchical Meta-Analysis over Related Non-parametric Bayesian Models, *Journal of the Royal Statistical Society, Series B* **66**: 735–749.

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

Nieto-Barajas, L. and Walker, S. G. (2002). Markov beta and gamma processes for modelling hazard rates, *Scandinavian Journal of Statistics* **29**: 413–424.

O'Hagan, A. (1992). Some Bayesian numerical analysis, *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, Oxford: Clarendon Press, pp. 345–365.

Paddock, S., Ruggeri, F., Lavine, M. and West, M. (2003). Randomised Polya Tree Models for Nonparametric Bayesian Inference, *Statistica Sinica* **13**: 443–460.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials, *Canadian Journal of Statistics* **27**: 105–126.

Petrone, S. (1999b). Random Bernstein Polynomials, *Scandinavian Journal of Statistics* **26**: 373–393.

Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme, *in* T. S. Ferguson, L. S. Shapeley and J. B. MacQueen (eds), *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, Haywar, California: IMS Lecture Notes - Monograph Series, pp. 245–268.

Pitman, J. and Yor, M. (1987). The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator, *The Annals of Probability* **25**: 855–900.

Quintana, F. A. (2006). A predictive view of Bayesian clustering, *Journal of Statistical Planning and Inference* **136**: 2407–2429.

Quintana, F. A. and Müller, P. (2004). Nonparametric Bayesian Assessment of the Order of Dependence for Binary Sequences, *Journal of Computational and Graphical Statistics* **13**: 213–231.

Quintana, F. A. and Newton, M. A. (2000). Computational aspects of Nonparametric Bayesian analysis with applications to the modeling of multiple binary sequences, *Journal of Computational and Graphical Statistics* **9**(4): 711–737.

Regazzini, E., Guglielmi, A. and Di Nunno, G. (2002). Theory and numerical analysis for exact distributions of functionals of a Dirichlet process, **30**: 1376–1411.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *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.

Rolin, J.-M. (1992). Some useful properties of the Dirichlet process, *Technical Report 9207*, Center for Operations Research & Econometrics, Université Catholique de Louvain.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica* **4**: 639–650.

Sinha, S., Mukherjee, B. and Ghosh, M. (2004). Bayesian semiparametric modeling for matched case-control studies with multiple disease states, *Biometrics* **60**: 41–49.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - alternative to reversible jump methods, *Annals of Statistics* **28**: 40–74.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion), *Journal of the American Statistical Association* **82**: 528–550.

Tiwari, R., Salinas-Torres, V. H. and Pereira, C. A. B. (1997). Convergence of Dirichlet measures arising in context of Bayesian analysis of competing risks models, *Journal of Multivariate Analysis* **62**: 24–35.

van der Merwe, A. J. and Pretorius, A. L. (2003). Bayesian estimation in animal breeding using the Dirichlet process prior for correlated random effects, *Genetics Selection Evolution* **35**: 137–158.

Walker, S. and Muliere, P. (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme, *The Annals of Statistics* **25**: 1762–1780.

Walker, S., Damien, P., Laud, P. and Smith, A. (1999). Bayesian nonparametric inference for distributions and related functions (with discussion), *Journal of the Royal Statistical Society, Series B* **61**: 485–527.

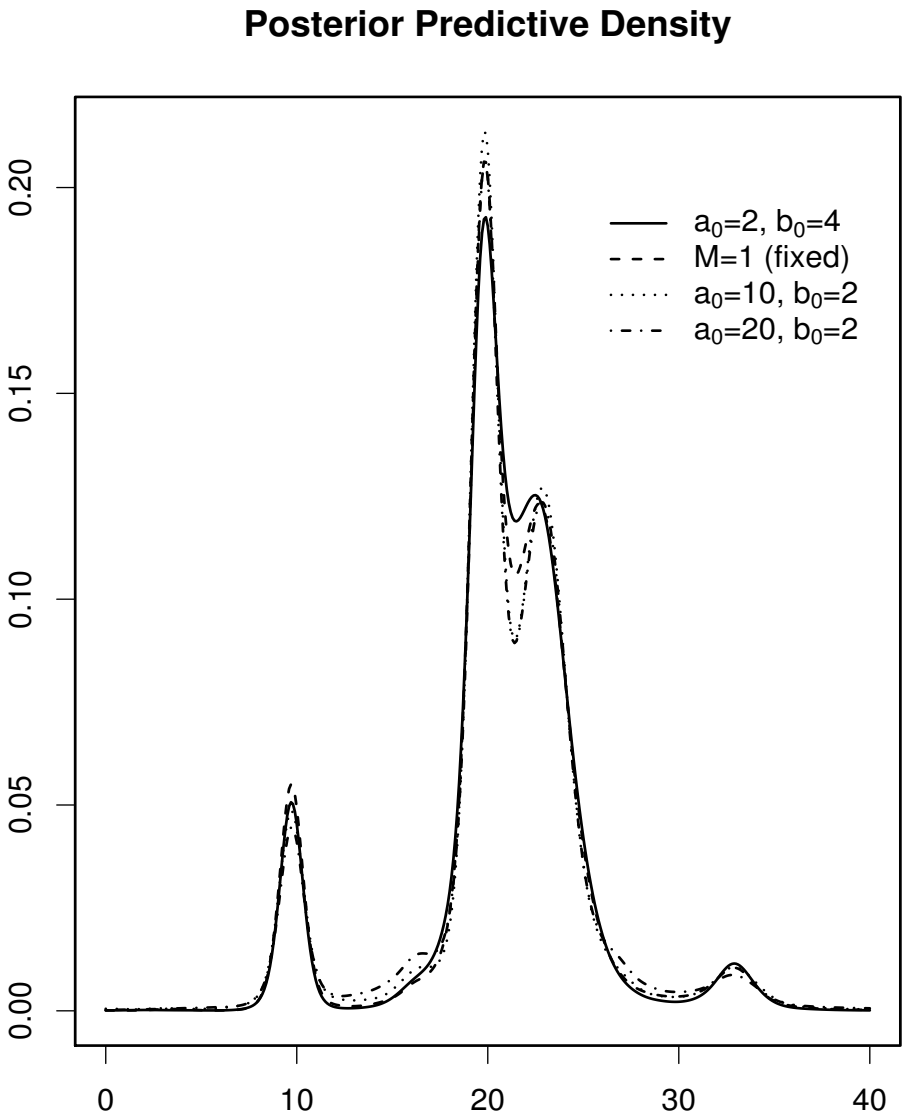


Figure 1: Posterior predictive distribution for each of four prior specifications.

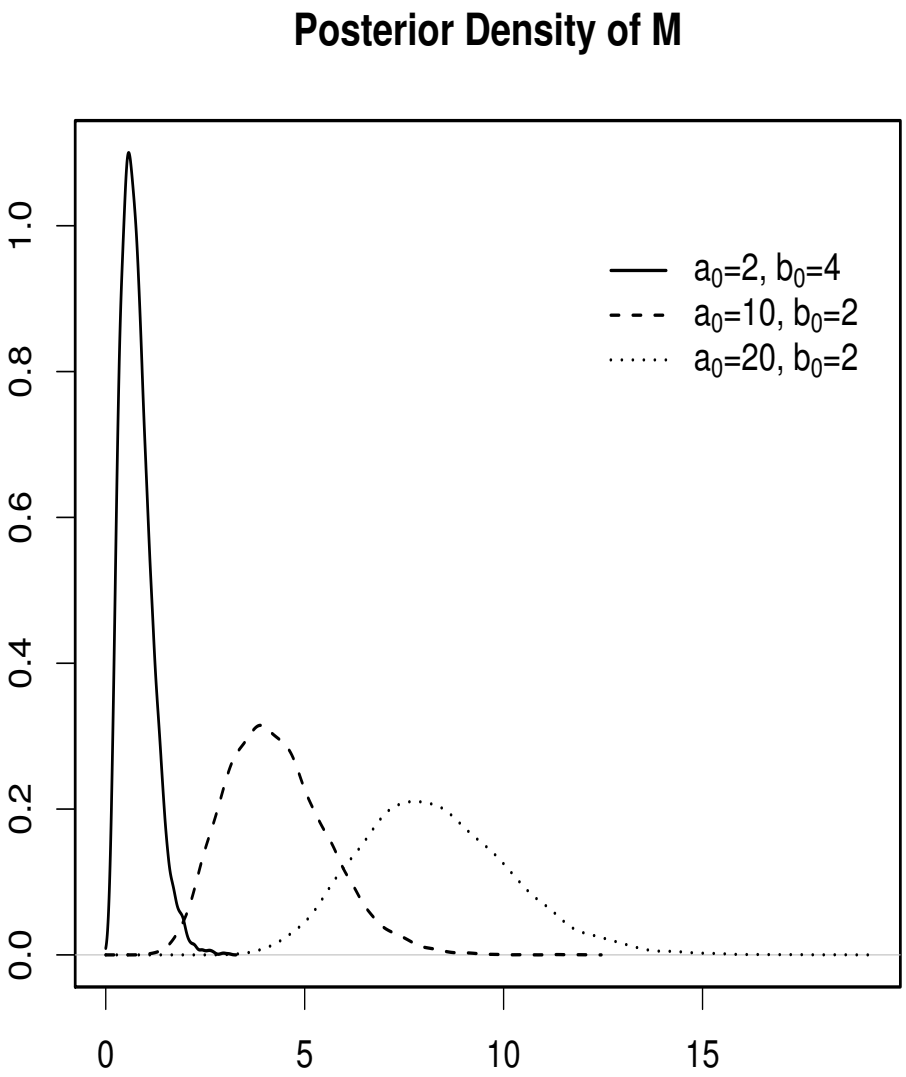
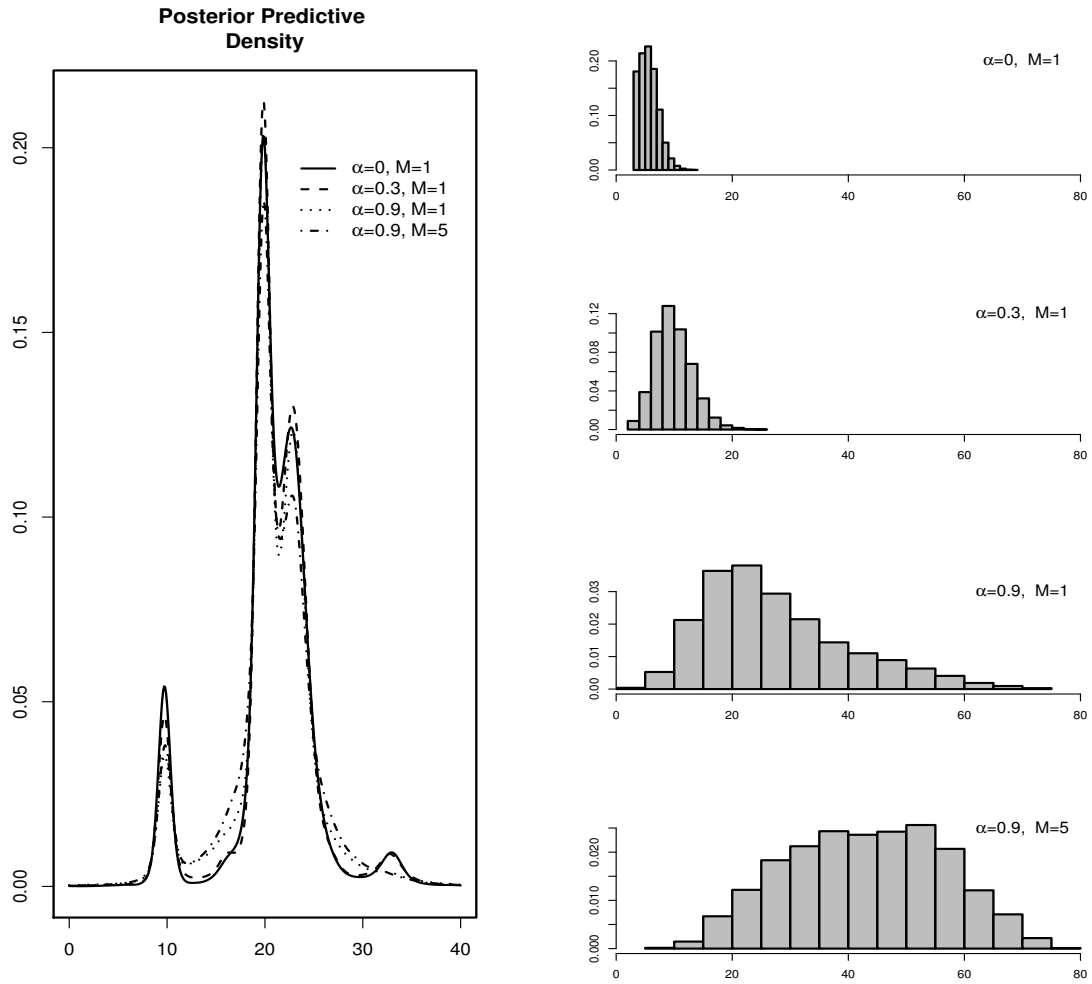
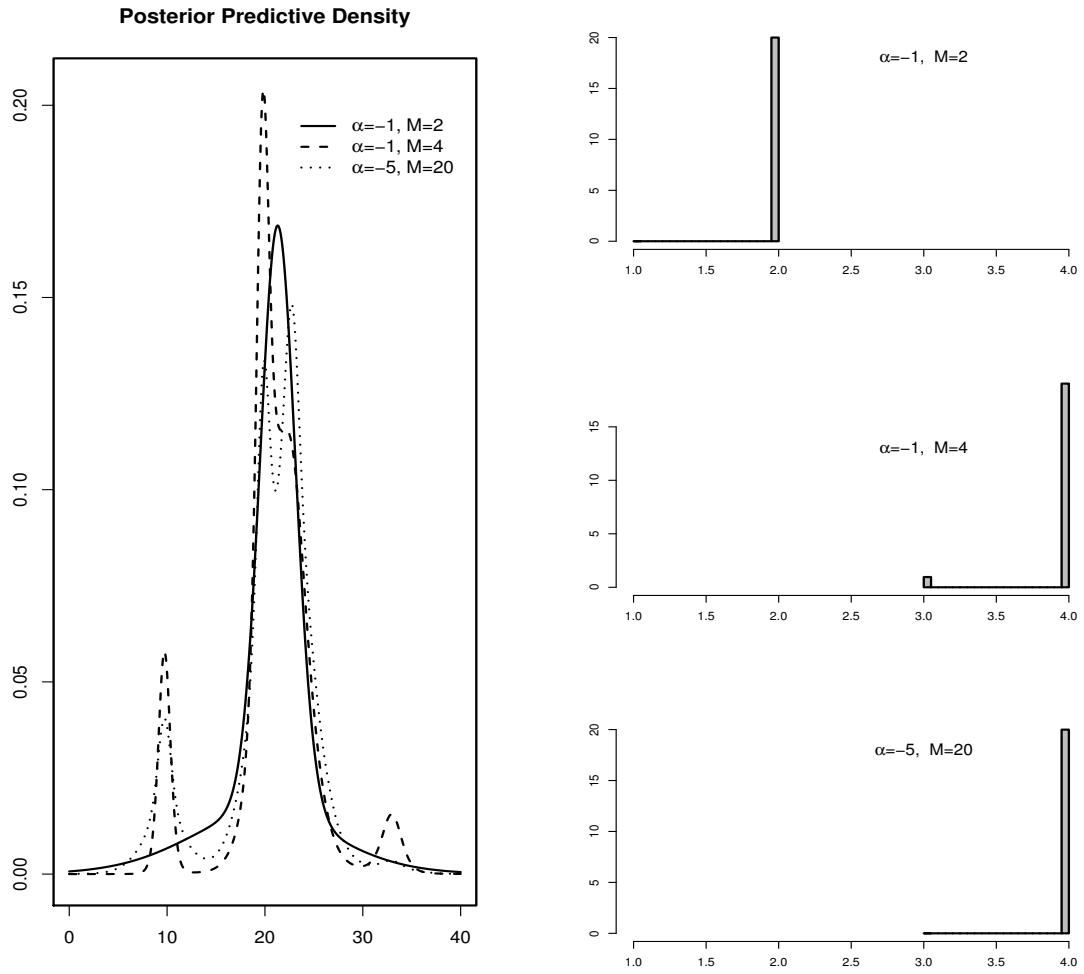


Figure 2: Posterior distribution of M for each of three prior specifications.



(a) (b)

Figure 3: (a) Posterior predictive distribution and (b) posterior distribution of the number of clusters for four prior specifications of the PY process. Note that for $\alpha = 0$ and $M = 1$ the PY process becomes the DP.



(a) (b)

Figure 4: (a) Posterior predictive distribution and (b) posterior distribution of the number of clusters for three prior specifications of the PY process.