Submitted to Statistical Science

Sufficientness postulates for Gibbs-type priors and hierarchical generalizations

S. Bacallado, M. Battiston, S. Favaro* and L. Trippa†

University of Cambridge, University of Oxford, University of Torino and Dana-Farber Cancer Institute

Abstract. A fundamental problem in Bayesian nonparametrics consists of selecting a prior distribution by assuming that the corresponding predictive probabilities obey certain properties. An early discussion of such a problem, although in a parametric framework, dates back to the seminal work by English philosopher W. E. Johnson, who introduced a noteworthy characterization for the predictive probabilities of the symmetric Dirichlet prior distribution. This is typically referred to as Johnson's "sufficientness" postulate. In this paper we review some nonparametric generalizations of Johnson's postulate for a class of nonparametric priors known as species sampling models. In particular we revisit and discuss the "sufficientness" postulate for the two parameter Poisson-Dirichlet prior within the more general framework of Gibbs-type priors and their hierarchical generalizations.

Key words and phrases: Bayesian nonparametrics, Dirichlet and two parameter Poisson-Dirichlet process, discovery probability, Gibbs-type species sampling models, hierarchical species sampling models, Johnson's "sufficientness" postulate, Pólya-like urn scheme, predictive probabilities.

1. INTRODUCTION

At the heart of Bayesian nonparametric inference lies the fundamental concept of discrete random probability measure, whose distribution acts as a nonparametric prior, the most notable example being the Dirichlet process by Ferguson [25]. Species sampling models, first introduced by Pitman [52], form a very gen-

Department of Pure Mathematics and Mathematical Statistics, CB3 0WB Cambridge, UK (e-mail: sergiobacallado@gmail.com); Department of Statistics, OX1 3LB Oxford, UK (e-mail: marco.battiston@stats.ox.ac.uk); Department of Economics and Statistics, Corso Unione Sovietica 218/bis, 10134 Torino, IT (e-mail: stefano.favaro@unito.it); Department of Biostatistics and Computational Biology, 3 Blackfan Cir, Boston, MA 02215, USA (e-mail: ltrippa@jimmy.harvard.edu).

*Also affiliated to Collegio Carlo Alberto, Moncalieri, Italy.

†Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, USA

eral class of discrete random probability measures $P = \sum_{i\geq 1} p_i \delta_{X_i^*}$ defined by the sole requirements that $(p_i)_{i\geq 1}$ are nonnegative random weights such that $\sum_{i\geq 1} p_i = 1$ almost surely, and $(X_i^*)_{i\geq 1}$ are random locations independent of $(p_i)_{i\geq 1}$ and independent and identically distributed as a nonatomic base distribution ν_0. The term "species sampling" refers to the fact that the distribution $\mathscr{P}$ of P has a natural interpretation as a (prior) distribution for the unknown species composition $(p_i)_{i\geq 1}$ of a population of individuals $(X_i)_{i\geq 1}$ belonging to species X_i^*'s. As discussed in Pitman [52] and Lee et al. [40], the definition of species sampling models provides some insights on the structural sampling properties of these discrete random probability measures. However, for being usable as nonparametric priors, a distribution for the random probability $(p_i)_{i\geq 1}$ has to be specified. Among the various approaches for specifying such a distribution, the most common are the stick-breaking approach by Ishwaran and James [34] and the normalization approach by James [35], Pitman [53] and Regazzini et al. [60]. These approaches lead to popular species sampling models such as the Dirichlet process, the generalized Dirichlet process (Hjort [31] and Ishwaran and James [34]), the two parameter Poisson-Dirichlet process (Perman et al. [50] and Pitman and Yor [55]) and the normalized generalized Gamma process (James [35], Prünster [56] and Pitman [53]) to name a few. The reader is referred to Lijoi and Prünster [44] for a comprehensive and stimulating account of species sampling models, as well as generalizations thereof, with applications to Bayesian nonparametrics.

A common building block in Bayesian nonparametrics, either at the level of observed data or at the latent level of hierarchical models, consists of a sample from a species sampling model P with distribution $\mathscr{P}$. According to de Finetti's representation theorem, such a sample is part of an exchangeable sequence $(X_i)_{i\geq 1}$ with directing (de Finetti) measure $\mathscr{P}$, i.e. $\lim_{n\to+\infty} n^{-1} \sum_{1\leq i\leq n} \delta_{X_i} = P$ almost surely. In particular, due to the discreteness of species sampling models, a sample of size n from P features $K_n = k \leq n$ distinct species, labelled by $X_1^*, \ldots, X_{K_n}^*$, with corresponding frequencies $\mathbf{N}_n = (N_{1,n}, \ldots, N_{K_n,n}) = \mathbf{n} = (n_1, \ldots, n_k)$ such that $\sum_{1\leq i\leq K_n} N_{i,n} = n$. More formally, if $(X_1, \ldots, X_n)$ is a random sample from P, namely

$$\begin{aligned} X_i \,|\, P &\overset{\text{iid}}{\sim} P \qquad i = 1, \ldots, n, \\ P &\sim \mathscr{P}, \end{aligned} \tag{1}$$

then the sample induces a random partition Π_n of $\{1, \ldots, n\}$ whose blocks corresponds to the equivalence classes for the random equivalence relations $i \sim j \iff X_i = X_j$ almost surely. The random partition Π_n is exchangeable, namely the distribution of Π_n is a symmetric function of the frequencies $\mathbf{n}$. This function, denoted $p_{n,k}(\mathbf{n})$, is known as the exchangeable partition probability function (EPPF), a concept introduced in Pitman [51] as a development of earlier results in Kingman [38].

The notion of exchangeable random partition of $\{1, \ldots, n\}$ can be extended to the natural numbers $\mathbb{N}$. In particular, the infinite exchangeable sequence $(X_i)_{i\geq 1}$ induces an exchangeable random partition Π of $\mathbb{N}$, where exchangeable means that the distribution of Π is invariant under finite permutations of its elements. This partition can be described by the sequence $(\Pi_n)_{n\geq 1}$ of its restrictions to the first

n integer numbers, i.e., Π_n is obtained from Π by discarding all elements greater than n. Conversely, a sequence of random exchangeable partitions $(\Pi_n)_{n\geq 1}$ defines an exchangeable random partition of $\mathbb{N}$ provided that this sequence is consistent, i.e., Π_m is the restriction of Π_n to the first m elements, for all $m < n$. Consistency implies that

$$p_{n,k}(\mathbf{n}) = p_{n+1,k+1}(\mathbf{n}, 1) + \sum_{i=1}^{k} p_{n+1,k}(n_1, \mathbf{n} + \mathbf{e}_i) \tag{2}$$

for all $n \geq 1$, where $\mathbf{e}_i$ denotes a k-dimensional vector with all entries equal to zero but the i-th entry equal to 1. As a direct consequence of Kingman's theory of exchangeable random partitions of $\mathbb{N}$, the predictive probabilities of $(X_i)_{i\geq 1}$ are

$$\Pr[X_{n+1} \in \cdot \,|\, X_1, \ldots, X_n] = g(n, k, \mathbf{n})\nu_0(\cdot) + \sum_{i=1}^{k} f_i(n, k, \mathbf{n})\delta_{X_i^*}(\cdot), \tag{3}$$

for any $n \geq 1$, where ν_0 is a nonatomic distribution on the sample space and where $g(n, k, \mathbf{n}) := p_{n+1,k+1}(\mathbf{n}, 1)/p_{n,k}(\mathbf{n})$ and $f_i(n, k, \mathbf{n}) := p_{n+1,k}(n_1, \mathbf{n} + \mathbf{e}_i)/p_{n,k}(\mathbf{n})$ are nonnegative functions of $(n, k, \mathbf{n})$, respectively describing the probability that the X_{n+1} will be a new value and the probability that it will be equal to X_i^*. From (2), it follows that g and f_i must satisfy the following constraint: $g(n, k, \mathbf{n}) + \sum_{1\leq i\leq k} f_i(n, k, \mathbf{n}) = 1$. The functions g and f_i completely determine the distribution of $(X_i)_{i\geq 1}$ and, in turn, the distribution of Π. See Pitman [52] for a detailed account of exchangeable random partitions and species sampling models.

Within the Bayesian nonparametric framework (1), how to select the prior distribution $\mathscr{P}$ is an important issue. Of course one approach is to select $\mathscr{P}$ by appealing to prior information about P, and then attempt to incorporate this information into $\mathscr{P}$. This is often a difficult task for nonparametric priors, since P is an infinite dimensional object. Alternatively, one may select $\mathscr{P}$ by assuming that the predictive probabilities (3) obey or exhibit some characteristic or property. Indeed in practical applications it may be that the form of the functions g and f_i may be an adequate description of our current state of knowledge. An early discussion of this alternative approach, although in a parametric framework, dates back to the seminal work by English philosopher W. E. Johnson. Specifically, assuming $T < +\infty$ possible species that are known and equiprobable prior to observations, Johnson [37] characterized the T-dimensional symmetric Dirichlet distribution as the unique prior for which g depends only on n, k and T, and f_i depends only on n, n_i and T. As a direct consequence of the parametric assumption that $T < +\infty$, of course, $g = 0$ for all $k \geq T$. Using the terminology in Good [28], this characterization of the Dirichlet prior is referred to as Johnson's "sufficientness" postulate. We refer to the work of Zabell [71] and Zabell [73] for a review of Johnson's postulate. See also the monograph by Zabell [75] for a more comprehensive account of sufficientness, exchangeability and predictive probabilities.

In this paper we discuss and derive some generalizations of Johnson's postulate that arise by removing the parametric assumption of a prespecified number $T < +\infty$ of possible species in the population. We focus on species sampling

models that allow either for an infinite number of species or for a finite random number T of species, with T having unbounded support over $\mathbb{N}$. Regazzini [58], and later on Lo [48], provided a nonparametric counterpart of Johnson's postulate. Specifically, under the assumption of an infinite number of species in the population, they showed that the Dirichlet process is the unique species sampling model for which the function g depends only on n, and the function f_i depends only on n and n_i. A noteworthy extension of this nonparametric sufficientness postulate was presented in Zabell [74], and it characterizes the two parameter Poisson-Dirichlet process of Pitman [51] as the unique species sampling model for which g depends only on n and k, and f_i depends only on n and n_i. Here we revisit the seminal work of Zabell [74] within the more general framework of the Gibbs-type species sampling models introduced by Gnedin and Pitman [27], and nowadays widely used in Bayesian nonparametrics. Gibbs-type species sampling models, which include the Dirichlet process and two parameter Poisson-Dirichlet process as special cases, suggest for the formulation of a novel nonparametric sufficientness postulate in which the function g depends only on n and k, and the function f_i depends only on n, k and n_i. We present such a postulate and, in light of that, we show how the sufficientness postulates of Regazzini [58] and Zabell [74] may be rephrased in terms of an intuitive Pólya-like urn scheme for Gibbs-type species sampling models. Table 1 provides with a schematic summary of sufficientness postulates for species sampling models. Our study is completed with a discussion on the problem of formulating analogous nonparametric sufficientness postulates in the context of the hierarchical species sampling models introduced by Teh et al. [67].

TABLE 1
Sufficientness postulates for species sampling models (SSM): T-dimensional symmetric Dirichlet distribution (T-SD), Dirichlet process (DP), two parameter Poisson-Dirichlet process (2PD) and Gibbs-type SSM.

SSM	NUMBER T OF SPECIES	$g(n,k,\mathbf{n})$	$f_i(n,k,\mathbf{n})$
T-SD	Known $T < +\infty$	$g(n,k,T)$	$f(n,n_i)$
DP	$T = +\infty$	$g(n)$	$f(n,n_i)$
2PD	$T = +\infty$	$g(n,k)$	$f(n,n_i)$
Gibbs-type SSM	$T = +\infty$	$g(n,k)$	$f(n,k,n_i)$

The paper is structured as follows. Section 2 contains a brief review on the sampling properties of the class of Gibbs-type species sampling models. In Section 3 we review the sufficientness postulate of Zabell [74], we present its generalization within the more general framework of Gibbs-type species sampling models, and we introduce a Pólya-like urn scheme for describing the predictive probabilities of Gibbs-type species sampling models. In Section 4 we discuss how Johnson's sufficientness postulate can be extended to the framework of hierarchical species sampling models. Section 5 contains a discussion of the proposed characterizations and open questions. Proofs of our results are provided as online supplementary material.

2. A BRIEF REVIEW OF GIBBS-TYPE PRIORS

As recently discussed in De Blasi et al. [17], Gibbs-type species sampling models, or Gibbs-type priors, may be considered as the most "natural" generalization of the Dirichlet process. Indeed, apart of the well-known conjugacy of the Dirichlet process, Gibbs-type species sampling models share numerous properties that are appealing from both a theoretical and an applied point of view: i) they admit a simple and intuitive definition in terms of predictive probabilities, which is a generalization of the Blackwell and MacQueen [9] urn scheme; ii) they stand out in terms of mathematical tractability, which allows to study their distributional properties for finite sample sizes and asymptotically; iii) they admit a stick-breaking representation and a representation as normalized random measures, thus taking the advantages of both representations; iv) they are characterized by a flexible parameterization, thus including numerous interesting special cases, most of them still unexplored. All these properties have made the class of Gibbs-type priors a common choice in several contexts, such as in hierarchical mixture modeling, species sampling problems, feature and graph modeling, hidden Markov modeling, etc. In this section we briefly review Gibbs-type species sampling models, with emphasis towards their predictive probabilities and sampling properties. The reader is referred to the monographs by Pitman [54] and Bertoin [8] for a comprehensive account of Gibbs-type species sampling models, and to Lijoi and Prünster [44] and De Blasi et al. [17] for reviews on their use in Bayesian nonparametrics.

Among various possible definitions of Gibbs-type species sampling models, the most intuitive is given in terms of their predictive probabilities. See, e.g., Pitman [53] and Gnedin and Pitman [27]. These predictive probabilities are of the general form (3), for a suitable specification of the nonnegative functions g and f_i. In particular let $(X_1, \ldots, X_n)$ be a sample from an arbitrary species sampling model P, and assume that $(X_1, \ldots, X_n)$ features $K_n = k \leq n$ species, labelled by $X_1^*, \ldots, X_{K_n}^*$, with corresponding frequencies $\mathbf{N}_n = \mathbf{n}$. For $\alpha < 1$ and for ν_0 a nonatomic probability measure, P is a Gibbs-type species sampling model if

$$\Pr[X_1 \in \cdot] = \nu_0(\cdot) \tag{4}$$

and

$$\Pr[X_{n+1} \in \cdot \,|\, X_1, \ldots, X_n] = \frac{V_{n+1,k+1}}{V_{n,k}} \nu_0(\cdot) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{i=1}^{k} (n_i - \alpha) \delta_{X_i^*}(\cdot) \tag{5}$$

for any $n \geq 1$, where $(V_{n,k})_{1 \leq k \leq n, n \geq 1}$ are nonnegative weights satisfying the triangular recursion $V_{n,k} = V_{n+1,k}(n - \alpha k) + V_{n+1,k+1}$ with the proviso $V_{1,1} := 1$. By combining the predictive probabilities (5) with the nonparametric sufficientness postulate in Regazzini [58] and Lo [48], it follows that Gibbs-type species sampling models generalize the Dirichlet process by introducing the dependency on k in both the functions g and f_i. See also Zabell [71] and references therein for details.

Gnedin and Pitman [27] characterized the de Finetti measure of an exchangeable sequence $(X_i)_{i \geq 1}$ distributed as (4) and (5). Such a characterization relies on the notion of Poisson-Kingman model introduced by Pitman [53]. Specifically,

for any $\alpha \in (0,1)$ let $(J_i)_{i\geq 1}$ be decreasing ordered jumps of an α-stable subordinator, namely a subordinator with Lévy measure $\rho(\mathrm{d}x) = C_\alpha x^{-\alpha-1}\mathrm{d}x$ for some constant C_α. See Sato [63] and references therein for details. Furthermore, let $P_i = J_i/T_\alpha$ where $T_\alpha = \sum_{i\geq 1} J_i < +\infty$ almost surely, and let $\mathrm{PK}(\alpha; t)$ denote the conditional distribution of $(P_i)_{i\geq 1}$ given $T_\alpha = t$. In particular T_α is a positive α-stable random variable, and we denote by f_α its density function. If we denote by $T_{\alpha,h}$ a random variable with density function $f_{T_{\alpha,h}}(t) = h(t)f_\alpha(t)$, for any nonnegative function h, then an α-stable Poisson-Kingman model is defined as the discrete random probability measure $P_{\alpha,h} = \sum_{i\geq 1} P_{i,h}\delta_{X_i^*}$, where $(P_{i,h})_{i\geq 1}$ is distributed as $\int_{(0,+\infty)} \mathrm{PK}(\alpha; t) f_{T_{\alpha,h}}(t)\mathrm{d}t$ and $(X_i^*)_{i\geq 1}$ are random variables, independent of $(P_{i,h})_{i\geq 1}$, and independent and identically distributed as ν_0. According to Gnedin and Pitman [27], if $(X_i)_{i\geq 1}$ is an exchangeable sequence distributed as (4) and (5) then the de Finetti measure of $(X_i)_{i\geq 1}$ is the law of: i) an α-stable Poisson-Kingman model, for $\alpha \in (0,1)$; ii) the Dirichlet process, for $\alpha = 0$; iii) an M-dimensional symmetric Dirichlet distribution, with M being a nonnegative discrete random variable on $\mathbb{N}$, for $\alpha < 0$. In other terms $(X_i)_{i\geq 1}$ distributed as (4) and (5) admits a finite number M of species for $\alpha < 0$, and an infinite number of species for $\alpha \in [0,1)$.

The characterization of Gnedin and Pitman [27] leads to identify explicit expressions for the $V_{n,k}$'s in (5). In particular, for the class of α-stable Poisson-Kingman models an expression for $V_{n,k}$ was provided in Pitman [53], and further investigated by Ho et al. [32]. See also James [36] and references therein. Let $\Gamma(\cdot)$ denote the Gamma function. For any $\alpha \in (0,1)$ and $c > 0$ let $S_{\alpha,c}$ be a polynomially tilted α-stable random variable, i.e. $f_{S_{\alpha,c}}(s) = \Gamma(c\alpha+1)s^{-\alpha c} f_\alpha(s)/\Gamma(c+1)$, and let $B_{a,b}$ be a Beta random variable with parameter (a,b) independent of $S_{\alpha,c}$. Then,

$$V_{n,k} = \frac{\alpha^k \Gamma(k)}{\Gamma(n)} \mathbb{E}\left[h\left(\frac{S_{\alpha,k}}{B_{\alpha k, n-\alpha k}}\right)\right]. \tag{6}$$

We refer to Chapter 4 of Pitman [54] for additional details on (6). For the Dirichlet process, the expression of $V_{n,k}$ is well-known from the seminal work of Ewens [20], i.e.

$$V_{n,k} = \frac{\theta^k}{(\theta)_n} \tag{7}$$

for any $\theta > 0$. See also Antoniak [2] for an alternative derivation of (7) in terms of the urn scheme description of the Dirichlet process in Blackwell and MacQueen [9]. For the M-dimensional symmetric Dirichlet distributions, for any $\alpha < 0$ one has

$$V_{n,k} = \frac{\prod_{i=0}^{k-1}(M|\alpha| + i\alpha)}{(M|\alpha|)_n}. \tag{8}$$

Conditionally to $M = m$, the expression (8) dates back to the seminal work of Fisher et al. [26]. In particular they derived (8) and they also considered the passage to the limit as $m \to +\infty$ and $-\alpha \to 0$ for fixed $\theta = m\alpha > 0$, which leads to the weight in (7). See also Johnson [37], Watterson [70] and Engen [19] for a detailed account of the M-dimensional symmetric Dirichlet species sampling model.

Among Gibbs-type species sampling models with $\alpha \in (0,1)$, the two parameter Poisson-Dirichlet process certainly stands out. See, e.g., Perman et al. [50], Pitman [51], Pitman and Yor [55] and Pitman [53]. Another noteworthy example is the normalized generalized Gamma process, introduced in Pitman [53] and further investigated in Bayesian nonparametrics, e.g., James [35], Lijoi et al. [43], Lijoi et al. [45] and James [36]. For $\alpha \in (0,1)$ and $\theta > -\alpha$, the two parameter Poisson-Dirichlet process is a Gibbs-type species sampling model with $h(t) = \alpha\Gamma(\theta)t^{-\theta}/\Gamma(\theta/\alpha)$. In particular, by replacing this function in (6), one obtains

$$V_{n,k} = \frac{\prod_{i=0}^{k-1}(\theta + i\alpha)}{(\theta)_n}, \tag{9}$$

where $(\theta)_n$ is the ascending factorial, i.e., $(\theta)_n := \prod_{0\leq i\leq n-1}(\theta+i)$ with the proviso $(\theta)_0 = 1$. For $\alpha \in (0,1)$ and $\tau \geq 0$ the normalized generalized Gamma process is a Gibbs-type species sampling model with $h(t) = \exp\{\tau - \tau^{1/\alpha}t\}$. By replacing this function in (6),

$$V_{n,k} = \frac{\alpha^k \mathrm{e}^{\tau}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k - \frac{i}{\alpha}, \tau\right), \tag{10}$$

where $\Gamma(\cdot,\cdot)$ is the incomplete Gamma function. Note that (9) may be viewed as a suitable mixture of (10). That is, if $G_{\theta/\alpha,1}$ is a Gamma random variable with parameter $(\theta/\alpha, 1)$ then (9) can be written as (10) where τ is replaced by $G_{\theta/\alpha,1}$. In general, for any $\theta > 0$ the two parameter Poisson-Dirichlet process may be viewed as hierarchical generalization of the normalized generalized Gamma process, with a Gamma prior over τ. See Section 5 in Pitman and Yor [55] for details.

The predictive probabilities (5) lead to the distribution of the exchangeable random partition Π_n induced by a sample $(X_1, \ldots, X_n)$ from a Gibbs-type species sampling model. In particular, let $p_k^{(n)}(\mathbf{n})$ denote the EPPF of Π_n, that is probability of any particular partition of the set $\{1, \ldots, n\}$ induced by $(X_1, \ldots, X_n)$ and featuring $K_n = k$ distinct blocks with frequencies $\mathbf{N}_n = \mathbf{n}$, for any $n \geq 1$. Then, by a direct application of the predictive probabilities (5), one may easily verify that

$$p_k^{(n)}(\mathbf{n}) = V_{n,k} \prod_{i=1}^{k} (1-\alpha)_{(n_i-1)}. \tag{11}$$

Moreover, by marginalizing $\Pr[K_n = k, \mathbf{N}_n = \mathbf{n}] = (k!)^{-1}\binom{n}{n_1,\ldots,n_k} p_k^{(n)}(\mathbf{n})$ with respect to the frequencies $\mathbf{n}$, one obtains the distribution of K_n. In particular, one has

$$\Pr[K_n = k] = V_{n,k} \frac{\mathscr{C}(n,k;\alpha)}{\alpha^k}, \tag{12}$$

where $\mathscr{C}(n,k;\alpha)$ is the generalized factorial coefficient, namely $\mathscr{C}(n,k;a) := (k!)^{-1}\sum_{1\leq i\leq k}(-1)^i\binom{k}{i}(-ai)_n$. As discussed in Gnedin and Pitman [54] and De Blasi et al. [17], the mathematical tractability of Gibbs-type species sampling models originates from the product form of the EPPF (11). Such a product form

is closely related to the notion of product partition model in Quintana and Iglesias [57].

The role of the parameter α in the distribution (11) is easily interpreted. In particular, a first interpretation of α follows from the predictive probabilities (5). Indeed, $\alpha > 0$ acts an interesting reinforcement mechanism in the empirical part of the predictive probability (5). Note that the probability that X_{n+1} coincides with the species X_i^*, for any $i = 1, \ldots, k$, is a function of the frequency n_i and α. In particular, the ratio of the probabilities assigned to any pair of species (X_i^*, X_j^*) is

$$\frac{n_i - \alpha}{n_j - \alpha} \tag{13}$$

If $\alpha \to 0$ the ration (13) reduces to the ratio of the frequencies of the two species, and therefore the coincidence probability is proportional to the frequency of the species. On the other hand if $\alpha > 0$ and $n_i > n_j$ then the ratio is an increasing function of α. Accordingly, as α increases the mass is reallocated from the species X_j^* to the species X_i^*. In other terms the sampling procedure tends to reinforce, among the observed species, those having higher frequencies. See De Blasi et al. [17] and references therein for a detailed discussion on such a reinforcement mechanism. If $\alpha < 0$, the reinforcement mechanism works in the opposite way in the sense that the coincidence probabilities are less than proportional to the species frequencies.

A further interpretation of the parameter α arises from the large n asymptotic behavior of the random variable K_n with distribution (12). This behaviour was first investigated by Korwar and Hollander [39] for the Dirichlet process, and then extended by Pitman [53] to the general framework of Gibbs-type species sampling model. See also Gnedin and Pitman [27] and Pitman [54] for details. The parameter α determines the rate at which K_n increases, as the sample size n increases. Three different rates may be identified for Gibbs-type species sampling models. Let

$$c_n(\alpha) := \begin{cases} n^\alpha & \text{if } \alpha \in (0,1) \\ \log(n) & \text{if } \alpha = 0 \\ 1 & \text{if } \alpha \in (-\infty, 0), \end{cases}$$

for any $n \geq 1$. Then there exists a random variable S_α, positive and finite almost surely, such that

$$\frac{K_n}{c_n(\alpha)} \to S_\alpha \tag{14}$$

almost surely, as $n \to +\infty$. Using the terminology in Pitman [53], S_α is referred to as the α-diversity of the the Gibbs-type species sampling model. More precisely: i) for $\alpha \in (0,1)$ the α-diversity coincides, in distribution, with $T_{\alpha,h}^{-\alpha}$; ii) for $\alpha = 0$ the α-diversity is a random variable whose distribution degenerates at $\theta > 0$; iii) for $\alpha < 0$ the α-diversity coincides, in distribution, with the random number M of species in the population. The larger α, the faster the rate of increase of K_n or, in other terms, the more new species are generated from the sampling mechanism described in (5).

Gibbs-type species sampling models have been extensively used in the context of Bayesian nonparametric inference for species sampling problems. See, e.g., Lijoi et al. [42], Lijoi et al. [45], Favaro et al. [21], Favaro et al. [22], Bacallado et al. [5, 6] and Arbel et al. [3]. Species sampling problems are arguably the field in which the mathematical tractably of Gibbs-type species sampling models can be most appreciated. In the last few years a plethora of posterior properties of Gibbs-type priors, for finite sample sizes and asymptotically, have been derived and applied for estimating population's features and predicting features of additional unobservable samples. Gibbs-type species sampling models have been also applied in the context of mixture modeling, thus generalizing the seminal work by Lo [47]. See, e.g., Ishwaran and James [34], Lijoi et al. [41], Lijoi et al. [42], Favaro and Walker [23] and Lomeli et al. [49]. While maintaining the same computational tractability of the Dirichlet process mixture model, the availability of the additional parameter α allows for a better control of the clustering behaviour. Most recently, Gibbs-type species sampling models have been proposed for Bayesian nonparametric inference for ranked data in Caron et al. [12], sparse exchangeable random graphs and networks in Caron and Fox [12] and Herlau [30], feature allocations in Teh and Görür [66], Broderick et al. [10], Heaukulani and Roy [29], Roy [62] and Battiston et al. [7], reversible Markov chains in Bacallado et al. [4], dynamic textual data in Chen et al. [14] and Chen et al. [15], and bipartite graphs in Caron [11].

3. SUFFICIENTNESS POSTULATES AND URN SCHEMES FOR GIBBS-TYPE PRIORS

A noteworthy generalization of Johnson's sufficientness postulate was first discussed in the work of Zabell [74]. Specifically, let P be an arbitrary species sampling model with predictive probabilities (3), and consider the following assumptions: A1) $\Pr[\Pi_n = \pi_n] > 0$ for all the partitions π_n of $\{1, \ldots, n\}$, that is no scenario is deemed, a priori, to be impossible; A2) $g(n, k, \mathbf{n}) = g(n, k)$, that is the probability of observing a new species depends only on n and k; A3) $f_i(n, k, \mathbf{n}) = f(n, n_i)$, that is the probability of observing the species X_i^* depends only on n and n_i. Zabell [74] showed that if just these three assumptions are imposed, then there exist three parameters $\alpha \in (0, 1)$, $\theta > -\alpha$ and $c_n \geq 0$ such that

i) if $k \geq 2$ then

$$g(n, k) = \frac{\theta + k\alpha}{\theta + n}; \qquad f(n, n_i) = \frac{n_i - \alpha}{\theta + n}; \tag{15}$$

ii) if $k = 1$ then

$$g(n, k) = \frac{\theta + \alpha}{\theta + n} - c_n; \qquad f(n, n) = \frac{n - \alpha}{\theta + n} + c_n. \tag{16}$$

In other words, if a species sampling model satisfies the assumptions A1)-A3), then the functions g and f_i in the predictive probabilities (3) must have the expressions (15) and (16). Zabell's sufficientness postulate may be viewed as a nonparametric counterpart of the classical Johnson's postulate, in the sense that it allows to remove the assumption of a prespecified number $T < +\infty$ of possible

species in the population. See Zabell [71], Zabell [73] and references therein for details.

As discussed in Zabell [74], the parameters $(c_n)_{n\geq 1}$ represent adjustments of the predictive probabilities that arise when only one species is observed in an exchangeable sequence $(X_i)_{i\geq 1}$ of trials. That is a partition consisting of a single block is observed. Accordingly one may set $c_n = 0$, for any $n \geq 1$, by imposing the following additional assumption: A4) $\Pr[K_n > 1] = 1$ almost surely for any $n \geq 1$. In particular, let $(X_1, \ldots, X_n)$ be a sample of size n from an arbitrary species sampling model P, such that $(X_1, \ldots, X_n)$ features $K_n = k \leq n$ species $X_1^*, \ldots, X_{K_n}^*$ with corresponding frequencies $\mathbf{N}_n = \mathbf{n}$. Then under A1)-A4) one has

$$\Pr[X_{n+1} \in \cdot \,|\, X_1, \ldots, X_n] = \frac{\theta + k\alpha}{\theta + n}\nu_0(\cdot) + \frac{1}{\theta + n}\sum_{i=1}^{k}(n_i - \alpha)\delta_{X_i^*}(\cdot), \tag{17}$$

for any $n \geq 1$, which are precisely the predictive probabilities of the two parameter Poisson-Dirichlet process. An intuitive description of (17) was proposed by Zabell [74] in terms of the following Pólya-like urn scheme. Consider an urn containing both colored and black balls, where colored balls may be interpreted as the individuals with their associated species (color). Balls are drawn and then replaced, in such a way that the probability of a particular ball being drawn at any stage is proportional to its selection weight. Initially the urn contains a black ball with weight $\theta > 0$, and at the n-th draw: i) if we pick a colored ball then it is returned to the urn with ball of the same color with weight 1; ii) if we pick a black ball, then it is returned to the urn with a black ball of weight $\alpha \in (0, 1)$ and a ball of a new color with weight $1 - \alpha$. If X_n is the color of the ball returned in the urn after the n-th draw, and such a color is generated according to the nonatomic distribution ν_0, then it can be easily verified that the predictive probabilities $\Pr[X_{n+1} \in \cdot \,|\, X_1, \ldots, X_n]$ coincides with (17), for any $\alpha \in (0, 1)$ and $\theta > 0$.

According to Zabell's sufficientness postulate, the two parameter Poisson-Dirichlet process is the unique species sampling model for which the function g depends only on n and k, and the function f_i depends only on n and n_i, for any $i = 1, \ldots, k$. As a limiting special case of Zabell's characterization, for $\alpha \to 0$ the Dirichlet process is the unique species sampling model for which the function g depends only on n, and the function f_i depends only on n and n_i, for any $i = 1, \ldots, k$. The predictive probabilities (5) of a Gibbs-type species sampling model generalize those of the two parameter Poisson-Dirichlet process by introducing the dependency on k in the function f. In particular, within the class of Gibbs-type species sampling model one may rephrase Zabell's sufficientness postulated as follows: for any index $\alpha \in (0, 1)$ the two parameter Poisson-Dirichlet process is the unique Gibbs-type species sampling model for which the ratio $V_{n+1,k}/V_{n,k}$ in (5) simplifies in such a way to remove the dependency on the number k of observed species. The normalized generalized Gamma process, whose predictive probabilities are expressed in terms of the $V_{n,k}$'s in (10), is a representative example of a Gibbs-type species sampling model for which such a simplification does not occur. See, e.g., Lijoi et al. [43] and Lijoi et al. [45] for details. The predictive probabilities of Gibbs-type species sampling models thus suggest for a generalization of the Zabell's sufficientness postulate, where the as-

sumption A3) is replaced by the assumption $f_i(n,k,\mathbf{n}) = f(n,k,n_i)$, that is the probability of observing the species X_i^* depends only on n, k and n_i, for any $i = 1,\ldots,k$. The following generalization of the Zabell's sufficientness postulate can be proved.

PROPOSITION 1. *Let P be a pecies sampling model with predictive probabilities of the form (3), and allowing either for an infinite number of species or for a finite random number T of species, with T being supported on $\mathbb{N}$. Furthermore, assume that*

A1) $Pr[\Pi_n = \pi_n] > 0$ *for all the partitions* π_n *of the set* $\{1,\ldots,n\}$;
A2) $g(n,k,\mathbf{n}) = g(n,k)$;
A3) $f_i(n,k,\mathbf{n}) = f(n,k,n_i)$ *for any* $i = 1,\ldots,k$.

Under A1)-A3) there exists a parameter $\alpha < 1$ and a collection of nonnegative weights $(V_{n,k})_{1\leq k\leq n,n\geq 1}$ with $V_{1,1} = 1$ and satisfying $V_{n,k} = V_{n+1,k}(n-\alpha k) + V_{n+1,k+1}$ such that

$$g(n,k) = \frac{V_{n+1,k+1}}{V_{n,k}}; \qquad f(n,k,n_i) = \frac{V_{n+1,k}}{V_{n,k}}(n_i - \alpha)$$

for any $i = 1,\ldots,k$. In other terms, if a species sampling model P satisfies the assumptions A1)-A3) then P is a Gibbs-type species sampling model with parameter $\alpha < 1$.

As we pointed out in the Introduction, Zabell's sufficientness postulate generalizes the original framework of Regazzini [58] and Lo [48] by introducing the dependency on k in the function g. Proposition 1 provides an even more general framework by introducing the dependency on k in both the function g and the function f_i, for any $i = 1,\ldots,k$, while maintaining the same structure with respect to the dependency on the frequencies counts $\mathbf{n}$. The proof of Proposition 1 is rather long and technical, although along lines similar to the proof of Zabell's sufficientness postulate. In particular it consists of verifying the following main steps:

i) showing that the function $f(n,k,n_i)$ is a linear with respect to n_i, for any $n \geq 1$, $1 \leq k \leq n$, i.e., there exist parameters $a_{n,k}$ and $b_{n,k}$ such that $f(n,k,m) = a_{n,k} + b_{n,k}m$;
ii) showing that the parameter $b_{n,k}$ is different from zero, for any $n \geq 1$ and $1 \leq k \leq n$; this allows us to introduce an additional parameter $\alpha_{n,k} = -a_{n,k}/b_{n,k}$, which we show to be independent of n and k and to be strictly less than 1;
iii) introducing the new parametrization $V_{n,k}$, showing that it satisfies the recursion specific of the Gibbs-type prior and finally recovering the f_i and g of a generic Gibbs-type prior.

See Section 1 of the supplementary material for the proof of Proposition 1. Note that Proposition 1 does not characterize the entire class of Gibbs-type species sampling models. Indeed we confined ourself to species sampling models allowing either for an infinite number of species or for a finite random number T of species, with T being supported on $\mathbb{N}$. According to the characterization of Gnedin and Pitman [27], this restriction excludes Gibbs-type species sampling models with

$\alpha < 0$ and with M being a distribution with finite support. It remains an open problem to check whether it is possible to characterize the entire class of Gibbs-type priors by relaxing A1).

One can derive an intuitive urn scheme that describes the predictive probabilities for the class of Gibbs-type species sampling models. Let $(V_{n,k})_{1\leq k\leq n, n\geq 1}$ be a collection of nonnegative weights such that $V_{1,1} = 1$ and $V_{n,k} = V_{n+1,k}(n-\alpha k) + V_{n+1,k+1}$. Consider an urn containing both colored and black balls, where colored balls may be interpreted as the individuals with their associated species (color). The urn initially contains only a black ball with an arbitrary weight. Balls are drawn successively from the urn with probabilities proportional to their weights, and the drawing mechanism is described by the following Pólya-like urn scheme. Assuming that at the i-th draw black balls have weight M, and that there are k distinct colors in the urn with weights $M_1, \ldots, M_k$, respectively, at the $(i+1)$-th draw:

i) if we pick a black ball, then it is returned to the urn together with a black ball of weight

$$B^*_{i+1} = M\frac{V_{i+2,k+2}V_{i+1,k}}{V_{i+2,k+1}V_{i+1,k+1}} - M, \tag{18}$$

and a ball of a new color with weight

$$A^*_{i+1} = (1-\alpha)M\frac{V_{i+1,k}}{V_{i+1,k+1}}; \tag{19}$$

ii) if we pick a non-black ball, then it is returned to the urn together with a black ball of weight

$$\tilde{B}_{i+1} = M\frac{V_{i+2,k+1}V_{i+1,k}}{V_{i+2,k}V_{i+1,k+1}} - M, \tag{20}$$

and an additional ball of the same color with weight

$$\tilde{A}_{i+1} = M\frac{V_{i+1,k}}{V_{i+1,k+1}}. \tag{21}$$

If X_n is the color the non-black ball returned in the urn after the n-th draw, then it can be verified that $\Pr[X_{n+1} \in \cdot \,|\, X_1, \ldots, X_n]$ coincides with the predictive probabilities (5) of the class of Gibbs-type species sampling models. We refer to Section 2 of the supplementary material for details on formulae (18), (19), (20) and (21). Hereafter we denote by $\mathbf{X}_{n,k}$ a sample of size n from the above urn scheme and featuring $K_n = k \leq n$ distinct colors, labelled by $X^*_1, \ldots, X^*_{K_n}$, with frequencies $\mathbf{N}_n = \mathbf{n}$.

Note that the urn scheme proposed in Zabell [74] is recovered from (18), (19), (20) and (21) by setting $V_{n,k}$ of the form (9) and $M = \theta + k\alpha$. Note that under this assumptions the black ball is updated only when the black ball is drawn. This is indeed a feature of Zabell's urn scheme. Differently, in our urn scheme the weight of the black ball is updated when the black ball is drawn (18) and also when a non-black ball is drawn (20). According to (20), in order to update the black ball only when the black ball is drawn we must assume the following constrain

$$\frac{V_{n+2,k+1}V_{n+1,k}}{V_{n+2,k}V_{n+1,k+1}} = 1. \tag{22}$$

By means of (5), it can be easily verified that the assumption (22) is equivalent to

$$\begin{aligned}(23)\qquad &\Pr[X_{n+2} \text{ is of color } X_i^* \mid \boldsymbol{X}_{n+1,k+1}] \\ &\quad = \Pr[X_{n+2} \text{ is of color } X_i^* \mid \boldsymbol{X}_{n+1,k}],\end{aligned}$$

for $i = 1, \ldots, k$, i.e. the probability of observing at the next step a species of type i is independent of k. By Zabell's sufficientness postulate and Proposition 1 together, we know that, for any $\alpha \in (0,1)$ and $\theta > -\alpha$, the two parameter Poisson-Dirichlet process is the unique Gibbs-type species sampling model for which (23) holds true. In Section 2 of the supplementary material we present a direct proof of this fact, which is stated here as a proposition. This proposition holds true for all $\alpha < 1$ and does not rely on Zabell's characterization but only on Proposition 1.

PROPOSITION 2. *The two parameter Poisson-Dirichlet process is the unique Gibbs-type species sampling model for which the assumptions* (22) *hold true.*

Now, let us consider the alternative scenario in which the weight of the black ball is not updated neither when the black ball is drawn, nor when a non-black ball is drawn. Recall that, from the Pólya-like urn scheme of Zabell [74], this scenario is obtained by letting $\alpha \to 0$. In other terms we are considering the predictive probabilities characterizing the Dirichlet process. According to (18) and (20), in order to never update the black ball we must assume condition (22) together with

$$(24)\qquad \frac{V_{n+2,k+2} V_{n+1,k}}{V_{n+2,k+1} V_{n+1,k+1}} = 1.$$

By means of (5), it can be easily verified that the two assumptions are equivalent to assuming (23) and

$$\begin{aligned}(25)\qquad &\Pr[X_{n+2} \text{ is a new color } \mid \boldsymbol{X}_{n+1,k+1}] \\ &\quad = \Pr[X_{n+2} \text{ is a new color } \mid \boldsymbol{X}_{n+1,k}].\end{aligned}$$

According to Regazzini [58], for any $\theta > 0$ the Dirichlet process is the unique species sampling model for which (23) and (25) hold true. The next proposition provides an alternative proof of this result, not relying on the results of Regazzini [58], by only on Proposition 1. Its proof is presented in Section 2 of the supplementary material.

PROPOSITION 3. *The Dirichlet process is the unique Gibbs-type species sampling model for which the assumptions* (23) *and* (25) *hold true.*

So far we discussed the relationship between: i) the dependency on k of the ratio $V_{n+1,k+1}/V_{n,k}$ and $V_{n+1,k}/V_{n,k}$, which appear in the predictive probabilities (5); ii) the updates of the black ball in the above Pólya-like urn scheme. According to Proposition 2 the weight of the black ball is updated only when the black-ball is drawn if and only if $V_{n+1,k+1}/V_{n,k}$ depends on k and $V_{n+1,k}/V_{n,k}$ does not depend on k. The opposite scenario consists of updating the weight of the

black ball when a non-black ball is drawn, and not updating it when the black ball is drawn. According to (18) and (20), this scenario is obtained by assuming only (24). In the next proposition we show that this constraint alone implies that $\alpha = 0$. That is, imposing not to reinforce the black ball when a black ball is picked leads to the trivial scenario in which the weight of the black ball is actually never updated. See Section 2 of the supplementary material for the proof of the next proposition.

PROPOSITION 4. *The Dirichlet process is the unique Gibbs-type species sampling model for which the assumption* (25) *holds true.*

If $\alpha \to 0$ then the urn scheme of Zabell [74] reduces to the Pólya-like urn scheme introduced by Hoppe [33]. Hoppe [33] showed that the configuration of the colored balls after n draws from the urn is distributed as the sampling formula of Ewens [20], i.e, the distribution of the number of different gene types (alleles) and their frequencies at a selectively neutral locus under the infinitely-many-alleles model of mutation with rate $\theta > 0$. Hence, the following natural genetic interpretation for the Hoppe's urn scheme: colors are mutations and the black ball, which is ignored in describing the urn configuration, is a device for introducing new mutations. See Crane [16] for detailed account of the interplay between Hoppe's urn and Ewens sampling formula, as well as for their genetic interpretations. Under the Zabell's urn scheme, as well as under our general urn scheme, the distribution of the configuration of the colored balls after n draws from the urn can be easily derived from (11). See, e.g., Pitman [51] and Pitman [53]. However, despite explicit generalized Ewens sampling formulae are available, we are not aware of any genetic interpretation of them. Even for the simplest case of the Zabell's urn scheme, a natural genetic interpretation seems missing from the literature. While the parameter θ might be still interpreted as a mutation parameter, it is not clear a natural genetic interpretation for the parameter $\alpha \in (0, 1)$. See Feng and Hoppe [24] for a discussion.

4. SUFFICIENTNESS POSTULATES AND HIERARCHICAL DIRICHLET PROCESSES

The hierarchical Dirichlet process was introduced in Teh et al. [67], while its two parameter generalization is due to Teh [65]. Let P be a species sampling model with nonatomic base distribution ν_0. Hierarchical species sampling models are defined as collections of species sampling models, say $P_1, \ldots, P_r$, with the same random base distribution P. Due to the discreteness of P, the support of the P_j's is contained in that of P and, hence, all the P_j's share the same random support of P. A sample from a hierarchical species sampling model is then part of a random array $(X_{j,i})_{i\geq 1, 1\leq j\leq r}$, which is partially exchangeable in the sense de Finetti [18] originally attached to this term, i.e., each sequence $(X_{j,i})_{i\geq 1}$ is exchangeable for all $j \leq r$. The distribution of a sample $(X_{j,i})_{1\leq i\leq n_j, 1\leq j\leq r}$ from a hierarchical species sampling model can be expressed in the following hierarchical form

$$X_{j,i} \,|\, P_j \overset{\text{ind}}{\sim} P_j \qquad i = 1, \ldots, n_j, \quad j = 1, \ldots, r, \tag{26}$$

$$P_j \,|\, P \overset{\text{ind}}{\sim} \mathscr{P}_j(P) \qquad j = 1, \ldots, r,$$

$$P \quad \sim \quad \mathscr{P},$$

where n_j is the size of the sample from P_j and, in the second line, $\mathscr{P}_j$ is indexed by j because the conditional distribution may depend on additional population-specific parameters and P. One may think of the sample $(X_{j,i})_{1\leq i\leq n_j, 1\leq j\leq r}$ as a collection of samples from r different populations. Within population, observations are exchangeable, but across populations their dependence becomes weaker. Consideration of hierarchical models defined by layers of species sampling models raises the interesting problem of whether there exists sufficientness postulates that characterize the resulting models. In this section we discuss such a problem with respect to the two parameter hierarchical Poisson-Dirichlet process, namely: i) the P_j's are two parameter Poisson-Dirichlet process with parameters $\alpha_j \in [0,1)$, $\theta_j > -\alpha_j$ and with common base distribution P; ii) P is a two parameter Poisson-Dirichlet process with parameters $\alpha \in [0,1)$, $\gamma > -\alpha$ and nonatomic base distribution ν_0.

To describe the two parameter hierarchical Poisson-Dirichlet process we adopt the notation of Teh and Jordan [68]. Let $X_1^{**}, \ldots, X_K^{**}$ be the K distinct species observed in the joint sample from the r populations. Observations in population j are grouped in clusters. We remark that there may be two or more clusters in the population j composed of individuals of the same species. We therefore denote by $m_{j,k}$ the number of clusters in population j sharing species X_k^{**} and by $n_{j,t,k}$ the number of observations in population j, belonging to the t-th cluster and having species X_k^{**}. Within cluster t observations belong to the same species. We use dots in the subscripts to denote that we are summing over indexes, e.g. $n_{j\cdot\cdot}$ and $m_{j\cdot}$ are the number of observations and the number of clusters in population j respectively. Finally, we denote by $(X_{j,1}^*, \ldots, X_{j,m_{j\cdot}}^*)$ the species of the $m_{j\cdot}$ clusters in population j. Given a sample $(X_{j,i})_{1\leq j\leq r, 1\leq i\leq n_{j\cdot\cdot}}$, the predictive probability of $X_{j,n_{j\cdot\cdot}+1}$ is

$$\frac{\theta_j + m_{j\cdot}\alpha_j}{\theta_j + n_{j\cdot\cdot}} P(\cdot) + \frac{1}{\theta_j + n_{j\cdot\cdot}} \sum_{t=1}^{m_{j\cdot}} (n_{jt\cdot} - \alpha_j)\delta_{X_{j,t}^*}(\cdot) \tag{27}$$

for any $j = 1, \ldots, r$, whereas the predictive probability for a new cluster $X_{j,m_{j\cdot}+1}^*$ is

$$\frac{\gamma + K\alpha}{\gamma + m_{\cdot\cdot}} \nu_0(\cdot) + \frac{1}{\gamma + m_{\cdot\cdot}} \sum_{k=1}^{K} (m_{\cdot k} - \alpha)\delta_{X_k^{**}}(\cdot), \tag{28}$$

These two formulae should be understood as follows. $X_{j,i+1}$ joins the t-th cluster in population j and belongs to species $X_{j,t}^*$ with probability proportional to $(n_{jt\cdot} - \alpha_j)$, or it forms a new cluster with probability proportional to $(\theta_j + m_{j\cdot}\alpha_j)$. In this latter case, the species of this new cluster, $X_{j,m_{j\cdot}+1}^*$ is sampled from (28). Such a species is one of those already observed among all populations, say X_k^*, with probability proportional to $(m_{\cdot k} - K\alpha)$, or it is a new species with probability proportional to $(\gamma + K\alpha)$. The parameters α_j and θ_j have the same interpretation as for the predictive probabilities (17). Instead, α and γ control the total number and the sharing of cluster values among populations: the lower γ the lower is the average total number of different species observed K; the larger α the lower is the number of species shared across populations. We refer to Teh and Jordan [68] for further details.

The predictive probabilities of the hierarchical Dirichlet process arises from (27) and (28) by setting $\alpha = 0$ and $\alpha_j = 0$ for any $j = 1, \ldots, r$. We refer to Teh et al. [67] for a detailed account on this predictive probabilities, with a description in terms of the so-called Chinese restaurant franchise process. We assume the $\theta_j = \theta$ for any $j = 1, \ldots$. Now, let $(X_i)_{i\geq 1}$ be an exchangeable sequence directed by a Dirichlet process P with parameter γ and base distribution ν_0. Given P, or equivalently given $(X_i)_{i\geq 1}$, let $(X_{j,i})_{i\geq 1, j\geq 1}$ be a collection of conditionally independent exchangeable sequences, the j-th sequence being directed by a Dirichlet process P_j with parameter θ and base distribution P. We observe that in order to implement the second level of the hierarchy, and generate a finite sample of observations from multiple populations, it is not necessary to resort to P, but it is sufficient to have a truncated version of the Pólya urn sequence $(X_i)_{i\geq 1}$. In particular, it is enough to have at hand $(X_i)_{i\leq n_{\cdots}}$, because the exchangeable sequences at the second level of the hierarchy needs at most $n_{\cdots}$ conditional independent samples from P. In the next proposition we introduce a sufficientness postulate for the hierarchical Dirichlet process. Our postulate thus extends the characterization of Regazzini [58] and reveals some limitations of the hierarchical Dirichlet process.

PROPOSITION 5. *Let $(X_{j,i})_{i\geq 1, j\geq 1}$ be a partially exchangeable array directed by a hierarchical species sampling model, and assume that its predictive probabilities are such that the conditional probability of $X_{\ell,n_\ell+1}$ given the sample $(X_{j,i})_{1\leq i\leq n_j, j\leq r}$ is*

$$w_{n_\ell}\hat{F}_{\ell,n_\ell} + (1 - w_{n_\ell})F[(X_{j,i})_{1\leq i\leq n_j, j\leq r}] \tag{29}$$

where

i) $\hat{F}_{\ell,n_\ell}$ is the empirical distribution of $X_{\ell,1}, \ldots, X_{\ell,n_\ell}$;
ii) w_{n_ℓ} varies only with the population specific sample sizes n_ℓ;
iii) $F[(X_{j,i})_{1\leq i\leq n_j, j\leq r}]$ does not depend on ℓ and can include point masses that coincide with the $(X_{j,1}, \ldots, X_{j,n_j})$ values.

Then the hierarchical Dirichlet process is the directing measure of the array $(X_{j,i})_{i\geq 1, j\geq 1}$.

The proof of Proposition 5 is presented in Section 3 of the supplementary material. Proposition (5) imposes some constraints on the predictive probabilities of the partially exchangeable array $(X_{j,i})_{i\geq 1, 1\leq j\leq r}$. In particular, the constraint on the form (29) for the predictive probabilities, with the function $F[(X_{j,i})_{1\leq i\leq n_j, j\leq r}]$ not depending on ℓ and the weights w_{n_ℓ} depending only on the sample size n_ℓ, is the most relevant in practice. Indeed this constraint requires that the conditional probability of discovering a new species in an additional sample from the population j, given the sample $(X_{j,i})_{1\leq i\leq n_j, 1\leq j\leq r}$, depends only on the size of the sample from the population j, in a way that is homogeneous across populations. More formally, probabilities of discovering a new species by sampling from one of the populations are proportional to the vector of weights $[(1 - w_{n_1}), \ldots, (1 - w_{n_r})]$. This assumption is violated in numerous real-world examples, as evidenced in Figure 1. This figure shows an estimate of the Shannon entropy for the distribution of bacterial species in 900 samples of the vaginal microbiome taken from the

work of Ravel et al. [59]. Note that the predictive probabilities of a hierarchical Dirichlet process, conditioned on these data, would assign an equal probability to the event of discovering a new species from each of these populations, because the sample sizes n_j's are equal, despite the evident disparity in the diversity of species.

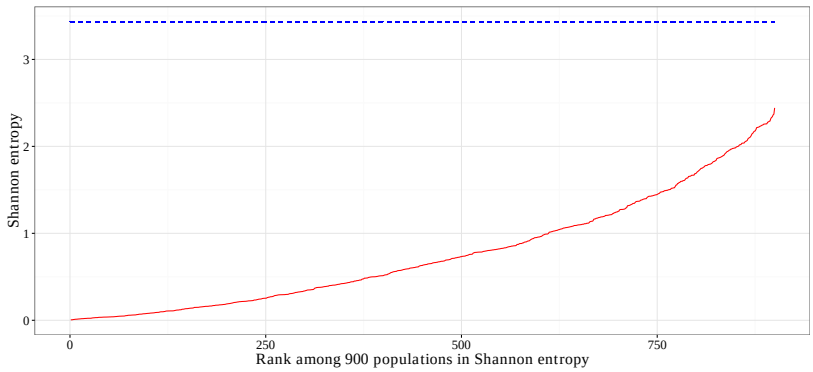

FIG 1. *The empirical Shannon entropy in the microbial distribution across 900 samples of the vaginal microbiome (Ravel et al. [59]), which are ranked according to the level of diversity. The dashed blue line shows the Shannon entropy of the Uniform distribution for the same number of species.*

Proposition 5 does not extend easily to the two parameter hierarchical Poisson-Dirichlet process. In fact, we believe it may not be trivial to provide a suffientness postulate for this model, unless one makes use of latent variables. Specifically, consider the predictive probabilities (27) and (28). If we condition on a set of variables that determine the steps in $(X_{j,i})_{i\geq 1, 1\leq j\leq r}$ in which P is sampled, then it is not difficult to formulate sufficientness conditions that characterize the exchangeable sequences. In particular the sufficientness characterization of Zabell [74] could be applied to each layer of the process. However, conditioning on this set of variables is not in the spirit of Johnson sufficientness postulate because, first, the variables that determine when P is sampled are not observable since the species observed at those steps are not necessarily "new", and second, unlike the exchangeability of a random partition the hierarchical structure assumed does not have an apparent subjective motivation. We also note that model interpretability, in this case, is provided by the overall probability construction, rather than by characteristics of the joint distribution of dependent random partitions, which in most cases presents analytic expressions that are far from trivial. With the exception of the correlations between the random probabilities $P_1, \ldots, P_r$, results to quantify and understand the degree of dependence among $(X_{j,i})_{i\geq 1, 1\leq j\leq r}$ remain limited.

5. DISCUSSION

In this paper we reviewed and discussed some nonparametric counterparts of the celebrated Johnson's "sufficientness" postulate. In particular we presented a general framework for "sufficientness" which extends previous characterizations by Regazzini [58], Zabell [71], Lo [48] and Zabell [75] for the Dirichlet process

and two parameter Poisson-Dirichlet process. The reader is referred to the works of Zabell [72], Walker and Muliere [69], Rolles [61] and Bacallado et al. [4] for related "sufficientness" characterizations in the context of neutral to the right random probability measures and Markov chains. Following the parallel with the "sufficientness" postulates for the Dirichlet process and the two parameter Poisson-Dirichlet process, we paired our postulate with a simple Pólya-like urn scheme for describing the predictive probabilities of Gibbs-type species sampling priors. Such a scheme provides a novel and intuitive interpretation of these predictive probabilities in terms of the updates of a sequence of balls drawn for a Pólya-like urn. We find this interpretation particularly useful in order to highlights the fundamental differences between the Dirichlet process, the two parameter Poisson-Dirichlet process, and the more general class of Gibbs-type species sampling models. In particular we show how the sufficientness postulates originally proposed by Zabell [74] and Regazzini [58] may be rephrased in terms of our Pólya-like urn scheme.

The Pólya-like urn schemes for the Dirichlet process and the two-parameter Poisson-Dirichlet process are often applied in hierarchical constructions. While hierarchical species sampling priors had a tremendous impact on several applied fields, it still remains difficult to guide a selection of the prior distribution with subjective arguments, such as the number of species and their variability across populations. On the other hand it also remains challenging to tune hierarchical constructions to optimize the performance of the resulting tools, quantified by classification and prediction error metrics. Our hope, and a motivation for our work, is that "sufficientness" postulates and urn schemes contribute to a better understanding and interpretability of hierarchical constructions [67] and dependent random distributions [64] that combine layers of exchangeable random partitions. In particular the study of Gibbs-type exchangeable random partitions has the potential of contributing to the critical evaluation of hierarchical constructions for data analysis. When, for example, heterogeneous populations, say in ecology of microbiome studies, are modeled using dependent random partitions embedded in hierarchical constructions, how can we use the imputed layers of partitions generated through Markov chain Monte Carlo algorithms or other approaches to evaluate the construction of the prior model? When can we say that the use of hierarchical species sampling priors appears appropriate? Which type of assumption can we leverage on to tackle this type of problems? The theoretical characterization and classification of random partitions will allow the statistical and machine learning communities to approach these problems.

SUPPLEMENTARY MATERIAL

Online supplementary material includes the proofs of Proposition 1, 2, 3, 4 and 5, and the derivation of the Pólya-like urn scheme for Gibbs-type species sampling models.

ACKNOWLEDGEMENTS

The authors are grateful to the Associate Editor and to three anonymous referees, whose comments and suggestions helped to improve the paper substantially. Stefano Favaro is supported by the European Research Council through

StG N-BNP 306406. Marco Battiston's research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC grant agreement number 617071.

REFERENCES

[1] Aldous, D. (1985). *Exchangeability and related topics* Ecole d'Eté de Probabilités de Saint-Flour XIII. Lecture notes in mathematics, Springer - Heidelberg.
[2] Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, **2**, 1152–1174.
[3] Arbel, J., Favaro, S., Nipoti, B. and Teh, Y.W. (2017). Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statist. Sinica*, in press.
[4] Bacallado, S., Favaro, S. and Trippa, L. (2013). Bayesian nonparametric analysis of reversible Markov chains. *Ann. Statist.*, **41**, 870–896.
[5] Bacallado, S., Favaro, S. and Trippa, L. (2015). Looking-backward probabilities for Gibbs-type exchangeable random partitions, *Bernoulli*, **21**, 1–37
[6] Bacallado, S., Favaro, S. and Trippa, L. (2015). Bayesian nonparametric inference for shared species richness in multiple populations, *Journal of Statistical Planning and Inference*, **166**, 14–23
[7] Battiston, M, Favaro, S. Roy, D.M. and Teh, Y.W. (2016). A characterization of product-form exchangeable feature probability functions. *Preprint arXiv:1607.02066*.
[8] Bertoin, J. (2006). *Random fragmentation and coagulation processes.* Cambridge University Press.
[9] Blackwell, D. and MacQueen, J.B. (1973). Ferguson Distributions via Pólya urn schemes. *Ann. Statist.*, **1**, 353–355.
[10] Broderick, T., Pitman, J. and Jordan, M. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Anal.*, **8**, 1–22.
[11] Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. *Adv. Neur. Inf. Proc. Sys.*
[12] Caron, F. and Fox, E.B. (2015). Sparse graphs with exchangeable random measures. *Preprint arXiv:1401.1137*.
[13] Caron, F., Teh, Y.W. and Murphy, T.B. (2014). Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes. *Ann. Appl. Statist.* **8**, 1145–1181.
[14] Chen, C., Ding, N. and Buntine, W (2012). Dependent hierarchical normalized random measures for dynamic topic modeling. *Int. Conf. Mach. Learn.*
[15] Chen, C., Rao, V.A., Buntine, W. and Teh, Y.W. (2013). Dependent normalized random measures. *Int. Conf. Mach. Learn.*
[16] Crane, H. (2016). The ubiquitous Ewens sampling formula. *Statist. Sci.*, **31**, 1–19.
[17] De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I. and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *EEE Trans. Pattern Anal. Mach. Intell.*, **37**, 212–229.
[18] de Finetti, B. (1938). Sur la condition d'"équivalence partielle". In *VI Colloque Gèneve: "Act. Sc. Ind."* Herman, Paris.
[19] Engen, S. (1999). *Stochastic abundance models with emphasis on biological communities and species diversity.* Chapman and Hall.
[20] Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112.
[21] Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B*, **71**, 993–1008.
[22] Favaro, S., Lijoi, A. and Prünster, I. (2012). A new estimator of the discovery probability. *Biometrics*, **68**, 1188–1196.
[23] Favaro, S. and Walker, S.G. (2013). Slice sampling σ-stable Poisson-Kingman mixture models. *J. Comput. Graph. Statist.*, **22**, 830–847.
[24] Feng, S. and Hoppe, F.M. (1998). Large deviation principles for some random com-

binatorial structures in population genetics and Brownian motion. *Ann. Appl. Probab.*, **8**, 975–994.
[25] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.
[26] Fisher, R.A., Corbet, A.S., Williams, C.B (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Animal Ecol.* **12**, 42–58.
[27] Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, **138**, 5674–5685.
[28] Good, I.J. (1965). *The estimation of probabilities: an essay on modern Bayesian methods.* MIT press.
[29] Heaukulani, C. and Roy, D.M. (2015). Gibbs-type Indian buffet processes. *Preprint arXiv:1512.02543*.
[30] Herlau, T. (2015). Completely random measures for modeling block-structured sparse networks. *Preprint arXiv:1507.02925*.
[31] Hjort, N.L. (2000). Bayesian analysis for a generalised Dirichlet process prior. *Technical Report, Matematisk Institutt, Universitetet i Oslo*.
[32] Ho, M., James, L.F. and Lau, J.W. (2007). Gibbs partitions (EPPF's) derived from a stable subordinator are Fox H and Meijer G transforms. *Preprint arXiv:0708.0619*.
[33] Hoppe, F.H. (1984). Pólya-like urns and the Ewens sampling formula. *J. Math. Biol.*, **20**, 91–94.
[34] Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick- breaking priors. *J. Amer. Stat. Ass.*, **96** 161–173.
[35] James, L.F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *Preprint arXiv:math/0205093*.
[36] James, L.F. (2013). Stick-breaking PG(α, ζ)-generalized Gamma processes. *Preprint arXiv:1308.6570*.
[37] Johnson, W.E. (1932). Probability: the deductive and inductive problems. *Mind*, **41**, 409–423.
[38] Kingman, J.F.C. (1978). The representation of partiton structure. *J. London. Math. Soc.*, **18**, 374–380.
[39] Korwar, R.M. and Hollander, M. (1973). Contribution to the theory of Dirichlet processes. *Ann. Probab.*, **1**, 705–711.
[40] Lee, J., Quintana, F.A., Müller, P. and Trippa, L. (2013). Defining predictive probability functions for species sampling models. *Statist. Sci.*, **28**, 209–222.
[41] Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse-Gaussian priors. *J. Amer. Stat. Assoc.* **100** 1278–1291.
[42] Lijoi, A., Mena, R.H. and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, **94**, 769–786.
[43] Lijoi, A., Mena, R.H. and Prünster, I. (2007a). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. Roy. Statist. Soc. Ser. B*, **69**, 769–786.
[44] Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, Hjort, N.L., Holmes, C.C. Müller, P. and Walker, S.G. Eds. Cambridge University Press.
[45] Lijoi, A., Prünster, I. and Walker, S.G. (2008). Investigating nonparametric priors with Gibbs structure. *Statist. Sinica*, **18**, 1653–1668.
[46] Lijoi, A., Prünster, I. and Walker, S.G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.*, **18**, 1519–1547.
[47] Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. *Ann. Statist.*, **12**, 351–357.
[48] Lo, A.Y. (1991). A characterization of the Dirichlet process. *Statist. Probab. Lett.*, **12**, 185–187.
[49] Lomeli, M., Favaro, S and Teh, Y.W. (2017). A marginal sampler for σ-stable Poisson-Kingman mixture models. *J. Comput. Graph. Statist.*, **26**, 44–53.
[50] Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields.* **92**, 21–39.
[51] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, **102**, 145–158.
[52] Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statis-*

tics, Probability and Game Theory, Ferguson, T.S., Shapley, L.S. and MacQueen, J.B. Eds., Institute of Mathematical Statistics.

[53] Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: A Festschrift for Terry Speed*, Goldstein, D.R. Eds. Institute of Mathematical Statistics.

[54] Pitman, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture notes in mathematics, Springer - New York.

[55] Pitman, J. and Yor, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.

[56] Prünster, I. (2002). *Random probability measures derived from increasing additive processes and their application to Bayesian statistics.* Ph.d thesis, University of Pavia.

[57] Quintana, F.A. and Iglesias, P.L. (2003). Bayesian clustering and product partition models. *J. Roy. Statist. Soc. Ser. B*, **65**, 557–574

[58] Regazzini, E. (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilià. *Giornale dell'Istituto Italiano degli Attuari*, **41**, 77–89.

[59] Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., Brotman, R.M., Davis, C.C., Ault, K., Peralta, L., Forney, L.J. (2011). Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. USA*, **108**, 4680–4687

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

[61] Rolles, S. (2003). How edge-reinforced random walk arises naturally. *Probab. Theory Related Fields*, **126**, 243–260.

[62] Roy, D.M. (2014). The continuum-of-urns scheme, generalized beta and Indian buffet processes, and hierarchies thereof. *Preprint arXiv:1501.00208.*

[63] Sato, K. (1999). *Lévy processes and infinitely divisible distributions.* Cambridge University Press.

[64] Jo, S., Lee, J., Müller, P., Quintana, F. and Trippa, L. (2016). *Dependent species sampling models for spatial density estimation. Bayesian Analysis*

[65] Teh, Y.W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. *Proceedings of the 21st International Conference on Computational Linguistic*

[66] Teh, Y.W. and Görür (2010). Indian buffet processes with power law behavior. *Adv. Neur. Inf. Proc. Sys.*

[67] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566–1581.

[68] Teh, Y.W. and Jordan, M.I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*, Hjort, N.L., Holmes, C.C. Müller, P. and Walker, S.G. Eds., Cambridge University Press, Cambridge.

[69] Walker, S.G. and Muliere, P. (1999). A characterization of a neutral to the right prior via an extension of Johnson's sufficientness postulate. *Ann. Statist.*, **27**, 589–599

[70] Watterson, G.A. (1976). The stationary distribution of the infinitely-many neutral alleles diffusion model. *J. Appl. Probab.*, **13**, 639–651

[71] Zabell, S.L. (1982). W. E. Johnson's "sufficientness" postulate. *Ann. Statist.*, **10**, 1090–1099

[72] Zabell, S.L. (1985). Characterizing Markov exchangeable sequences. *J. Theoret. Probab*, **8**, 175–178

[73] Zabell, S.L. (1992). Predicting the unpredictable. *Synthese*, **90**, 205–232

[74] Zabell, S.L. (1997). The continuum of inductive methods revisited. In *The cosmos of science: essays in exploration*, Earman, J. and Norton, J.D. Eds. Universty of Pittsburgh Press.

[75] Zabell, S.L. (2005). The continuum of inductive methods revisited. In *Symmetry and its discontents: essays on the history of inductive probability.* Cambridge Univ. Press, New York.