# 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

Department of Biostatistics, M. D. Anderson Cancer Center

pmueller@mdanderson.org

Lorenzo Trippa

Department of Biostatistics

Havard University

ltrippa@jimmy.harvard.edu

August 16, 2012

**Abstract**

We review 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). It is straightforward to define a PPF from an EPPF, but the converse is not necessarily true. In this paper, we introduce the notion of putative PPFs and show novel conditions for a putative PPF to define an EPPF. We show that all possible PPFs 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 putative PPFs to define an EPPF. Finally we show posterior inference for a large class of SSMs with a PPF that is not linear in cluster size and discuss a numerical method to derive its PPF.

*Key words and phrases:* Species sampling Prior, Exchangeable partition probability functions, Prediction probability functions. [1]

# 1 Introduction

The status of the Dirichlet process (Ferguson, 1973) (DP) among nonparametric priors is comparable to 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 (Blackwell and McQueen, 1973). Markov chain Monte Carlo 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 model (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

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

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 (2008), Navarrete, Quintana and Müller (2008), James, Lijoi and Prünster (2009) and Jang, Lee and Lee (2010).

Suppose $(X_1, X_2, \ldots)$ is a sequence of random variables. In a traditional application the sequence arises as a random sample from a large population of units, and $X_i$ records the species of the $i$-th individual in the sample. This explains the name SSM. Let $\tilde{X}_j$ be the $j$th distinct species to appear. Let $n_{jn}$ be the number of times the $j$th species $\tilde{X}_j$ appears in $(X_1, \ldots, X_n)$, $j = 1, 2, \ldots$, and

$$\mathbf{n}_n = (n_{jn},\ j = 1, \ldots, k_n),$$

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 $\{i \leq n :\ X_i = \tilde{X}_j\}$ 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 partition probability function.

**PPF:** Let $\nu$ be a diffuse (or nonatomic) probability measure on a complete separable metric space $\mathcal{X}$ equipped with Borel $\sigma$-field. 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 $\delta_x$ is the degenerate probability measure at $x$. Examples of SSS include the Pólya urn sequence $(X_1, X_2, \ldots)$ whose distribution is the same as the marginal distribution of independent observations from a Dirichlet random distribution $F$, i.e. $X_1, X_2, \ldots | F \overset{iid}{\sim} F$ with $F \sim DP(\alpha\nu)$, where $\alpha > 0$. The conditional distribution of the Pólya urn sequence

is

$$X_{n+1} \mid X_1, \ldots, X_n \sim \sum_{j=1}^{k_n} \frac{n_j}{n+\alpha} \delta_{\tilde{X}_j} + \frac{\alpha}{n+\alpha} \nu.$$

This marginalization property has been a central tool for posterior simulation in DP mixture models, which benefit from the fact that one can integrate out $F$ using the marginalization property. The posterior distribution becomes then free of the infinite dimensional object $F$. Thus, Markov chain Monte Carlo algorithms for DP mixtures do not pose bigger difficulties than the usual parametric Bayesian models (MacEachern 1994, MacEachern and Müller 1998). Similarly, alternative discrete random distributions have been considered in the literature and proved computationally attractive due to analogous marginalization properties, see for example Lijoi et al. (2005) and Lijoi et al. (2007) .

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

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

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). Note 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 $\alpha$. The use of $p_j$ in (1) implies

$$\begin{aligned} p_j(\mathbf{n}) &= \mathbb{P}(X_{n+1} = \tilde{X}_j \mid X_1, \ldots, X_n), \quad j = 1, \ldots, k_n, \\ p_{k_n+1}(\mathbf{n}) &= \mathbb{P}(X_{n+1} \notin \{\tilde{X}_1, \ldots, \tilde{X}_{k_n}\} \mid X_1, \ldots, X_n). \end{aligned}$$

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

An important point in the above definition is that a sequence $X_i$ can be a SSS only if it is exchangeable.

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

$$G = \sum_{h=1}^{\infty} P_h \delta_{m_h} + R\nu, \tag{3}$$

for some sequence of positive random variables $(P_h)$ and $R$ such that $1 - R = \sum_{h=1}^{\infty} P_h \leq 1$ with probability 1, $(m_h)$ is a sequence of independent variables with distribution $\nu$, 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 (3) the *SSM* of the SSS $(X_i)$.

**EPPF:** A third alternative definition of a SSS and corresponding SSM is in terms of the implied probability model on a sequence of random partitions.

Suppose a SSS $(X_1, X_2, \ldots)$ is given. Since the de Finetti measure (3) is partly discrete, there are ties among $X_i$'s. The ties among $(X_1, X_2, \ldots X_n)$ for a given $n$ induce an equivalence relation in the set $[n] = \{1, 2, \ldots, n\}$, i.e., $i \sim j$ if and only if $X_i = X_j$. This equivalence relation on $[n]$, in turn, induces the partition $\Pi_n$ of $[n]$. Due to the exchangeability of $(X_1, X_2, \ldots)$, it can be easily seen that the random partition $\Pi_n$ is an exchangeable random partition on $[n]$, i.e., for any partition $\{A_1, A_2, \ldots, A_k\}$ of $[n]$, the probability $P(\Pi_n = \{A_1, A_2, \ldots, A_k\})$ is invariant under any permutation on $[n]$ and can be expressed as a function of $\mathbf{n} = (n_1, n_2, \ldots, n_k)$, where $n_i$ is the cardinality of $A_i$ for $i = 1, 2, \ldots, k$. Extending the above argument to the entire SSS, we can get an exchangeable random partition on the natural numbers $\mathbb{N}$ from the SSS. Kingman (1978, 1982) showed a remarkable result, called Kingman's representation theorem, that in fact every exchangeable random partition can be obtained by a SSS.

For any partition $\{A_1, A_2, \ldots, A_k\}$ of $[n]$, we can represent $P(\Pi_n = \{A_1, A_2, \ldots, A_k\}) =$

$p(\mathbf{n})$ for a symmetric function $p : \mathbb{N}^* \to [0,1]$ satisfying

$$\begin{aligned} p(1) &= 1, \\ p(\mathbf{n}) &= \sum_{j=1}^{k(\mathbf{n})+1} p(\mathbf{n}^{j+}), \quad \text{for all } \mathbf{n} \in \mathbb{N}^*, \end{aligned} \tag{4}$$

where $\mathbf{n}^{j+}$ is the same as $\mathbf{n}$ except that the $j$th element is increased by 1. This function is called an exchangeable partition probability function (EPPF) and characterizes the distribution of an exchangeable random partition on $\mathbb{N}$.

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

$$p_j(\mathbf{n}) = \frac{p(\mathbf{n}^{j+})}{p(\mathbf{n})}, \quad j = 1, 2, \ldots, k+1. \tag{5}$$

The converse is not true. Not every putative $p_j(\mathbf{n})$ defines an EPPF and thus a SSM and a SSS. For example, it is easy to show that $p_j(\mathbf{n}) \propto n_j^2 + 1$, $j = 1, \ldots, k(\mathbf{n})$ does not. In Bayesian 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 crucial for posterior computation, applied Bayesians tend to focus on it to specify 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 (5) 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 some conditions to address this gap. We consider methods to define PPFs in two different directions. First we give an easily verifiable necessary condition for a putative PPF to arise from an EPPF (Lemma 1) and a necessary and sufficient condition for a putative PPF to arise from an EPPF. A consequence of this result is an elementary proof of the characterization of all possible PPFs with form $p_j(\mathbf{n}) \propto f(n_j)$. This result has been proved earlier by Gnedin and Pitman

(2006). Although the result in section 2 gives necessary and sufficient conditions for a putative PPF to be a PPF, the characterization is not constructive. It does not give any guidance in how to create a new PPF for a specific application. In section 3, we propose an alternative approach to define a SSM based on directly defining a joint probability model for the $P_h$ in (3). We develop a numerical algorithm to derive the corresponding PPF. This facilitates the use of such models for nonparametric Bayesian data analysis. This approach can naturally create PPFs with very different features than the well known PPF under the DP.

The literature reports some PPFs with closed form analytic expressions other than the PPF under the DP prior. There are a few directions which have been explored for constructing extensions of the DP prior and deriving PPFs. The normalization of complete random measures (CRM) has been proposed in Kingman (1975). A CRM such as the generalized gamma process (Brix, 1999), after normalization, defines a discrete random distribution, and, under mild assumptions, a SSM. Developments and theoretical results on this approach have been discussed in a series of papers; see for example Perman et al. (1992), Pitman (2003), and Regazzini et al. (2003). Normalized CRM models have been also studied and applied in Lijoi et al. (2005), Nieto et al. (2004), and more recently in James et al. (2009). A second related line of research considered the so called Gibbs models. In these models the analytic expressions of the PPFs share similarities with the DP model. An important example is the Pitman-Yor process. Contributions include Gnedin and Pitman (2006), Lijoi, Mena, and Prünster (2007), Lijoi, Prünster, and Walker (2008ab), and Gnedin et al. (2010). Lijoi and Prünster (2010) provide a recent overview on major results from the literature on normalized CRM and Gibbs-type partitions.

## 2 When Does a PPF Imply an EPPF?

Suppose we are given a putative PPF $(p_j)$. Using equation (5), one can attempt to define a function $p : \mathbb{N}^* \to [0,1]$ inductively by the following mapping:

$$\begin{aligned} p(1) &= 1 \\ p(\mathbf{n}^{j+}) &= p_j(\mathbf{n})p(\mathbf{n}), \quad \text{for all } \mathbf{n} \in \mathbb{N} \text{ and } j = 1, 2, \ldots, k(\mathbf{n}) + 1. \end{aligned} \tag{6}$$

In general, equation (6) does not lead to a unique definition of $p(\mathbf{n})$ 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 $\{\{1,3\},\{2\}\}$ and $\{\{1,2\},\{3\}\}$, respectively. If $p_2(1)p_1(1,1) \neq p_1(1)p_2(2)$, equation (6) does not define a function $p : \mathbb{N}^* \to [0,1]$. The following lemma shows a condition for a PPF for which equation (6) 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 $[n]$ with clusters indexed in the order of appearance. For $1 \leq m \leq n$, let $\Pi_m$ be the restriction of $\Pi$ on $[m]$. Let $\mathbf{n}(\Pi) = (n_1, \ldots, n_k)$ where $n_i$ is the cardinality of $A_i$, let $\Pi(i)$ be the class index of element $i$ in partition $\Pi$ and $\Pi([n]) = (\Pi(1), \ldots, \Pi(n))$.

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

$$p_i(\mathbf{n})p_j(\mathbf{n}^{i+}) = p_j(\mathbf{n})p_i(\mathbf{n}^{j+}), \quad \text{for all } \mathbf{n} \in \mathbb{N}^*, i, j = 1, 2, \ldots, k(\mathbf{n}) + 1, \tag{7}$$

then $p$ defined by (6) is a function from $\mathbb{N}^*$ to $[0,1]$, i.e., $p$ in (6) is uniquely defined.

**Proof.** Let $\mathbf{n} = (n_1, \ldots, n_k)$ with $\sum_{i=1}^k n_i = n$ and $\Pi$ and $\Omega$ 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 (7).

The opposite is easy to show. Assume $p_j$ defines a unique $p(\mathbf{n})$. Consider (7) 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 exchangeability, i.e., invariance of $p$ with respect to permutations of the group indices $j = 1, \ldots, k(\mathbf{n})$. When the function $p$, recursively defined by expression (6) satisfies the balance imposed by equation (7) it is called *partially exchangeable probability function* (Pitman, 1995, 2006) and the resulting random partition of $\mathbb{N}$ is termed partially exchangeable. In Pitman (1995), it is proved that a $p : \mathbb{N}^* \to [0, 1]$ is a *partially exchangeable probability function* if and only if it exists a sequence of non negative random variables $P_i$, $i = 1, \ldots$, with $\sum_i P_i \leq 1$ such that

$$p(n_1, \ldots, n_k) = E\left[\prod_{i=1}^{k} P_i^{n_i - 1} \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} P_i\right)\right], \tag{8}$$

where the expectation is with respect to the distribution of the sequence $(P_i)$. We refer to Pitman (1995) for an extensive study of partially exchangeable random partitions.

It is easily checked whether or not a given PPF satisfies the condition of Lemma 1. Corollary 1 describes all possible PPFs that have the probability of cluster memberships depend on a function of the cluster size only. This result is part of a theorem in Gnedin and Pitman (2006), but we give here a more straightforward proof.

**Corollary 1.** Suppose a putative PPF $(p_j)$ satisfies (7) and

$$p_j(n_1, \ldots, n_k) \propto \begin{cases} f(n_j), & j = 1, \ldots, k \\ \theta, & j = k + 1, \end{cases} \tag{9}$$

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 (7) 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 $k \in \mathbb{N}$

$$f(m) = am + b, \tag{10}$$

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 (10) implies $b = 0$. Thus, we have $f(k) = ak$ for some $a > 0$. ■

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. This is the EPPF for a DP with total mass $\theta/a$. Thus, Corollary 1 implies that the EPPF under the DP is the only EPPF that satisfies (9). The Corollary shows that it is not an entirely trivial matter to come up with a putative PPF that leads to

a valid EPPF. A version of Corollary 1 is also well known as Johnson's Sufficientness postulate (Good, 1965). See also the discussion in Zabell (1982).

We now give a necessary and sufficient condition for the function $p$ defined by (5) to be an EPPF, without any constraint on the form of $p_j$ (as were present in the earlier results). Suppose $\sigma$ 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, $\sigma$ 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 (7) as well as the following condition: for all $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$, and permutations $\sigma$ on $[k]$ and $i = 1, \ldots, k$,

$$p_i(n_1, \ldots, n_k) = p_{\sigma^{-1}(i)}(n_{\sigma(1)}, n_{\sigma(2)}, \ldots, n_{\sigma(k)}). \tag{11}$$

Then, $p$ defined by (6) is an EPPF. The condition is also necessary; if $p$ is an EPPF then (11) holds.

**Proof.** Fix $\mathbf{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$ and a permutation on $[k]$, $\sigma$ . We wish to show that for the function $p$ defined by (6)

$$p(n_1, \ldots, n_k) = p(n_{\sigma(1)}, n_{\sigma(2)}, \ldots, n_{\sigma(k)}). \tag{12}$$

Let $\Pi$ 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 $\Omega$ 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 (11). This completes the proof of the sufficient direction.

Finally, we show that every EPPF $p$ satisfies (6) and (11). By Lemma 1 every EPPF satisfies (6). Condition (12) is true by the definition of an EPPF, which includes the condition of symmetry in its arguments. And (12) implies (11). ■

Fortini et al. (2000) prove results related to Theorem 1. They provide sufficient conditions for a system of predictive distributions $p(X_n \mid X_1, \ldots, X_{n-1})$, $n = 1, \ldots$, of a sequence of random variables $(X_i)$, that imply exchangeability. The relation between these conditions and Theorem 1 becomes apparent by constructing a sequence $(X_i)$ that induces a $p$-distributed random partition of $\mathbb{N}$. Here, it is implicitly assumed the mapping of $(X_i)$ to the only partition such that $i, j \in \mathbb{N}$ belong to the same subset if and only if $X_i = X_j$.

A second more general example, which extends the predictive structure considered in Corollary 1, includes the so called Gibbs random partitions. Within this class of models

$$p(n_1, n_2, \ldots, n_k) = V_{n,k} \prod_{i=1}^{k} W_{n_i}, \tag{13}$$

where $(V_{n,k})$ and $(W_{n_i})$ are sequences of positive real numbers. In this case the predictive probability of a novel species is a function of the sample size $n$ and of the number of observed species $k$. See Lijoi et al. (2007) for related distributional results on Gibbs type

models. Gnedin and Pitman (2006) obtained sufficient conditions for the sequences $(V_{n,k})$ and $(W_{n_i})$ which imply that $p$ is an EPPF.

# 3 SSM's Beyond the DP

## 3.1 The $SSM(p, \nu)$

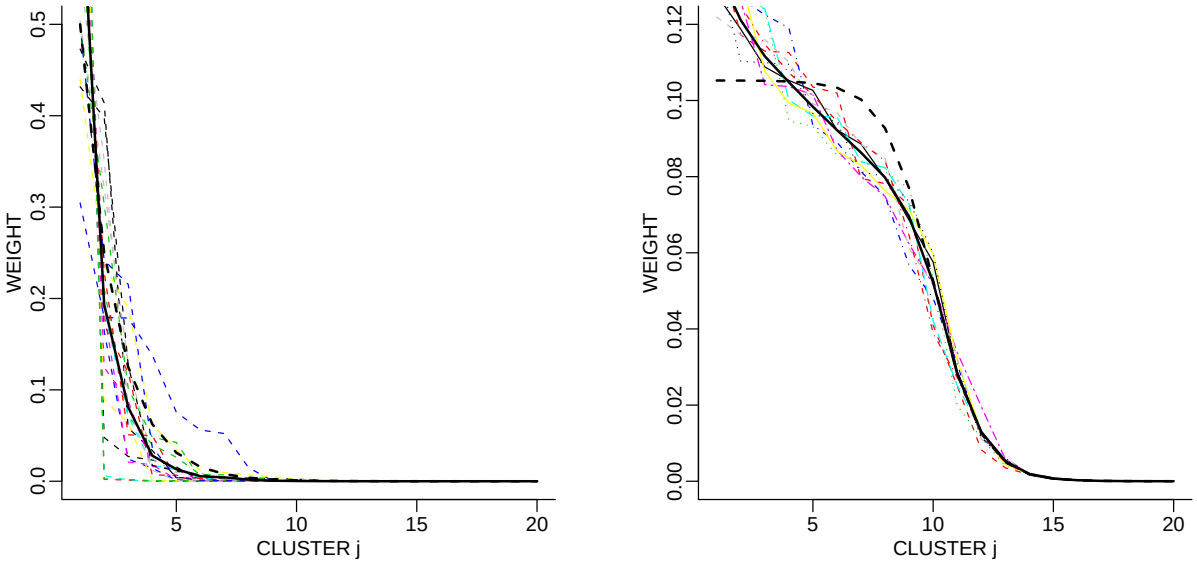
We know that an SSM with a non-linear PPF, i.e., $p_j$ different from the PPF of a DP, can not be described as a function $p_j \propto f(n_j)$ of $n_j$ only. It must be a more complicated function $f(\mathbf{n})$. Alternatively one could try to define an EPPF, and deduce the implied PPF. But directly specifying a symmetric function $p(\mathbf{n})$ such that it complies with (4) is difficult. As a third alternative we propose to consider the weights $\mathbf{P} = \{P_h,\ h = 1, 2, \ldots\}$ in (3). 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. The stick breaking construction for a DP prior with total mass $\theta$ implies $E(P_h) = \theta^{h-1}(1+\theta)^{-h}$. Such geometrically decreasing mean weights are inappropriate to describe prior information in many applications. The weights can be interpreted as asymptotic relative cluster sizes. A typical application of the DP prior is, for example, a partition of patients in a clinical study into clusters. However, if clusters correspond to disease subtypes defined by variations of some biological process, then one would rather expect a number of clusters with a priori comparable size. Many small clusters with very few patients are implausible, and would also be of little clinical use. This leads us to propose the use of alternative SSM's.

Figure 1b shows an alternative probability model $p(\mathbf{P})$. There are many ways to define $p(\mathbf{P})$; we consider, for $h = 1, 2, \ldots$,

$$P_h \propto u_h \qquad \text{or} \qquad P_h = \frac{u_h}{\sum_{i=1}^{\infty} u_i},$$

where $u_h$ are independent and nonnegative random variables with

$$\sum_{i=1}^{\infty} u_i < \infty, \qquad \text{a.s.} \tag{14}$$




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

Figure 1: The lines in each panel show 10 draws $\mathbf{P} \sim p(\mathbf{P})$ for the DP (left) and for the SSM defined in (16) below (right). 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.

A sufficient condition for (14) is

$$\sum_{i=1}^{\infty} E(u_i) < \infty, \tag{15}$$

by the monotone convergence theorem. Note that when the unnormalized random variables $u_h$ are defined as the sorted atoms of a non-homogeneous Poisson process on the positive real line, under mild assumptions, the above $(P_h)$ construction coincides with the Poisson-Kingman models. Ferguson and Klass (1972) provide a detailed discussion on the outlined mapping of a Poisson process into a sequence of unnormalized positive weights. In this particular case the mean of the Poisson process has to satisfy minimal requirements (see, for example, Pitman, 2003) to ensure that the sequence $(P_i)$ is well defined.

As an illustrative example in the following discussion, we define for, $h = 1, 2, \ldots,$

$$P_h \propto e^{X_h} \qquad \text{with} \qquad X_h \sim N\left(\log(1 - \{1 + e^{b-ah}\}^{-1}), \sigma^2\right), \tag{16}$$

where $a, b, \sigma^2$ are positive constants. The existence of such random probabilities is guaranteed by (15), which is easy to check.

The S-shaped nature of the random distribution (16), when 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) of Figure 1 the prior mean of the sorted and unsorted weights is almost indistinguishable, because the prior already implies strong stochastic ordering of the weights.

The prior in Figure 1b reflects prior information of an investigator who believes that there should be around 5 to 10 clusters of comparable size in the population. This is in sharp contrast to the (often implausible) assumption of one large dominant cluster and geometrically smaller clusters that is reflected in panel (a). Prior elicitation can exploit such readily interpretable implications of the prior choice to propose models like (16).

We use SSM$(p, \nu)$ to denote a SSM defined by $p(\mathbf{P})$ for the weights $P_h$ and $m_h \overset{iid}{\sim} \nu$. The attraction of defining the SSM through $\mathbf{P}$ is that by (3) any joint probability model $p(\mathbf{P})$ such that $P(\sum_h P_h = 1)$ defines a proper SSM. There are no additional constraints

as for the PPF $p_j(\mathbf{n})$ or the EPPF $p(\mathbf{n})$. 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 second approach requires an algorithm to derive the PPF starting from an arbitrarily defined $p(\mathbf{P})$.

## 3.2 An Algorithm to Determine the PPF

Recall definition (3) for an SSM random probability measure. Assuming a proper SSM we have

$$G = \sum_{h=1}^{\infty} P_h \delta_{m_h}. \tag{17}$$

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 $\pi_j$ 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 $\pi_j$ are deterministic. After marginalizing with respect to the $m_h$ or with respect to 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 $p(X_{n+1} = \tilde{X}_j \mid \mathbf{n}, \pi, \mathbf{P})$. The expectation is with respect to $p(\mathbf{P}, \pi \mid \mathbf{n})$. Next we approximate the integral with respect to $p(\mathbf{P}, \pi \mid \mathbf{n})$ 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 $\pi$ and $\mathbf{P}$ together define the size-biased permutation of $(P_j)$,

$$\tilde{P}_j = P_{\pi_j},\ j = 1, 2, \ldots$$

The size-biased permutation $(\tilde{P}_j)$ of $(P_j)$ is a resampled version of $(P_j)$ where sampling is done with probability proportional to $P_j$ and without replacement. Once the sequence $(P_j)$ is simulated, it is computationally straightforward to get $(\tilde{P}_j)$. Note also 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 size-biased probabilities $\tilde{P}_j$ we write the desired PPF as

$$\begin{aligned} p_j(\mathbf{n}) &= p(s_{n+1} = j \mid \mathbf{n}) \\ &= \int p(s_{n+1} = j \mid \mathbf{n}, \tilde{\mathbf{P}})\, p(\tilde{\mathbf{P}} \mid \mathbf{n})\, \mathrm{d}\tilde{\mathbf{P}} \\ &\propto \int p(s_{n+1} = j \mid \mathbf{n}, \tilde{\mathbf{P}})\, p(\mathbf{n} \mid \tilde{\mathbf{P}})\, p(\tilde{\mathbf{P}})\, \mathrm{d}\tilde{\mathbf{P}} \\ &\approx \frac{1}{L} \sum p(s_{n+1} = j \mid \mathbf{n}, \tilde{\mathbf{P}}^{(\ell)})\, p(\mathbf{n} \mid \tilde{\mathbf{P}}^{(\ell)}). \end{aligned} \tag{18}$$

The Monte Carlo sample $\tilde{\mathbf{P}}^{(\ell)}$, or equivalently $(\mathbf{P}^{(\ell)}, \pi^{(\ell)})$, is obtained 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 (18) are easily evaluated. The first factor is given as predictive cluster membership probabilities

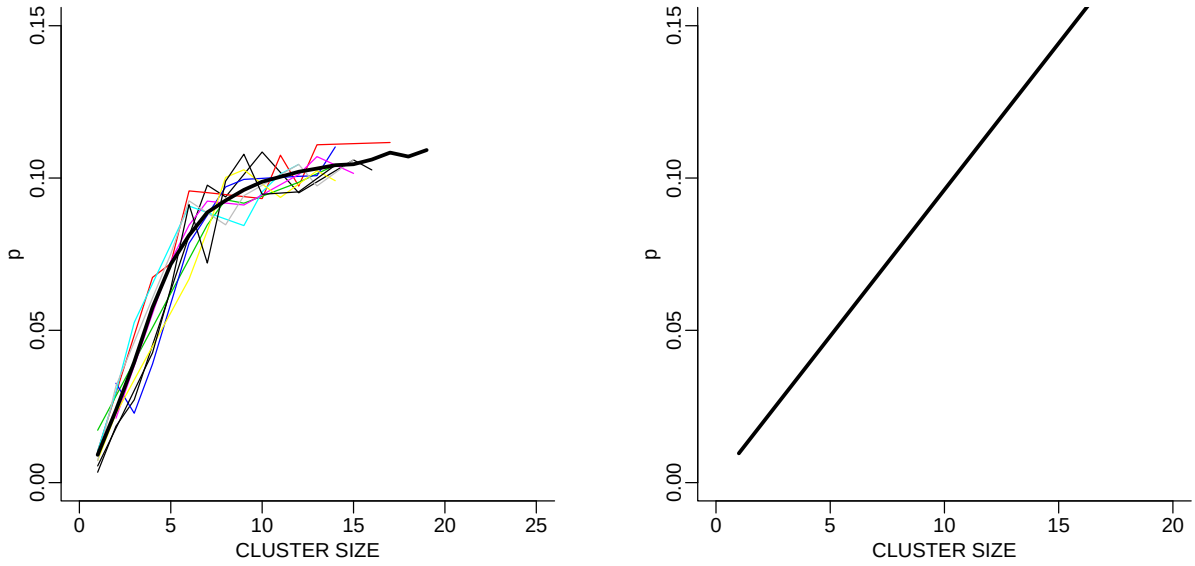
$$p(s_{n+1} = j \mid \mathbf{n}, \tilde{\mathbf{P}}) = \begin{cases} \tilde{P}_j, & j = 1, \ldots, k_n \\ (1 - \sum_{j=1}^{k_n} \tilde{P}_j), & j = k_n + 1. \end{cases} \tag{19}$$

The second factor is evaluated as

$$p(\mathbf{n}|\tilde{\mathbf{P}}) = \prod_{j=1}^{k} \tilde{P}_j^{n_j - 1} \prod_{j=1}^{k-1} (1 - \sum_{i=1}^{j} \tilde{P}_i).$$

Note that the second factor coincides with the previously mentioned (cf. expression 8 ) Pitman's representation result for partially exchangeable partitions.

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 with respect to all other features of $\mathbf{s}$, in particular the multiplicity of cluster sizes etc. The figure also shows probabilities (19) for specific simulations.




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

Figure 2: Panel (a) shows the PPF (19) for a random probability measure $G \sim \text{SSM}(p, \nu)$, with $P_h$ as in (16). 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 A Simulation 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 mimic typical features of such models.

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

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

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,

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

The distribution $p$ of the weights for the $\mathrm{SSM}(p, \cdot)$ prior is defined as in (16). The total mass parameter $M$ in the DP prior is fixed to match the prior mean number of clusters, $E(k_n)$, implied by (16). We find $M = 2.83$. Let $\mathrm{Ga}(x;\, a, b)$ indicate that the random variable $x$ has a Gamma distribution with shape parameter $a$ and inverse scale parameter $b$. For both prior models we use

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

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 $SSM(p, \nu)$. It is evaluated using (18).

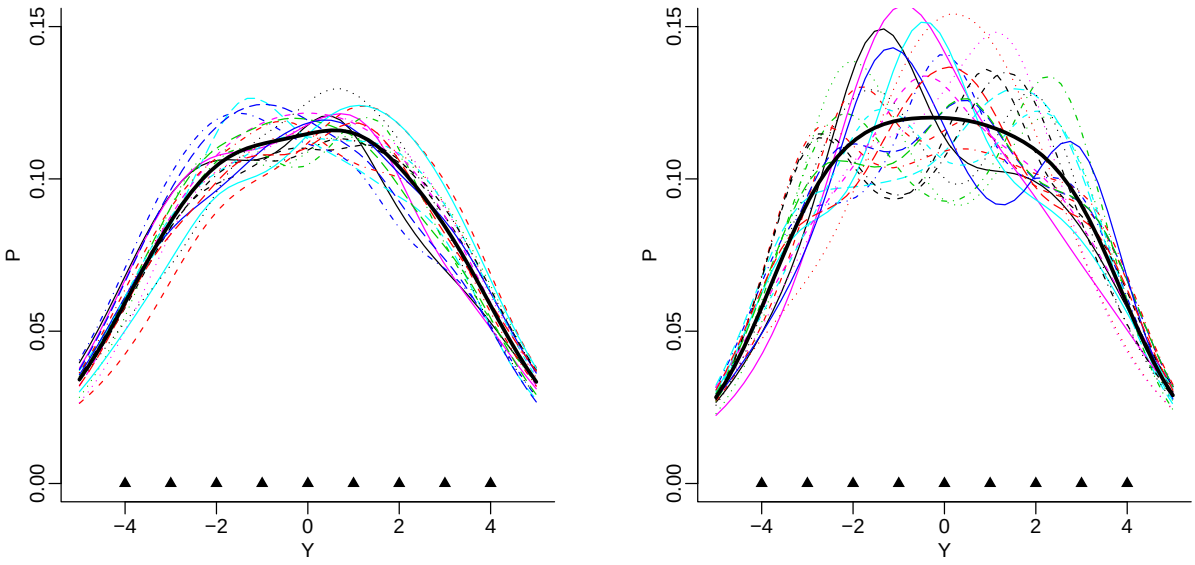
| Sarcoma | LEI | LIP | MFH | OST | Syn | Ang | MPNST | Fib |
|---|---|---|---|---|---|---|---|---|
| | 6/28 | 7/29 | 3/29 | 5/26 | 3/20 | 2/15 | 1/5 | 1/12 |

Table 1: Sarcoma data. For each disease subtype (top row) we report the total number of patients and the number of treatment successes. See León-Novelo at al. (2012) for a discussion of disease subtypes.

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 $E(k_n \mid \text{data}) = 6.4$ under the SSM model versus $E(k_n \mid \text{data}) = 5.1$ 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.

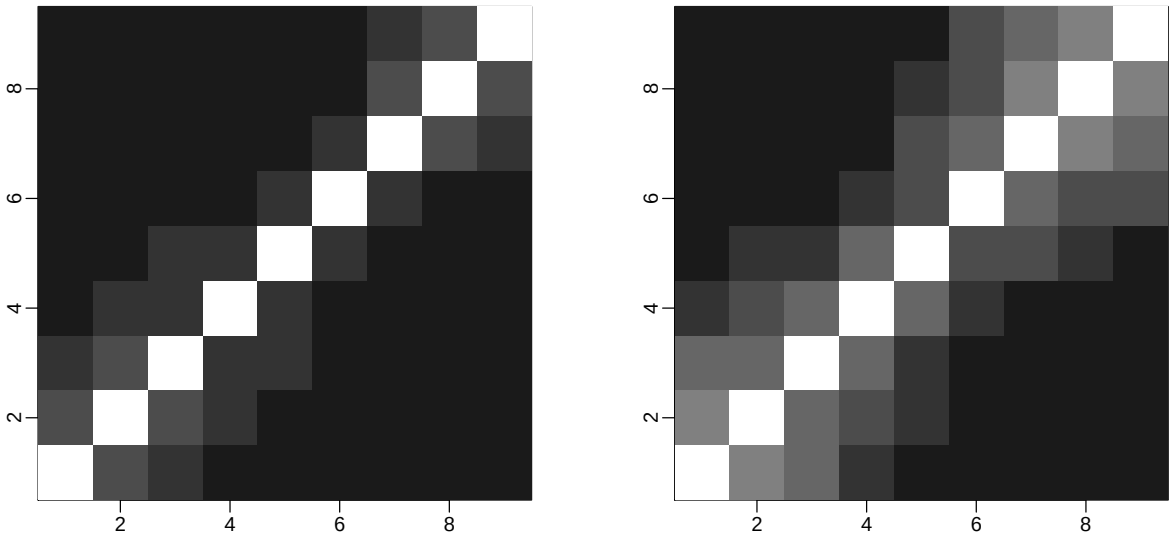
## 3.4 Analysis of Sarcoma Data

We analyze data from of a small phase II clinical trial for sarcoma patients that was carried out in M. D. Anderson Cancer Center. The study was designed to assess efficacy of a treatment for sarcoma patients across different subtypes. We consider the data accrued for 8 disease subtypes that were classified as having overall intermediate prognosis, as




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

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




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

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

presented in Table 1. Each table entry indicates the total number of patients for each sarcoma subtype, and the number of patients who reported a treatment success. See further discussion in León-Novelo et al. (2012).

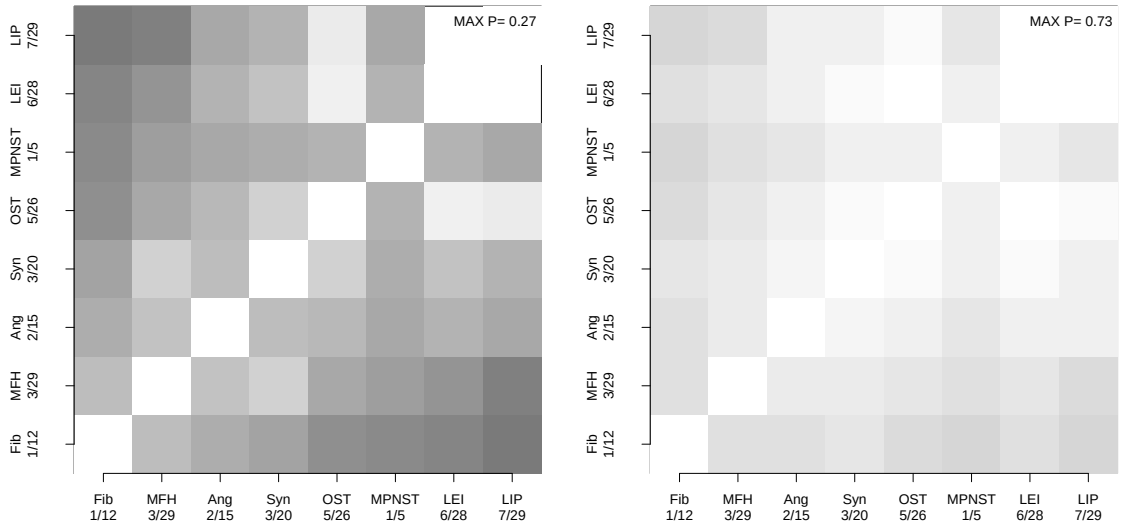
One limitation of these data is the small sample size, which prevents separate analysis for each disease subtype. On the other hand, it is not clear that we should simply treat the subtypes as exchangeable. We deal with these issues by modeling each table entry as a binomial response and adopt a hierarchical framework for the success probabilities. The hierarchical model includes a random partition of the subtypes. Conditional on a given partition, data across all subtypes in the same cluster are pooled, thus allowing more precise inference on the common success probabilities for all subtypes in this cluster. We consider two alternative models for the random partition, based on a DP$(M, \nu)$ prior versus a SSM$(p, \nu)$ prior. Specifically, we consider the following models:

$$\begin{aligned} y_i|\pi_i &\sim \text{Bin}(n_i, \pi_i) \\ \pi_i|G &\sim G \\ G &\sim DP(M,\nu) \quad \text{or} \quad SSM(p,\nu), \end{aligned}$$

where $\nu$ is a diffuse probability measure on $[0, 1]$, and $p$ is again defined as in (16).

The hierarchical structure of the data and the aim of clustering subpopulations in order to achieve borrowing of strength is in continuity with a number of applied contributions. Several of these, for instance, are meta analyses of medical studies (Berry and Christensen, 1979), with subpopulations defined by medical institutions or by clinical trials. In most cases the application of the DP is chosen for computational advantages and (in some cases) due to the easy implementation of strategies for prior specification (Liu, 1996). With a small number of studies, as in our example, ad hoc construction of alternative SSM combines hierarchical modeling with advantageous posterior clustering. The main advantage is the possibility of avoiding the exponential decrease typical of the ordered DP atoms.

In this particular analysis, we used $M = 2.83$, and chose $\nu$ to be the $Beta(0.15, 0.85)$ distribution, which was designed to match the prior mean of the observed data, and has



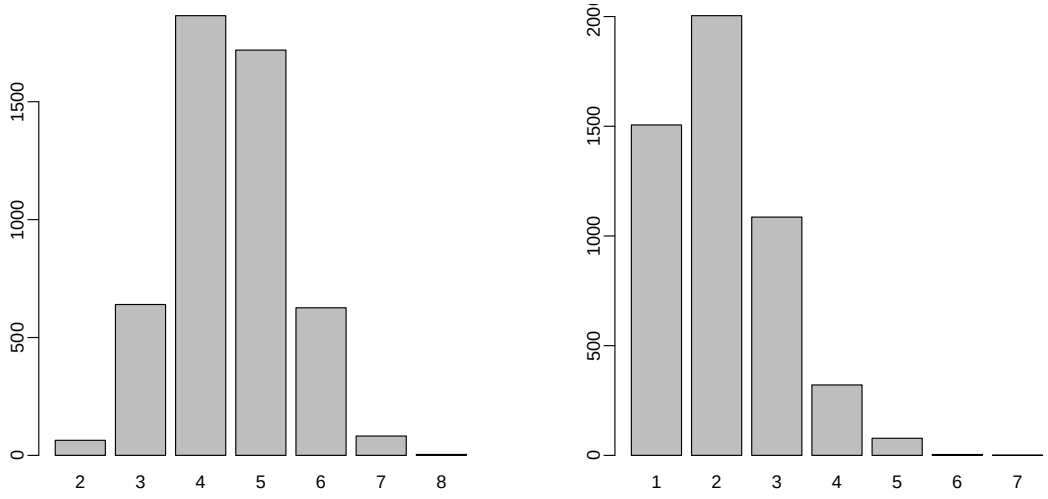

$p(s_i = s_j \mid y)$ SSM DP

Figure 5: Posterior probabilities of pairwise co-clustering, $p_{ij} = p(s_i = s_j \mid y)$. The grey scales in the two panels are scaled as black for $p_{ij} = 0$ to white for $p_{ij} = \max_{r,s} p_{rs}$. The maxima are indicated in the right top of the plots.

prior equivalent sample size of 1. The total mass $M = 2.83$ for the DP prior was selected to achieve matching prior expected number of clusters under the two models. The DP prior on $G$ favors the formation of large clusters (with matched prior mean number of clusters) which leads to less posterior shrinkage of cluster-specific means. In contrast, under the SSM prior the posterior puts more weight on several smaller clusters.

Figure 5 shows the estimated posterior probabilities of pairwise co-clustering for model (16) in the left panel, and for the DP case (right panel). Clearly, compared to the the DP model, the chosen SSM induces a posterior distribution with more clusters, as reflected in the lower posterior probabilities $p(s_i = s_j \mid y)$ for all $i, j$.

Figure 6 shows the posterior distribution of the number of clusters under the SSM and DP mixture models. Under the DP (right panel) includes high probability for a single cluster, $k = 1$, with $n_1 = 8$. The high posterior probability for few large clusters also implies high posterior probabilities $\widehat{p}_{ij}$ of co-clustering. Under the SSM (left panel) the




$p(k \mid y)$ SSM DP

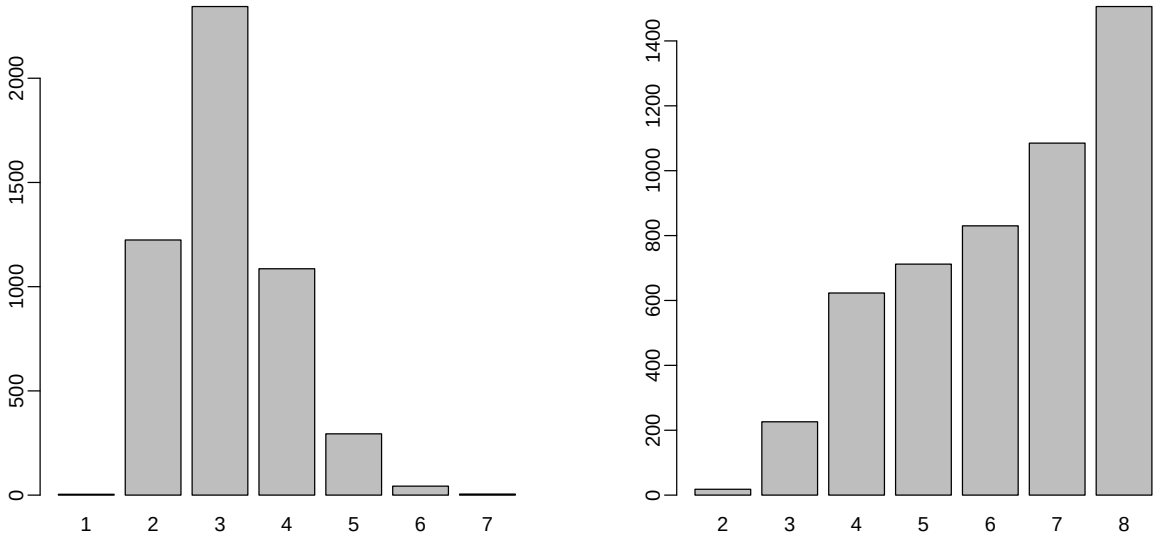
Figure 6: Posterior distribution on the number of clusters.

posterior distribution on $\rho$ retains substantial uncertainty. Finally, the same pattern is confirmed in the posterior distribution of sizes of the largest cluster, $p(n_1 \mid y)$, shown in Figure 7. The high posterior probability for a single large cluster of all $n = 8$ sarcoma subtypes seems unreasonable for the given data.

# 4 Discussion

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

We provide an alternative class of models that allows for more general PPFs. These models are obtained by directly specifying the distribution of unnormalized weights $u_h$. The proposed approach for defining SSMs allows the incorporation of the desired quali-




$p(n_1 \mid y)$ SSM DP

Figure 7: Posterior distribution on the size of the largest cluster.

tative properties concerning the decrease of the ordered clusters cardinalities. This flexibility comes at the cost of additional computation required to implement the algorithm described in Section 3.2, compared to the standard approaches under DP-based models. Nevertheless, the benefits obtained in the case of data sets that require more flexible models, compensate the increase in computational effort. A different strategy for constructing discrete random distributions has been discussed in Trippa and Favaro (2012). In several applications, the scope for which SSMs are to be used suggests these *desired qualitative properties*. Nonetheless, we see the definition of a theoretical framework supporting the selection of a SSM as an open problem.

R code for an implementation of posterior inference under the proposed new model is available at `http://math.utexas.edu/users/pmueller/`.

# Acknowledgments

Jaeyong Lee was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (20090075171). Fernando Quintana was supported by grant FONDECYT 1100010. Peter Müller was partially funded by grant NIH/NCI CA075981.

# References

[1] Donald A. Berry and Ronald Christensen. Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.*, 7(3):558–568, 1979.

[2] David Blackwell and James B. MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

[3] Anders Brix. Generalized gamma measures and shot-noise Cox processes. *Adv. in Appl. Probab.*, 31(4):929–953, 1999.

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

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

[6] Thomas S. Ferguson and Michael J. Klass. A representation of independent increment processes without gaussian components. *Ann. Math. Statist.*, 43:1634–1643, 1972.

[7] S. Fortinelli, L. Ladelli, and E. Regazzini. Exchangeability, predictive distributions and parametric models. *Sankhyā*, 62:86–109, 2000.

[8] A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Rossiĭskaya Akademiya Nauk. Sankt-Peterburgskoe Otdelenie. Matematicheskiĭ Institut im. V. A. Steklova. Zapiski Nauchnykh Seminarov (POMI)*, 325(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 12):83–102, 244–245, 2005.

[9] Alexander Gnedin, Chris Haulk, and Jim Pitman. Characterizations of exchangeable partitions and random discrete distributions by deletion properties. In *Probability and mathematical genetics*, volume 378 of *London Math. Soc. Lecture Note Ser.*, pages 264–298. Cambridge Univ. Press, Cambridge, 2010.

[10] Alexander Gnedin and Jim Pitman. Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, 138(3):5674–5685, 2006.

[11] Irving John Good. *The estimation of probabilities. An essay on modern Bayesian methods*. Research Monograph, No. 30. The M.I.T. Press, Cambridge, Mass., 1965.

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

[13] Lancelot James, Antonio Lijoi, and Igor Prünster. Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.*, 36(1):76–97, 2009.

[14] Lancelot F. James. Large sample asymptotics for the two-parameter Poisson-Dirichlet process. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 187–199. Inst. Math. Statist., Beachwood, OH, 2008.

[15] Gunho Jang, Jaeyong Lee, and Sangyeol Lee. Posterior consistency of species sampling priors. *Statist. Sinica*, 20:581–593, 2010.

[16] J. F. C. Kingman. Random discrete distribution. *J. Roy. Statist. Soc. Ser. B*, 37:1–22, 1975. With a discussion by S. J. Taylor, A. G. Hawkes, A. M. Walker, D. R. Cox, A. F. M. Smith, B. M. Hill, P. J. Burville, T. Leonard and a reply by the author.

[17] J. F. C. Kingman. The representation of partition structures. *J. London Math. Soc. (2)*, 18(2):374–380, 1978.

[18] J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, 1982.

[19] Luis León-Novelo, B. Nebiyou Bekele, Peter Müller, and Fernando A. Quintana. Borrowing Strength with Non-Exchangeable Priors over Subpopulations. *Biometrics*, 2012. In press.

[20] 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.

[21] Antonio Lijoi, Ramsés H. Mena, and Igor Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007.

[22] Antonio Lijoi and Igor Prünster. Models beyond the Dirichlet Process. In N.L. Hjort, C. Holmes, P. Müller, and S.G. Walker, editors, *Bayesian Nonparametrics*, pages 80–136. Cambridge Univ. Press, Cambridge, 2010.

[23] 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.

[24] Antonio Lijoi, Igor Prünster, and Stephen G. Walker. Bayesian nonparametric estimators derived from conditional Gibbs structures. *The Annals of Applied Probability*, 18(4):1519–1547, 2008.

[25] Antonio Lijoi, Igor Prünster, and Stephen G. Walker. Investigating nonparametric priors with Gibbs structure. *Statistica Sinica*, 18(4):1653–1668, 2008.

[26] Jun S. Liu. Nonparametric Hierarchical Bayes via Sequential Imputations. *The Annals of Statistics*, 24(3):911–930, 1996.

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

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

[29] Carlos Navarrete, Fernando A. Quintana, and Peter Müller. Some issues on nonparametric bayesian modeling using species sampling models. *Statistical Modelling. An International Journal*, 8(1):3–21, 2008.

[30] L.E. Nieto-Barajas, I. Prünster, and S.G. Walker. Normalized random measures driven by increasing additive processes. *Ann. Statist.*, 32:2343–2360, 2004.

[31] Mihael Perman, Jim Pitman, and Marc Yor. Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields*, 92(1):21–39, 1992.

[32] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

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

[34] 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.

[35] Jim Pitman. Poisson-Kingman partitions. In *Statistics and science: a Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes Monogr. Ser.*, pages 1–34. Inst. Math. Statist., Beachwood, OH, 2003.

[36] E. Regazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *Ann. Statist.*, 31:560–585, 2003.

[37] Lorenzo Trippa and Stefano Favaro. A Class of Normalized Random Measures with an Exact Predictive Sampling Scheme. *Scandinavian Journal of Statistics*, 2012. In press.

[38] Sandy L. Zabell. W. E. Johnson's "sufficientness" postulate. *Ann. Statist.*, 10(4):1090–1099 (1 plate), 1982.