Gibbs Sampling Methods for Stick-Breaking Priors

Hemant Ishwaran and Lancelot F. James

A rich and flexible class of random probability measures, which we call stick-breaking priors, can be constructed using a sequence of independent beta random variables. Examples of random measures that have this characterization include the Dirichlet process, its two-parameter extension, the two-parameter Poisson–Dirichlet process, finite dimensional Dirichlet priors, and beta two-parameter processes. The rich nature of stick-breaking priors offers Bayesians a useful class of priors for nonparametric problems, while the similar construction used in each prior can be exploited to develop a general computational procedure for fitting them. In this article we present two general types of Gibbs samplers that can be used to fit posteriors of Bayesian hierarchical models based on stick-breaking priors. The first type of Gibbs sampler, referred to as a Pólya urn Gibbs sampler, is a generalized version of a widely used Gibbs sampling method currently employed for Dirichlet process computing. This method applies to stick-breaking priors with a known Pólya urn characterization, that is, priors with an explicit and simple prediction rule. Our second method, the blocked Gibbs sampler, is based on an entirely different approach that works by directly sampling values from the posterior of the random measure. The blocked Gibbs sampler can be viewed as a more general approach because it works without requiring an explicit prediction rule. We find that the blocked Gibbs avoids some of the limitations seen with the Pólya urn approach and should be simpler for nonexperts to use.

KEY WORDS: Blocked Gibbs sampler; Dirichlet process; Generalized Dirichlet distribution; Pitman–Yor process; Pólya urn Gibbs sampler; Prediction rule; Random probability measure; Random weights; Stable law.

1. INTRODUCTION

This article presents two Gibbs sampling methods for fitting Bayesian nonparametric and semiparametric hierarchical models that are based on a general class of priors that we call *stick-breaking priors*. The two types of Gibbs samplers are quite different in nature. The first method is applicable when the prior can be characterized by a generalized Pólya urn mechanism and it involves drawing samples from the posterior of a hierarchical model formed by marginalizing over the prior. Our *Pólya urn Gibbs sampler* is a direct extension of the widely used Pólya urn sampler developed by Escobar (1988, 1994), MacEachern (1994), and Escobar and West (1995) for fitting the Ferguson (1973, 1974) Dirichlet process. Although here we focus on its application to stick-breaking priors (such as the Dirichlet process), in principle, the Pólya urn Gibbs sampler can be applied to any random probability measure with a known *prediction rule*. The prediction rule characterizes the Pólya urn description of the prior and is defined as the conditional distribution of a future observation Y_{n+1} given previous sampled values $Y_1, \ldots, Y_n$ from the prior (as illustration, the prediction rule for the Dirichlet process leads to the famous Blackwell–MacQueen Pólya urn; see Blackwell and MacQueen 1973 for more discussion).

Our second method, the *blocked Gibbs sampler*, works in greater generality in that it can be applied when the Pólya urn characterization is unknown. Thus, this method can be used for a stick-breaking measure without needing to know the prediction rule (in fact, we argue that sometimes even if one knows the prediction rule, the blocked Gibbs might still be preferable to Pólya urn sampling). A key aspect of the blocked Gibbs approach is that it avoids marginalizing over the prior, thus allowing the prior to be directly involved in the Gibbs sampling scheme. This allows direct sampling of the nonparametric posterior, leading to several computational and inferential advantages (see the discussion in Sections 5.3 and 5.4). Including the prior in the update is an approach that is not always exploited when fitting nonparametric hierarchical models. This is especially true for models using Dirichlet process priors. Some notable exceptions were discussed by Doss (1994), who showed how to directly update the Dirichlet process in censored data problems, and by Ishwaran and Zarepour (2000a), who updated approximate Dirichlet processes in hierarchical models. Also see the literature regarding the beta process (Hjort 1990) in survival analysis for further examples of Bayesian computational methods that include the prior in the update (for example, see Damien, Laud, and Smith 1996 and Laud, Damien, and Smith 1998).

1.1 Stick-Breaking Priors

A more precise definition will be given shortly, but for now we note that stick-breaking priors are almost surely discrete random probability measures $\mathcal{P}$ that can be represented generally as

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

where we write $\delta_{Z_k}(\cdot)$ to denote a discrete measure concentrated at Z_k. In (1), the p_k are random variables (called random weights) chosen to be independent of Z_k and such that $0 \le p_k \le 1$ and $\sum_{k=1}^{N} p_k = 1$ almost surely. It is assumed that Z_k are iid random elements with a distribution H over a measurable Polish space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, where it is assumed that H is nonatomic (i.e., $H\{y\} = 0$ for each $y \in \mathcal{Y}$). Stick-breaking priors can be constructed using either a finite or an infinite number of terms, $1 \le N \le \infty$; in some cases, we may even choose N to depend upon the sample size n.

The method of construction for the random weights is what sets stick-breaking priors apart from general random measures

Hemant Ishwaran is Associate Staff, Department of Biostatistics and Epidemiology/Wb4, Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195 (E-mail: *ishwaran@bio.ri.ccf.org*). Lancelot James is Assistant Professor, Department of Mathematical Sciences, Johns Hopkins University, Baltimore, MD 21218-2692 (E-mail: *james@brutus.mts.jhu.edu*).
The authors are grateful to the reviewers of a first draft of this article, who provided some key insights that greatly improved the final version. This work was partially supported by the Acheson J. Duncan Fund for the Advancement of Research in Statistics, award 00-1, Department of Mathematical Sciences, Johns Hopkins University.

$\mathcal{P}$ expressible as (1). Call $\mathcal{P}$ a $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ random probability measure, or a stick-breaking random measure, if it is of the form (1) and

$$p_1 = V_1 \quad \text{and} \quad p_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k, \quad k \geq 2, \quad (2)$$

where V_k are independent Beta(a_k, b_k) random variables for $a_k, b_k > 0$, and where $\mathbf{a} = (a_1, a_2, \ldots)$ and $\mathbf{b} = (b_1, b_2, \ldots)$. Informally, construction (2) can be thought of as a stick-breaking procedure, where at each stage we independently and randomly, break what is left of a stick of unit length and assign the length of this break to the current p_k value.

The stick-breaking notion for constructing random weights has a very long history. For example, see Halmos (1944), Freedman (1963), Fabius (1964), Connor and Mosimann (1969), and Kingman (1974). However, our definition of a $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measure appears to be a new concept (to the best of our knowledge) and it provides a unified way to connect a collection of seemingly unrelated measures scattered throughout the literature. These include (a) the Ferguson Dirichlet process (Ferguson 1973, 1974), (b) the two-parameter Poisson–Dirichlet process (Pitman and Yor 1997), (c) Dirichlet-multinomial processes (Muliere and Secchi 1995), m-spike models (Liu 1996), finite dimensional Dirichlet priors (Ishwaran and Zarepour 2000b,c) and (d) beta two-parameter processes (Ishwaran and Zarepour 2000a). (This list is by no means complete, and we anticipate that more measures will eventually be recognized as being stick-breaking in nature.)

1.1.1 The Case $N < \infty$ (Finite dimensional priors). The class of $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measures can be broken up into two groups, depending on whether a finite or infinite number of beta random variables are used in its construction. When $N < \infty$, we necessarily set $V_N = 1$ to ensure that $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ is well defined. In this case, we have $\mathbf{a} = (a_1, \ldots, a_{N-1})$, $\mathbf{b} = (b_1, \ldots, b_{N-1})$, and

$$p_1 = V_1 \quad \text{and} \quad p_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k, \quad k = 2, \ldots, N. \quad (3)$$

Setting $V_N = 1$ guarantees that $\sum_{k=1}^{N} p_k = 1$ with probability 1, because

$$1 - \sum_{k=1}^{N-1} p_k = (1 - V_1) \cdots (1 - V_{N-1}). \quad (4)$$

Section 3 shows that random weights defined in this manner have the generalized Dirichlet distribution. One important consequence of this is that all finite dimensional Dirichlet priors (Ishwaran and Zarepour 2000c) can be seen to be $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measures. Another attribute of the generalized Dirichlet distribution is its conjugacy to multinomial sampling. This property is a key ingredient to implementing the blocked Gibbs sampler in Section 5.

1.1.2 The Case $N = \infty$ (Infinite dimensional priors). We write $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ to emphasize the case when $N = \infty$. A $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ infinite dimensional prior is only well defined if its random weights sum to 1 with probability 1. The following lemma provides a simple method for checking this condition (see the Appendix for a proof).

Lemma 1. For the random weights in the $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ random measure,

$$\sum_{k=1}^{\infty} p_k = 1 \quad \text{a.s. iff} \quad \sum_{k=1}^{\infty} E\big(\log(1 - V_k)\big) = -\infty. \quad (5)$$

Alternatively, it is sufficient to check that $\sum_{k=1}^{\infty} \log(1 + a_k/b_k) = +\infty$.

Infinite dimensional priors include the Dirichlet process, its two-parameter extension, the two-parameter Poisson–Dirichlet process (Pitman and Yor 1997), and beta two-parameter processes (Ishwaran and Zarepour 2000a).

1.2 Outline of the Text

The layout of this article is as follows. In Section 2, we discuss the two-parameter Poisson–Dirichlet process, or what we refer to as the Pitman–Yor process (Pitman and Yor 1997). We recall its characterization in terms of the Poisson process and compare this construction to the simpler stick-breaking construction that will be used in our blocked Gibbs sampler. We also review its prediction rule and, thus, its characterization as a Pólya urn which is needed for its computation via the Pólya urn sampler. Section 3 focuses on finite dimensional $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measures and presents several important examples, including finite dimensional Dirichlet priors (Ishwaran and Zarepour 2000c) and almost sure truncations of $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ measures (see Theorems 1 and 2 for methods for selecting truncation levels).

The generalized Pólya urn Gibbs sampler is presented in Section 4, including an acceleration step to improve mixing of the Markov chain. Theorem 3 characterizes the posterior under the Pitman–Yor prior in semiparametric hierarchical models and can be used to estimate mean posterior functionals and their laws based on output from the Pólya urn sampler. Section 5 presents the blocked Gibbs sampler for fitting hierarchical models based on finite dimensional stick-breaking priors (these include the priors mentioned in Section 3, and truncations and approximations of the Pitman–Yor process as examples). We indicate some of the important properties of the blocked Gibbs sampler, including its ease in handling non-conjugacy and its good mixing behavior. Its good mixing properties are attributed to the manner in which it blocks parameters and are due to its behavior as a data augmentation procedure (see Section 5.3). Section 6 presents extensive simulations that compare the mixing behavior of our two Gibbs samplers under different priors.

2. THE PITMAN–YOR PROCESS, $\mathcal{PY}(a, b)$

The recently developed two-parameter Poisson–Dirichlet process of Pitman and Yor (1997) has been the subject of a considerable amount of research interest. (see Pitman 1995, 1996a,b, 1997, 1999; Kerov 1995; Mekjian and Chase 1997;

Zabell 1997; Tsilevich 1997; Carlton 1999). However, because most of this literature has appeared outside of statistics, it has gone largely unnoticed that the process possesses several properties that make it potentially useful as a Bayesian nonparametric prior. One such key property is its characterization as a stick-breaking random measure by Pitman (1995, 1996a). As shown there, the *size-biased random permutation* for the ranked random weights from the measure produces a sequence of random weights derived using a *residual allocation scheme* (Pitman 1996b), or what we are calling a stick-breaking scheme. This then identifies the process as a two-parameter stick-breaking random measure with parameters $a_k = 1 - a$, $b_k = b + ka$, where $0 \le a < 1$ and $b > -a$ (Pitman 1995, 1996a). Another key property of the process is its characterization as a generalized Pólya urn. As we will see, this characterization follows from an explicit description of its prediction rule (see Section 2.2).

For convenience, we refer to the two-parameter Poisson–Dirichlet process as the Pitman–Yor process, writing it as $\mathcal{PY}(a, b)$ to indicate its two shape parameters. The Ferguson (1973, 1974) Dirichlet process is one example of a $\mathcal{PY}(a, b)$ process, corresponding to the measure $\mathcal{PY}(0, \alpha)$ with parameters $a = 0$ and $b = \alpha$. For notation, we usually write $\mathrm{DP}(\alpha H)$ for this measure to indicate a Dirichlet process with finite measure αH. Another important example that we consider is the $\mathcal{PY}(\alpha, 0)$ process with parameters $a = \alpha$ and $b = 0$. This selection of shape parameters yields a measure whose random weights are based on a stable law with index $0 < \alpha < 1$. The $\mathrm{DP}(\alpha H)$ and stable law $\mathcal{PY}(\alpha, 0)$ processes are key processes because they represent the canonical measures of the Pitman–Yor process (Pitman and Yor 1997, Corollary 21). In the following subsection, we look at these two processes in more detail and review their stick-breaking and Poisson process characterizations.

2.1 Poisson Process Characterization

Earlier representations for the $\mathrm{DP}(\alpha H)$ measure and other measures derived from infinitely divisible random variables were constructed using Lévy measures applied to the Poisson point process (see Ferguson and Klass 1972). An unpublished thesis by McCloskey (1965) appears to be the first work that drew comparisons between the Dirichlet process and beta random variable stick-breaking procedures. However, it was not until Sethuraman (1994) that these connections were formalized (also see Sethuraman and Tiwari 1982, Donnelly and Joyce 1989, and Perman, Pitman, and Yor 1992). Let $\Gamma_k = E_1 + \cdots + E_k$, where E_k are iid exp(1) random variables. Sethuraman (1994) established the following remarkable identity showing that the Dirichlet process defined by Ferguson (1973) is a $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ measure:

$$\mathrm{DP}(\alpha H)(\cdot) = \sum_{k=1}^{\infty} \frac{\nu^{-1}(\Gamma_k)}{\sum_{k=1}^{\infty} \nu^{-1}(\Gamma_k)} \delta_{Z_k}(\cdot)$$
$$\overset{\mathcal{D}}{=} V_1 \delta_{Z_1}(\cdot) + \sum_{k=2}^{\infty} \big[(1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k\big] \delta_{Z_k}(\cdot),$$

where $V_k \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$, $\alpha > 0$, and ν^{-1} is the inverse of the Lévy measure for a Gamma(α) random variable,

$$\nu(x) = \alpha \int_x^{\infty} \exp(-u) u^{-1}\, du, \qquad 0 < x < \infty.$$

Pitman and Yor (1997) established a parallel identity for the $\mathcal{PY}(a, b)$ process using its characterization in terms of a subordinator of the Poisson process. In particular, for the $\mathcal{PY}(\alpha, 0)$ process, Pitman and Yor (1997, Proposition 9) and Perman et al. (1992) proved the remarkable fact

$$\mathcal{PY}(\alpha, 0)(\cdot) = \sum_{k=1}^{\infty} \frac{\Gamma_k^{-1/\alpha}}{\sum_{k=1}^{\infty} \Gamma_k^{-1/\alpha}} \delta_{Z_k}(\cdot)$$
$$\overset{\mathcal{D}}{=} V_1 \delta_{Z_1}(\cdot) + \sum_{k=2}^{\infty} \big[(1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k\big] \delta_{Z_k}(\cdot), \tag{6}$$

where $V_k \overset{\text{ind}}{\sim} \mathrm{Beta}(1 - \alpha, k\alpha)$ for $0 < \alpha < 1$.

Note that both Poisson process constructions rely on random weights constructed using infinitely divisible random variables. In the $\mathrm{DP}(\alpha H)$ process, $p_k = J_k / J$ is the value $J_k = \nu^{-1}(\Gamma_k)$ normalized by $J = \sum_{k=1}^{\infty} J_k$, a Gamma($\alpha$) random variable. In the $\mathcal{PY}(\alpha, 0)$ process, $p_k = J_k / J$ is the value of $J_k = \Gamma_k^{-1/\alpha}$ normalized by the random variable J with a stable law with index $0 < \alpha < 1$.

It is instructive to compare the complexity of the Poisson based random weights to their much simpler stick-breaking counterparts. For example, even trying to sample values from the Poisson process construction is difficult, because the denominator of each random weight involves an infinite sum (even the numerator can be hard to compute with certain Lévy measures, like in the case of the Dirichlet process). The stick-breaking characterization thus represents a tremendous innovation, making the the Pitman–Yor process more amenable to nonparametric applications in much the same way the Sethuraman (1994) construction does for the Dirichlet processs. A perfect example to illustrate its utility is our blocked Gibbs sampler, which is able to exploit such stick-breaking constructions to fit hierarchical models based on approximate Pitman–Yor priors (see Sections 5 and 6 for more details).

2.2 Generalized Pólya Urn Characterization

As shown by Pitman (1995, 1996a), the $\mathcal{PY}(a, b)$ process also can be characterized in terms of a generalized Pólya urn mechanism. This important connection presents us with another Gibbs sampling method for fitting nonparametric models based on this prior (see Section 4).

Call $Y_1, \ldots, Y_n$ *a sample from the* $\mathcal{PY}(a, b)$ *process* if

$$Y_i | P \overset{\text{iid}}{\sim} P, \qquad i = 1, \ldots, n, \qquad P \sim \mathcal{PY}(a, b). \tag{7}$$

As shown in Pitman (1995, 1996a), the prediction rule for the $\mathcal{PY}(a, b)$ process is

$$\mathbb{P}\{Y_i \in \cdot | Y_1, \dots, Y_{i-1}\} = \frac{b + am_i}{b + i - 1} H(\cdot) + \sum_{j=1}^{m_i} \frac{n^*_{j,i} - a}{b + i - 1} \delta_{Y^*_{j,i}}(\cdot), \qquad i = 2, 3, \dots, n,$$

where $\{Y^*_{1,i}, \dots, Y^*_{m_i,i}\}$ denotes the unique set of values in $\{Y_1, \dots, Y_{i-1}\}$, each occurring with frequency $n^*_{j,i}$ for $j = 1, \dots, m_i$ (notice that $n^*_{1,i} + \cdots + n^*_{m_i,i} = i - 1$).

From this, it follows that the joint distribution for $(Y_1, \dots, Y_n)$ can be characterized equivalently by the following generalized Pólya urn scheme. Let $\zeta_1, \dots, \zeta_n$ be iid H. If $0 \le a < 1$ and $b > -a$, then

$$Y_1 = \zeta_1,$$
$$(Y_i \mid Y_1, \dots, Y_{i-1}) = \begin{cases} \zeta_i & \text{with probability} \\ & \quad (b + am_i)/(b + i - 1), \\ Y^*_{j,i} & \text{with probability} \\ & \quad (n^*_{j,i} - a)/(b + i - 1), \end{cases} \tag{8}$$

for $i = 2, 3, \dots, n$.

2.2.1 Exchangeability and Full Conditionals. The sampled values $\{Y_1, \dots, Y_n\}$ produced from the urn scheme (8) are exchangeable because they have the same law as values drawn using (7). Thus, by exchangeability, we can determine the full conditional distribution for any Y_i by knowing the full conditional of a specific Y; the most convenient to describe being Y_n, which we already have seen:

$$\mathbb{P}\{Y_n \in \cdot | Y_1, \dots, Y_{n-1}\} = \frac{b + am_n}{b + n - 1} H(\cdot) + \sum_{j=1}^{m_n} \frac{n^*_{j,n} - a}{b + n - 1} \delta_{Y^*_{j,n}}(\cdot).$$

This prediction rule is the key component in the Pólya urn Gibbs sampler described in Section 4, as it allows us to work out the full conditionals for each Y_i needed to run the sampler.

Note that the special case when $a = 0$ and $b = \alpha > 0$ corresponds to the $\mathrm{DP}(\alpha H)$ process and produces the Blackwell–MacQueen prediction rule

$$\mathbb{P}\{Y_n \in \cdot | Y_1, \dots, Y_{n-1}\} = \frac{\alpha}{\alpha + n - 1} H(\cdot) + \frac{1}{\alpha + n - 1} \sum_{j=1}^{n-1} \delta_{Y_j}(\cdot).$$

2.2.2 A Finite Dimensional Dirichlet Prior. Another important example of an exchangeable urn sequence (8) is derived using values $a = -\alpha/N$, $N \ge n$, and $b = \alpha > 0$ (see Pitman 1995, 1996a). This yields

$$\mathbb{P}\{Y_n \in \cdot \, | Y_1, \dots, Y_{n-1}\} = \frac{\alpha(1 - m_n/N)}{\alpha + n - 1} H(\cdot) + \sum_{j=1}^{m_n} \frac{n^*_{j,n} + \alpha/N}{\alpha + n - 1} \delta_{Y^*_{j,n}}(\cdot),$$

although the $Y_1, \dots, Y_n$ generated from this urn scheme *are not* a sample from the $\mathcal{PY}(a, b)$ process, but instead are a sample from the random measure $\mathcal{P}$ defined with symmetric Dirichlet random weights

$$\mathcal{P}(\cdot) = \sum_{k=1}^{N} \frac{G_{k,N}}{\sum_{k=1}^{N} G_{k,N}} \delta_{Z_k}(\cdot), \quad G_{k,N} \overset{\text{iid}}{\sim} \text{Gamma}\left(\frac{\alpha}{N}\right). \tag{9}$$

In the following section, we show that (9) is an example of a $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measure. It is also a mixture of Dirichlet processes. Writing $\mathcal{P} = \mathrm{DP}_N(\alpha H)$ to emphasize this connection, we can express (9) in the notation of Antoniak (1974) as

$$\mathrm{DP}_N(\alpha H)(\cdot) \overset{\mathcal{D}}{=} \int \mathrm{D}P(\alpha \xi_N(\mathbf{Z}, \cdot)) H^N(d\mathbf{Z}),$$

where $\xi_N(\mathbf{Z}, \cdot) = \sum_{k=1}^{N} \delta_{Z_k}(\cdot)/N$ is the empirical measure based on $\mathbf{Z} = (Z_1, \dots, Z_N)$.

As shown in Ishwaran and Zarepour (2000b), the $\mathrm{DP}_N(\alpha H)$ random measure can be used to approximate integrable functionals of the Dirichlet process; that is,

$$\mathrm{DP}_N(\alpha H)(g) \xrightarrow{\text{d}} \mathrm{DP}(\alpha H)(g)$$

for each real-valued measurable function g that is integrable with respect to H. Also see Kingman (1974), Muliere and Secchi (1995), Liu (1996), Walker and Wakefield (1998), Green and Richardson (1999), Neal (2000), and Ishwaran and Zarepour (2000a,c), who discussed this measure in different contexts and under different names. For example, Muliere and Secchi (1995) referred to (9) as a Dirichlet-multinomial processes, Liu (1996), in the context of importance sampling, dubbed it an m-spike model, and Ishwaran and Zarepour (2000b,c) referred to it as a finite dimensional Dirichlet prior.

3. EXAMPLES OF $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ MEASURES

3.1 Generalized Dirichlet Random Weights

By considering Connor and Mosimann (1969), it follows that the law for the random weights $\mathbf{p} = (p_1, \dots, p_N)$ defined by (3) is a *generalized Dirichlet distribution.* Write $\mathcal{GD}(\mathbf{a}, \mathbf{b})$ for this distribution, where $\mathbf{a} = (a_1, \dots, a_{N-1})$ and $\mathbf{b} = (b_1, \dots, b_{N-1})$. The density for $\mathbf{p}$ is

$$\left(\prod_{k=1}^{N-1} \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)}\right) p_1^{a_1 - 1} \cdots p_{N-1}^{a_{N-1}-1} p_N^{b_{N-1}-1} \times (1 - P_1)^{b_1 - (a_2 + b_2)} \cdots (1 - P_{N-2})^{b_{N-2} - (a_{N-1} + b_{N-1})}, \tag{10}$$

where $P_k = p_1 + \cdots + p_k$ and $\Gamma(\cdot)$ is the gamma function.

From this fact, we can make the nice connection that all random measures based on Dirichlet random weights are $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measures. More precisely, let $\mathcal{P}$ be a random measure (1) with random weights $\mathbf{p}$, where

$$\mathbf{p} = (p_1, \dots, p_N) \sim \text{Dirichlet}(a_1, \dots, a_N).$$

Then, by (10), it easily follows that $\mathbf{p}$ has a $\mathcal{GD}(\mathbf{a}, \mathbf{b})$ distribution with $\mathbf{a} = (a_1, \dots, a_{N-1})$ and $\mathbf{b} = (\sum_{k=2}^{N} a_k, \sum_{k=3}^{N} a_k,$

$\dots, a_N)$. In other words, $\mathcal{P}$ is a $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measure. Ishwaran and Zarepour (2000c) called such measures finite dimensional Dirichlet priors and showed that they can be used as sieves in finite normal mixture problems. A special case of a finite dimensional Dirichlet prior is the $\mathrm{DP}_N(\alpha H)$ process discussed in Section 2.2.2. In this case, $\mathbf{a} = (\alpha/N, \dots, \alpha/N)$ for some $\alpha > 0$.

3.2 Almost Sure Truncations of $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ Measures

Another useful class of $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measures are constructed by applying a truncation to the $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ measure. The truncation is applied by discarding the $N+1$, $N+2, \dots$ terms in the $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ random measure and replacing p_N with $1 - p_1 - \cdots - p_{N-1}$. Notice that this also corresponds to setting $V_N = 1$ in the stick-breaking procedure (3).

Determination of an appropriate truncation level can be based on the moments of the random weights. Consider the following theorem, which can be used for truncating the $\mathcal{PY}(a, b)$ measure.

Theorem 1. Let p_k denote the random weights from a given $\mathcal{PY}(a, b)$ measure. For each positive integer $N \geq 1$ and each positive integer $r \geq 1$, let

$$T_N(r, a, b) = \left(\sum_{k=N}^{\infty} p_k\right)^r, \qquad U_N(r, a, b) = \sum_{k=N}^{\infty} p_k^r.$$

Then

$$\mathrm{E}\big(T_N(r, a, b)\big) = \prod_{k=1}^{N-1} \frac{(b+ka)^{(r)}}{(b+(k-1)a+1)^{(r)}}, \qquad N \geq 2,$$

and

$$\mathrm{E}\big(U_N(r, a, b)\big) = \mathrm{E}\big(T_N(r, a, b)\big) \frac{(1-a)^{(r-1)}}{(b+(N-1)a+1)^{(r-1)}},$$

where $\gamma^{(r)} = \gamma(\gamma+1)\cdots(\gamma+r-1)$ for each $\gamma > 0$ and $\gamma^{(0)} = 1$.

See the Appendix for a proof. Several simplifications occur for the moments of the $\mathrm{DP}(\alpha H)$ process, which corresponds to the case $a = 0$ and $b = \alpha$. We have

$$\mathrm{E}\big(T_N(r, a, b)\big) = \left(\frac{\alpha}{\alpha + r}\right)^{N-1}$$

and

$$\mathrm{E}\big(U_N(r, a, b)\big) = \left(\frac{\alpha}{\alpha + r}\right)^{N-1} \frac{\Gamma(r)\Gamma(\alpha+1)}{\Gamma(\alpha+r)}.$$

Notice that both expressions decrease exponentially fast in N and, thus, for a moderate N, we should be able to achieve an accurate approximation. A precise bound is given in Theorem 2. For some $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ measures a very large value for N may be necessary to achieve reasonable accuracy. For example, in the $\mathcal{PY}(\alpha, 0)$ process based on a stable law random variable [see (6)], we have

$$\mathrm{E}\big(T_N(1, a, b)\big) = \frac{\alpha^{N-1}(N-1)!}{(\alpha+1)\cdots((N-2)\alpha+1)}, \qquad 0 < \alpha < 1.$$

Note that the value of N needed to keep this value small rapidly increases as α approaches 1. Thus it may not be feasible to approximate the $\mathcal{PY}(\alpha, 0)$ process over all α values.

If the $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measure is applied in a Bayesian hierarchical model as a prior, then an appropriate method for selecting N is to choose a value that yields a Bayesian marginal density that is nearly indistinguishable from its limit. Suppose that $\mathbf{X} = (X_1, \dots, X_n)$ is the observed data derived from the Bayesian nonparametric hierarchical model

$$\begin{aligned}(X_i|Y_i) &\overset{\text{ind}}{\sim} \pi(X_i|Y_i), \qquad i = 1, \dots, n,\\ (Y_i|P) &\overset{\text{iid}}{\sim} P,\\ P &\sim \mathcal{P}_N(\mathbf{a}, \mathbf{b}).\end{aligned}$$

Then the Bayesian marginal density under the truncation $\mathcal{P}_N = \mathcal{P}_N(\mathbf{a}, \mathbf{b})$ equals

$$\mu_N(\mathbf{X}) = \int \left(\prod_{i=1}^n \int_{\mathcal{Y}} f(X_i|Y_i) P(dY_i)\right) \mathcal{P}_N(dP),$$

where $f(x|y)$ is the density for x given y. Thus to properly select N, the marginal density μ_N should be close to its limit μ_∞ under the prior $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$. Consider the following theorem, the proof of which follows by arguments given in Ishwaran and James (2000).

Theorem 2. Let p_k denote the random weights from a given $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ measure. If $\|\cdot\|_1$ denotes the $\mathcal{L}_1$ distance, then

$$\|\mu_N - \mu_\infty\|_1 \leq 4\left(1 - \mathrm{E}\left[\left(\sum_{k=1}^{N-1} p_k\right)^n\right]\right).$$

In particular, for the $\mathcal{PY}(a, b)$ process, we have

$$\|\mu_N - \mu_\infty\|_1 \leq 4(1 - \mathrm{E}[1 - T_N(1, a, b)]^n)$$

and, for the $\mathrm{DP}(\alpha H)$ process,

$$\|\mu_N - \mu_\infty\|_1 \sim 4n \exp(-(N-1)/\alpha).$$

For example, the last approximation shows that the sample size makes almost no dent on the $\mathcal{L}_1$ distance in a Dirichlet process approximation. For example, if $n = 10^5$, $N = 150$, and $\alpha = 5$, then we get an $\mathcal{L}_1$ bound of 4.57×10^{-8}. Therefore, even for huge sample sizes, a mere truncation of $N = 150$ leads to an approximating hierarchical model that is *virtually indistinguishable* from one based on the $\mathrm{DP}(\alpha H)$ prior. See Ishwaran and James (2000) for more discussion and for applications of this truncation to estimate finite mixtures of normals. Also see Muliere and Tardella (1998), who used an "ϵ-truncation" to sample Dirichlet prior functionals.

4. PÓLYA URN GIBBS SAMPLERS

Stick-breaking measures can be used as practical and versatile priors in Bayesian nonparametric and semiparametric hierarchical models. We discuss the slightly more general semiparametric setting, which corresponds to the problem where we observe data $\mathbf{X} = (X_1, \ldots, X_n)$, derived from the hierarchical model

$$\begin{aligned}(X_i|Y_i, \theta) &\overset{\text{ind}}{\sim} \pi(X_i|Y_i, \theta), \qquad i = 1, \ldots, n,\\ (Y_i|P) &\overset{\text{iid}}{\sim} P,\\ \theta &\sim \pi(\theta),\\ P &\sim \mathcal{P}, \end{aligned} \tag{11}$$

where $\pi(X_i|Y_i, \theta)$ denotes the conditional distribution of X_i given Y_i and θ. Here $\theta \in \Re^d$ represents a finite dimensional parameter, whereas the sequence $Y_1, \ldots, Y_n$ is unobserved random elements with conditional distribution P sampled from our stick-breaking prior $\mathcal{P}$. Further extensions to (11) are also possible, such as extensions to include hyperparameters for θ. The model also can be extended to include a mixture of random measures by extending the distribution H for the Z_k to include hyperparameters. See West, Müller, and Escobar (1994), Escobar and West (1998), and MacEachern (1998) for several examples of semiparametric hierarchical models of the form (11) based on the Dirichlet process.

4.1 Marginalized Hierarchical Models

A powerful Gibbs sampling approach for sampling the posterior of (11) can be developed when the Y_i drawn from $\mathcal{P}$ can be characterized in terms of a generalized Pólya urn scheme. This method works for both the $\mathcal{PY}(a, b)$ and $\text{DP}_N(\alpha H)$ processes described in Section 2, which both generate Y_i values from the urn mechanism (8). This approach generalizes the method discovered by Escobar (1988, 1994), MacEachern (1994), and Escobar and West (1995) for Dirichlet process computing.

The method works by first integrating over P in (11) to create a marginalized semiparametric hierarchical model

$$\begin{aligned}(X_i|Y_i, \theta) &\overset{\text{ind}}{\sim} \pi(X_i|Y_i, \theta), \qquad i = 1, \ldots, n,\\ (Y_1, \ldots, Y_n) &\sim \pi(Y_1, \ldots, Y_n),\\ \theta &\sim \pi(\theta), \end{aligned} \tag{12}$$

where $\pi(Y_1, \ldots, Y_n)$ denotes the joint distribution for $\mathbf{Y} = (Y_1, \ldots, Y_n)$ defined by the underlying Pólya urn.

The proposed Gibbs sampler exploits two key facts about the joint distribution for $\mathbf{Y}$. First, that the Y_i are exchangeable and, second, that the full conditional distribution for at least one Y_i can be written in closed form (by exchangeability we can then automatically deduce the full conditional for any other Y).

4.2 Gibbs Sampling Algorithm for $\mathcal{PY}(a, b)$ and $DP_N(\alpha H)$ Processes

For simplicity, we describe the algorithm for the case when $\mathcal{P}$ is the $\mathcal{PY}(a, b)$ or $\text{DP}_N(\alpha H)$ process, but, in principle, the method can be used for any almost surely discrete measure with an explicit prediction rule. Let $\mathbf{Y}_{-i}$ denote the subvector of $\mathbf{Y}$ formed by removing the i-th coordinate. Then to draw values from the posterior distribution $\pi(\mathbf{Y}, \theta|\mathbf{X})$ of (12), we iteratively draw values from the conditional distributions of

$$(Y_i|\mathbf{Y}_{-i}, \theta, \mathbf{X}), \qquad i = 1, \ldots, n, \qquad (\theta|\mathbf{Y}, \mathbf{X}).$$

In particular, each iteration of the Gibbs sampler draws the following samples:

(a) $(Y_i|\mathbf{Y}_{-i}, \theta, \mathbf{X})$ for each $i = 1, \ldots, n$: The required conditional distributions are defined by

$$\mathbb{P}\{Y_i \in \cdot \,|\, \mathbf{Y}_{-i}, \theta, \mathbf{X}\} = q_0^* \,\mathbb{P}\{Y_i \in \cdot \,|\, \theta, X_i\} + \sum_{j=1}^{m} q_j^* \,\delta_{Y_j^*}(\cdot), \tag{13}$$

where

$$q_0^* \propto (b + am) \int_{\mathcal{Y}} f(X_i|Y, \theta)\, H(dY),$$

$$q_j^* \propto (n_j^* - a)\, f(X_i|Y_j^*, \theta),$$

and these values are subject to the constraint that they sum to 1, that is, $\sum_{j=0}^m q_j^* = 1$. Here we are dropping the dependence on i for notational simplicity and we write $\{Y_1^*, \ldots, Y_m^*\}$ for the set of unique values in $\mathbf{Y}_{-i}$, where each value occurs with frequency n_j^* for $j = 1, \ldots, m$. The expression $f(x|y, \theta)$ refers to the density of x given y and θ, and is assumed to be jointly measurable in x, y, and θ.

(b) $(\theta|\mathbf{Y}, \mathbf{X})$: By the usual application of Bayes theorem, this is the density

$$f(\theta|\mathbf{Y}, \mathbf{X}) \propto \pi(d\theta) \prod_{i=1}^{n} f(X_i|Y_i, \theta).$$

Conjugacy and Methods for Accelerating Mixing. As discussed in Escobar (1988, 1994), computing the value for q_0^* in (13) is simplified in the $\text{DP}(\alpha H)$ setting when there is conjugacy, and the same holds true in the more general setting as well [i.e., by conjugacy, we mean that the distribution H is conjugate for Y in $f(X|Y, \theta)$]. Without conjugacy, the problem of computing q_0^* becomes a more delicate issue. In this case, the various solutions that have been proposed for the $\text{DP}(\alpha H)$ process can be applied here as well. For example, see West et al. (1994), MacEachern and Müller (1998), Walker and Damien (1998), or Neal (2000) for different approaches.

Like the Escobar (1988, 1994) Pólya urn sampler, the generalized Pólya urn sampler has a tendency to mix slowly if the values of q_j^* become much larger than the value of q_0^* in (13). When this occurs, the sampler can get stuck at the current unique values $Y_1^*, \ldots, Y_m^*$ of $\mathbf{Y}$ and it may take many iterations before any new Y^* values are generated. As a method to circumvent this problem for the $\text{DP}(\alpha H)$ process, West et al. (1994) proposed resampling the unique values $Y_1^*, \ldots, Y_m^*$ at the end of each iteration of the Gibbs sampler. A general version of this method was described by MacEachern (1994).

In this *acceleration method*, the value for $\mathbf{Y}$ is reexpressed in terms of the unique values $Y_1^*, \ldots, Y_m^*$ and the cluster membership $\mathbf{C} = (C_1, \ldots, C_n)$, where each C_i records which Y_j^*

corresponds to the value for Y_i; that is, $Y_i = Y_j^*$ iff $C_i = j$. In the original method proposed by MacEachern (1994), the Gibbs sampler was modified to include a step to generate the cluster membership $\mathbf{C}$, which is then followed by an update for the unique Y_j^* values given the current value for $\mathbf{C}$. However, a simpler approach is to run the original Gibbs sampler to generate a current value for $\mathbf{Y}$, and from this, compute the current membership $\mathbf{C}$, and then update the unique Y_j^* values given $\mathbf{C}$. Details can be found in West et al. (1994), Bush and MacEachern (1996), and Escobar and West (1998). In the generalized Pólya urn sampler, the method works simply by adding the following third step:

(c) Draw samples from the conditional distributions for $(Y_j^*|\mathbf{C}, \theta, \mathbf{X})$ for $j = 1, \ldots, m$. In particular, for each j, the required conditional density is

$$f(Y_j^*|\mathbf{C}, \theta, \mathbf{X}) \propto H(dY_j^*) \prod_{\{i:C_i=j\}} f(X_i|Y_j^*, \theta).$$

Now use the newly sampled Y_j^* and the current value for the cluster membership $\mathbf{C}$ to determine the new updated value for $\mathbf{Y}$.

4.3 Limitations With Pólya Urn Gibbs Sampling

Although the generalized Pólya urn Gibbs sampler is a versatile method for fitting Bayesian models, there are several limitations with this approach that are worth highlighting:

(a) The method relies on the full conditional distribution of $(Y_i|\mathbf{Y}_{-i}, \theta, \mathbf{X})$, which results in a Gibbs sampler that uses a *one-coordinate-at-a-time update* for parameters. This can produce large q_j^* coefficients, which result in a slowly mixing Markov chain. The simple acceleration method discussed above can be applied to enhance mixing in general, but in some cases these methods still suffer from slow mixing, because they implicitly rely on one-at-a-time updates in some form or another.

(b) Calculating the q_0^* coefficient in (13) is problematic in the nonconjugate case. Although there are many proposed solutions, these tend to complicate the description of the algorithm and may make them less accessible to nonexperts.

(c) A third deficiency arises from the effect of marginalizing over P. Although marginalizing is the key that underlies the Pólya urn approach, it has the undesirable side effect that it allows inference for the posterior of P to be based *only on the posterior Y_i values.*

4.4 Posterior Inference in the $\mathcal{PY}(a, b)$ Process

Theorem 3 can be used in the $\mathcal{PY}(a, b)$ process to approximate posterior mean functionals using the values of $\mathbf{Y}$ drawn from our Pólya urn Gibbs sampler. This provides a partial solution to the third deficiency highlighted in the previous section.

We assume that all relevant distributions are defined over measurable Polish spaces. Thus H is a distribution over the measurable Polish space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ and the prior $\pi(\theta)$ for θ is defined over the Borel space $(\Re^d, \mathcal{B}(\Re^d))$. We also write $\mathcal{M}$ for the space of probability measures over $\mathcal{Y}$ and $\mathcal{B}(\mathcal{M})$ for the corresponding σ algebra induced by weak convergence.

Theorem 3 provides a characterization for $\mathcal{P}(\cdot|\mathbf{X})$, the posterior distribution of P given $\mathbf{X}$, in the semiparametric model (11). It is a generalization of the result for the Dirichlet process in Lo (1984, Theorem 1) to the Pitman–Yor process. Also see Antoniak (1974, Theorem 3). A proof is given in the Appendix.

Theorem 3. Let ψ be a nonnegative or integrable function over $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$. If $\mathcal{P}$ is the $\mathcal{PY}(a, b)$ process, then $\mathcal{P}(\cdot|\mathbf{X})$ in (11) is characterized by

$$\int_{\mathcal{M}} \psi(P)\, \mathcal{P}(dP|\mathbf{X}) = \int_{\mathcal{Y}^n} \int_{\mathcal{M}} \psi(P)\, \mathcal{P}(dP|\mathbf{Y})\, \pi(d\mathbf{Y}|\mathbf{X}), \quad (14)$$

where

$$\pi(d\mathbf{Y}|\mathbf{X}) = \frac{\pi(d\mathbf{Y}) \int_{\Re^d} \prod_{i=1}^n f(X_i|Y_i, \theta)\, \pi(d\theta)}{\int_{\mathcal{Y}^n} \int_{\Re^d} \prod_{i=1}^n f(X_i|Y_i, \theta)\, \pi(d\theta)\, \pi(d\mathbf{Y})}$$

and $\pi(\mathbf{Y})$ is the joint distribution for $\mathbf{Y}$ determined by the generalized Pólya urn (8).

Furthermore, $\mathcal{P}(\cdot|\mathbf{Y})$ is the posterior for the Pitman–Yor process given $\mathbf{Y}$, a sample drawn from it, and, therefore, by considering Pitman (1996a, Corollary 20), it is the random probability measure

$$\mathcal{P}(\cdot|\mathbf{Y}) = \sum_{j=1}^m p_j^* \delta_{Y_j^*}(\cdot) + p_{m+1}^* \mathcal{P}^*(\cdot), \quad (15)$$

where $\{Y_1^*, \ldots, Y_m^*\}$ are the unique set of Y_i values occurring with frequencies n_j^* and

$$(p_1^*, \ldots, p_m^*, p_{m+1}^*) \sim \text{Dirichlet}(n_1^* - a, \ldots, n_m^* - a, b + am)$$

is independent of $\mathcal{P}^*$, which is a $\mathcal{PY}(a, b + ma)$ process.

Estimating Functionals. We can use (14) to estimate various posterior mean functionals. Several popular choices for ψ are

$$\psi(P) = P(A) = \int_A P(dy),$$

which could be used to estimate the mean posterior cdf by choosing A to be a cylinder set, and

$$\psi(P) = \int_{\mathcal{Y}} k(x_0, y)\, P(dy),$$

which could be used to estimate the mean density at the point x_0 in density estimation problems where $k(\cdot, \cdot)$ is a kernel.

To estimate (14) we average

$$\int_{\mathcal{M}} \psi(P)\, \mathcal{P}(dP|\mathbf{Y}) = \sum_{j=1}^m \frac{n_j^* - a}{b + n} \psi(\delta_{Y_j^*}) + \frac{b + am}{b + n} \psi(H)$$

over the different values of $\mathbf{Y}$ drawn using the Gibbs sampler. The expression on the right-hand side follows from (15).

The preceding method works for mean functionals because of the simplification that occurs from integrating. However, if our interest is in drawing values directly from the law $\mathcal{L}(\psi(P)|\mathbf{X})$, then we have to randomly sample a measure P from the distribution of $\mathcal{P}(\cdot|\mathbf{Y})$ to produce our posterior draw $\psi(P)$. However, implementing this scheme cannot be done without some form of approximation, because it is not possible to get an exact draw from $\mathcal{P}(\cdot|\mathbf{Y})$. To get around this, we can approximate (15) with

$$\sum_{j=1}^{m} p_j^* \delta_{Y_j^*}(\cdot) + p_{m+1}^* \mathcal{P}_N^*(\cdot),$$

where $\mathcal{P}_N^*(\cdot)$ is some approximation to $\mathcal{P}^*$. One such approximation could be based on the almost sure truncations discussed in Section 3.2, with an appropriate truncation level chosen using either Theorem 1 or Theorem 2. Another method for approximating $\mathcal{P}^*$ is to draw simulated data from its urn (8) with values $a := a$ and $b := b + am$. The random measure based on a large number of these simulated values then closely approximates $\mathcal{P}^*$.

5. BLOCKED GIBBS SAMPLER

The limitations with the Pólya urn Gibbs sampler highlighted in Section 4.3 can be avoided by using what we call the blocked Gibbs sampler. This method applies to Bayesian nonparametric and semiparametric models similar to (11), but where the prior $\mathcal{P}$ is assumed to be a finite dimensional $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ measure (such as the finite dimensional Dirichet priors discussed in Section 3 and truncations to priors such as the Pitman–Yor process). The finite dimensionality of such priors is a key to the success of the method because it allows us to express our model entirely in terms of a finite number of random variables. This then allows the blocked Gibbs sampler to update *blocks of parameters*, which, because of the nature of the prior, are drawn from simple multivariate distributions (in particular, the update for $\mathbf{p}$ is straightforward because of its connection to the generalized Dirichlet distribution established in Section 3.1).

With a finite dimensional prior $P \sim \mathcal{P}_N(\mathbf{a}, \mathbf{b})$, the Bayesian semiparametric model (11) can be rewritten as

$$\begin{aligned}
(X_i|\mathbf{Z}, \mathbf{K}, \theta) &\overset{\text{ind}}{\sim} \pi(X_i|Z_{K_i}, \theta), \quad i = 1, \dots, n, \\
(K_i|\mathbf{p}) &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} p_k \, \delta_k(\cdot), \\
(\mathbf{p}, \mathbf{Z}) &\sim \pi(\mathbf{p}) \times H^N(\mathbf{Z}), \\
\theta &\sim \pi(\theta), \qquad (16)
\end{aligned}$$

where $\mathbf{K} = (K_1, \dots, K_n)$, $\mathbf{Z} = (Z_1, \dots, Z_N)$, $\mathbf{p} = (p_1, \dots, p_N) \sim \mathcal{GD}(\mathbf{a}, \mathbf{b})$, and Z_k are iid H. The key to expression (16) is to notice that Y_i equals Z_{K_i}, where the K_i act as classification variables to identify the Z_k associated with each Y_i.

5.1 Direct Posterior Inference

Rewriting the model in the form (16) allows the blocked Gibbs sampler to sample the posterior distribution $\mathcal{P}(\cdot|\mathbf{X})$ directly. The method works by iteratively drawing values from the conditional distributions of the blocked variables

$$\begin{aligned}
&(\mathbf{Z}|\mathbf{K}, \theta, \mathbf{X}), \\
&(\mathbf{K}|\mathbf{Z}, \mathbf{p}, \theta, \mathbf{X}), \\
&(\mathbf{p}|\mathbf{K}), \\
&(\theta|\mathbf{Z}, \mathbf{K}, \mathbf{X}). \qquad (17)
\end{aligned}$$

Doing so eventually produces values drawn from the distribution of $(\mathbf{Z}, \mathbf{K}, \mathbf{p}, \theta|\mathbf{X})$. Thus, each draw $(\mathbf{Z}, \mathbf{K}, \mathbf{p}, \theta)$ defines a random probability measure

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

which (eventually) gives us the draw from the posterior $\mathcal{P}(\cdot|\mathbf{X})$ that we seek. This overcomes one of the obstacles observed with Pólya urn Gibbs samplers, where inference for $\mathcal{P}(\cdot|\mathbf{X})$ requires special methods like those discussed in Section 4.4. Notice that the blocked Gibbs also can be used to sudy the posterior distribution $\pi(\mathbf{Y}|\mathbf{X})$ using $Y_i = Z_{K_i}$.

5.2 Blocked Gibbs Algorithm

The work in Ishwaran and Zarepour (2000a) can be extended straightforwardly to derive the required conditional distributions (17). Let $\{K_1^*, \dots, K_m^*\}$ denote the set of current m unique values of $\mathbf{K}$. To run the blocked Gibbs, draw values in the following order:

(a) Conditional for $\mathbf{Z}$: Simulate $Z_k \overset{\text{iid}}{\sim} H$ for each $k \in \mathbf{K} - \{K_1^*, \dots, K_m^*\}$. Also, draw $(Z_{K_j^*}|\mathbf{K}, \theta, \mathbf{X})$ from the density

$$f(Z_{K_j^*}|\mathbf{K}, \theta, \mathbf{X}) \propto H(dZ_{K_j^*}) \prod_{\{i: K_i = K_j^*\}} f(X_i|Z_{K_j^*}, \theta), \qquad j = 1, \dots, m. \quad (18)$$

(b) Conditional for $\mathbf{K}$: Draw values

$$(K_i|\mathbf{Z}, \mathbf{p}, \theta, \mathbf{X}) \overset{\text{ind}}{\sim} \sum_{k=1}^{N} p_{k,i} \, \delta_k(\cdot), \qquad i = 1, \dots, n,$$

where

$$(p_{1,i}, \dots, p_{N,i}) \propto \big(p_1 f(X_i|Z_1, \theta), \dots, p_N f(X_i|Z_N, \theta)\big).$$

(c) Conditional for $\mathbf{p}$: By the conjugacy of the generalized Dirichlet distribution to multinomial sampling, it follows that our draw is

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

where

$$V_k^* \overset{\text{ind}}{\sim} \text{Beta}\Big(a_k + M_k, b_k + \sum_{l=k+1}^{N} M_l\Big), \qquad \text{for } k = 1, \dots, N - 1,$$

and M_k records the number of K_i values that equal k.

(d) Conditional for θ: As before, draw θ from the density (remembering that $Y_i = Z_{K_i}$)

$$f(\theta|\mathbf{Z},\mathbf{K},\mathbf{X}) \propto \pi(d\theta)\prod_{i=1}^{n} f(X_i|Y_i,\theta).$$

5.3 Mixing Properties: Blocking and Data Augmentation

The good mixing properties of the blocked Gibbs sampler can be attributed to its ability to update the blocked parameters $\mathbf{Z}$, $\mathbf{K}$, and $\mathbf{p}$ using simple multivariate draws. This encourages good mixing for these parameters, which in turn encourages good mixing for the unobserved Y_i values, which are updated simultaneously from these values.

The success of the blocked Gibbs over the standard Pólya urn sampler also can be attributed to the effect of data augmentation. The blocked Gibbs does a type of data augmentation by augmenting the parameters from the urn scheme to include the prior. Consider the $\mathrm{DP}_N(\alpha H)$ measure defined in Section 2.2.2. In the Pólya urn Gibbs sampler for fitting (16) under this prior, we sample $(\mathbf{Y},\theta)$ from the density proportional to

$$\begin{aligned}\prod_{i=1}^{n} f(X_i|Y_i,\theta)\prod_{i=1}^{n}\Bigg(&\frac{\alpha(1-m_i/N)}{\alpha+i-1}H(dY_i)\\ &+\sum_{j=1}^{m_i}\frac{n^*_{j,i}+\alpha/N}{\alpha+i-1}\delta_{Y^*_{j,i}}(dY_i)\Bigg)\pi(d\theta)\\ =\prod_{i=1}^{n} f(X_i|Y_i,\theta)\Bigg(&\iint\prod_{i=1}^{n}P(dY_i)\mathcal{P}_{\alpha\xi_N}(dP)\\ &\times H^N(d\mathbf{Z})\Bigg)\pi(d\theta),\end{aligned}$$

where $\mathcal{P}_{\alpha\xi_N}$ is the Dirichlet process with finite measure $\alpha\xi_N(\mathbf{Z},\cdot)$ for a fixed value of $\mathbf{Z}$.

If we augment the parameters to include $\mathbf{Z}$ and the random measure P, we now sample $(\mathbf{Y},\theta,\mathbf{Z},P)$ from the density proportional to

$$\begin{aligned}&\prod_{i=1}^{n} f(X_i|Y_i,\theta)\prod_{i=1}^{n}P(dY_i)\,\mathcal{P}_{\alpha\xi_N}(dP)\,H^N(d\mathbf{Z})\,\pi(d\theta)\\ &=\prod_{i=1}^{n} f(X_i|Y_i,\theta)\prod_{i=1}^{n}\Bigg(\sum_{k=1}^{N}p_k\,\delta_{Z_K}(dY_i)\Bigg)\\ &\quad\times\pi(d\mathbf{p})\,H^N(d\mathbf{Z})\,\pi(d\theta),\end{aligned}$$

where $\pi(d\mathbf{p})$ is the density for a Dirichlet$(\alpha/N,\ldots,\alpha/N)$ distribution. Notice that after transforming Y_i to Z_{K_i}, this density is exactly the same density used by the blocked Gibbs sampler.

Thus in the $\mathrm{DP}_N(\alpha H)$ case, the blocked Gibbs sampler is doing an exact augmentation. However, because the $\mathrm{DP}_N(\alpha H)$ measure is a good approximation to the Dirichlet process, our argument also shows that the blocked Gibbs sampler under the $\mathrm{DP}_N(\alpha H)$ prior acts like an approximate data augmentation procedure in the Dirichlet process Pólya urn Gibbs sampler. Thus, at least for the Dirichlet process, we can anticipate that the blocked Gibbs sampler will perform as well or better than the Pólya urn method. Section 6 offers some empirical evidence to support this claim.

5.4 Nonconjugacy and Hierarchical Extensions

Another key feature of the blocked Gibbs sampler is that it easily handles the issue of conjugacy, a problem that arises only in the context of drawing values from (18). With conjugacy, the draw from (18) is done exactly, whereas the nonconjugate case can be handled easily by using standard Markov Chain Monte Carlo methods such as Metropolis–Hastings, for example. We note that the draw for (18) is the same draw used in the acceleration step for the Pólya urn sampler, so it adds no further complexity to the blocked Gibbs than already is seen in urn samplers.

The blocked Gibbs sampler can be extended to accommodate hierarchical extensions to (16). For example, the distribution for H can be expanded to include hyperparameters (such as a mean and variance parameter in the case that H is a normal distribution). It is also possible to place priors on the shape parameters $\mathbf{a}$ and $\mathbf{b}$ in the $\mathcal{P}_N(\mathbf{a},\mathbf{b})$ prior; see Ishwaran and Zarepour (2000a), who developed an exact update for the Dirichlet process cluster parameter α and updates for some other beta two-parameter processes.

6. COMPARISON OF ALGORITHMS

In this section, we present some empirical evidence that compares the mixing performance for the Pólya urn Gibbs sampler (PG sampler) of Section 4.2, the accelerated version (PG_a sampler), and the blocked Gibbs sampler (BG sampler) described in Section 5.2. We study each of these methods by applying them to the nonparametric hierarchical model

$$\begin{aligned}(X_i|Y_i)&\overset{\text{ind}}{\sim}\mathrm{N}(Y_i,\sigma),\quad i=1,\ldots,n,\\ (Y_i|P)&\overset{\text{ind}}{\sim}P,\\ P&\sim\mathcal{P},\end{aligned}\tag{19}$$

with the distribution H of the Z_k used in $\mathcal{P}$ chosen to be the normal distribution $\mathrm{N}(0,A)$.

In the Pólya urn Gibbs samplers, the value of q_0^* for each Y_i and its X_i observation is proportional to

$$\frac{b+am}{\sqrt{2\pi(\sigma+A)}}\exp\left(\frac{-X_i^2}{2(\sigma+A)}\right).\tag{20}$$

When A is very large compared to $b+am$, this value becomes very small and suppresses the ability of the PG sampler to mix well. We look at this case, which is an example similar to the one studied by MacEachern (1994), and illustrates why the PG_a sampler improves mixing. We see that the BG sampler also performs well in this example.

We set $\sigma=1$ and $A=25$, and simulated $n=50$ observations X_i from a standard normal distribution. The Pólya urn Gibbs samplers were applied to three different priors $\mathcal{P}$: (1) the $\mathrm{DP}(\alpha H)$ process with $\alpha=1$, (2) the $\mathrm{DP}_N(\alpha H)$ process with $N=50$, $\alpha=1$, and (3) the $\mathcal{PY}(\alpha,0)$ process with $\alpha=0.25$, that is, the random measure (6) based on a stable random variable with index $\alpha=0.25$. The BG sampler also was applied to the $\mathrm{DP}_N(\alpha H)$ prior, but we substituted $\mathcal{P}_N(\mathbf{a}, \mathbf{b})$ almost sure truncations for priors 1 and 3 by using the methods of Section 3.2. For the $\mathrm{DP}(\alpha H)$ prior we used a truncation value of $N = 50$ and for the $\mathcal{PY}(0.25, 0)$ process we used $N = 100$. These values produce almost identical models by Theorem 2.

In each case we used a 2,000 iteration burn-in. After this we sequentially collected 1,000 batches of sampled values, each of size B. For each batch we computed the 0.05, 0.25, 0.75, and 0.95 percentile, as well as the mean, for each of the Y_i values. We then computed the mean value and standard deviation over the 1,000 batches for each of the summary statistics for each Y_i value. The averaged value over i is the (averaged) mean value and (averaged) standard deviation value we report in Tables 1–3, which correspond to $B = 50, 100, 250$. A small standard deviation is evidence of good mixing.

As expected, the PG sampler performance is poor in all examples, with standard deviations much larger than those seen in the PG_a and BG sampler. Both the PG_a and BG sampler perform well in all the examples; both exhibit fairly similar behavior, but with the slight edge going to the PG_a sampler for the Dirichlet process priors 1 and 2, and to the BG sampler for the Pitman–Yor process (prior 3). We suspect that the PG_a sampler did not do as well as the BG sampler in problem 3 because of the way q_0^* depends on the number of clusters m. Notice that by (20), the value for q_0^* is determined by $am = m/4$ ($a = 0.25, b = 0$), which becomes as small as $1/4$ when $m = 1$.

Our overall experiences with the three Gibbs sampling methods lead us to conclude that the accelerated PG_a sampler always should be applied in place of the PG sampler. The extra computations required are minimal and are more than offset by the improved mixing. We suspect that the PG_a sampler might be slightly more efficient than the BG sampler with small sample sizes in conjugate nonparametric models with the Dirichlet process, but with other types of priors and in nonconjugate models, we expect the BG sampler to mix more efficiently.

APPENDIX: PROOFS

A.1 PROOF OF LEMMA 1

To establish (5), take the limit of (4) as $N \to \infty$ and take logs to see that,

$$\sum_{k=1}^{\infty} p_k = 1 \quad \text{a.s.} \quad \text{iff} \quad \sum_{k=1}^{\infty} \log(1 - V_k) = -\infty \quad \text{a.s.}$$

The expression on the right-hand side is a sum of independent random variables and, therefore, by the Kolmogorov three series theorem equals $-\infty$ almost surely iff $\sum_{k=1}^{\infty} \mathrm{E}\big(\log(1 - V_k)\big) = -\infty$.

Alternatively, by (4),

$$\prod_{N=1}^{\infty} \frac{\mathrm{E}(\sum_{k=N+1}^{\infty} p_k)}{\mathrm{E}(\sum_{k=N}^{\infty} p_k)} = \prod_{N=1}^{\infty} \mathrm{E}(1 - V_N) = \prod_{N=1}^{\infty} \frac{b_N}{a_N + b_N}.$$

If $\sum_{N=1}^{\infty} \log(1 + a_N/b_N) = +\infty$, then the right-hand side equals zero and we must have that $\mathrm{E}(\sum_{k=1}^{N} p_k) \to 1$. However, because $\sum_{k=1}^{N} p_k$ is positive and increasing, it follows that $\sum_{k=1}^{\infty} p_k = 1$ almost surely.

A.2 PROOF OF THEOREM 1

Let $\mathcal{P}$ be a specific $\mathcal{PY}(a, b)$ random measure. Then

$$\mathcal{P}(\cdot) = V_1\, \delta_{Z_1}(\cdot) + (1 - V_1)\Big(V_1^*\, \delta_{Z_1^*}(\cdot) + (1 - V_1^*) V_2^*\, \delta_{Z_2^*}(\cdot) + (1 - V_1^*)(1 - V_2^*) V_3^*\, \delta_{Z_3^*}(\cdot) + \cdots\Big),$$

where $V_k^* = V_{k+1}$ are independent $\mathrm{Beta}(1 - a, b + a + ka)$ random variables and $Z_k^* = Z_{k+1}$ are iid H. From this, deduce that

$$\mathcal{P}(\cdot) \overset{\mathcal{D}}{=} V_1\, \delta_{Z_1}(\cdot) + (1 - V_1)\, \mathcal{P}^*(\cdot),$$

where, on the right-hand side, V_1, Z_1, and $\mathcal{P}^*$ are mutually independent, and $\mathcal{P}^*$ is a $\mathcal{PY}(a, b + a)$ process.

Table 1. Mean Value and Standard Deviation Over 1,000 Batches, Each of Size B = 50

		PG		PG_a		BG	
Prior	*Statistic*	*Mean*	*s.d.*	*Mean*	*s.d.*	*Mean*	*s.d.*
$\mathrm{DP}(\alpha H)$	.05 percentile	−.512	.253	−.711	.126	−.686	.136
	.25 percentile	−.388	.199	−.488	.051	−.474	.057
	mean	−.313	.149	−.323	.055	−.321	.056
	.75 percentile	−.259	.223	−.178	.071	−.189	.079
	.95 percentile	−.048	.308	.133	.178	.112	.183
$\mathrm{DP}_N(\alpha H)$	.05 percentile	−.513	.242	−.714	.128	−.690	.129
	.25 percentile	−.399	.187	−.490	.052	−.474	.053
	mean	−.326	.144	−.322	.056	−.322	.055
	.75 percentile	−.281	.215	−.174	.074	−.190	.077
	.95 percentile	−.060	.315	.138	.183	.112	.179
$\mathcal{PY}(0.25, 0)$	.05 percentile	−.336	.171	−.601	.081	−.582	.082
	.25 percentile	−.305	.129	−.444	.038	−.433	.038
	mean	−.287	.122	−.323	.042	−.322	.039
	.75 percentile	−.285	.149	−.224	.056	−.226	.047
	.95 percentile	−.193	.206	.009	.125	−.027	.128

NOTE: Summary statistics are evaluated for each Y_i ($i = 1, \ldots, n = 50$), for each of the 1,000 batches. The mean and standard deviations over the 1,000 batches for each Y_i are computed. The average value over i is the reported (averaged) mean and (averaged) standard deviation. Output is based on the model (19) via the Pólya urn Gibbs sampler (PG sampler) of Section 4.2, the accelerated version (PG_a sampler), and the blocked Gibbs sampler (BG sampler) of Section 5.2.

Table 2. Mean Value and Standard Deviation Over 1,000 batches, Each of Size B = 100, for Different Summary Statistics From Model (19)

		PG		*PG$_a$*		*BG*	
Prior	*Statistic*	*Mean*	*s.d.*	*Mean*	*s.d.*	*Mean*	*s.d.*
DP(αH)	.05 percentile	−.527	.235	−.726	.102	−.704	.111
	.25 percentile	−.411	.179	−.490	.036	−.478	.040
	mean	−.326	.130	−.322	.040	−.323	.041
	.75 percentile	−.271	.205	−.172	.052	−.187	.060
	.95 percentile	−.029	.293	.151	.143	.129	.150
DP$_N$(αH)	.05 percentile	−.528	.227	−.721	.098	−.704	.106
	.25 percentile	−.409	.173	−.490	.036	−.477	.038
	mean	−.328	.126	−.322	.039	−.322	.039
	.75 percentile	−.276	.198	−.174	.051	−.187	.057
	.95 percentile	−.041	.290	.149	.142	.130	.142
$\mathcal{PY}(0.25, 0)$	.05 percentile	−.347	.182	−.603	.059	−.586	.065
	.25 percentile	−.322	.154	−.444	.027	−.435	.028
	mean	−.298	.142	−.322	.030	−.322	.029
	.75 percentile	−.294	.179	−.225	.041	−.229	.034
	.95 percentile	−.191	.226	.019	.088	−.011	.102

NOTE: Mean values and standard deviations are computed as in Table 1.

Similar reasoning shows that

$$U_1(r,a,b) \overset{\mathcal{D}}{=} V_1^r + (1-V_1)^r\, U_1(r,a,b+a),$$

where, on the right-hand side, V_1 and $U_1(r,a,b+a)$ are mutually independent. Therefore, taking expectations,

$$\mathrm{E}\big(U_1(r,a,b)\big) = \frac{(1-a)^{(r)}}{(b+1)^{(r)}} + \frac{(b+a)^{(r)}}{(b+1)^{(r)}}\mathrm{E}\big(U_1(r,a,b+a)\big).$$

It is easy to check that the solution to this is

$$\mathrm{E}\big(U_1(r,a,b)\big) = \frac{(1-a)^{(r-1)}}{(b+1)^{(r-1)}}.$$

Furthermore, for $N \geq 2$, we have that

$$U_N(r,a,b) \overset{\mathcal{D}}{=} \Big((1-V_1)^r \cdots (1-V_{N-1})^r\Big)\, U_1(r,a,b+(N-1)a),$$

where all the variables on the right-hand side are mutually independent. Taking expectations, we have

$$\mathrm{E}\big(U_N(r,a,b)\big) = \left(\prod_{k=1}^{N-1} E\big(1-V_k\big)^r\right)\frac{(1-a)^{(r-1)}}{(b+(N-1)a+1)^{(r-1)}},$$

where the product is identified as $\mathrm{E}\big(T_N(r,a,b)\big)$ by noting that

$$T_N(r,a,b) \overset{\mathcal{D}}{=} \Big((1-V_1)^r \cdots (1-V_{N-1})^r\Big)\, T_1(r,a,b+(N-1)a)$$
$$= (1-V_1)^r \cdots (1-V_{N-1})^r.$$

A.3 LEMMA FOR PROOF OF THEOREM 3

The following lemma is a key fact that we need in the proof of Theorem 3. It is an extension of Lo's (1984, Lemma 1) work to the Pitman–Yor case.

Table 3. Mean Value and Standard Deviation Over 1,000 Batches, Each of Size B = 250, for Different Summary Statistics From Model (19)

		PG		*PG$_a$*		*BG*	
Prior	*Statistic*	*Mean*	*s.d.*	*Mean*	*s.d.*	*Mean*	*s.d.*
DP(αH)	.05 percentile	−.553	.205	−.728	.065	−.710	.081
	.25 percentile	−.421	.165	−.491	.023	−.478	.028
	mean	−.326	.111	−.323	.026	−.322	.027
	.75 percentile	−.256	.183	−.171	.033	−.185	.040
	.95 percentile	−.011	.253	.157	.094	.136	.106
DP$_N$(αH)	.05 percentile	−.548	.208	−.728	.064	−.708	.073
	.25 percentile	−.417	.161	−.491	.023	−.476	.024
	mean	−.324	.106	−.322	.025	−.322	.025
	.75 percentile	−.257	.179	−.171	.032	−.186	.037
	.95 percentile	−.017	.252	.157	.092	.137	.096
$\mathcal{PY}(0.25, 0)$	.05 percentile	−.377	.158	−.607	.038	−.588	.042
	.25 percentile	−.358	.142	−.444	.019	−.433	.018
	mean	−.325	.125	−.322	.019	−.321	.019
	.75 percentile	−.319	.169	−.227	.025	−.228	.022
	.95 percentile	−.199	.199	.026	.050	−.002	.067

NOTE: Mean values and standard deviations are computed as in Table 1.

Lemma 2. Let g be any nonnegative or quasiintegrable function on the measurable Polish space $(\mathcal{Y}^n \times \mathcal{M}, \mathcal{B}(\mathcal{Y}^n) \otimes \mathcal{B}(\mathcal{M}))$. Let $Y_1, \ldots, Y_n$ be a sample drawn from the $\mathcal{PY}(a, b)$ process $\mathcal{P}$. Then

$$\int_{\mathcal{M}} \int_{\mathcal{Y}^n} g(\mathbf{Y}, P)\, P(dY_1) \cdots P(dY_n)\, \mathcal{P}(dP) = \int_{\mathcal{Y}^n} \left[\int_{\mathcal{M}} g(\mathbf{Y}, P)\, \mathcal{P}(dP|\mathbf{Y}) \right] \pi(d\mathbf{Y}),$$

where $\mathcal{P}(\cdot|\mathbf{Y})$ is the posterior for $\mathcal{P}$ based on $\mathbf{Y} = (Y_1, \ldots, Y_n)$ as defined by (15) and $\pi(\mathbf{Y})$ is the joint distribution for $\mathbf{Y}$ as defined by the generalized Pólya urn (8).

Proof. The lemma is a special case of the Fubini–Tonelli theorem and the result follows because we have identified the appropriate disintegrations for the product measure

$$P(dY_1) \cdots P(dY_n)\, \mathcal{P}(dP)$$

of $(\mathbf{Y}, P)$.

A.4 PROOF OF THEOREM 3

We follow the style of proof used in Lo (1984, Theorem 1). By Bayes rule, $\mathcal{P}(P|\mathbf{X})$ is characterized by

$$\int_{\mathcal{M}} \psi(P) \mathcal{P}(dP|\mathbf{X}) = \frac{\int_{\mathcal{M}} \psi(P) \int_{\Re^d} \prod_{i=1}^n \int_{\mathcal{Y}} f(X_i|Y_i, \theta)\, P(dY_i)\, \pi(d\theta)\, \mathcal{P}(dP)}{\int_{\mathcal{M}} \int_{\Re^d} \prod_{i=1}^n \int_{\mathcal{Y}} f(X_i|Y_i, \theta)\, P(dY_i)\, \pi(d\theta)\, \mathcal{P}(dP)}. \quad (21)$$

By Fubini's theorem, we can rewrite the numerator as

$$\int_{\Re^d} \int_{\mathcal{M}} \int_{\mathcal{Y}^n} \left[\psi(P) \prod_{i=1}^n f(X_i|Y_i, \theta) \right] P(dY_1) \cdots P(dY_n)\, \mathcal{P}(dP)\, \pi(d\theta).$$

By applying Lemma 2 to the integrand where $g(\mathbf{Y}, P)$ corresponds to the expression in square brackets, the previous expression now becomes

$$\int_{\Re^d} \int_{\mathcal{Y}^n} \prod_{i=1}^n f(X_i|Y_i, \theta) \int_{\mathcal{M}} \psi(P)\, \mathcal{P}(dP|\mathbf{Y})\, \pi(d\mathbf{Y})\, \pi(d\theta) = \int_{\mathcal{Y}^n} \int_{\mathcal{M}} \psi(P)\, \mathcal{P}(dP|\mathbf{Y}) \left(\int_{\Re^d} \prod_{i=1}^n f(X_i|Y_i, \theta)\, \pi(d\theta) \right) \pi(d\mathbf{Y}).$$

The last rearrangement follows from the Fubini–Tonelli theorem.

Applying this same argument to the denominator of (21) with $\psi(P) = 1$ results in the expression (14).

[Received February 2000. Revised July 2000.]

REFERENCES

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152–1174.

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

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

Carlton, M. A. (1999), "Applications of the Two-Parameter Poisson–Dirichlet Distribution," unpublished Ph.D. thesis, University of California, Los Angeles, Dept. of Statistics.

Connor, R. J., and Mosimann, J. E. (1969), "Concepts of Independence for Proportions With a Generalization of the Dirichlet Distribution," *Journal of the American Statistical Association*, 64, 194–206.

Damien, P., Laud, P. W., and Smith, A. F. M. (1996), "Implementation of Bayesian Non-Parametric Inference Based on Beta Processes," *Scandinavian Journal of Statistics*, 23, 27–36.

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

Doss, H. (1994), "Bayesian Nonparametric Estimation for Incomplete Data via Successive Substitution Sampling," *The Annals of Statistics*, 22, 1763–1786.

Escobar, M. D. (1988), "Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means," unpublished Ph.D. thesis, Yale University, Dept. of Statistics.

——— (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.

——— (1998), "Computing Nonparametric Hierarchical Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Mueller, and D. Sinha, New York: Springer-Verlag, 1–22.

Fabius, J. (1964), "Asymptotic Behavior of Bayes Estimates," *Annals of Mathematical Statistics*, 35, 846–856.

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

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

Ferguson, T. S., and Klass, M. J. (1972), "A Representation of Independent Increment Processes Without Gaussian Components," *The Annals of Mathematical Statistics*, 43, 1634–1643.

Freedman, D. (1963), "On the Asymptotic Behavior of Bayes Estimates in the Discrete Case," *Annals of Mathematical Statistics*, 34, 1386–1403.

Green, P., and Richardson, S. (1999), "Modelling Heterogeneity With and Without the Dirichlet Process," unpublished manuscript.

Halmos, P. (1944), "Random Alms," *The Annals of Mathematical Statistics*, 15, 182–189.

Hjort, N. L. (1990), "Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data," *The Annals of Statistics*, 18, 1259–1294.

Ishwaran, H., and James, L. F. (2000), "Approximate Dirichlet Process Computing for Finite Normal Mixtures: Smoothing and Prior Information," unpublished manuscript.

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

——— (2000b), "Exact and Approximate Sum-Representations for the Dirichlet Process," unpublished manuscript.

——— (2000c), "Dirichlet Prior Sieves in Finite Normal Mixtures," unpublished manuscript.

Kerov, S. (1995), "Coherent Random Allocations and the Ewens–Pitman Formula," PDMI Preprint, Steklov Institute of Mathematics, St. Petersburg.

Kingman, J. F. C. (1974), "Random Discrete Distributions," *Journal of the Royal Statistical Society*, Series B, 37, 1–22.

Laud, P. W., Damien, P., and Smith, A. F. M. (1998), "Bayesian Nonparametric and Covariate Analysis of Failure Time Data," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Mueller, and D. Sinha, New York: Springer-Verlag, pp. 213–225.

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

——— (1998), "Computational Methods for Mixture of Dirichlet Process Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Mueller, and D. Sinha, New York: Springer-Verlag.

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

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

Mekjian A. Z., and Chase, K. C. (1997), "Disordered Systems, Power Laws and Random Processes," *Physics Letters A*, 229 340–346.

Muliere, P., and Secchi, P. (1995), "A Note on a Proper Bayesian Bootstrap," Technical Report 18, Università degli Sudi di Pavia, Dipartimento di Economia Politica e Metodi Quantitativi.

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

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

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

Pitman, J. (1995), "Exchangeable and Partially Exchangeable Random Partitions," *Probability Theory amd Related Fields*, 102, 145–158.

——— (1996a), "Some Developments of the Blackwell–MacQueen Urn Scheme," in *Statistics, Probability and Game Theory*, eds. T. S. Ferguson, L. S. Shapley, and J. B. MacQueen, IMS Lecture Notes—Monograph Series (Vol. 30), Hayward, CA: Institute of Mathematical Statistics, pp. 245–267.

——— (1996b), "Random Discrete Distributions Invariant Under Size-Biased Permutation," *Advances in Applied Probability* 28, 525–539.

——— (1997), "Partition Structures Derived From Brownian Motion and Stable Subordinators," *Bernoulli*, 3, 79–86.

——— (1999), "Coalescents With Multiple Collisions," *The Annals of Probability*, 27, 1870–1902.

Pitman, J. and Yor, M. (1997), "The Two-Parameter Poisson–Dirichlet Distribution Derived From a Stable Subordinator," *The Annals of Probability*, 25, 855–900.

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

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

Tsilevich, N. V. (1997), "Distribution of the Mean Value for Certain Random Measures," POMI Preprint 240, Steklov Institute of Mathematics, St. Petersburg.

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

Walker, S., and Wakefield, J. (1998), "Population models With a Nonparametric Random Coefficient Distribution," *Sankhy*a, Ser. B, 60, 196–214.

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

Zabell, S. L. (1997), "The Continuum of Inductive Methods Revisited," in *The Cosmos of Science*, eds. J. Earman and J. D. Norton, Pittsburgh–Konstanz Series in the Philosophy and History of Science, Pittsburgh, PA: University of Pittsburgh Press, pp. 351–385.