On a class of random probability measures with general predictive structure

Scandinavian Journal of Statistics
Theory and Applications

Scandinavian Journal of Statistics, Vol. 38: 359–376, 2011
doi: 10.1111/j.1467-9469.2010.00702.x

On a Class of Random Probability Measures with General Predictive Structure

STEFANO FAVARO

Department of Statistics and Applied Mathematics and Collegio Carlo Alberto, University of Turin

IGOR PRÜNSTER

Department of Statistics and Applied Mathematics, Collegio Carlo Alberto and ICER, University of Turin

STEPHEN G. WALKER

Institute of Mathematics, Statistics and Actuarial Science, University of Kent

ABSTRACT. In this study, we investigate a recently introduced class of non-parametric priors, termed generalized Dirichlet process priors. Such priors induce (exchangeable random) partitions that are characterized by a more elaborate clustering structure than those arising from other widely used priors. A natural area of application of these random probability measures is represented by species sampling problems and, in particular, prediction problems in genomics. To this end, we study both the distribution of the number of distinct species present in a sample and the distribution of the number of new species conditionally on an observed sample. We also provide the Bayesian Non-parametric estimator for the number of new species in an additional sample of given size and for the discovery probability as function of the size of the additional sample. Finally, the study of its conditional structure is completed by the determination of the posterior distribution.

Key words: Bayesian Non–parametrics, Dirichlet process, exchangeable random partitions, generalized gamma process, Lauricella hypergeometric function, species sampling models

1. Introduction

Let $(X_n)_{n\geq 1}$ be a sequence of exchangeable observations defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with values in a complete and separable metric space $\mathbb{X}$ equipped with the Borel σ-field $\mathscr{X}$. Then, by de Finetti's representation theorem, there exists a random probability measure (r.p.m.) $\tilde{P}$ such that given $\tilde{P}$, a sample $X_1, \ldots, X_n$ from the exchangeable sequence is independent and identically distributed (i.i.d.) with distribution $\tilde{P}$. That is, for every $n \geq 1$ and any $A_1, \ldots, A_n \in \mathscr{X}$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n \mid \tilde{P}) = \prod_{i=1}^{n} \tilde{P}(A_i).$$

The law of the r.p.m. $\tilde{P}$ acts as a non-parametric prior for Bayesian inference. In this study, we focus on r.p.m.s that are almost surely discrete and with non-atomic prior guess at the shape $P_0(\cdot) := E[\tilde{P}(\cdot)]$. By the almost sure discreteness, we expect ties in the sample, namely that $X_1, \ldots, X_n$ contain $K_n \leq n$ distinct observations $X_1^*, \ldots, X_{K_n}^*$ with frequencies $\mathbf{N}_{K_n} = (N_1, \ldots, N_{K_n})$ such that $\sum_{j=1}^{K_n} N_j = n$. The joint distribution of K_n and $\mathbf{N}_{K_n}$:

$$\mathbb{P}[\{K_n = k\} \cap \{\mathbf{N}_{K_n} = \mathbf{n}\}] = p_k^{(n)}(n_1, \ldots, n_k), \tag{1}$$

provides the partition distribution, that is the probability that a sample of size n exhibits k distinct observations with frequencies $\mathbf{n}$. Such a distribution is known in the literature as *exchangeable partition probability function* (EPPF; Pitman, 1995) and uniquely determines the probability law of an exchangeable random partition. Almost sure discrete r.p.m.s and the exchangeable random partitions they induce have always played an important role in a variety of research areas such as population genetics, machine learning, combinatorics, excursion theory, statistical physics and Bayesian Non-parametrics. In particular, in Bayesian Non-parametric inference, the use of random partitions dates back to the seminal work of Lo (1984): his approach consists of exploiting a discrete r.p.m. as a basic building block in hierarchical mixture models. In this way, the discrete r.p.m. induces an exchangeable random partition for the latent variables, providing an effective tool for inference on the clustering structure of the observations. See e.g. Lo & Weng (1989), James (2002), Ishwaran & James (2003), Lijoi *et al.* (2007a), Navarrete *et al.* (2008) and Müller & Quintana (2010b), for extensions in various directions. Since the introduction of the Dirichlet process in Ferguson (1973), other classes of almost surely discrete r.p.m.s have been proposed in the literature. Among them, we mention species sampling models (Pitman, 1996), stick-breaking r.p.m.s (Ishwaran & James, 2001), normalized random measures with independent increments (NRMI; Regazzini *et al.*, 2003) and Poisson–Kingman models (Pitman, 2003). Within these classes, all specific r.p.m.s, which enjoy sufficient mathematical tractability, represent valid alternatives to the Dirichlet process: the most notable are the two parameter Poisson–Dirichlet process (Pitman, 1995, 1996) and the normalized generalized gamma process (James, 2002; Pitman, 2003; Lijoi *et al.*, 2007a); both recover the normalized stable process (Kingman, 1975) and the Dirichlet process as limiting cases and the latter also contains the normalized inverse-Gaussian process. By close inspection of these tractable processes, one can observe that they all generate samples $X_1, \ldots, X_n$, for $n \geq 1$, which are characterized by a system of predictive distributions of the type:

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1, \ldots, X_n) = g_0(n,k)P_0(\cdot) + g_1(n,k)\sum_{j=1}^{k}(n_j - \sigma)\delta_{X_j^*}(\cdot), \tag{2}$$

where $\sigma \in [0, 1)$ and g_0 and g_1 are suitable non-negative functions satisfying $g_0(n,k) + g_1(n,k) \times (n - \sigma k) = 1$ for any $n \geq 1$ and $k \leq n$. An almost surely discrete r.p.m. generating a predictive distributions as the above is termed Gibbs-type r.p.m. The class of Gibbs-type r.p.m.s has been recently introduced and studied by Gnedin & Pitman (2005), where also a characterization of its members is provided: indeed, Gibbs-type r.p.m.s are Dirichlet process mixtures when $\sigma = 0$ and Poisson–Kingman models based on the stable subordinators when $\sigma \in (0, 1)$ (see theorem 12 in Gnedin & Pitman, 2005). Further investigations related to Bayesian Non-parametrics can be found in Ho *et al.* (2007) and Lijoi *et al.* (2008c).

Recently, Bayesian Non-parametric methods have found a fertile ground of applications in biostatistics. Interesting reviews can be found in Dunson (2010) and Müller & Quintana (2004, 2010a). One of such recent applications concerns species sampling problems, which gained a renewed interest due to their importance in genomics. In Lijoi *et al.* (2007b), properties of samples generated by Gibbs-type r.p.m.s have been analysed. In particular, given a sample $(X_1, \ldots, X_n)$ consisting in a collection of k distinct species with labels $(X_1^*, \ldots, X_k^*)$ and frequencies $\mathbf{n}$, interest is in the distributional properties of an additional sample of size m and, especially, in the distribution of the new distinct species. In genomics, the population is typically a cDNA library and the species are unique genes that are progressively sequenced; see Lijoi *et al.* (2007a,c, 2008a) and references therein. Bayesian estimators for this and related problems have been derived under the hypothesis that the exchangeable sequence is governed by a Gibbs-type prior. It is to be noted that the number of distinct species in the given sample

K_n turns out to be a sufficient statistic for prediction of the number of new distinct species (and other interesting quantities) to be observed in a future sample (Lijoi *et al.*, 2008b). This implies that the information arising from the frequencies $\mathbf{n}$ has to be incorporated into the parameters of the model, as, otherwise, prediction of new species would not depend at all on $\mathbf{n}$. For instance, if the species are exchangeable with a two-parameter Poisson–Dirichlet prior, then, given a sample of size n, the $(n+1)$th observation is a new species with probability $(\theta+\sigma k)/(\theta+n)$, where $\theta > -\sigma$ and $\sigma \in [0,1)$. Such a probability depends on the distinct observed species k but not on their frequencies $\mathbf{n}$, whose conveyed information can be summarized through the selection of θ and σ. In principle, one would like priors that lead to richer predictive structures, in which the probability of sampling a new species depends explicitly on both K_n and $\mathbf{N}_{K_n}$. However, by dropping the Gibbs structure assumption, serious issues of mathematical tractability arise.

In this study, we consider a class of r.p.m.s, which is not of Gibbs-type, and show that one can still derive analytic expressions for the quantities of interest leading to prediction schemes that incorporate all the sample information. In pursuing this goal, we will also highlight some nice connections between Bayesian Non-parametrics and exchangeable random partitions on one side and the theory of special functions on the other. Other examples of this close connection can be found in Regazzini (1998), Lijoi & Regazzini (2004) and James *et al.* (2008) where functionals of the Dirichlet process are considered. The specific class we consider is represented by the generalized Dirichlet process introduced in Regazzini *et al.* (2003) and further investigated in Lijoi *et al.* (2005). In particular, the generalized Dirichlet process is an NRMI obtained by normalization of superposed independent gamma processes with increasing integer-valued scale parameter and gives rise to a system of predictive distributions of the type:

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1, \ldots, X_n) = w_0(n,k,\mathbf{n})P_0(\cdot) + \sum_{j=1}^{k} n_j w_j(n,k,\mathbf{n})\delta_{X_j^*}(\cdot), \tag{3}$$

where the weights $w_0(n,k,\mathbf{n})$ and $w_j(n,k,\mathbf{n})$, for $j=1,\ldots,k$, now explicitly depend on $\mathbf{n}$ thus conveying the additional information provided by the frequencies $n_1,\ldots,n_k$ directly into the prediction mechanism. To our knowledge, the generalized Dirichlet process represents the first example in the literature of almost surely discrete r.p.m., which is not of Gibbs-type and still leads to a closed-form predictive structure.

The study is structured as follows. In section 2, we recall the concept of exchangeable random partition of Gibbs-type together with some distributional results related to samples generated by a Gibbs-type r.p.m. In section 3, we provide distributional results related to the prior and posterior probability distribution of discovering a certain number of new species in a sample generated by a generalized Dirichlet process. Moreover, a characterization of the posterior distribution is obtained. Section 4 illustrates the roles of the prior parameters and the predictive behaviour of the generalized Dirichlet process. The Appendix contains a short review on completely random measures (CRMs) and Bell polynomials. Proofs of the results can be found in the Supporting Information, which may be found in the online version of this article.

2. Gibbs-type r.p.m.s

A random partition of the set of natural numbers $\mathbb{N}$ is defined as a consistent sequence $\Pi := \{\Pi_n, n \geq 1\}$ of random elements, with Π_n taking values in the set of all partitions of $\{1,\ldots,n\}$ into some number of disjoint blocks. Consistency implies that each Π_n is obtained from Π_{n+1} by discarding, from the latter, the integer $n+1$. A random partition Π is exchange-

able if, for each n, the probability distribution of Π_n is invariant under all permutations of $\{1,\ldots,n\}$. Consider now the set $\mathcal{D}_{k,n} := \{(n_1,\ldots,n_k) \in \{1,\ldots,n\}^k : \sum_{i=1}^k n_i = n\}$ and let $\{p_k^{(n)}, n \geq 1\}$ be a sequence of functions such that $p_k^{(n)} : \mathcal{D}_{k,n} \to [0,1]$ satisfies the properties:

(i) $p_1^{(1)}(1) = 1$;
(ii) for any $(n_1,\ldots,n_k) \in \mathcal{D}_{k,n}$ with $n \geq 1$ and $k \in \{1,\ldots,n\}$,

$$p_k^{(n)}(n_1,\ldots,n_k) = p_k^{(n)}(n_{\rho(1)},\ldots,n_{\rho(k)}),$$

where ρ is an arbitrary permutation of the indices $(1,\ldots,k)$;
(iii) for any $(n_1,\ldots,n_k) \in \mathcal{D}_{k,n}$ with $n \geq 1$ and $k \in \{1,\ldots,n\}$, the following addition rule holds true:

$$p_k^{(n)}(n_1,\ldots,n_k) = \sum_{j=1}^{k} p_k^{(n+1)}(n_1,\ldots,n_j+1,\ldots,n_k) + p_{k+1}^{(n+1)}(n_1,\ldots,n_k,1).$$

Then, $p_k^{(n)}$ is an EPPF (Pitman, 1995). In particular, the EPPF uniquely determines the probability law of an exchangeable random partition according to the equality:

$$\mathbb{P}(\Pi_n = (A_1,\ldots,A_k)) = p_k^{(n)}(|A_1|,\ldots,|A_k|) \quad \text{for any } n \geq 1 \text{ and } k \leq n,$$

where $|A|$ stands for the cardinality of the set A. For a comprehensive account, the reader is referred to Pitman (2006). As already seen in (1), an exchangeable sequence of random variable (r.v.) governed by an almost sure discrete r.p.m. always yields an EPPF, which corresponds to the samples' partition distribution. In the following, let

$$(a)_n = \prod_{j=1}^{n} (a+j-1)$$

be the ascending factorial of a with the convention that $(a)_0 \equiv 1$. In light of the above considerations, Gibbs-type random partitions can be defined via their EPPF as follows.

Definition 1 (Gnedin & Pitman, 2005)
An exchangeable random partition Π *of the set of natural numbers is said to be of Gibbs form if, for all* $1 \leq k \leq n$ *and for any* $(n_1,\ldots,n_k)$ *in* $\mathcal{D}_{k,n}$, *the EPPF of* Π *can be represented as*

$$p_k^{(n)}(n_1,\ldots,n_k) = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{(n_j-1)} \tag{4}$$

for some $\sigma \in [0,1)$ *and some set of non-negative real numbers* $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ *satisfying the recursion* $V_{n,k} = V_{n+1,k+1} + (n-\sigma k) V_{n+1,k}$ *with* $V_{1,1} = 1$.

Recall that, according to Pitman (1996), a species sampling model is an almost surely discrete r.p.m. $\tilde{P}(\cdot) = \sum_{i \geq 1} \tilde{w}_i \delta_{Y_i}$ such that the masses $\tilde{w}_i$s are independent from the locations Y_is, which are i.i.d. from a non-atomic distribution P_0. Then, one can define Gibbs-type r.p.m.s as the class of species sampling models that induce exchangeable random partitions of Gibbs-type, i.e. the EPPF corresponding to a sample of size n generated by a Gibbs-type r.p.m. is of the form (4). It then follows that the predictive distributions associated with a Gibbs-type r.p.m. are of the form (2) with

$$g_0(n,k) := \frac{V_{n+1,k+1}}{V_{n,k}} \quad g_1(n,k) := \frac{V_{n+1,k}}{V_{n,k}}.$$

Before recalling the distributional results for samples drawn from Gibbs-type priors, we introduce some useful notation to be used throughout the study. We denote by $X_k^{(1,n)} :=$

$(X_1,\ldots,X_n)$ a 'basic sample' of size n containing $k\in\{1,\ldots,n\}$ distinct species, which corresponds to the typically available information. Analogously, we denote by $X^{(2,m)}:=(X_{n+1},\ldots,X_{n+m})$ the additional, unobserved, sample of size m, whose distinctive characteristics have to be predicted based on $X_k^{(1,n)}$. Moreover, let $K_m^{(n)}:=K_{n+m}-K_n$ be the number of new species in $X^{(2,m)}$ and denote by $X_j^{(2,m)}$ a new m-sample featuring $K_m^{(n)}=j$. By $\mathfrak{S}(n,k,\sigma)$, we denote the generalized factorial coefficient, which, for any $n\geq 1$ and $k=1,\ldots,n$ and $\sigma\in\mathbb{R}$, can be represented as

$$\mathfrak{S}(n,k;\sigma)=\frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(-j\sigma)_n$$

with the proviso that $\mathfrak{S}(0,0,\sigma)=1$ and $\mathfrak{S}(n,0,\sigma)=0$ for all $n\geq 1$. See Charalambides (2005) for a review of generalized factorial coefficients. The probability distribution of the number of distinct observations K_n in the 'basic sample', derived in Gnedin & Pitman (2005), corresponds to

$$\mathbb{P}(K_n=k)=\frac{V_{n,k}}{\sigma^k}\mathfrak{S}(n,k,\sigma). \tag{5}$$

In Lijoi *et al.* (2007a), the probability distribution (5) is reinterpreted as the prior probability distribution of the number of species to be observed in the 'basic sample' and represents the starting point for the determination of the following relevant quantities:

- the probability distribution, and the expected value, of the number of new species in the second sample conditionally on the number of species in the 'basic sample' $X_k^{(1,n)}$;
- the probability of discovering a new species at the $(n+m+1)$th draw, without actually observing the second sample $X^{(2,m)}$.

Evaluating the probability in (1) is equivalent to determining $\mathbb{P}(K_m^{(n)}=j\,|\,X_k^{(1,n)})$ for any $j=0,1,\ldots,m$ and for any $k=1,2,\ldots,n$, which can be interpreted as the 'posterior' probability distribution of the number of species to be observed in a sample of size m. As shown in Lijoi *et al.* (2007a), in the Gibbs case, such a distribution is given by

$$\mathbb{P}(K_m^{(n)}=j\,|\,X_k^{(1,n)})=\frac{V_{n+m,k+j}}{V_{n,k}\sigma^j}\sum_{s=j}^{m}\binom{m}{s}\mathfrak{S}(s,j,\sigma)(n-k\sigma)_{(m-s)}. \tag{6}$$

From (6) one immediately recovers $\mathbb{E}[K_m^{(n)}\,|\,X_k^{(1,n)}]$, the Bayes estimator of $K_m^{(n)}$ given $X_k^{(1,n)}$ under a quadratic loss function. Moreover, (6) implies that K_n is sufficient for predicting the number of new distinct species.

The determination of the probability in (2) corresponds to estimating the probability $\mathbb{P}(K_1^{(n+m)}=1\,|\,K_n=k)$ without observing $K_m^{(n)}$. The corresponding Bayes estimator, given in Lijoi *et al.* (2007a) for Gibbs-type prior driven exchangeable sequences, is of the form:

$$\hat{D}_m^{(n:k)}=\sum_{j=0}^{m}\frac{V_{n+m+1,k+j+1}}{V_{n,k}\sigma^j}\sum_{s=j}^{m}\binom{m}{s}\mathfrak{S}(s,j,\sigma)(n-k\sigma)_{(m-s)},$$

and automatically provides a solution to the important problem of determining the sample size such that the probability of discovering a new species falls below a given threshold.

It is worth noting that the above estimators provide Bayesian Non-parametric counterparts to frequentist non-parametric procedures, which suffer from serious drawbacks when the size of the additional sample m is larger than the size n of the initial sample. See Lijoi *et al.* (2007c) for a comparison with the popular Good–Toulmin estimator (Good & Toulmin, 1956). We conclude this section by specializing the above *formulae* to the Dirichlet case. This is useful in view of section 3 where a generalization of the Dirichlet process will be investigated.

Example 1. The Dirichlet process with parameter measure α is a Gibbs-type r.p.m. with $\sigma=0$. Setting $a:=\alpha(\mathbb{X})$, its EPPF is of the form:

$$p_k^{(n)}(n_1,\ldots,n_k)=\frac{a^k}{(a)_n}\prod_{j=1}^{k}\Gamma(n_j), \tag{7}$$

which represents a version of the celebrated Ewens' sampling formula (Ewens, 1972). The prior distribution of the number of distinct species within a sample of size n, due to Ewens (1972) and Antoniak (1974), is obtained by letting $\sigma\to 0$ in (5), which yields

$$\mathbb{P}(K_n=k)=\frac{a^k}{(a)_n}|s(n,k)|,$$

where $|s(n,k)|:=\lim_{\sigma\to 0}\mathfrak{S}(n,k,\sigma)/\sigma^k$ stands for the signless or absolute Stirling number of the first kind. Moreover, the 'posterior' distribution of the number of distinct species to be observed in the additional sample becomes

$$\mathbb{P}(K_m^{(n)}=j\,|\,X_k^{(1,n)})=\frac{a^j(a)_n}{(a)_{(n+m)}}|s(m,j,n)| \tag{8}$$

for any $j=0,1,\ldots,m$, where $|s(m,j,n)|$ denotes a non-central signless Stirling number of the first kind (see Charalambides, 2005). The following proposition provides expressions for the species sampling estimators by greatly simplifying those obtained in Lijoi *et al.* (2007a).

Proposition 1
Let $\{X_n,n\geq 1\}$ be an exchangeable sequence governed by a Dirichlet process with non-atomic parameter measure α. Then

$$\mathbb{E}(K_m^{(n)}=j\,|\,X_k^{(1,n)})=\sum_{i=1}^{m}\frac{a}{a+n+i-1}, \tag{9}$$

$$\hat{D}_m^{(n:k)}=\frac{a}{a+n+m}. \tag{10}$$

Note that (9) admits an interesting probabilistic interpretation. In fact, it can be rewritten as $\frac{a}{a+n}\sum_{i=1}^{m}\frac{a+n}{a+n+i-1}$, which is equal to $\mathbb{P}(X_{n+1}=\text{new}\,|\,X_k^{(1,n)})\mathbb{E}_{a+n}(K_m)$, where $\mathbb{E}_{a+n}(K_m)$ is the unconditional expected number of species in a sample of size m corresponding to a Dirichlet process with total mass parameter $a+n$. Moreover, (10) is simply equal to the probability of observing a new species given a sample of size $n+m$. Interestingly, (8), and consequently the corresponding estimators in (9) and (10), solely depend on the sample size: prediction does not depend on K_n and $\mathbf{N}_{K_n}$ and so all this information has to be summarized by a single parameter a. This, which is a characterizing property of the Dirichlet process (Zabell, 1982), represents a severe limitation for predictive purposes.

3. A class of r.p.m.s without Gibbs structure

We first recall the definition of the generalized Dirichlet process and then provide solutions to the species sampling problems described in (1) and (2) of section 2, when the exchangeable sequence is governed by a generalized Dirichlet process.

Let us start by defining a CRM $\tilde{\xi}$ (i.e. a random measure such that for any disjoint sets $A_1,\ldots,A_n\in\mathcal{X}$, the r.v.s $\tilde{\mu}(A_1),\ldots,\tilde{\mu}(A_n)$ are mutually independent) characterized by its Lévy intensity:

$$v(\mathrm{d}s,\mathrm{d}x)=\frac{\exp\{-s\}}{s(1-\exp\{-s\})}(1-\exp\{-\gamma s\})\,\mathrm{d}s\,\alpha(\mathrm{d}x)\quad s\geq 0,\ \ \gamma>0, \tag{11}$$

where α is a finite measure on $\mathbb{X}$ with $a:=\alpha(\mathbb{X})>0$. See the Appendix for the basic facts about CRM and Lijoi & Prünster (2010) for their central role in Bayesian Non-parametrics. The Lévy intensity (11) implies that, for any $A\in\mathscr{X}$, $\tilde{\xi}(A)$ is the negative logarithm transform of a Beta d.f. with parameters $(\alpha(A),\gamma)$ and belongs to the class of generalized convolutions of mixtures of exponential d.f.s, introduced in Bondesson (1981). Note that, if $\mathbb{X}=\mathbb{R}^+$ and $\alpha(\mathrm{d}x)=\mathrm{d}x$, the corresponding Lévy process represents an interesting special case of the class of subordinators with logarithmic singularity deeply investigated in von Renesse *et al.* (2008).

We are now in a position to recall the definition of the generalized Dirichlet process.

Definition 2 (Regazzini *et al.*, 2003)
Given the CRM $\tilde{\xi}$ identified by (11)*, the generalized Dirichlet process with parameters α and γ is defined as*

$$\tilde{P}(\cdot)\overset{d}{=}\frac{\tilde{\xi}(\cdot)}{\tilde{\xi}(\mathbb{X})}.$$

If $\gamma=1$, the intensity in (11) reduces to the intensity of a gamma CRM and, hence, $\tilde{P}$ becomes a simple Dirichlet process. The fact that the generalized Dirichlet process, which represents a subclass of NRMIs, is not of Gibbs-type follows immediately from Gnedin & Pitman (2005) and Lijoi *et al.* (2008c): for $\sigma=0$, the only Gibbs-type NRMI is the Dirichlet process, whereas for $\sigma>0$ the only NRMIs of Gibbs-type are normalized generalized gamma processes. This implies that the weights in the predictive distribution (3) necessarily depend, not only on n and k, but also on all frequencies $\mathbf{n}=(n_1\ldots,n_k)$ and this corresponds precisely to the general dependence structure we are aiming at.

For various reasons in the following we always assume $\gamma\in\mathbb{N}$. From an interpretational point of view, $\gamma\in\mathbb{N}$ allows to see $\tilde{\xi}$ as arising from the superposition of γ independent gamma CRMs: specifically, $\tilde{\xi}=\sum_{i=1}^{\gamma}\tilde{\xi}^{(i)}$, where $\tilde{\xi}^{(i)}$ is a gamma CRM with scale parameter i and shape parameter α. Moreover, the marginal distribution is now a member of the important class of generalized gamma convolutions due to Thorin (1977). In fact, $\tilde{\xi}(A)$, for some $A\in\mathbb{X}$, is then distributed as the convolution of γ independent r.v.s with parameters $(i,\alpha(A))$, for $i=1,\ldots,\gamma$, i.e.

$$\mathbb{E}[\exp\{-\lambda\tilde{\xi}(A)\}]=\prod_{i=1}^{\gamma}\left(\frac{i}{i+\lambda}\right)^{\alpha(A)}\qquad\lambda\geq 0.$$

See James *et al.* (2008) for an interesting account on the connections between generalized gamma convolutions and Bayesian Non-parametrics. In terms of mathematical tractability, it is well-known that convolutions of gamma r.v.s can be represented in terms of Lauricella functions (Exton, 1976), which will represent the key quantities in terms of which the relevant closed-form expressions will be provided. In the following, we also assume α to be a non-atomic measure, which is tantamount to requiring the prior guess at the shape $P_0(\cdot)=\mathbb{E}[\tilde{P}(\cdot)]$ to be non-atomic given that $P_0(\cdot)=\alpha(\cdot)/a$.

A first treatment of the generalized Dirichlet process in this set-up was provided in Lijoi *et al.* (2005), where its finite-dimensional distributions, moments and linear functionals were studied. Moreover, its EPPF, interpretable as the joint distribution of the number of species and their frequencies according to (1), is given by

$$p_k^{(n)}(\mathbf{n}) = \frac{(\gamma!)^a a^k \prod_{j=1}^k \Gamma(n_j)}{\gamma^{\gamma a}(\gamma a)_n} \mathfrak{F}(k, n, \mathbf{n}, a, \gamma) \tag{12}$$

with

$$\mathfrak{F}(k, n, \mathbf{n}, a, \gamma) := \sum_{\mathbf{r}^k} F_{\mathrm{D}}^{(\gamma-1)}\left(\gamma a, \boldsymbol{\alpha}^*(\mathbf{n}, \mathbf{r}^k); \gamma a + n; \frac{\mathbf{J}\gamma - 1}{\gamma}\right),$$

where the sum is over $\mathbf{r}^k := (r_1, \ldots, r_k) \in \{1, \ldots, \gamma\}^k$ and $F_{\mathrm{D}}^{(\gamma-1)}(\cdot, \cdot; \cdot; \cdot)$ is the fourth form of the Lauricella multiple hypergeometric function. The vectors appearing in the arguments of $F_{\mathrm{D}}^{(\gamma-1)}$ are defined as $\boldsymbol{\alpha}^*(\mathbf{n}, \mathbf{r}^k) := (\alpha_1^*(\mathbf{n}, \mathbf{r}^k), \ldots, \alpha_{\gamma-1}^*(\mathbf{n}, \mathbf{r}^k))$ with $\alpha_l^*(\mathbf{n}, \mathbf{r}^k) := a + \sum_{i=1}^k n_i \mathbb{1}_{\{l=r_i\}}$ for $l = 1, \ldots, \gamma - 1$ and $\mathbf{J}_{\gamma-1} := (1, \ldots, \gamma - 1)$. By setting $\gamma = 1$, from (12) one recovers the Ewens' sampling formula (7). The predictive distributions associated with $\tilde{P}$ are then of the form (3) with

$$w_0(n, k, \mathbf{n}) = \frac{a\mathfrak{F}(k+1, n+1, \mathbf{n}^+, a, \gamma)}{(\gamma a + n)\mathfrak{F}(k, n, \mathbf{n}, a, \gamma)}, \quad w_j(n, k, \mathbf{n}) = \frac{\mathfrak{F}(k, n+1, \mathbf{n}_j^+, a, \gamma)}{(\gamma a + n)\mathfrak{F}(k, n, \mathbf{n}, a, \gamma)}, \tag{13}$$

where we have set $\mathbf{n}^+ := (n_1, \ldots, n_k, 1)$ and $\mathbf{n}_i^+ := (n_1, \ldots, n_i + 1, \ldots, n_k)$ for $i = 1, \ldots, k$.

Before proceeding, a comparison of the predictive structures of Gibbs-type r.p.m.s and the generalized Dirichlet process is in order. In the Gibbs case, the predictive distributions (2) are a linear combination of the prior guess P_0 and a weighted empirical distribution. So X_{n+1} is new with probability $g_0(n, k)$, whereas it coincides with X_j^* with probability $g_1(n, k)(n_j - \sigma)$, for $j = 1, \ldots, k$. The predictive distributions associated with a generalized Dirichlet process are characterized by a more elaborate structure, which exploits all available information in the sample $X_1, \ldots, X_n$: they are still a linear combination of the prior guess P_0 and a weighted empirical distribution, but now X_{n+1} is new with probability $w_0(n, k, \mathbf{n})$ and coincides with X_j^* with probability $n_j w_j(n, k, \mathbf{n})$, for $j = 1, \ldots, k$. Therefore, from (13), we observe that both the weight assigned to each X_j^* and the weight assigned to a new observation depend on the number of distinct observations k as well as on their frequencies $\mathbf{n}$. Moreover, the balance between new and old observations depends on k and $\mathbf{n}$. To the authors knowledge, it is the only r.p.m. not of Gibbs-type, which admits closed-form expressions for the EPPF and the predictive distributions. Hence, it is definitely worth looking for a solution to the problems (1) and (2) described in section 2 in the generalized Dirichlet process case.

The first aim is to derive the distribution of the number of distinct species K_n, which would represent the analogue in the generalized Dirichlet case of (5) for Gibbs-type r.p.m.s. To this end, we resort to the definition of the (n, k)th partial Bell polynomial associated with a non-negative sequence of real numbers $w_\bullet := \{w_i, i \geq 0\}$. A brief account on partial Bell polynomials is given in the Appendix.

Proposition 2

Let $\{X_n, n \geq 1\}$ be an exchangeable sequence governed by a generalized Dirichlet process with non-atomic parameter measure α and parameter $\gamma \in \mathbb{N}$. Then

$$\mathbb{P}(K_n = k) = \frac{((\gamma)!)^a a^k}{\gamma^{\gamma a}\Gamma(n)} \int_0^1 \frac{z^{\gamma a - 1}(1-z)^{n-1}}{\prod_{l=1}^{\gamma-1}(1 - zl/\gamma)^a} B_{n,k}(w_\bullet(z, \gamma))\, \mathrm{d}z, \tag{14}$$

where by convention $\prod_{l=1}^0 (1 - zl/\gamma)^a = 1$ and $B_{n,k}(w_\bullet(z, \gamma))$ is the (n, k)th partial Bell polynomial with $w_\bullet(z, \gamma) := \{w_i(z, \gamma), i \geq 1\}$ such that

$$w_i(z,\gamma)=(i-1)!\left(\sum_{l=1}^{\gamma-1}(1-zl/\gamma)^{-i}+1\right) \tag{15}$$

with the proviso $\sum_{l=1}^{0}(\gamma-zl)^{-i}=0$.

As for the evaluation of (14), it is important to remark that, for fixed n and k, $B_{n,k}(w_\bullet)$ is a polynomial of degree n in the variable $(1-zi/\gamma)^{-1}$, for $i=1,\ldots,\gamma-1$, with a particular set of coefficients specified according to the coefficients of the (n,k)th partial Bell polynomial $B_{n,k}(w_\bullet)$. Therefore, (14) can be easily evaluated using the generalized Picard integral representation of the fourth-type Lauricella multiple hypergeometric function $F_{\mathrm{D}}^{(\gamma-1)}$. For instance, if $\gamma=2$, $F_{\mathrm{D}}^{(1)}$ corresponds to the Gauss hypergeometric function ${}_2F_1$ (see Exton, 1976) and (14) reduces to a weighted linear combination of Gauss hypergeometric functions.

Now, we provide a solution to the species sampling problem (1) described in section 2.

Proposition 3

Let $\{X_n, n\geq 1\}$ *be an exchangeable sequence governed by a generalized Dirichlet process with non-atomic parameter measure* α *and parameter* $\gamma\in\mathbb{N}$. *Then*

$$\mathbb{P}(K_m^{(n)}=j\mid X_k^{(1,n)})=\frac{a^j(\gamma a)_n}{\Gamma(n+m)\mathfrak{F}(k,n,\mathbf{n},a,\gamma)}\sum_{s=j}^{m}\binom{m}{s}(n)_{(m-s)}$$
$$\times\sum_{\mathbf{r}^k}\int_0^1\frac{z^{\gamma a-1}(1-z)^{n+m-1}B_{s,j}(w_\bullet(z,\gamma))F_{\mathrm{D}}^{(k-1)}(-(m-s),\mathbf{n}_{k-1};n;\mathbf{W})}{\prod_{l=1}^{\gamma-1}(1-zl/\gamma)^{a+\sum_{i=1}^{k}n_i\mathbb{1}_{\{l=r_i\}}+(m-s)\mathbb{1}_{\{l=r_k\}}}}\,\mathrm{d}z, \tag{16}$$

where $\mathbf{n}_{k-1}=(n_1,\ldots,n_{k-1})$ *and* $\mathbf{W}=(w_k-w_1/w_k,\ldots,w_k-w_{k-1}/w_k)$ *with* $w_i=\prod_{l=1}^{\gamma-1}(1-zl/\gamma)^{-\mathbb{1}_{\{l=r_i\}}}$.

It is important to remark that, for the generalized Dirichlet process, the conditional distribution of the number of new distinct species exhibits the desired dependence on both K_n and $\mathbf{N}_{K_n}$. This is in contrast to Gibbs-type r.p.m.s, where we have dependence solely on K_n and not even on K_n in the Dirichlet case. Hence, although the distribution in (16) is quite complicated, such a property makes the generalized Dirichlet process appealing for practical purposes. Moreover, having (16) at hand, the computation of the Bayes estimate of $K_m^{(n)}$, given the basic sample $X_k^{(1,n)}$, namely $\mathbb{E}[K_m^{(n)}\mid X_k^{(1,n)}]$, is straightforward.

With reference to problem (2), we now derive a Bayesian estimator of the probability of discovering a new species at the $(n+m+1)$th draw, given an initial observed sample of size n with k distinct species and frequencies $\mathbf{n}$ and without observing the intermediate sample of size m. The next result provides a solution to this problem.

Proposition 4

Let $\{X_n, n\geq 1\}$ *be an exchangeable sequence governed by a generalized Dirichlet process with non-atomic parameter measure* α *and parameter* $\gamma\in\mathbb{N}$. *Then, the Bayes estimate, with respect to a squared loss function, of the probability of observing a new species at the* $(n+m+1)$*th draw, conditional on an initial sample of size n with k distinct species and frequencies* $\mathbf{n}$, *is given by*

$$\hat{D}_m^{n:k:\mathbf{n}}=\sum_{j=0}^{m}\frac{a^{j+1}(\gamma a+n)^{-1}(\gamma a)_n(\gamma a)_{n+m+1}}{\mathfrak{F}(k,n,\mathbf{n},a,\gamma)\Gamma(n+m+1)(\gamma a)_{n+m}}\sum_{s=j}^{m}\binom{m}{s}(n)_{(m-s)}$$
$$\times\sum_{\mathbf{r}^{k+1}}\int_0^1\frac{z^{\gamma a-1}(1-z)^{n+m-1}B_{s,j}(w_\bullet(z,\gamma))F_{\mathrm{D}}^{(k-1)}(-(m-s),\mathbf{n}_{k-1};n;\mathbf{W})}{\prod_{l=1}^{\gamma-1}(1-zl/\gamma)^{a+\sum_{i=1}^{k+1}(n_i\mathbb{1}_{\{i\leq k\}}+\mathbb{1}_{\{i>k\}})+(m-s)\mathbb{1}_{\{l=r_k\}}}}\,\mathrm{d}z, \tag{17}$$

where $\mathbf{n}_{k-1}=(n_1,\ldots,n_{k-1})$ *and* $\mathbf{W}=(w_k - w_1/w_k,\ldots,w_k - w_{k-1}/w_k)$ *with* $w_i=\prod_{l=1}^{\gamma-1}(1 - zl/\gamma)^{-\mathbb{1}_{\{l=r_i\}}}$.

The Bayes estimator in (17), together with $\mathbb{E}[K_m^{(n)} \mid (K_n, \mathbf{N}_{K_n})]$, represent new Bayesian counterparts to the celebrated Good–Toulmin estimator (Good & Toulmin, 1956) and represent alternatives to Bayesian estimators derived from Gibbs-type r.p.m.s (Lijoi *et al.*, 2007a, 2008b). With respect to the latter, these estimators have the advantage of incorporating all the information conveyed by the sample at the cost of a higher computational complexity.

In order to complete the description of the conditional structure of generalized Dirichlet processes, we now derive the posterior distribution that is the conditional distribution of $\tilde{P}$ given a sample $\mathbf{X}=(X_1,\ldots,X_n)$ featuring K_n distinct observations, denoted by $(X_1^*,\ldots,X_{K_n}^*)$, with frequencies $\mathbf{N}_{K_n}$. We will use the symbol $Z^{(W)}$ to denote a random element whose distribution coincides with a regular conditional distribution of Z, given W. By adapting the general results for NRMIs of James *et al.* (2009), in the next proposition we provide the desired posterior characterization of both the unnormalized CRM $\tilde{\xi}$ with intensity (11) and the generalized Dirichlet process $\tilde{P}(\cdot)=\tilde{\xi}(\cdot)/\tilde{\xi}(\mathbb{X})$.

Proposition 5

Let $\tilde{P}$ *be a generalized Dirichlet process with non-atomic parameter measure* α *and parameter* $\gamma\in\mathbb{N}$. *Then, the distribution of* $\tilde{\xi}$, *given the observations* $\mathbf{X}$ *and suitable latent variable* U_n (*see* 20), *coincides with*

$$\tilde{\xi}^{(U_n,\mathbf{X})} \stackrel{d}{=} \tilde{\xi}^{(U_n)} + \sum_{j=1}^{K_n} J_i^{(U_n,\mathbf{X})}\delta_{X_j^*},$$

where

(i) $\tilde{\xi}^{(U_n)}$ *is a CRM with intensity measure*

$$\nu^{(U_n)}(\mathrm{d}x,\mathrm{d}v)=\sum_{l=1}^{\gamma}\frac{\exp\{-v(l+U_n)\}}{v}\,\mathrm{d}v\,\alpha(\mathrm{d}x); \tag{18}$$

(ii) X_j^* *are fixed points of discontinuity, for* $j=1,\ldots,K_n$, *and the* $J_j^{(U_n,\mathbf{X})}$*s are the corresponding jumps that are absolutely continuous w.r.t. to the Lebesgue measure with density*

$$f_{J_j^{(U_n,\mathbf{X})}}(v)\propto v^{n_i-1}\sum_{l=1}^{\gamma}\exp\{-v(l+U_n)\}\quad j=1,\ldots,K_n; \tag{19}$$

(iii) *the jumps* $J_j^{(U_n,\mathbf{X})}$, *for* $j=1,\ldots,K_n$, *are mutually independent and independent from* $\tilde{\xi}^{(U_n)}$.

Moreover, the latent variable U_n, *given* $\mathbf{X}$, *is absolutely continuous w.r.t. the Lebesgue measure with density*

$$f_{U_n^{(\mathbf{X})}}(u)\propto u^{n-1}\prod_{l=1}^{\gamma}(l+u)^{-a}\prod_{j=1}^{K_n}\Gamma(n_j)(\zeta(n_j,1+u)-\zeta(n_j,1+\gamma+u)), \tag{20}$$

where $\zeta(x,y)$ *stands for the generalized Riemann Zeta function* (*or Hurwitz function*) *with parameters* x *and* y.

Finally, the posterior distribution of $\tilde{P}$, *given* $\mathbf{X}$ *and* U_n, *is again an NRMI* (*with fixed points of discontinuity*) *and coincides in distribution with*

$$w\frac{\tilde{\xi}^{(U_n)}}{\tilde{\xi}^{(U_n)}(\mathbb{X})}+(1-w)\frac{\sum_{j=1}^{K_n}J_j^{(U_n,\mathbf{X})}\delta_{X_j^*}}{\sum_{j=1}^{K_n}J_j^{(U_n,\mathbf{X})}}, \tag{21}$$

where $w=\tilde{\xi}^{(U_n)}(\mathbb{X})[\tilde{\xi}^{(U_n)}(\mathbb{X})+\sum_{j=1}^{K_n}J_j^{(U_n,\mathbf{X})}]^{-1}$.

The previous result completes the theoretical analysis of the conditional structure induced by generalized Dirichlet processes and is also useful for practical purposes. Indeed, large values of the parameter γ combined with large additional samples m make the numerical computation of the distributions and estimators derived in propositions 3 and 4 cumbersome. If this is the case, then one can devise a simulation algorithm relying on the posterior characterization of proposition 5. By combining an inverse Lévy measure algorithm, such as the Ferguson–Klass method (see Ferguson & Klass, 1972; Walker & Damien, 2000), for simulating trajectories of $\tilde{\xi}^{(U_n)}$ with a Metropolis–Hastings step for drawing samples from $U_n^{\mathbf{X}}$, one easily obtains realizations of the posterior distribution of the generalized Dirichlet process. Then one can sample a new value X_{n+1}, update the posterior according to proposition 5 and sample a realization of the posterior given $(\mathbf{X}, X_{n+1})$. Proceeding along these lines up to step m, one obtains a realization of the additional sample $X_{n+1}, \ldots, X_{n+m}$. By repeating the procedure N times, one obtains a collection of future scenarios $\{(X_{n+1}^{(i)}, \ldots, X_{n+m}^{(i)}) : i = 1, \ldots, N\}$, which can be used in order to evaluate the quantities of interest. For instance, if $j^{(i)}$ is the number of new distinct species observed in $X_{n+1}^{(i)}, \ldots, X_{n+m}^{(i)}$, $\mathbb{E}[K_m^{(n)} \mid K_n]$ can be evaluated as $1/N \sum_{i=1}^{N} j_m^{(i)}$. Finally note that proposition 5 is also important in the context of mixture modelling, where inference is necessarily simulation-based given the complexity of the models: in fact, it allows to derive conditional sampling schemes, which, in the case of the generalized Dirichlet process, are simpler to implement than algorithms based on the marginal or predictive distributions.

4. Illustration

In this section, we illustrate the behaviour of the generalized Dirichlet process. We start by considering the role of the parameters (a, γ) in terms of prior specification. Then, we show how predictions based on the generalized Dirichlet process adapt to the information conveyed by the data, whereas those derived from the Dirichlet process are not sensitive to it.

With reference to the prior specification of the generalized Dirichlet process, we focus on the qualitative behaviour of the distribution of K_n in (14) as the parameters (a, γ) vary. The parameter a has the same role as in the Dirichlet case in that it controls the location of the distribution of K_n: by increasing a (with γ fixed), the distribution of K_n moves to the right and, consequently, the *a priori* expected number of species becomes larger. On the contrary, the parameter γ allows to tune the flatness of the distribution of K_n: indeed, by increasing γ (with a fixed), the distribution of K_n becomes more flat and obviously moves also to the right. Hence, in some sense, one can say that a large value of γ yields a less informative prior for K_n. This role of γ is illustrated in Fig. 1, where, for $a = 1$ and with $n = 30$, the distributions of K_n are depicted as γ varies (see also Fig. 2).

In order to highlight the posterior behaviour of the generalized Dirichlet process, suppose one is considering the following experiment: a dataset of n observations is going to be collected and, based on the available prior information, a certain number of distinct observations within these data are expected. Once n observations are collected and the number of distinct ones $K_n = k$ recorded, a prediction on the number of new distinct observations within another dataset of m observations has to be provided. Let us assume that the number of observations to be collected at the first stage is $n = 30$ and that the prior guess on the number of distinct ones is its central value 15, i.e. we need to consider a prior specification such that $\mathbb{E}[K_{30}] = 15$. For the Dirichlet process this is achieved by imposing $a = 11.26$, whereas for the generalized Dirichlet process with parameter $\gamma = 5, 10, 15$ one needs to set $a = 2.61, 1.45, 1.04$, respectively. Figure 2 displays the prior distribution of K_{30} corresponding to the four considered processes.

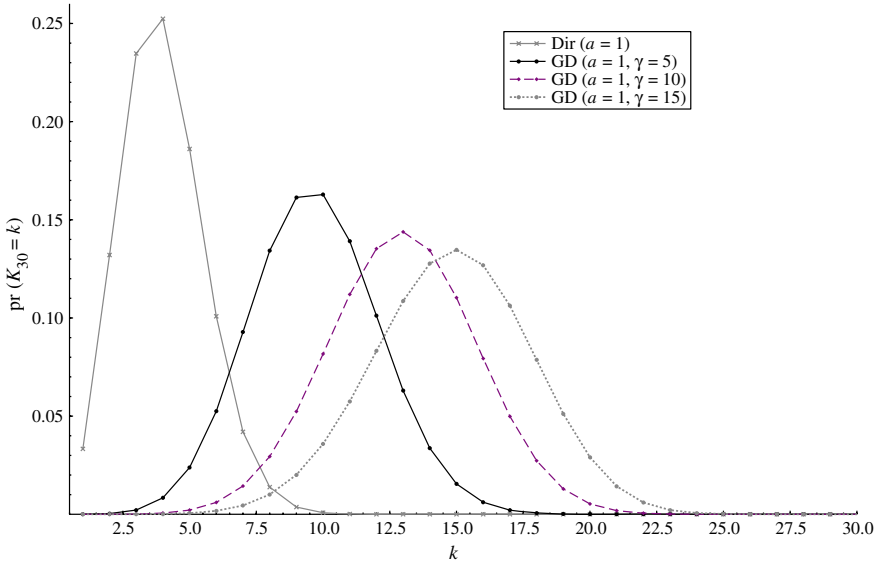

Fig. 1. Distributions of K_{30} corresponding to the Dirichlet process and the three choices of the generalized Dirichlet process with parameters $a=1$ and $\gamma=5, 10, 15$, respectively.

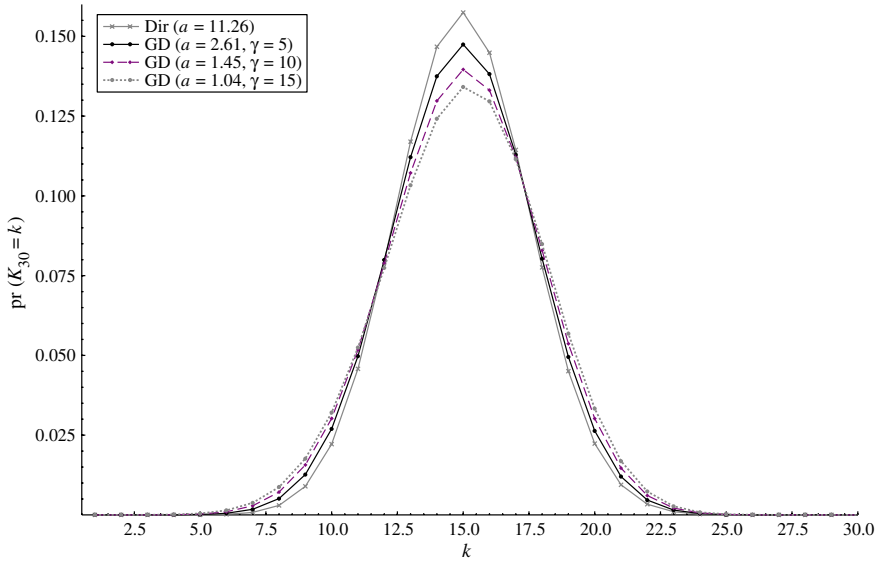

Fig. 2. Distributions of K_{30} corresponding to the Dirichlet process and the three choices of the generalized Dirichlet process such that $\mathbb{E}[K_{30}]=15$.

For facing the prediction problem at issue, we have to make some assumptions on the observed sample of size $n=30$, namely on the number of distinct observed species K_{30} and on their frequencies $N_1,\ldots,N_{K_{30}}$. For illustrative purposes, it seems best to consider the two extreme cases that are given by: (i) all species are distinct, i.e. $K_{30}=30$ implying $N_1=\cdots=N_{30}=1$; (ii) only one species has been observed, i.e. $K_{30}=1$ implying $N_1=30$. A first interesting quantity to look at is given by the sample coverage, i.e. the proportion of distinct species repre-

Table 1 *Estimated sample coverage* $\hat{C}$ *and posterior expected number of species* $\mathbb{E}[K_{30}^{(30)} \mid X_k^{(1,n)}]$ *for the Dirichlet process and the three choices of generalized Dirichlet process corresponding to basic samples given by* $K_{30} = 30$ [*case* (*i*)] *and* $K_{30} = 1$ [*case* (*ii*)]*, respectively*

	Case (i)		Case (ii)	
$m = n = 30$	$\hat{C}$	$\mathbb{E}[K_{30}^{(30)} \mid X_k^{(1,n)}]$	$\hat{C}$	$\mathbb{E}[K_{30}^{(30)} \mid X_k^{(1,n)}]$
Dir($a = 11.26$)	0.727	6.211	0.727	6.211
GD($a = 2.61, \gamma = 5$)	0.702	6.831	0.760	5.703
GD($a = 1.45, \gamma = 10$)	0.683	7.309	0.781	5.343
GD($a = 1.04, \gamma = 15$)	0.669	7.688	0.795	5.100

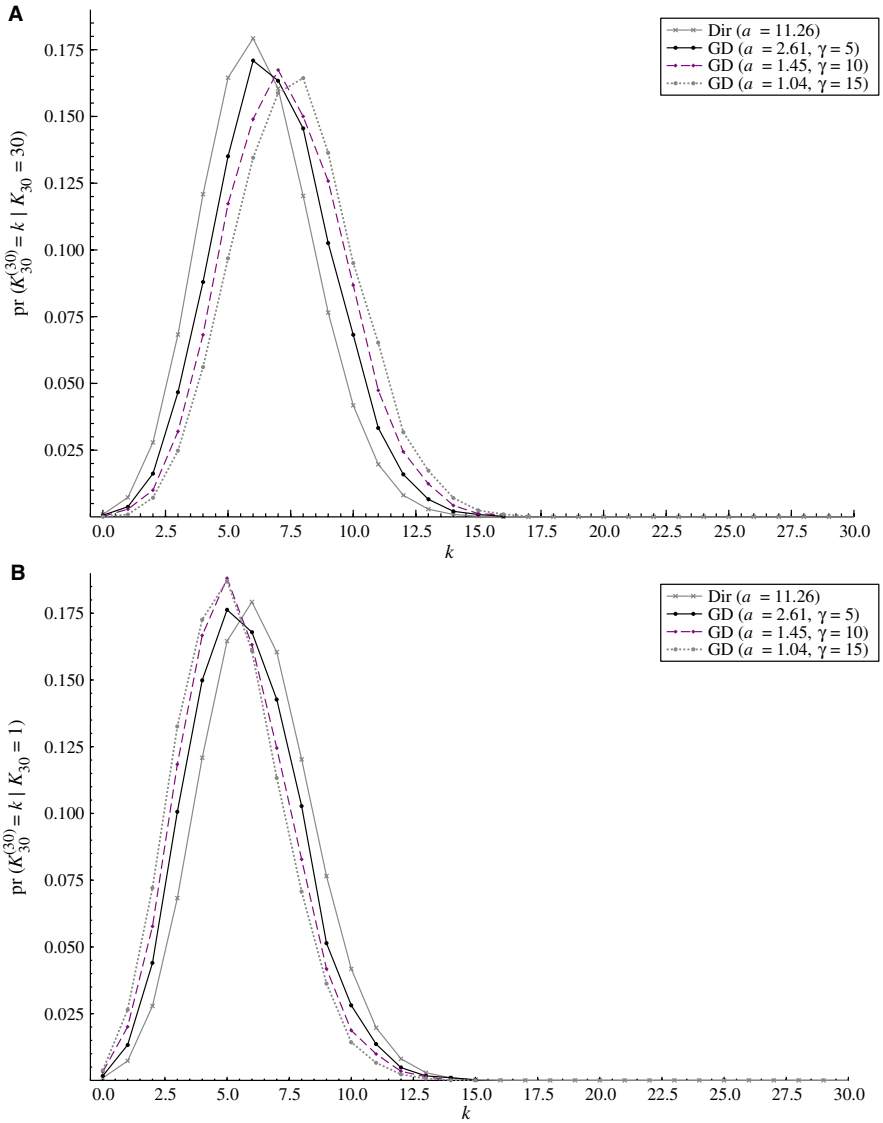

Fig. 3. Distributions of $K_{30}^{(30)}$ corresponding to the Dirichlet process and the three choices of generalized Dirichlet process conditional on $K_{30} = 30$ and $K_{30} = 1$, respectively.

sented in the observed sample, which in a Bayesian non-parametric framework coincides with $\hat{C} = \mathbb{P}(X_{n+1} = \text{already observed species} \mid X_1, \ldots, X_n)$. See Lijoi *et al.* (2007c) for a discussion. In the Dirichlet case such an estimate depends on the observed sample only through its size: consequently, the estimated coverage is 0.727 for both cases (i) and (ii). In contrast, for the generalized Dirichlet process, such an estimate heavily depends on the observed sample and is lower for case (i) and higher for case (ii). This is in agreement with intuition which would suggest that observation of many distinct species implies that there are many more unobserved ones. Table 1 displays the estimates corresponding to the different choices of parameters mentioned above and for both cases (i) and (ii).

Turning attention to prediction, and still considering the same parameter specifications and the two extreme cases (i) and (ii) outlined before, suppose one is interested in predicting the number of new distinct species to be observed in an additional sample of size $m = 30$. Clearly the Dirichlet process does not distinguish between the two cases (i) and (ii), whereas the generalized Dirichlet process does. Table 1 reports the corresponding estimates. Fig. 3 displays the posterior distributions of the number of new species given the observed samples (i) and (ii), i.e. $K_{30}^{(30)} \mid K_{30} = 1, N_1 = 30$ and $K_{30}^{(30)} \mid K_{30} = 30, N_1 = 1, \ldots, N_{30} = 1$. It is apparent how the generalized Dirichlet process nicely adapts to the information conveyed by the data by shifting either to the left or to the right of the distribution corresponding to the Dirichlet case.

Acknowledgements

The authors are grateful to an associate editor and two anonymous referees for their valuable comments and suggestions. Special thanks are also due to Ramsés H. Mena for some useful discussions. Stefano Favaro and Igor Prünster are partially supported by the Italian Ministry of University and Research, Grant 2008MK3AFZ.

Supporting Information

Additional Supporting Information may be found in the Online version of this article.

Data S1. Proofs of propositions 1–5.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.

Bell, E. T. (1927). Partition polynomials. *Ann. Math.* **29**, 38–46.

Bondesson, L. (1981). Classes of infinitely divisible distributions and densities. *Z. Wahrsch. verw. Geb.* **57**, 39–71.

Charalambides, C. A. (2005). *Combinatorial methods in discrete distributions*. Wiley, Hoboken, NJ.

Comtet, L. (1974). *Advanced combinatorics.* D, Reidel Publishing Company, Boston.

Dunson, D. B. (2010). Non-parametric Bayes applications to biostatistics. In *Bayesian nonparametrics* (eds N. L. Hjort, C. C. Holmes, P. Müller & S. G. Walker). pp. 223–268. Cambridge University Press, in press.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* **3**, 87–112.

Exton, H. (1976). *Multiple hypergeometric functions and application*, Ellis Horwood, Chichester.

Favaro, S., Prünster, I. & Walker, S. G. (2009). On a generalized Chu-Vandermonde identity. Technical Report, University of Turin.

Ferguson, T. S. (1973). A Bayesian analysis of some non-parametric problems. *Ann. Statist.* **1**, 209–230.
Ferguson, T. S. & Klass, M. J. (1972). A representation of independent increments processes without Gaussian components. *Ann. Math. Statist.* **43**, 1634–1643.
Gnedin, A. & Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S. Peterburg. Otdel. Mat. Inst. Steklov.* (*POMI*) **325**, 83–102.
Good, I. J. & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63.
Ho, M.-W., James, L. F. & Lau, J. W. (2007). Gibbs partitions (EPPF's) derived from a stable Subordinator are Fox H and Meijer G transforms, Preprint. *MatharXiv* arXiv:0708.0619v2.
Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96**, 161–173.
Ishwaran, H. & James, L. F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica* **13**, 1211–1235.
James, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. Manuscript. *MatharXiv* arXiv:math/0205093v1.
James, L. F., Roynette, B. & Yor, M. (2008). Generalized gamma convolutions, Dirichlet means, Thorin measures, with explicit examples. *Probab. Surv.* **5**, 346–415.
James, L. F., Lijoi, A. & Prünster, I. (2009). Posterior analysis for normalized random measure with independent increments. *Scand. J. Statist.* **36**, 76–97.
Kingman, J. F. C. (1975). Random discrete distributions. *J. Roy. Statist. Soc.* B **37**, 1–22.
Kingman, J. F. C. (1993). *Poisson processes*. Oxford University Press, Oxford.
Lijoi, A. & Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian nonparametrics* (eds N. L. Hjort, C. C. Holmes, P. Müller & S. G. Walker). pp 80–130. Cambridge University Press, in press.
Lijoi, A. & Regazzini, E. (2004). Means of a Dirichlet process and multiple hypergeometric functions. *Ann. Probab.* **32**, 1469–1495.
Lijoi, A., Mena, R. H. & Prünster, I. (2005). Bayesian nonparametric analysis for a generalized Dirichlet process prior. *Statist. Inf. Stoc. Proc.* **8**, 283–309.
Lijoi, A., Mena, R. H. & Prünster, I. (2007a). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. Roy. Statist. Soc.* B **69**, 715–740.
Lijoi, A., Mena, R. H. & Prünster, I. (2007b). Bayesian Nonparametric estimation of the probability of discovering a new species. *Biometrika* **94**, 769–786.
Lijoi A., Mena, R.H. & Prünster, I. (2007c). A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinform.* **8**, 339.
Lijoi A., Mena, R. H. & Prünster, I. (2008a). A Bayesian nonparametric approach for comparing clustering structures in EST libraries. *J. Comput. Biol.* **15**, 1315–1327.
Lijoi, A., Prünster, I. & Walker, S. G. (2008b). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18**, 1519–1547.
Lijoi, A., Prünster, I. & Walker, S. G. (2008c). Investigating nonparametric priors with Gibbs structure. *Statist. Sinica* **18**, 1653–1668.
Lo, A. I. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351–357.
Lo, A. I. & Weng, C. S. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Ann. Inst. Statist. Math.* **41**, 227–245.
Müller, P. & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19**, 95–110.
Müller, P. & Quintana, F. A. (2010a). More nonparametric Bayesian models for biostatistics. In *Bayesian nonparametrics* (eds N. L. Hjort, C. C. Holmes, P. Müller & S. G. Walker). pp 274–290. Cambridge University Press, in press.
Müller, P. & Quintana, F. A. (2010b). Random Partition models with regression on covariates. *J. Stat. Plann. Inf.*, doi:10.1016/j.jspi-2010.03.002.
Navarrete, C., Quintana, F. A. & Müller, P. (2008). Some issues on nonparametric Bayesian modeling using species sampling models. *Stat. Model.* **8**, 3–21.
Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Rel. Fields* **102**, 145–158.
Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, probability and game theory. Papers in honor of David Blackwell* (eds T. S. Ferguson, S. L. Shapley & J. B. MacQueen). *Lecture Notes Monograph* Series **30**, 245–267. IMS, Hayward.

Pitman, J. (2003). Poisson-Kingman partitions. In *Science and statistics: a Festschrift for Terry speed* (ed. D. R. Goldstein). *Lecture Notes Monograph Series*, **40**, 1–34. IMS, Beachwood, OH.
Pitman, J. (2006). *Combinatorial stochastic processes*. Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics No. 1875, Springer, New York.
Regazzini, E. (1998). An example of the interplay between statistics and special functions. In *Tricomi's idea and contemporary applied mathematics* Acc. Naz. Lincei e Acc. Scienze Torino, Roma, Atti dei Convegni Lincei, Vol. **147**, pp. 303–320.
Regazzini, E., Lijoi, A. & Prünster, I. (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.* **31**, 560–585.
von Renesse, M. K., Yor, M. & Zambotti, L. (2008). Quasi-invariance properties of a class of subordinators. *Stoch. Proc. Appl.* **118**, 2038–2057.
Thorin, O. (1977). On the infinite divisibility of the Pareto distribution, *Scand Actuar J.* **1**, 31–40.
Walker, S. G. & Damien, P. (2000). Representation of Lévy processes without Gaussian components. *Biometrika* **87**, 477–483.
Zabell, S. L. (1982). W. E. Johnson's "sufficientness" postulate. *Ann. Statist.* **10**, 1090–1099.

Received February 2009, in final form February 2010

Stephen G. Walker, Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Kent CT2 7NZ, UK.
E-mail: s.g.walker@kent.ac.uk

Appendix

Completely random measures

Denote by $(\mathbb{M}, \mathscr{M})$ the space of finite measures on $(\mathbb{X}, \mathscr{X})$ equipped with the corresponding Borel σ-algebra. Let $\tilde{\mu}$ be a random element defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\mathbb{M}, \mathscr{M})$ such that, for any $A_1, \ldots, A_n$ in $\mathscr{X}$, with $A_i \cap A_j = \emptyset$ for any $i \neq j$, the r.v.s $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are mutually independent. Then $\tilde{\mu}$ is a *completely random measure* (CRM). CRMs can always be represented as linear functionals of a Poisson random measure and, therefore, $\tilde{\mu}$ is characterized by the *Lévy–Khintchine* representation:

$$\mathbb{E}[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(dx)}] = \exp\left\{-\int_{\mathbb{R}^+ \times \mathbb{X}} [1 - e^{-sf(x)}] \nu(ds, dx)\right\},$$

where $f : \mathbb{X} \to \mathbb{R}$ is a measurable function such that $\int |f| \, d\tilde{\mu} < \infty$ a.s. and ν, which uniquely determines $\tilde{\mu}$, is the intensity measure of the underlying Poisson random measure. See Kingman (1993) for an exhaustive account on CRMs. If $\tilde{\mu}$ is defined on $\mathbb{X} = \mathbb{R}$, one can also consider the càdlàg random distribution function induced by $\tilde{\mu}$, namely $\{\tilde{\mu}((-\infty, x]) : x \in \mathbb{R}\}$. Such a random function defines an *increasing additive process*, that is a process with positive independent increments.

Bell polynomials

The partition polynomials, introduced by Bell (1927), are multivariable polynomials that are defined by a sum extended over all partitions of their index. Partition polynomials have found many applications in combinatorics, probability theory and statistics, as well as in number theory. A particular type of partition polynomials are the so-called Bell polynomials (see Comtet, 1974).

Definition 2
Let $w_\bullet := \{w_i, i \geq 1\}$ *be a sequence of real numbers. Then the* (n, k)*th partial Bell polynomial* $B_{n,k}(w_\bullet)$ *is defined by the expansion:*

$$\exp\{xw(t)\}=\sum_{n=0}^{+\infty}\sum_{k=0}^{+\infty}B_{n,k}(w_\bullet)x^k\frac{t^n}{n!},$$

where $w(t)$ *is the exponential generating function of the sequence* $w_\bullet$ *and* $w_0=w(0)=0$.

From definition 2 it is possible to isolate $B_{n,k}(w_\bullet)$ by differentiating the appropriate number of times and then setting $x=t=0$, i.e.

$$B_{n,k}(w_\bullet)=\frac{\partial^n}{\partial t^n}\frac{1}{k!}\frac{\partial^k}{\partial x^k}\exp\{xw(t)\}|_{x=0,t=0}$$

for all $n\geq 0$ and $k\geq 0$. This shows that $B_{n,k}(w_\bullet)$ corresponds to the nth Taylor coefficient of $(1/k!)w^k(t)$ or $w^k(t)/k!=\sum_{n=0}^{+\infty}B_{n,k}(w_\bullet)t^n/n!$. By setting $k=0$, one gets $B_{0,0}=1$ and $B_{n,0}=0$ for $n\geq 1$, whereas for $k=1$ one has $B_{n,1}=w_n$ for all $n\geq 0$. Also, since $w_0=0$, one has

$$\frac{1}{k!}w^k(t)=\frac{1}{k!}\left(w_1t+w_2\frac{t^2}{2!}+\cdots\right)^k=w_1^k\frac{t^k}{k!}+\cdots \tag{22}$$

so that $B_{n,k}(w_\bullet)=0$ whenever $k>n$ and $B_{n,n}(w_\bullet)=w_1^n$ for all $n\geq 0$. By expanding (22) and examining the coefficient of $t^n/n!$, one obtains the following explicit expression for $B_{n,k}(w_\bullet)$

$$B_{n,k}(w_\bullet)=\sum_{\substack{i_1,i_2,\ldots\geq 0\\ i_1+i_2+\cdots=k\\ i_1+2i_2+3i_3+\cdots=n}}\frac{n!}{i_1!i_2!\cdots(1!)^{i_1}(2!)^{i_2}\cdots}w_1^{i_1}w_2^{i_2}\cdots.$$

Note that, if the variable w_s occurs in $B_{n,k}(w_\bullet)$, then the summation conditions imply that, for some $i_1\geq 0, i_2\geq 0,\ldots, i_s\geq 1,\ldots$, we have $s-1\leq i_2+2i_3+\cdots+(s-1)i_s+\cdots=n-k$, giving $s\leq n-k$. Moreover, $B_{n,k}(w_\bullet)$ is homogeneous of degree k and it can be shown by a combinatorial argument that all of the coefficients are actually integers.

Definition 3
Let $w_\bullet:=\{w_i, i\geq 1\}$ *be a sequence of real numbers. Then the Bell polynomial* $B_n(x,w_\bullet)$ *is a polynomial in* x *defined by*

$$B_n(x,w_\bullet)=\sum_{k=0}^{n}x^kB_{n,k}(w_\bullet).$$

A combinatorial lemma

Here we recall a generalization of the well-known multivariate Chu–Vandermonde convolution formula (see Charalambides, 2005) derived in Favaro *et al.* (2009), which is of great help in proving propositions 3 and 4. Recall that $\mathcal{D}_{k,n}=\{(n_1,\ldots,n_k)\in\{1,\ldots,n\}^k:\sum_{i=1}^k n_i=n\}$ and define $\mathcal{D}_{k,n}^{(0)}=\{(n_1,\ldots,n_k)\in\{0,\ldots,n\}^k:\sum_{i=1}^k n_i=n\}$.

Lemma 1 (Favaro *et al.*, 2009)
For any $r\geq 1, k\geq 1$ *and* $a_i>0$, *with* $i=1,\ldots k$,

$$\sum_{(r_1,\ldots,r_k)\in\mathcal{D}_{k,r}^{(0)}}\binom{r}{r_1,\ldots,r_k}\prod_{i=1}^{k}w_i^{r_i}(a_i)_{(n_i+r_i-1)}$$
$$=w_k^r\left(n+\sum_{i=1}^{k}a_i-k\right)_r\prod_{i=1}^{k}(a_i)_{n_i-1}F_{\mathrm{D}}^{(k-1)}\left(-r,\mathbf{a};n+\sum_{i=1}^{k}a_i-k;\mathbf{W}\right),$$

where $(n_1,\ldots,n_k) \in \mathcal{D}_{k,n}, w_i \in \mathbb{R}^+$ *for* $i=1,\ldots,k, \mathbf{a}:=(n_1+a_1-1,\ldots,n_{k-1}-a_{k-1}-1)$ *and* $\mathbf{W}:=(w_k-w_1/w_k,\ldots,w_k-w_{k-1}/w_k)$.

Note that the usual Chu–Vandermonde formula is recovered by setting $w_i=1$ and $a_i=1$, for $i=1,\ldots,k$. Given the importance of this result to our purposes, we also provide a new proof in the Supporting Information, which is based on probabilistic arguments, instead of relying on the theory of special functions and inductive reasoning as in Favaro *et al.* (2009).