Defining Predictive Probability Functions for Species Sampling Models

Jaeyong Lee

Department of Statistics, Seoul National University

leejyc@gmail.com

Fernando A. Quintana

Departamento de Estadísica, Pontificia Universidad Católica de Chile

quintana@mat.puc.cl

Peter Müller and Lorenzo Trippa

Department of Biostatistics, M. D. Anderson Cancer Center

{pmueller,ltrippa}@mdanderson.org

November 17, 2008

Abstract

We study the class of species sampling models (SSM). In particular, we investigate the relation between the exchangeable partition probability function (EPPF) and the predictive probability function (PPF). An EPPF defines a PPF, but the converse is not necessarily true. In this paper, we show novel conditions for a PPF to define an EPPF. We show that all possible PPF's in a certain class have to define (unnormalized) probabilities for cluster membership that are linear in cluster size. We give a new necessary and sufficient condition for arbitrary PPFs to define an EPPF. Finally we construct a new class of SSM's with PPF that is not linear in cluster size. [1]

1 Introduction

We study the nature of predictive probability functions (PPF) that define species sampling models (SSMs) (Pitman, 1996). Almost all known SSMs are characterized by a PPF that is essentially a linear function of cluster size. We study conditions for more general PPFs and propose a large class of such SSMs. The PPF for the new model class is not easily described in closed form anymore. We provide instead a numerical algorithm that allows easy posterior simulation.

By far, the most popular example of SSM is the Dirichlet process (Ferguson, 1973) (DP). The status of the DP among nonparametric priors has been that of the normal distribution among finite dimensional distributions. This is in part due to the marginalization property: a random sequence sampled from a random probability measure with a Dirichlet process prior forms marginally a Polya urn sequence. Markov chain Monte Carlo

[1] *AMS 2000 subject classifications:* Primary 62C10; secondary 62G20

Key words and phrases: Species sampling Prior, Exchangeable partition probability functions, Prediction probability functions

Jaeyong Lee was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-070-C00021). Fernando Quintana was supported by grants FONDECYT 1060729 and Laboratorio de Análisis Estocástico PBCT-ACT13.

simulation based on the marginalization property has been the central computational tool for the DP and facilitated a wide variety of applications. See MacEachern (1994), Escobar and West (1995) and MacEachern and Müller (1998), to name just a few. In Pitman (1995,1996), the species sampling prior (SSM) is proposed as a generalization of the DP. SSMs can be used as flexible alternatives to the popular DP model in nonparametric Bayesian inference. The SSM is defined as the directing random probability measure of an exchangeable species sampling sequence which is defined as a generalization of the Polya urn sequence. The SSM has a marginalization property similar to the DP. It therefore enjoys the same computational advantage as the DP while it defines a much wider class of random probability measures. For its theoretical properties and applications, we refer to Ishwaran and James (2003), Lijoi, Mena and Prünster (2005), Lijoi, Prünster and Walker (2005), James (2006), Jang, Lee and Lee (2007), and Navarrete, Quintana and Müller (2008).

Suppose $(X_1, X_2, \ldots)$ is a sequence of random variables. This sequence is typically considered a random sample from a large population of species, i.e. X_i is the species of the ith individual sampled. Let M_j be the index of the first observed individual of the j-th species. We define $M_1 = 1$ and $M_j = \inf\{n : n > M_{j-1}, X_n \notin \{X_1, \ldots, X_{n-1}\}\}$ for $j = 2, 3, \ldots$, with the convention $\inf \emptyset = \infty$. Let $\tilde{X}_j = X_{M_j}$ be the jth distinct species to appear which is defined conditional on the event $M_j < \infty$. Let n_{jn} be the number of times the jth species $\tilde{X}_j$ appears in $(X_1, \ldots, X_n)$, i.e.,

$$n_{jn} = \sum_{m=1}^{n} I(X_m = \tilde{X}_j), \quad j = 1, 2, \ldots \text{ and } \mathbf{n}_n = (n_{1n}, n_{2n}, \ldots) \text{ or } (n_{1n}, n_{2n}, \ldots, n_{k_n,n}),$$

where $k_n = k_n(\mathbf{n}_n) = \max\{j : n_{jn} > 0\}$ is the number of different species to appear in $(X_1, \ldots, X_n)$. The sets $\{i : X_i = \tilde{X}_j\}$ define clusters that partition the index set $\{1, \ldots, n\}$. When n is understood from the context we just write n_j, $\mathbf{n}$ and k or $k(\mathbf{n})$.

We now give three alternative characterizations of species sampling sequences (i) by the predictive probability function, (ii) by the driving measure of the exchangeable sequence, and (iii) by the underlying exchangeable exchangeable partition probability function.

PPF: Let ν be a diffuse (or nonatomic) probability measure on $\mathcal{X}$. An exchangeable sequence $(X_1, X_2, \ldots)$ is called a species sampling sequence (SSS) if $X_1 \sim \nu$ and

$$X_{n+1} \mid X_1, \ldots, X_n \sim \sum_{j=1}^{k_n} p_j(\mathbf{n}_n)\delta_{\tilde{X}_j} + p_{k_n+1}(\mathbf{n}_n)\nu, \tag{1}$$

where δ_x is the degenerate probability measure at x. The sequence of functions $(p_1, p_2, \ldots)$ in (1) is called a sequence of predictive probability functions (PPF). It is defined on $\mathbb{N}^* = \cup_{k=1}^{\infty}\mathbb{N}^k$, where $\mathbb{N}$ is the set of natural numbers, and satisfies the conditions

$$p_j(\mathbf{n}) \geq 0 \text{ and } \sum_{j=1}^{k_n+1} p_j(\mathbf{n}) = 1, \text{ for all } \mathbf{n} \in \mathbb{N}^*. \tag{2}$$

Motivated by these properties of PPFs, we define a sequence of *putative PPFs* as a sequence of functions $(p_j, j = 1, 2, \ldots)$ defined on $\mathbb{N}^*$ which satisfies (2). Note that not all putative PPFs are PPFs, because (2) does not guarantee exchangeability of $(X_1, X_2, \ldots)$ in (1). An important feature in this defining property is that the weights $p_j(\cdot)$ depend on the data only indirectly through the cluster sizes $\mathbf{n}_n$. The widely used DP is a special case of a species sampling model, with $p_j(\mathbf{n}_n) \propto n_j$ and $p_{k+1}(\mathbf{n}_n) \propto \alpha$ for a DP with total mass parameter α. The use of p_j in (1) implies

$$\begin{aligned} p_j(\mathbf{n}) &= \mathbb{P}(X_{n+1} = \tilde{X}_j \mid X_1, \ldots, X_n), \quad j = 1, \ldots, k_n, \\ p_{k_n+1}(\mathbf{n}) &= \mathbb{P}(X_{n+1} \notin \{X_1, \ldots, X_n\} \mid X_1, \ldots, X_n). \end{aligned}$$

In words, p_j is the probability of the next observation being the j-th species (falling into the j-th cluster) and p_{k+1} is the probability of a new species (starting a new cluster).

An important point in the above definition is that the sequence X_i be exchangeable. The implied sequence X_i is an SSS only if it is exchangeable. With only one exception (Lijoi et al 2005), all known SSSs have a PPF of the form

$$p_j(\mathbf{n}) \propto \begin{cases} a + bn_j & j = 1, \ldots, k_n, \\ \theta(\mathbf{n}) & j = k_n + 1. \end{cases} \tag{3}$$

In Section 2 we show that all PPFs of the form $p_j(\mathbf{n}) = f(n_j)$ must be of the form (3). A corollary of these results is that a PPF p_j that depends on n_j other than linearly must be of a more complicated form. We define such PPF's in Section 3.

SSM: Alternatively a SSS can be characterized by the following defining property. Let δ_x denote a point mass at x. An exchangeable sequence of random variables (X_n) is a species sampling sequence if and only if $X_1, X_2, \ldots \mid G$ is a random sample from G where

$$G = \sum_{h=1}^{\infty} P_h \delta_{m_h} + R\nu, \tag{4}$$

for some sequence of positive random variables (P_h) and R such that $1-R = \sum_{i=h}^{\infty} P_h \leq 1$, (m_h) is a random sample from ν, and (P_i) and (m_h) are independent. See Pitman (1996). The result is an extension of the de Finetti's Theorem and characterizes the directing random probability measure of the species sample sequence. We call the directing random probability measure G in equation (4) the *SSM (or species sampling process)* of the SSS (X_i).

EPPF: A third alternative definition of an SSS and corresponding SSM is in terms of the implied probability model on a sequence of nested random partitions. Let $[n] = \{1, 2, \ldots, n\}$ and $\mathbb{N}$ be the set of natural numbers. A symmetric function $p : \mathbb{N}^* \to [0, 1]$ satisfying

$$\begin{aligned} p(1) &= 1, \\ p(\mathbf{n}) &= \sum_{j=1}^{k(\mathbf{n})+1} p(\mathbf{n}^{j+}), \text{ for all } \mathbf{n} \in \mathbb{N}^*, \end{aligned} \tag{5}$$

where $\mathbf{n}^{j+}$ is the same as $\mathbf{n}$ except that the jth element is increased by 1, is called an exchangeable partition probability function (EPPF). An EPPF $p(\mathbf{n})$ can be interpreted as a joint probability model for the vector of cluster sizes $\mathbf{n}$ implied by the configuration of ties in a sequence $(X_1, \ldots, X_n)$. The following result can be found in Pitman (1995). The joint probability $p(\mathbf{n})$ of cluster sizes defined from the ties in an exchangeable sequence (X_n) is an EPPF, i.e., satisfies (5). Conversely, for every symmetric function $p : \mathbb{N}^* \to [0, 1]$ satisfying (5) there is an exchangeable sequence that gives rise to p.

We are now ready to pose the problem for the present paper. It is straightforward to verify that any EPPF defines a PPF by

$$p_j(\mathbf{n}) = \frac{p(\mathbf{n}^{j+})}{p(\mathbf{n})}, \quad j = 1, 2, \ldots, k+1. \tag{6}$$

The converse is not true. Not every putative $p_j(\mathbf{n})$ defines an EPPF and thus an SSM and SSS. For example, it is easy to show that $p_j(\mathbf{n}) \propto n_j^2$ does not. In this paper we address two related issues. We characterize all possible PPF's with $p_j(\mathbf{n}) \propto f(n_j)$ and we provide a large new class of SSM's with PPF's that are not restricted to this format. Throughout this paper we will use the term PPF for a sequence of probabilities $p_j(\mathbf{n})$ only if there is a corresponding EPPF, i.e., $p_j(\cdot)$ characterizes a SSM. Otherwise we refer to $p_j(\cdot)$ as a putative PPF.

The questions are important for nonparametric Bayesians data analysis. It is often convenient or at least instructive to elicit features of the PPF rather than the joint EPPF. Since the PPF is the crucial property for posterior computation, applied Bayesians tend to focus on the PPF to generalize the species sampling prior for a specific problem. For example, the PPF defined by a DP prior implies that the probability of joining an existing cluster is proportional to the cluster size. This is not always desirable. Can the user define an alternative PPF that allocates new observations to clusters with probabilities proportional to alternative functions $f(n_j)$, and still define a SSS? In general, the simple answer is no. We already mentioned that a PPF implies a SSS if and only if it arises as in (6) from an EPPF. But this result is only a characterization. It is of little use for data analysis and modeling since it is difficult to verify whether or not a given PPF arises from an EPPF. In this paper we develop several novel conditions to address this gap. First we give an easily verifiable necessary condition for an PPF to arise from an EPPF (Lemma 1). We then exhaustively characterize PPFs of certain forms (Corollaries 1 and 2). Next we give a necessary and sufficient condition for a PPF to arise from an EPPF. Finally we propose an alternative approach to define an SSM based on directly defining a joint probability model for the P_h in (4). We develop a numerical algorithm to derive the corresponding PPF. This facilitates the use of such models for nonparametric Bayesian data analysis.

2 When does a PPF imply an EPPF?

Suppose we are given a putative PPF (p_j). Using equation (6), one can attempt to define a function $p : \mathbb{N}^* \to [0,1]$ inductively by the following mapping:

$$\begin{aligned} p(1) &= 1 \\ p(\mathbf{n}^{j+}) &= p_j(\mathbf{n})p(\mathbf{n}), \quad \text{for all } \mathbf{n} \in \mathbb{N} \text{ and } j = 1, 2, \ldots, k(\mathbf{n}) + 1. \end{aligned} \tag{7}$$

In general equation (7) does not lead to a unique definition $p(\mathbf{n})$ for each $\mathbf{n} \in \mathbb{N}^*$. For example, let $\mathbf{n} = (2,1)$. Then, $p(2,1)$ could be computed in two different ways as $p_2(1)p_1(1,1)$ and $p_1(1)p_2(2)$ which correspond to partitions $\{\{1,2\},\{3\}\}$ and $\{\{1,3\},\{2\}\}$, respectively. If $p_2(1)p_1(1,1) \neq p_1(1)p_2(2)$, equation (7) does not define a function $p : \mathbb{N}^* \to [0,1]$. The following lemma shows a condition for a PPF for which equation (7) leads to a valid unique definition of $p : \mathbb{N}^* \to [0,1]$.

Suppose $\Pi = \{A_1, A_2, \ldots, A_k\}$ is a partition of $[n]$ in the order of appearance. For $1 \leq m \leq n$, let Π_m be the restriction of Π on $[m]$. Let $\mathbf{n}(\Pi) = (n_1, \ldots, n_k)$ where n_i is the cardinality of A_i, $\Pi(i)$ be the class index of element i in partition Π and $\Pi([n]) = (\Pi(1), \ldots, \Pi(n))$.

Lemma 1. If and only if a putative PPF (p_j) satisfies

$$p_i(\mathbf{n})p_j(\mathbf{n}^{i+}) = p_j(\mathbf{n})p_i(\mathbf{n}^{j+}), \quad \text{for all } \mathbf{n} \in \mathbb{N}^*, i, j = 1, 2, \ldots, k(\mathbf{n}) + 1, \tag{8}$$

then p defined by (7) is a function from $\mathbb{N}^*$ to $[0,1]$, i.e., p in (7) is uniquely defined. Any permutation $\Pi([n])$ leads to the same value.

Proof. Let $\mathbf{n} = (n_1, \ldots, n_k)$ with $\sum_{i=1}^k n_i = n$ and Π and Ω be two partitions of $[n]$ with $\mathbf{n}(\Pi) = \mathbf{n}(\Omega) = \mathbf{n}$. Let $p^{\Pi}(\mathbf{n}) = \prod_{i=1}^{n-1} p_{\Pi(i+1)}(\mathbf{n}(\Pi_i))$ and $p^{\Omega}(\mathbf{n}) = \prod_{i=1}^{n-1} p_{\Omega(i+1)}(\mathbf{n}(\Omega_i))$. We need to show that $p^{\Pi}(\mathbf{n}) = p^{\Omega}(\mathbf{n})$. Without loss of generality, we can assume $\Pi([n]) = (1, \ldots, 1, 2, \ldots, 2, \ldots, k, \ldots, k)$ where i is repeated n_i times for $i = 1, \ldots, k$. Note that $\Omega([n])$ is just a certain permutation of $\Pi([n])$ and by a finite times of swapping two consecutive elements in $\Omega([n])$ one can change $\Omega([n])$ to $\Pi([n])$. Thus, it suffices to show when $\Omega([n])$ is different from $\Pi([n])$ in only two consecutive positions. But, this is guaranteed by condition (8).

The opposite is easy to show. Assume p_j defines a unique $p(\mathbf{n})$. Consider (8) and multiply on both sides with $p(\mathbf{n})$. By assumption we get on either side $p(\mathbf{n}^{i+j+})$. This completes the proof. ■

Note that the conclusion of Lemma 1 is not (yet) that p is an EPPF. The missing property is symmetry, i.e., invariance of p with respect to permutations of the group indices $j = 1, \ldots, k(\mathbf{n})$. We have only established invariance with respect to permutations of the subject indices $i = 1, \ldots, n$. But the result is very useful. It is easily verified for any given family p_i. The following two straightforward corollaries exhaustively characterize all possible PPFs that depend on group sizes in certain ways that are natural choices when defining a probability model. Corollary 1 describes all possible PPFs that have the probability of cluster memberships depend on a function of the cluster size only. Corollary 2 generalizes slightly by allowing cluster membership probabilities to depend on the cluster size and the number of clusters.

Corollary 1. Suppose a putative PPF (p_j) satisfies (8) and

$$p_j(n_1, \ldots, n_k) \propto \begin{cases} f(n_j), & j = 1, \ldots, k \\ \theta, & j = k+1, \end{cases}$$

where $f(k)$ is a function from $\mathbb{N}$ to $(0, \infty)$ and $\theta > 0$. Then, $f(k) = ak$ for all $k \in \mathbb{N}$ for some $a > 0$.

Proof. Note that for any $\mathbf{n} = (n_1, \ldots, n_k)$ and $i = 1, \ldots, k+1$,

$$p_i(n_1, \ldots, n_k) = \begin{cases} \frac{f(n_i)}{\sum_{u=1}^k f(n_u)+\theta}, & i = 1, \ldots, k \\ \frac{\theta}{\sum_{u=1}^k f(n_u)+\theta}, & i = k+1. \end{cases}$$

Equation (8) with $1 \leq i \neq j \leq k$ implies

$$\frac{f(n_i)}{\sum_{u=1}^k f(n_u)+\theta} \frac{f(n_j)}{\sum_{u\neq i}^k f(n_u)+f(n_i+1)+\theta} = \frac{f(n_j)}{\sum_{u=1}^k f(n_u)+\theta} \frac{f(n_i)}{\sum_{u\neq j}^k f(n_u)+f(n_j+1)+\theta},$$

which in turn implies

$$f(n_i) + f(n_j + 1) = f(n_j) + f(n_i + 1)$$

or

$$f(n_j + 1) - f(n_j) = f(n_i + 1) - f(n_i).$$

Since this holds for all n_i and n_j, we have for all $k \in \mathbb{N}$

$$f(m) = am + b, \tag{9}$$

for some $a, b \in \mathbb{R}$.

Now consider $i = k + 1$ and $1 \leq j \leq k$. Then,

$$\frac{\theta}{\sum_{u=1}^{k} f(n_u) + \theta} \frac{f(n_j)}{\sum_{u=1}^{k} f(n_u) + f(1) + \theta} = \frac{f(n_j)}{\sum_{u=1}^{k} f(n_u) + \theta} \frac{\theta}{\sum_{u \neq j}^{k} f(n_u) + f(n_j + 1) + \theta},$$

which implies $f(n_j) + f(1) = f(n_j + 1)$ for all n_j. This together with (9) implies $b = 0$. Thus, we have $f(k) = ak$ for some $a > 0$. ■

Remark 1. For any $a > 0$, the putative PPF

$$p_i(n_1, \ldots, n_k) \propto \begin{cases} an_i, & i = 1, \ldots, k \\ \theta, & i = k + 1 \end{cases}$$

defines a function $p : \mathbb{N} \to [0, 1]$

$$p(n_1, \ldots, n_k) = \frac{\theta^{k-1} a^{n-k}}{[\theta + 1]_{n-1;a}} \prod_{i=1}^{k} (n_i - 1)!,$$

where $[\theta]_{k;a} = \theta(\theta + a) \ldots (\theta + (k - 1)a)$. Since this function is symmetric in its arguments, it is an EPPF.

Corollary 1 characterizes all valid PPFs with $p_j = c\, f(n_j)$ and $p_{k+1} = c\, \theta$. The result does not exclude possible valid PPFs with a probability for a new cluster p_{k+1} that depends on $\mathbf{n}$ and k in different ways.

Corollary 2. Suppose a putative PPF (p_j) satisfies (8) and

$$p_j(n_1, \ldots, n_k) \propto \begin{cases} f(n_j, k), & j = 1, \ldots, k \\ g(n, k), & j = k + 1, \end{cases}$$

where $f(m, k)$ and $g(m, k)$ are functions from $\mathbb{N}^2$ to $(0, \infty)$. Then, the following hold:

(a) $f(m,k) = a_k m + b_k$ for some constants $a_k \geq 0$ and b_k.

(b) If $f(m,k)$ depends only on m, then $f(m,k) = am + b$ for some constants $a \geq 0$ and b.

(c) If $f(m,k)$ depends only on m and $g(m,k)$ depends only on k, then $f(m,k) = am+b$ and $g(k) = g(1) - b(k-1)$ for some constants $a \geq 0$ and $b < 0$.

(d) If $g(n,k) = \theta > 0$ and $b_k = 0$, then $a_k = a$ for all k.

(e) If $g(n,k) = g(k)$ and $b_k = 0$, then $g(k)a_{k+1} = g(k+1)a_k$.

Proof. For all k and $1 \leq i \neq j \leq k$, using the similar argument as in the proof of Corollary 1, we get

$$f(n_i + 1, k) - f(n_i, k) = f(n_j + 1, k) - f(n_j, k).$$

Thus, we have $f(m,k) = a_k m + b_k$ for some a_k, b_k. If $a_k < 0$, for sufficiently large m, $f(m,k) < 0$. Thus, $a_k \geq 0$. This completes the proof of (a). (b) follows from (a). For (c), consider (8) with $i = k+1$ and $1 \leq j \leq k$. With some algebra with (b), we get

$$g(k+1) - g(k) = f(n_j + 1) - f(n_j) - f(1) = -b,$$

which implies (c). (d) and (e) follow from (8) with $i = k+1$ and $1 \leq j \leq k$. ■

A prominent example of PPFs of the above form is the PPF implied by the Pitman-Yor process (Pitman and Yor, 1997). Consider a Pitman-Yor process with discount, strength and baseline parameters d, θ and G_0. The PPF is as in (c) above with $a = 1$ and $b = -d$.

Corollaries 1 and 2 describe practically useful, but still restrictive forms of the PPF. The characterization of valid PPFs can be further generalized. We now give a necessary and sufficient conditions for the function p defined by (6) to be an EPPF, without any constraint on the form of p_j (as were present in the earlier results). Suppose σ is a permutation of $[k]$ and $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$. Define $\sigma(\mathbf{n}) = \sigma(n_1, \ldots, n_k) = (n_{\sigma(1)}, n_{\sigma(2)}, \ldots, n_{\sigma(k)})$. In words, σ is a permutation of group labels and $\sigma(\mathbf{n})$ is the corresponding permutation of the group sizes $\mathbf{n}$.

Theorem 1. Suppose a putative PPF (p_j) satisfies (8) as well as the following condition: for all $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$, and permutations σ on $[k]$ and $i = 1, \ldots, k$,

$$p_i(n_1, \ldots, n_k) = p_{\sigma^{-1}(i)}(n_{\sigma(1)}, n_{\sigma(2)}, \ldots, n_{\sigma(k)}). \tag{10}$$

Then, p defined by (7) is an EPPF. The condition is also necessary. If p is an EPPF then (7) and (10) hold.

Proof. Fix $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$ and a permutation on $[k]$, σ . We wish to show that for the function p defined by (7)

$$p(n_1, \ldots, n_k) = p(n_{\sigma(1)}, n_{\sigma(2)}, \ldots, n_{\sigma(k)}). \tag{11}$$

Let Π be the partition of $[n]$ with $\mathbf{n}(\Pi) = (n_1, \ldots, n_k)$ such that

$$\Pi([n]) = (1, 2, \ldots, k, 1, \ldots, 1, 2, \ldots, 2, \ldots, k, \ldots, k),$$

where after the first k elements $1, 2, \ldots, k$, i is repeated $n_i - 1$ times for all $i = 1, \ldots, k$. Then,

$$p(\mathbf{n}) = \prod_{i=2}^{k} p_i(\mathbf{1}_{(i-1)}) \times \prod_{i=k}^{n-1} p_{\Pi(i+1)}(\mathbf{n}(\Pi_i)),$$

where $\mathbf{1}_{(j)}$ is the vector of length j whose elements are all 1's.

Now consider a partition Ω of $[n]$ with $\mathbf{n}(\Omega) = (n_{\sigma(1)}, n_{\sigma(2)}, \ldots, n_{\sigma(k)})$ such that

$$\Omega([n]) = (1, 2, \ldots, k, \sigma^{-1}(1), \ldots, \sigma^{-1}(1), \sigma^{-1}(2), \ldots, \sigma^{-1}(2), \ldots, \sigma^{-1}(k), \ldots, \sigma^{-1}(k)),$$

where after the first k elements $1, 2, \ldots, k$, $\sigma^{-1}(i)$ is repeated $n_i - 1$ times for all $i = 1, \ldots, k$. Then,

$$\begin{aligned}
p(n_{\sigma(1)}, n_{\sigma(2)}, \ldots, n_{\sigma(k)}) &= \prod_{i=2}^{k} p_i(\mathbf{1}_{(i-1)}) \times \prod_{i=k}^{n-1} p_{\Omega(i+1)}(\mathbf{n}(\Omega_i)) \\
&= \prod_{i=2}^{k} p_i(\mathbf{1}_{(i-1)}) \times \prod_{i=k}^{n-1} p_{\sigma^{-1}(\Omega(i+1))}(\sigma(\mathbf{n}(\Omega_i))) \\
&= \prod_{i=2}^{k} p_i(\mathbf{1}_{(i-1)}) \times \prod_{i=k}^{n-1} p_{\Pi(i+1)}(\mathbf{n}(\Pi_i)) \\
&= p(n_1, \ldots, n_k),
\end{aligned}$$

where the second equality follows from (10). This completes the proof of the sufficient direction.

Finally, we show that every EPPF p satisfies (7) and (10). By Lemma 1 every EPPF satisfies (7). Condition (11) is true by the definition of an EPPF, which includes the condition of symmetry in its arguments. And (11) implies (10). ■

3 A New Class of SSM's

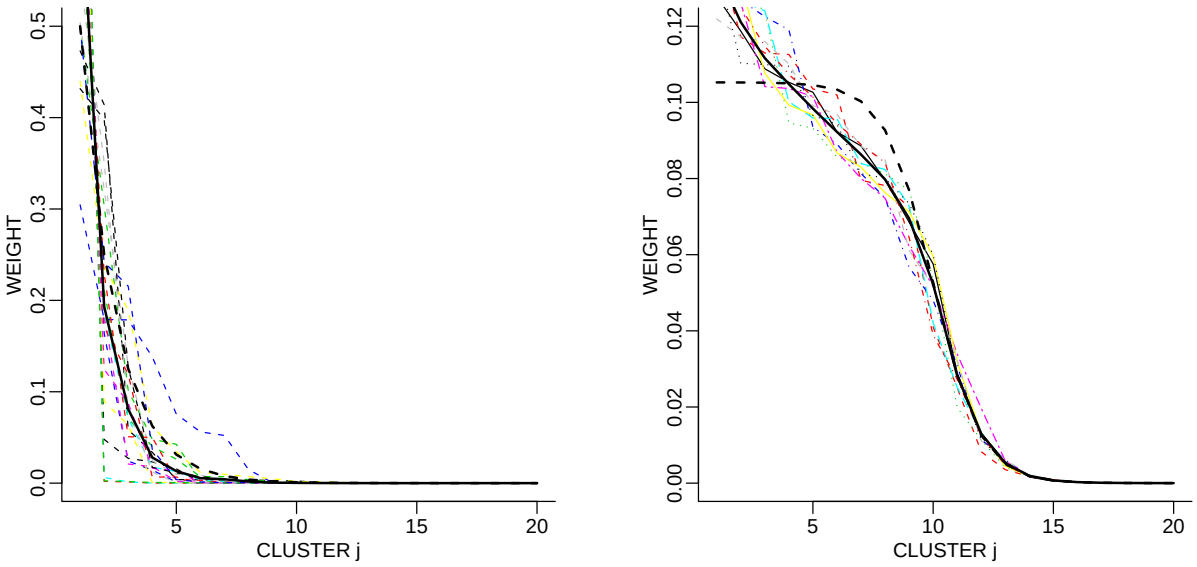
3.1 The $SSM(p, G_0)$

We now know that an SSM with a non-linear PPF, i.e., p_j different from (3), can not be described as a function $p_j \propto f(n_j)$ of n_j only. It must be a more complicated function $f(\mathbf{n})$. Alternatively one could try to define an EPPF, and deduce the implied PPF. But directly specifying a function $p(\mathbf{n})$ such that it complies with (5) is difficult. As a third alternative we propose to consider the weights $\mathbf{P} = \{P_h,\ h = 1, 2, \ldots\}$ in (4). Figure 1a illustrates $p(\mathbf{P})$ for a DP model. The sharp decline is typical. A few large weights account for most of the probability mass. Figure 1b shows an alternative probability model $p(\mathbf{P})$. There are many ways to define $p(\mathbf{P})$. In this example, we defined, for $h = 1, 2, \ldots$,

$$P_h \propto e^{X_h} \text{ with } X_h \sim N(log(1 - (1 + e^{b-ah})^{-1}), \sigma^2), \tag{12}$$

where a, b, σ^2 are positive constants. The S-shaped nature of the random distribution (plotted against h) distinguishes it from the DP model. The first few weights are *a priori* of equal size (before sorting). This is in contrast to the stochastic ordering of the DP and the Pitman-Yor process in general. In panel (a) the prior mean of the sorted and unsorted weights is almost indistinguishable, because the prior already implies strong stochastic ordering of the weights.

We use SSM(p, G_0) to denote a SSM defined by $p(\mathbf{P})$ for the weights P_h and $m_h \sim G_0$, i.i.d. The attraction of defining the SSM through $\mathbf{P}$ is that by (4) any joint probability model $p(\mathbf{P})$ defines an SSS, with the additional assumption of $P(R = 0) = 1$, i.e. a proper SSM (Pitman, 96). There are no additional constraints as for the PPF $p_j(\mathbf{n})$ or the EPPF



(a) $DP(M = 1, G_0)$ (b) SSM(p, G_0) (note the shorter y-scale).

Figure 1: The lines in each panel show 10 draws $\mathbf{P} \sim p(\mathbf{P})$. The P_h are defined for integers h only. We connect them to a line for presentation only. Also, for better presentation we plot the *sorted* weights. The thick line shows the prior mean. For comparison, a dashed thick line plots the prior mean of the *unsorted* weights. Under the DP the sorted and unsorted prior means are almost indistinguishable.

$p(\mathbf{n})$. However, we still need the implied PPF to implement posterior inference, and also to understand the implications of the defined process. Thus a practical use of this third approach requires an algorithm to derive the PPF starting from an arbitrarily defined $p(\mathbf{P})$. In this section we develop a numerical algorithm that allows to find $p_j(\cdot)$ for an arbitrary $p(\mathbf{P})$.

3.2 An Algorithm to Determine the PPF

Recall definition (4) for an SSM random probability measure. Assuming a proper SSM we have

$$G = \sum P_h \delta_{m_h}. \tag{13}$$

Let $\mathbf{P} = (P_h,\ h \in \mathbb{N})$ denote the sequence of weights. Recall the notation $\tilde{X}_j$ for the $j-$th unique value in the SSS $\{X_i,\ i = 1, \ldots, n\}$. The algorithm requires indicators that match the $\tilde{X}_j$ with the m_h, i.e., that match the clusters in the partition with the point masses of the SSM. Let $\pi_j = h$ if $\tilde{X}_j = m_h$, $j = 1, \ldots, k_n$. In the following discussion it is important that the latent indicators π_j are only introduced up to $j = k$. Conditional on m_h, $h \in \mathbb{N}$ and $\tilde{X}_j$, $j \in \mathbb{N}$ the indicators π_j are deterministic. After marginalizing w.r.t. the m_h or w.r.t. the $\tilde{X}_j$ the indicators become latent variables. Also, we use cluster membership indicators $s_i = j$ for $X_i = \tilde{X}_j$ to simplify notation. We use the convention of labeling clusters in the order of appearance, i.e., $s_1 = 1$ and $s_{i+1} \in \{1, \ldots, k_i, k_i + 1\}$.

In words the algorithm proceeds as follows. We write the desired PPF $p_j(\mathbf{n})$ as an expectation of the conditional probabilities $p(X_{n+1} = \tilde{X}_j \mid \mathbf{n}, \pi, \mathbf{P})$ w.r.t. $p(\mathbf{P}, \pi \mid \mathbf{n})$. Next we approximate the integral w.r.t. $p(\mathbf{P}, \pi \mid \mathbf{n})$ by a weighted Monte Carlo average over samples $(\mathbf{P}^{(\ell)}, \pi^{(\ell)}) \sim p(\mathbf{P}^{(\ell)})p(\pi^{(\ell)} \mid \mathbf{P}^{(\ell)})$ from the prior. Note that the properties of the random partition can be characterized by the distribution on $\mathbf{P}$ only. The point masses m_h are not required.

Using the cluster membership indicators s_i and the latent variables π_j to map clusters

with the point masses m_h of the SSM we write the desired PPF as

$$\begin{aligned} p_j(\mathbf{n}) &= p(s_{n+1} = j \mid \mathbf{s}) \\ &\propto \int p(s_{n+1} = j \mid \mathbf{s}, \mathbf{P}, \pi)\, p(\pi, \mathbf{P} \mid \mathbf{s})\, \mathrm{d}\pi \mathrm{d}\mathbf{P} \\ &\propto \int p(s_{n+1} = j \mid \mathbf{s}, \mathbf{P}, \pi)\, p(\mathbf{s} \mid \pi, \mathbf{P})\, p(\pi,\, \mathbf{P})\, \mathrm{d}\pi \mathrm{d}\mathbf{P} \\ &\approx \frac{1}{L} \sum p(s_{n+1} = j \mid \mathbf{s}, \mathbf{P}^{(\ell)}, \pi^{(\ell)})\, p(\mathbf{s} \mid \pi^{(\ell)}, \mathbf{P}^{(\ell)}). \end{aligned} \tag{14}$$

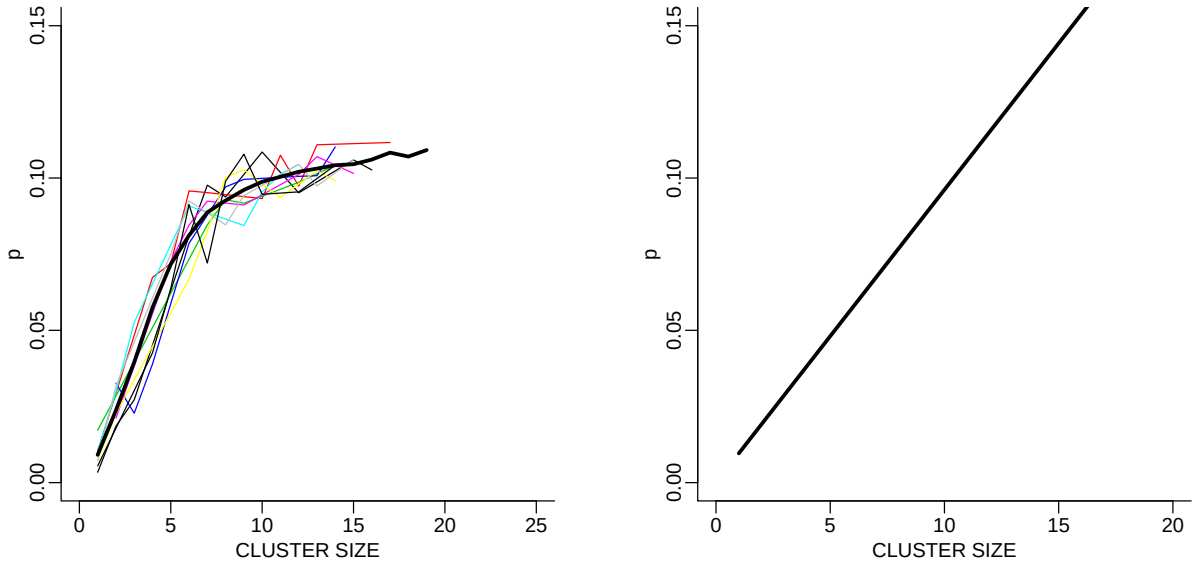
The Monte Carlo sample $(\mathbf{P}^{(\ell)}, \pi^{(\ell)})$ is generated by first generating $\mathbf{P}^{(\ell)} \sim p(\mathbf{P})$ and then $p(\pi_j^{(\ell)} = h \mid \mathbf{P}^{(\ell)}, \pi_1^{(\ell)}, \ldots, \pi_{j-1}^{(\ell)}) \propto P_h^{(\ell)}$, $h \notin \{\pi_1^{(\ell)}, \ldots, \pi_{j-1}^{(\ell)}\}$. In actual implementation the elements of $\mathbf{P}^{(\ell)}$ and $\pi^{(\ell)}$ are only generated as and when needed.

The terms in the last line of (14) are easily evaluated. Let $i_k = \min\{i : \ k_i = k_n\}$ denote the founding element of the last cluster. We use the predictive cluster membership probabilities

$$p(s_{i+1} = j \mid s_1, \ldots, s_i, \mathbf{P}, \pi) \propto \begin{cases} P_{\pi_j}, & j = 1, \ldots, k_i \\ P_{\pi_j}, & j = k_i + 1 \text{ and } i < i_k \\ (1 - \sum_{j=1}^{k_n} P_{\pi_j}), & j = k_n + 1 \text{ and } i \geq i_k. \end{cases} \tag{15}$$

The special case in the last line of (15) replaces P_{π_j} for $j = k_i + 1$ and $i \geq i_k$, i.e., for all i with $k_i = k_n$. The special case arises because π (in the conditioning set) only includes latent indicators π_j for $j = 1, \ldots, k_n$. The $(k_n + 1)$-st cluster can be mapped to any of the remaining probability masses. Note that $k_i = k_n$ for $i \geq i_k$. For the first factor in the last line of (14) we use (15) with $i = n$. The second factor is evaluated as $p(\mathbf{s} \mid \pi, \mathbf{P}) = \prod_{i=2}^{n} p(s_i \mid s_1, \ldots, s_{i-1}, \mathbf{P}, \pi)$.

Figure 2 shows an example. The figure plots $p(s_{i+1} = j \mid \mathbf{s})$ against cluster size n_j. In contrast, the DP Polya urn would imply a straight line. The plotted probabilities are averaged w.r.t. all other features of $\mathbf{s}$, in particular the multiplicity of cluster sizes etc. The figure also shows probabilities (15) for specific simulations.



(a) SSM$(p, \cdot)$ (b) DP$(M, \cdot)$

Figure 2: Panel (a) shows the PPF (15) for a random probability measure $G \sim \text{SSM}(p, G_0)$, with P_h as in (12). The thick line plots $p(s_{n+1} = j \mid \mathbf{s})$ against n_j, averaging over multiple simulations. In each simulation we used the same simulation truth to generate $\mathbf{s}$, and stop simulation at $n = 100$. The 10 thin lines show $p_j(\mathbf{n})$ for 10 simulations with different $\mathbf{n}$. In contrast, under the DP Polya urn the curve is a straight line, and there is no variation across simulations (panel b).

3.3 Example

Many data analysis applications of the DP prior are based on DP mixtures of normals as models for a random probability measure F. Applications include density estimation, random effects distributions, generalizations of a probit link etc. We consider a stylized example that is chosen to mimick typical features of such models.

In this section, we show posterior inference conditional on the data set $(y_1, y_2, \ldots, y_9) = (-4, -3, -2, \ldots, 4)$. We use this data because it highlights the differences in posterior inference between the SSM and DP priors. Assume $y_i \sim F$, *i.i.d.* with a semi-parametric mixture of normal prior on F,

$$y_i \overset{iid}{\sim} F, \text{ with } F(y_i) = \int N(y_i;\ \mu, \sigma^2)\, \mathrm{d}G(\mu, \sigma^2).$$

Here $N(x;\ m, s^2)$ denotes a normal distribution with moments (m, s^2) for the random variable x. We estimate F under two alternative priors,

$$G \sim \mathrm{SSM}(p, G_0) \text{ or } G \sim \mathrm{DP}(M, G_0).$$

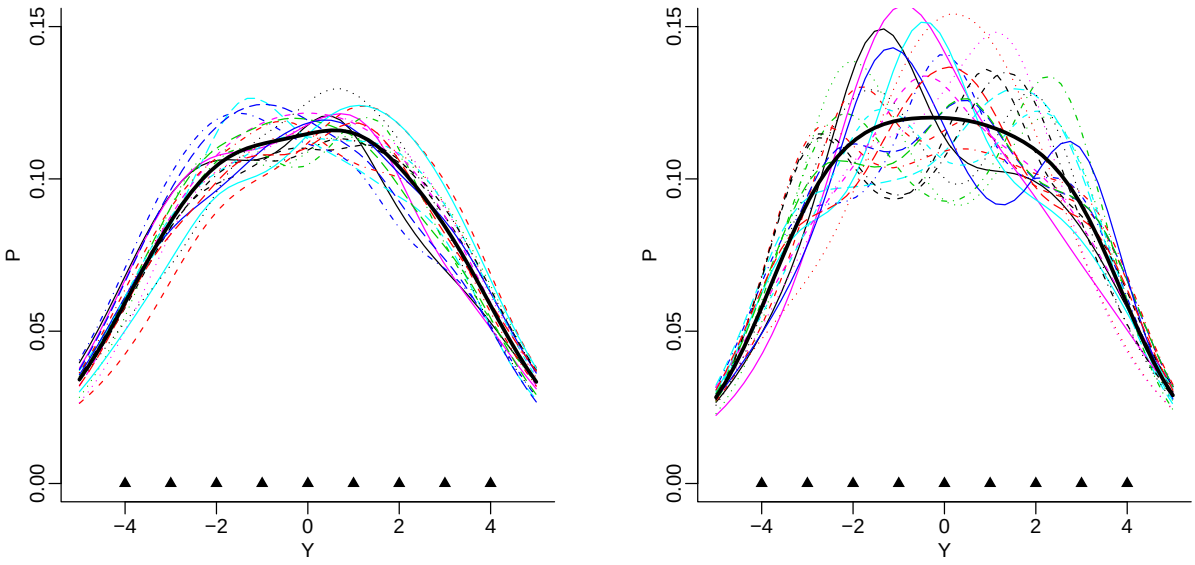
The distribution p of the weights for the $\mathrm{SSM}(p, \cdot)$ prior is defined as in (12). The total mass parameter M in the DP prior is fixed to match the prior mean number of clusters, $E(k_n)$, implied by (12). We find $M = 2.83$. Let $\mathrm{Ga}(x;\ a, b)$ indicate that the random variable x has a Gamma distribution with mean a/b. For both prior models we use

$$G_0(\mu, 1/\sigma^2) = N(x;\ \mu_0, c\,\sigma^2)\, \mathrm{Ga}(1/\sigma^2;\ a/2, b/2).$$

We fix $\mu_0 = 0$, $c = 10$ and $a = b = 4$. The model can alternatively be written as $y_i \sim N(\mu_i, \sigma_i^2)$ and $X_i = (\mu_i, 1/\sigma_i^2) \sim G$.

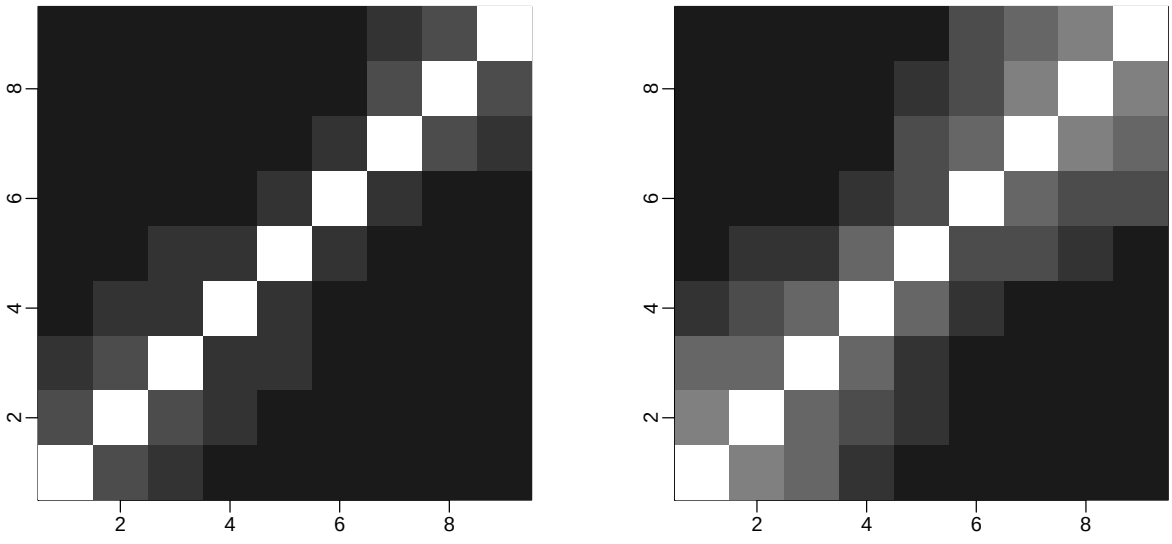
Figures 3 and 4 show some inference summaries. Inference is based on Markov chain Monte Carlo (MCMC) posterior simulation with 1000 iterations. Posterior simulation is for $(s_1, \ldots, s_n)$ only The cluster-specific parameters $(\tilde{\mu}_j, \tilde{\sigma}_j^2)$, $j = 1, \ldots, k_n$ are analytically marginalized. One of the transition probabilities (Gibbs sampler) in the MCMC requires the PPF under $SSM(p, G_0)$. It is evaluated using (14).

Figure 3 shows the posterior estimated sampling distributions F. The figure highlights a limitation of the DP prior. The single total mass parameter M controls both, the number of clusters and the prior precision. A small value for M favors a small number of clusters and implies low prior uncertainty. Large M implies the opposite. Also, we already illustrated in Figure 1 that the DP prior implies stochastically ordered cluster sizes, whereas the chosen SSM prior allows for many approximately equal size clusters. The equally spaced grid data $(y_1, \ldots, y_n)$ implies a likelihood that favors a moderate number of approximately equal size clusters. The posterior distribution on the random partition is shown in Figure 4. Under the SSM prior the posterior supports a moderate number of similar size clusters. In contrast, the DP prior shrinks the posterior towards a few dominant clusters. Let $n_{(1)} \equiv \max_{j=1,\ldots,k_n} n_j$ denote the leading cluster size. Related evidence can be seen in the marginal posterior distribution (not shown) of k_n and $n_{(1)}$. We find $E(k_n \mid data) = 6.4$ under the SSM model versus $E(k_n \mid data) = 5.1$ under the DP prior. The marginal posterior modes are $k_n = 6$ under the SSM prior and $k_n = 5$ under the DP prior. The marginal posterior modes for $n_{(1)}$ is $n_{(1)} = 2$ under the SSM prior and $n_{(1)} = 3$. under the DP prior.



(a) $G \sim \text{SSM}(p, G_0)$ (b) $G \sim \text{DP}(M, G_0)$

Figure 3: Posterior estimated sampling model $\overline{F} = E(F \mid data) = p(y_{n+1} \mid data)$ under the $\text{SSM}(p, G_0)$ prior and a comparable DP prior. The triangles along the x-axis show the data.



(a) $G \sim \text{SSM}(p, G_0)$ (b) $G \sim \text{DP}(M, G_0)$

Figure 4: Co-clustering probabilities $p(s_i = s_j \mid data)$ under the two prior models.

4 Discussion

We have reviewed alternative definitions of SSMs. We have shown that all SSMs with a PPF of the form $p_j(\mathbf{n}) = f(n_j)$ needs to necessarily be a linear function of n_j. In other words, the PPF depends on the current data only through the cluster sizes. The number of clusters and the multiplicities of cluster sizes do not change the prediction. This is an excessively simplifying assumption for most data analysis problems.

We provide an alternative class of models that allow for more general PPF. One of the important implications is the implied distribution of probability weights. The DP prior favors a priori a partition with stochastically ordered cluster sizes. The proposed new class allows any desired distribution of cluster sizes.

R code for an implementation of posterior inference under the proposed new model is available at `http://odin.mdacc.tmc.edu/~pm`.

References

[1] Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.*, 90(430):577–588, 1995.

[2] Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

[3] Hemant Ishwaran and Lancelot F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica*, 13(4):1211–1235, 2003.

[4] Lancelot James. Large sample asymptotics for the two parameter poisson dirichlet process. *Unpublished manuscript*, 2006.

[5] Gunho Jang, Jaeyong Lee, and Sangyeol Lee. Posterior consistency of species sampling priors. *Unpublished working paper*, 2007.

[6] Antonio Lijoi, Ramsés H. Mena, and Igor Prünster. Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc.*, 100(472):1278–1291, 2005.

[7] Antonio Lijoi, Igor Prünster, and Stephen G. Walker. On consistency of nonparametric normal mixtures for Bayesian density estimation. *J. Amer. Statist. Assoc.*, 100(472):1292–1296, 2005.

[8] Steven N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.*, 23(3):727–741, 1994.

[9] Steven N. MacEachern and Peter Müller. Estimating mixtures of dirichlet process models. *J. Comput. Graph. Statist.*, 7(1):223–239, 1998.

[10] Carlos Navarrete, A. Fernando Quintana, and Peter Müller. Some issues on nonparametric bayesian modeling using species sampling models. *Statistical Modelling. An International Journal*, 2008. To appear.

[11] Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, 102(2):145–158, 1995.

[12] Jim Pitman. Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, probability and game theory*, volume 30 of *IMS Lecture Notes Monogr. Ser.*, pages 245–267. Inst. Math. Statist., Hayward, CA, 1996.

[13] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.