Are Gibbs–type priors the most natural generalization of the Dirichlet process?

P. De Blasi[1], S. Favaro[1], A. Lijoi[2], R.H. Mena[3], I. Prünster[1] and M. Ruggiero[1]

[1] Università degli Studi di Torino and Collegio Carlo Alberto, Torino, Italy.
E-mail: pierpaolo.deblasi@unito.it; stefano.favaro@unito.it; igor.pruenster@unito.it; matteo.ruggiero@unito.it

[2] Università degli Studi di Pavia and Collegio Carlo Alberto, Torino, Italy.
E-mail: lijoi@unipv.it

[3] Universidad Autónoma de México, México
E-mail: ramses@sigma.iimas.unam.mx

Abstract

Discrete random probability measures and the exchangeable random partitions they induce are key tools for addressing a variety of estimation and prediction problems in Bayesian inference. Indeed, many popular nonparametric priors, such as the Dirichlet and the Pitman–Yor process priors, select discrete probability distributions almost surely and, therefore, automatically induce exchangeable random partitions. Here we focus on the family of *Gibbs–type priors*, a recent and elegant generalization of the Dirichlet and the Pitman–Yor process priors. These random probability measures share properties that are appealing both from a theoretical and an applied point of view: (i) they admit an intuitive characterization in terms of their predictive structure justifying their use in terms of a precise assumption on the learning mechanism; (ii) they stand out in terms of mathematical tractability; (iii) they include several interesting special cases besides the Dirichlet and the Pitman–Yor processes. The goal of our paper is to provide a systematic and unified treatment of Gibbs–type priors and highlight their implications for Bayesian nonparametric inference. We will deal with their distributional properties, the resulting estimators, frequentist asymptotic validation and the construction of time–dependent versions. Applications, mainly concerning hierarchical mixture models and species sampling, will serve to convey the main ideas. The intuition inherent to this class of priors and the neat results that can be deduced for it lead one to wonder whether it actually represents the most natural generalization of the Dirichlet process.

Key words and phrases: Bayesian Nonparametrics; Clustering; Consistency; Dependent process; Discrete nonparametric prior; Exchangeable partition probability function; Gibbs–type prior; Pitman–Yor process; Mixture model; Population Genetics; Predictive distribution; Species sampling.

1 Introduction and preliminaries

One of the main research lines within Bayesian Nonparametrics has been the proposal and study of classes of random probability measures whose laws act as nonparametric priors. Several such classes contain, as a special case, Ferguson's Dirichlet process [22], which still represents the cornerstone of the field. A recent review that covers many of these models and uses completely random measures as a unifying concept can be found in [45]. When going beyond the Dirichlet process one typically has to face a trade–off between the desire of generality (which, as far as inference is concerned, implies flexibility of the model) and tractability, both analytical and computational. Probably the most successful proposal is represented by the two–parameter Poisson–Dirichlet process introduced in [57] and further investigated in countless papers, most notably in [59, 63]. See [62] for a comprehensive review from a probabilistic perspective. Such a process is also known as Pitman–Yor (PY) process, especially in the Machine Learning community, according to a terminology introduced in [33] which we will also adopt in the present paper. For our purposes it is important to note that the PY process reduces to the Dirichlet process by setting one of its parameters equal to 0. Nonetheless, some important distributional features of the PY process are fundamentally different according as to whether the value of such a parameter is equal to 0 or not. A clear understanding of this aspect is possible by identifying a large class of priors, which embeds the PY process as a special case. Such a class is given by Gibbs–type priors, introduced in [28] and only briefly addressed in the above mentioned review of nonparametric priors [45], thus motivating the main focus of this paper. In fact, by close inspection of the predictive structure they lead to, it will become apparent that the variety of distributional characteristics can be actually traced back to crucially different assumptions on the learning mechanism. This leads to a novel classification of discrete nonparametric priors which also serves as motivation for the use of Gibbs–type priors. Moreover, Gibbs–type priors have the advantage of pinning down, in a neat way, the analytic tractability issue related to general classes of nonparametric priors: in fact, they allow to split the prediction rule in two stages and to highlight the key quantity allowing simplification of the relevant expressions. Indeed, throughout the following sections one can appreciate the beauty and simplicity of various analytical results that admit straightforward application to statistical inference. Finally, it is to be noted that Gibbs–type priors include other notable special cases of priors beyond the Dirichlet and the PY processes: for example, normalized inverse Gaussian processes [40] and their generalization given by normalized generalized gamma processes [43] as well as mixtures of symmetric Dirichlet distributions [28]. Given this, can one state with confidence that Gibbs–type priors are a natural generalization of the Dirichlet process, maybe the most natural?

The present paper aims at providing a survey on Gibbs–type priors that accounts for recent

findings both in the probabilistic and statistical literature. This will serve as an important opportunity for pointing out their analytical tractability, flexibility and suitability in a variety of inferential problems beyond current applications which include mixture models (see, e.g., [33, 43]), linguistics and information retrieval in document modeling ([73, 74]), species sampling ([42, 44, 55]) and survival analysis [37], among others.

1.1 Discrete random probability measures, exchangeable random partitions and predictive distributions

We first lay out the basics of Bayesian inference in an exchangeable framework and focus on some key concepts and tools. Suppose $(X_n)_{n\geq 1}$ is an (ideally) infinite sequence of observations, with each X_i taking values in some set $\mathbb{X}$. Moreover, $\mathbf{P}_{\mathbb{X}}$ is the set of all probability measures on $\mathbb{X}$. Assuming $(X_n)_{n\geq 1}$ to be *exchangeable* is equivalent to assuming the existence of a probability distribution Q on $\mathbf{P}_{\mathbb{X}}$ such that

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

for any $n \geq 1$. Hence, $\tilde{p}$ is a random probability measure on $\mathbb{X}$ and its probability distribution Q, also termed *de Finetti measure*, represents the prior distribution when (1) is used as a Bayesian model with an observed sample X_i, $i = 1, \ldots, n$. Whenever Q degenerates on a finite dimensional subspace of $\mathbf{P}_{\mathbb{X}}$, the inferential problem is usually called *parametric*. On the other hand, when the support of Q is infinite–dimensional then one typically speaks of a *nonparametric* inferential problem and it is generally agreed (see, e.g., [23]) that having a large topological support is a desirable property for a nonparametric prior. Given a sample $X_1, \ldots, X_n$ generated through (1), the (one–step ahead) predictive distribution coincides with the posterior expected value of $\tilde{p}$, that is

$$\mathrm{P}(X_{n+1} \in \cdot \,| X_1, \ldots, X_n) = \int_{\mathbf{P}_{\mathbb{X}}} p(\,\cdot\,)\, Q(\mathrm{d}p \,|\, X_1, \ldots, X_n), \tag{2}$$

where $Q(\,\cdot \,|\, X_1, \ldots, X_n)$ denotes the posterior distribution of $\tilde{p}$.

Discrete nonparametric priors, i.e. priors which select discrete distributions with probability 1, play a key role in most Bayesian nonparametric procedures. It is well–known that the Dirichlet and the PY process priors share this property and the same can be said for the broader class of Gibbs–type priors. In fact, any random probability measure associated to a discrete prior can be represented as

$$\tilde{p} = \sum_{j=1}^{\infty} \tilde{p}_j \, \delta_{Z_j} \tag{3}$$

where δ_c stands for the unit point mass concentrated at c, $(\tilde{p}_j)_{j\geq 1}$ is a sequence of non–negative random variables such that $\sum_{j\geq 1}\tilde{p}_j = 1$, almost surely, and $(Z_j)_{j\geq 1}$ is a sequence of $\mathbb{X}$–valued random variables. Henceforth, we further assume that $(\tilde{p}_j)_{j\geq 1}$ and $(Z_j)_{j\geq 1}$ are independent and that the Z_j's are iid from a diffuse probability measure P^* on $\mathbb{X}$ (or in other terms $\mathrm{P}(Z_i \neq Z_j) = 1$ for any $i \neq j$). Such a general subclass of discrete random probability measures has been called *species sampling models* by Pitman [59], a terminology that will be clarified in the following section.

As far as the observables X_i's are concerned, the discrete nature of Q implies that any sample $X_1, \ldots, X_n$ will feature ties with positive probability, therefore generating $K_n = k \leq n$ distinct observations $X_1^*, \ldots, X_k^*$ with frequencies $n_1, \ldots, n_k$ such that $\sum_{i=1}^k n_i = n$. When choosing and analyzing specific predictive structures, the key quantity to consider, from both a conceptual and a mathematical point of view, is the probability of observing a new distinct value not included in the sample $X_1, \ldots, X_n$, namely

$$\mathrm{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n), \tag{4}$$

which will appear throughout the paper. To be more concrete consider the Dirichlet and the PY processes. In the Dirichlet case, with parameters given by P^* and $\theta > 0$, one has

$$\mathrm{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n) = \frac{\theta}{\theta + n}.$$

In the PY case, in addition to P^*, one has two parameters (σ, θ) whose admissible values are $\sigma \in [0, 1)$ with $\theta > -\sigma$ or $\sigma < 0$ with $\theta = m|\sigma|$ for some positive integer m. One then has

$$\mathrm{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n) = \frac{\theta + \sigma k}{\theta + n}$$

from which it is apparent that the corresponding probability for the Dirichlet process is recovered by setting $\sigma = 0$.

Within such a framework, discrete random probability measures can be characterized in terms of the exchangeable random partition they imply, another key aspect of the paper for which we provide some essential background. Given the discreteness of Q, $\tilde{p}$ induces a partition of $X_1, \ldots, X_n$ that is well described by means of an extremely useful tool, namely the *exchangeable partition probability function* (EPPF) [59] given by

$$p_k^{(n)}(n_1, \ldots, n_k) = \int_{\mathbb{X}^k} \mathrm{E}\left(\tilde{p}^{n_1}(\mathrm{d}x_1) \cdots \tilde{p}^{n_k}(\mathrm{d}x_k)\right). \tag{5}$$

It is also of simple interpretability: it essentially corresponds to the probability, induced by $\tilde{p}$, of observing a sample of size n, $X_1, \ldots, X_n$, exhibiting $K_n = k$ distinct observations with

frequencies $n_1, \ldots, n_k$ or, equivalently, a specific partition into $K_n = k$ clusters with frequencies $n_1, \ldots, n_k$. See [59, 62] for details. Note also that an EPPF satisfies the addition rule

$$p_k^{(n)}(n_1, \ldots, n_k) = p_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1) + \sum_{j=1}^{k} p_k^{(n+1)}(n_1, \ldots, n_j + 1, \ldots, n_k). \tag{6}$$

For both Dirichlet and PY processes, the EPPF is available in closed form. In the former case it is given by

$$p_k^{(n)}(n_1, \ldots, n_k) = \frac{\theta^k}{(\theta)_n} \prod_{i=1}^{k} (n_i - 1)! \tag{7}$$

where $(\theta)_n = \theta(\theta + 1) \cdots (\theta + n - 1)$ for any $n \geq 1$. For the PY process, it coincides with

$$p_k^{(n)}(n_1, \ldots, n_k) = \frac{\prod_{i=1}^{k-1} (\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{i=1}^{k} (1 - \sigma)_{n_i - 1}. \tag{8}$$

The identification of the EPPF leads to the direct determination of the predictive distribution in (2). Indeed, if $X_1, \ldots, X_n$ is a sample featuring $k \leq n$ distinct values with respective frequencies $n_1, \ldots, n_k$, one has

$$\mathrm{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n) = \frac{p_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1)}{p_k^{(n)}(n_1, \ldots, n_k)} \tag{9}$$

and the predictive distribution in (2) is a linear combination of $P^*(\,\cdot\,) = \mathrm{E}(\tilde{p}(\,\cdot\,))$, which can be interpreted as the prior guess at the shape of $\tilde{p}$, and of a weighted measure of the observations, namely

$$\mathrm{P}(X_{n+1} \in \,\cdot\, | X_1, \ldots, X_n) = \frac{p_{k+1}^{(n+1)}(n_1, \ldots, n_k, 1)}{p_k^{(n)}(n_1, \ldots, n_k)} P^*(\,\cdot\,) + \sum_{j=1}^{k} \frac{p_k^{(n+1)}(n_1, \ldots, n_j + 1, \ldots, n_k)}{p_k^{(n)}(n_1, \ldots, n_k)} \delta_{X_j^*}(\,\cdot\,). \tag{10}$$

Note that the right hand side of (10) is guaranteed to sum up to 1 if evaluated over the whole space $\mathbb{X}$ by the addition rule (6). In the PY process case the predictive distribution takes on the form

$$\mathrm{P}(X_{n+1} \in \,\cdot\, | X_1, \ldots, X_n) = \frac{\theta + \sigma k}{\theta + n} P^*(\,\cdot\,) + \frac{1}{\theta + n} \sum_{j=1}^{k} (n_j - \sigma) \delta_{X_j^*}(\,\cdot\,) \tag{11}$$

which, for $\sigma = 0$, reduces to the well–known Dirichlet process predictive structure given by a linear combination of P^* and the empirical measure.

1.2 Applications to species sampling problems and mixture modeling

Discrete nonparametric priors in general, and Gibbs–type priors in particular, are suited for addressing inferential issues that arise in species sampling problems and in mixture modeling, among others. We now briefly sketch these frameworks.

Consider a discrete random probability measure (3) with the specifications as in Section 1.1. It is, then, apparent that (3) can be seen as a tool for describing the structure of a population made of different types or species with certain proportions, which are modeled through (3) as random proportions $\tilde{p}_j$. On the basis of this fact, one can equivalently use the Z_i's or the positive integers $\{1, 2, \ldots\}$ to label different species or types that can be sampled. Indeed, if $(\xi_n)_{n\geq 1}$ is an auxiliary integer–valued sequence such that $\mathrm{P}(\xi_n = j \,|\, \tilde{p}) = \tilde{p}_j$, for any n and j, model (1) corresponds to assuming that $X_i = Z_{\xi_i}$. Hence the X_n's can be interpreted as the *observed species labels* since, due to the diffuse nature of P^*, any two data points X_i and X_j, for $i \neq j$, differ if and only if ξ_i and ξ_j do. Moreover, one has that $\mathrm{P}(\xi_i = \xi_j) > 0$, for any $i \neq j$, and this entails that the i–th and the j–th observations may reveal the same species with positive probability. It is precisely this connection which motivates the terminology adopted in [60], *species sampling model*. Moreover, an exchangeable sequence $(X_n)_{n\geq 1}$ for which (1) holds true, with $\tilde{p}$ a species sampling model, takes on the name of *species sampling sequence*.

By virtue of this interpretation, there are a number of statistical problems one can face adopting a Bayesian nonparametric perspective. Indeed, in many statistical applications one typically observes a sample of species labels $X_1, \ldots, X_n$ and designs further sampling $X_{n+1}, \ldots, X_{n+m}$ on the basis of estimates of some quantities of interest such as, e.g.: the number of new distinct species that will be detected in a new sample of size m; the number of species with a given frequency, or with frequency below a certain threshold, in $X_1, \ldots, X_{n+m}$; the probability that the $(n+m+1)$–th draw will consist of a species having frequency $\ell \geq 0$ in $X_1, \ldots, X_{n+m}$. These, in turn, provide measures of overall and rare species diversity and are of interest in biological, ecological or linguistic studies, just to mention a few. In this respect, the predictive approach briefly sketched in Section 1.1 plays an important role and provides nice and elegant answers to these problems in the framework of Gibbs–type priors.

Discrete nonparametric priors are also basic building blocks for hierarchical mixture models that are typically used for density estimation and clustering but also in more complex dependent structures. To keep things simple consider the univariate density estimation case and let $f(\cdot \,|\, \cdot)$ denote a kernel defined on $\mathbb{R} \times \mathbb{X}$ and taking values in $\mathbb{R}^+$ such that $\int_{\mathbb{R}} f(y|x)\, \mathrm{d}y = 1$, for any x in $\mathbb{X}$. Hence, $f(\cdot\,|x)$ defines a density function on $\mathbb{R}$, for any x. The observations are then from

a sequence $(Y_n)_{n\geq 1}$ of real–valued random variables such that

$$\begin{aligned} Y_i \,|\, X_i &\overset{\text{ind}}{\sim} f(\cdot\,|X_i) \qquad i = 1, \ldots, n \\ X_i \,|\, \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p} \qquad\qquad\quad i = 1, \ldots, n \\ \tilde{p} &\sim Q. \end{aligned} \tag{12}$$

The typical choice for $\tilde{p}$ is represented by the Dirichlet process leading to the Dirichlet process mixture model introduced by Lo [49], which represents the most popular Bayesian nonparametric model to date. In addition to density estimation such model serves also clustering purposes. In fact, here $(X_n)_{n\geq 1}$ is a sequence of latent exchangeable random elements and the unobserved number K_n of distinct values among $X_1, \ldots, X_n$ is the number of clusters into which the observations $Y_1, \ldots, Y_n$ can be grouped. Posterior inferences for K_n are of great importance and the specification of a Gibbs–type prior $\tilde{p}$ in (12) allows for an effective detection of the number of clusters that have generated the data.

1.3 Outline of the paper

Section 2 first provides an intuitive characterization of Gibbs–type priors based on a suitable classification of species sampling models. This is, then, followed by a formal definition and an overview of their distributional properties that are of interest for applications to Bayesian inference. Particular emphasis is given to the role played by one of the parameters that characterizes them. Section 3 discusses the use of Gibbs–type priors within hierarchical mixture models for density estimation and clustering. Section 4 focuses on the application of Gibbs–type priors to prediction problems and Section 5 deals with their frequentist asymptotic properties. Section 6 concisely discusses extensions of Gibbs–type priors to dynamic contexts. Finally, Section 7 contains some concluding remarks trying to answer the question posed in the title of the paper.

2 Gibbs–type priors

An interesting and useful classification of species sampling models can be given in terms of the structure of the probability of generating a new value they induce. This leads to an intuitive characterization of Gibbs–type priors and represents also one of the main motivations for their use. Our result is somehow in the spirit of Zabell's [76] characterization of the Dirichlet process in terms of the so–called *Johnson's sufficientness postulate*. To this end, recall that the key quantity is (4) representing the probability of generating a new value given the past associated to a species sampling model as specified in Section 1.1. According to its structure one can classify the

underlying priors in three main categories. Denote by $\mathbf{\Theta}$ a finite–dimensional parameter possibly entering the specification of $\tilde{p}$ in (1). In general, one has $\mathrm{P}(X_{n+1} = \text{“new”} \,|\, X_1, \ldots, X_n) = f(n, k, n_1, \ldots, n_k, \mathbf{\Theta})$, which means that the probability of obtaining a new observation depends on the sample size n, the number of distinct values k, their frequencies $(n_1, \ldots, n_k)$ and the parameter Θ. We will denote $\mathrm{P}(X_{n+1} = \text{“new”} \,|\, X_1, \ldots, X_n)$ by $f(n, k, \mathbf{\Theta})$ if it does not depend on $(n_1, \ldots, n_k)$ and by $f(n, \mathbf{\Theta})$ if it depends neither on $(n_1, \ldots, n_k)$ nor on k.

Proposition 1 *Let $\tilde{p}$ be a species sampling model. Then the following classification in terms of the structure of the probability of generating a new value holds:*

(i) $\mathrm{P}(X_{n+1} = \text{“new”} \,|\, X_1, \ldots, X_n) = f(n, \mathbf{\Theta})$ *if and only if $\tilde{p}$ is a Dirichlet process;*

(ii) $\mathrm{P}(X_{n+1} = \text{“new”} \,|\, X_1, \ldots, X_n) = f(n, k, \mathbf{\Theta})$ *if and only if $\tilde{p}$ is of Gibbs–type;*

(iii) $\mathrm{P}(X_{n+1} = \text{“new”} \,|\, X_1, \ldots, X_n) = f(n, k, n_1, \ldots, n_k, \mathbf{\Theta})$ *otherwise.*

Even if the Dirichlet process has proven to perform well in several applied contexts, from a merely conceptual point of view it seems too restrictive to let the probability of generating new values depend solely on the sample size n and on its total mass parameter θ and to summarize all other information contained in the data by means of a suitable specification of the scalar parameter θ. One would like indeed such a probability to explicitly depend also on (at least) the number of distinct observed values, since it summarizes the heterogeneity in the sample. By virtue of (ii), this is tantamount to resorting to a Gibbs–type prior. According to the specific situation, one might want to model (4) as an increasing or decreasing function of K_n, which will be shown to correspond to Gibbs–type priors with a parameter, to be identified later, being either positive or negative, respectively. Case (iii), which corresponds to the most general setup and prediction of new values explicitly depends on all the information conveyed by the data, is in principle the most desirable prediction structure. However, there are two main operational problems that one needs to take into account. On the one hand, the general case (iii) gives rise to serious analytical hurdles and priors have to be studied on a case-by-case basis typically leading to quite complicated expressions (see [20]). On the other hand, it is not clear how one should explicitly specify the dependence of the probability of observing a new species on the observed frequencies $n_1, \ldots, n_k$ so that it reflects an opinion on the learning mechanism for the data. It is thus reasonable that such prior opinion be encoded through the finite–dimensional parameter $\mathbf{\Theta}$. Hence, the above classification neatly shows the origin of the mathematical tractability of Gibbs–type priors, which is due to a precise simplifying assumption on the prediction structure. Overall, such an assumption appears to be a satisfactory compromise between generality (or

flexibility) and tractability, and therefore motivates the attempt to study and understand the behavior of such priors.

After having stated and discussed a predictive characterization of Gibbs–type priors, we now provide a different, though equivalent, definition which is more useful when one wishes to analyze their distributional properties. As seen in Section 1, a discrete nonparametric prior $\tilde{p}$ associated to an exchangeable scheme of the type (1) can be characterized in terms of the associated EPPF $\{p_k^{(n)} : \ n \geq 1, \ 1, \leq k \leq n\}$ defined as in (5). Accordingly, one defines a *Gibbs–type prior* as a species sampling model such that

$$p_k^{(n)}(n_1, \ldots, n_k) = V_{n,k} \prod_{i=1}^{k} (1-\sigma)_{n_i - 1} \tag{13}$$

for any $n \geq 1$, $k \leq n$ and positive integers $n_1, \ldots, n_k$ such that $\sum_{i=1}^{k} n_i = n$, where $\sigma < 1$ and the set of non–negative weights $\{V_{n,k} : \ n \geq 1, \ 1 \leq k \leq n\}$ satisfies the forward recursive equation

$$V_{n,k} = (n - \sigma k) V_{n+1,k} + V_{n+1,k+1} \tag{14}$$

for any $k = 1, \ldots, n$ and $n \geq 1$, with $V_{1,1} = 1$. In light of (13) one can rephrase the reason for their tractability in more mathematical terms, namely the product form of their EPPFs which allows to handle conveniently the frequencies n_i. Given (13), the probability of obtaining a new distinct observation conditional on a sample $X_1, \ldots, X_n$ such that $K_n = k$ is

$$\mathrm{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n) = \frac{V_{n+1,k+1}}{V_{n,k}} = f(n, k, \boldsymbol{\Theta})$$

which is in accordance with the above characterization.

Remark 2 *According to the classification implied by Proposition 1, mixtures of the Dirichlet process, obtained by mixing with respect to the total mass θ of the base measure, are in class (ii). To see this, let π denote the prior on θ so that $\pi(\mathrm{d}\theta | X_1, \ldots, X_n) \propto \theta^k \pi(\mathrm{d}\theta)/(\theta)_n$, where $(\theta)_n$ is the n–th ascending factorial. Hence*

$$\mathbb{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n) = \int_{\mathbb{R}^+} \frac{\theta^{k+1}}{(\theta)_{n+1}} \, \pi(\mathrm{d}\theta)$$

will now depend on k. More generally, mixtures of Gibbs–type priors obtained by mixing with respect to a possible parameter entering the definition of $V_{n,k}$ are still of Gibbs–type and, thus, still lie in (ii). In contrast, Gibbs–type priors mixed with respect to σ are such that $\pi(\mathrm{d}\sigma | X_1, \ldots, X_n) \propto V_{n,k} \prod_{i=1}^{k} (1-\sigma)_{n_i - 1} \, \pi(\mathrm{d}\sigma)$ for some prior π on σ. This clearly implies that the resulting family of species sampling models is in (iii), although one still preserves a Gibbs structure conditionally

on σ*.*

The definition (13) implies that the induced predictive distributions are

$$\mathrm{P}\left(X_{n+1} \in \cdot \,\middle|\, X_1, \ldots, X_n\right) = \frac{V_{n+1,k+1}}{V_{n,k}}\, P^*(\cdot) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{i=1}^{k} (n_i - \sigma)\, \delta_{X_i^*}(\cdot). \tag{15}$$

Hence, the predictive distribution is a linear convex combination of the prior guess P^* at the shape of $\tilde{p}$ and of the weighted empirical distribution $\widehat{P}_n = (n - k\sigma)^{-1} \sum_{i=1}^{k} (n_i - \sigma)\, \delta_{X_i^*}$. The predictive structure (15) exhibits some appealing and intuitive features. In particular, the mechanism for allocating the predictive mass among "new" and previously observed data can be split into two stages. Given a sample $X_1, \ldots, X_n$, the first step consists in allocating the mass between a new value X_{k+1}^* sampled from P^* and the set of observed values $\{X_1^*, \ldots, X_k^*\}$. This first step depends only on n and k and not on the frequencies $n_1, \ldots, n_k$. The second step is the following: conditionally on X_{n+1} being a new value, it is sampled from the base measure P^*, whereas if X_{n+1} coincides with one of the previously observed values X_i^*, for $i = 1, \ldots, k$, the coincidence probabilities are determined by the size n_i of each cluster and by σ. Hence, even if the frequencies n_i do not affect the probability of allocating a predicted value between "new" and "old", they are explicitly involved conditional on the predicted value coinciding with a previously observed one: the more often a past observation is detected, the higher the probability of re–observing it. Also σ plays an interesting role in weighting the empirical measure since, for $\sigma > 0$, a reinforcement mechanism driven by σ takes place. Indeed, one can see that the ratio of the probabilities assigned to any pair of (X_i^*, X_j^*) is given by $(n_i - \sigma)/(n_j - \sigma)$. As $\sigma \to 0$, the previous quantity reduces to the ratio of the sizes of the two clusters and therefore the coincidence probability is proportional to the size of the cluster. On the other hand, if $\sigma > 0$ and $n_i > n_j$, the ratio is an increasing function of σ. Hence, as σ increases the mass is reallocated from X_j^* to X_i^*. This means that the sampling procedure tends to reinforce, among the observed clusters, those having higher frequencies, which represents an appealing feature in certain inferential contexts. See [43] for a discussion of such reinforcement mechanisms and their use in mixture models. If $\sigma < 0$, the reinforcement mechanism works in the opposite way in the sense that the coincidence probabilities are less than proportional to the cluster size.

Besides influencing the balancedness of the partition of the exchangeable random elements directed by a Gibbs–type prior, the parameter σ also determines the rate at which the number of clusters K_n increases, as the sample size n increases. As shown, e.g., in [61], if we introduce

$$c_n(\sigma) = \begin{cases} 1 & \sigma < 0 \\ \log n & \sigma = 0 \\ n^{\sigma} & \sigma \in (0, 1) \end{cases}$$

for any $n \geq 1$, then

$$\frac{K_n}{c_n(\sigma)} \xrightarrow{\text{a.s.}} S_\sigma \tag{16}$$

as $n \to \infty$. The limiting random variable S_σ is termed σ–*diversity*. See [62] for details. It is worth noting that if $\tilde{p}$ is the Dirichlet process with parameter measure θP^*, the σ–diversity is degenerate on the total mass $\theta > 0$ and $K_n \sim \theta \log n$, for n large enough, almost surely. This special case was pointed out in [39]. The larger σ, the faster the rate of increase of K_n or, in other terms, the more new values are generated. Clearly, the case where $\sigma < 0$ corresponds to a model accommodating for a finite number of distinct species in the population.

The combined effect of the reinforcement mechanism and the increase in the rate at which new values are generated, both driven by σ, is best visualized by looking at the special case of the PY process. By close inspection of their predictive distributions (11) one notes that a new value, thus with frequency 1, entering the conditioning sample produces two effects: it is assigned a mass proportional to $(1-\sigma)$, instead of 1, in the empirical component of the predictive and, correspondingly, a mass proportional to σ is added to the probability of generating a new value. Therefore, if $\sigma > 0$, new values are assigned a mass which is less than proportional to their cluster size (that is 1) and the remaining mass is added to the probability of generating a new value. The first phenomenon gives rise to the reinforcement mechanism described above: if the new value is, then, re–observed it increases the associated mass by a quantity which is now proportional to 1, and not less than proportional. The second effect implies that if X_{n+1} is new, the probability of generating yet another new value, which overall still decreases as a function of n, is increased by a factor of $\sigma/(\theta + n + 1)$. To sum up, the larger σ the stronger is the reinforcement mechanism and at the same time the higher is the probability of generating a new value, which intuitively explains why one then obtains a growth rate of n^σ for K_n. If $\sigma < 0$ things work the other way round and one sees that each new generated value decreases the probability of generating further new values, thus providing intuition for the fact that in the end only a finite number of values will be generated. If $\sigma = 0$, which corresponds to the Dirichlet process and mixtures of the Dirichlet process over the parameter θ, everything is proportional to the cluster sizes which do not alter the probability of generating new values. As for another instance of a Gibbs–type prior, namely the normalized generalized gamma process that will be discussed later, a mechanism analogous to the PY process with $\sigma \in (0, 1)$ can be identified though the proportionality constants that rescale the masses are different due to the difference of the underlying $V_{n,k}$'s.

2.1 Connections between Gibbs–type priors and product partition models

There is also a close connection between Gibbs–type priors, and in particular the random partitions they induce, and exchangeable product partition models. The latter were introduced by [32] and further studied, among others, by [2, 65]. If Π_n represents a random partition of the set of integers $\{1, \ldots, n\}$, a product partition model corresponds to a probability distribution for Π_n represented as follows

$$\mathrm{P}(\Pi_n = \{S_1, \ldots, S_k\}) \propto \prod_{i=1}^{k} \rho(S_i) \tag{17}$$

where $\rho(\,\cdot\,)$ is termed *cohesion function*. Now, let $|S| = \mathrm{card}(S)$ and impose the cohesion function $\rho(\,\cdot\,)$ to depend only on the cardinality of the set S, that is $\rho(S_i) := \rho(|S_i|) = \rho(n_i)$. This is a natural and reasonable choice for a cohesion function. Then the random partition in (17) is, for any $n \geq k \geq 1$, the random partition induced by an exchangeable sequence if and only if $\rho(n_i) = (1-\sigma)_{n_i-1}/n_i!$ for $i = 1, \ldots, k$ and $\sigma \in [-\infty, 1]$ with the proviso that $(1-\sigma)_{n_i-1} = 1$ when $\sigma = -\infty$ and that Π_n reduces to the singleton partition when $\sigma = 1$. This is equivalent to saying that Π_n is of Gibbs–type. Such a statement follows immediately from [28]. Therefore, random probability measures inducing exchangeable product partition models with cohesion function depending on the cardinality, i.e.

$$\begin{aligned} X_i^* | \Pi_n &\overset{\text{iid}}{\sim} P^* \qquad\qquad i = 1, \ldots, K_n \\ \Pi_n &\sim \text{ product partition distribution with } \rho(S) = \rho(|S|), \end{aligned}$$

coincide with the family of Gibbs–type priors.

2.2 Sub–classes of Gibbs–type priors

Many nonparametric priors currently used for Bayesian inference represent particular cases of Gibbs–type priors, such as the Dirichlet process and the PY family. Indeed, it can be verified that the set of weights

$$V_{n,k} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} \tag{18}$$

satisfies the recursive equation (14) if the pair (σ, θ) is such that $\sigma \in [0, 1)$ and $\theta > -\sigma$ or $\sigma < 0$ and $\theta = m|\sigma|$ for some positive integer m. These constraints identify the set of admissible values of the parameters (σ, θ). The corresponding Gibbs–type prior, identified by its EPPF (13), reduces to (7) for $\sigma = 0$ therefore leading to the Dirichlet process. For any admissible (σ, θ) the EPPF (13) coincides with (8), thus recovering the PY family. Another interesting special case of the PY process, and *a fortiori* of Gibbs–type priors, is represented by the normalized σ-stable process introduced by [38] which is obtained as a PY process with $\sigma \in (0, 1)$ and $\theta = 0$.

Before discussing other special cases of Gibbs–type priors, it is worth having a closer look at the PY family with $\sigma < 0$ and $\theta = m|\sigma|$. In this case the weights in (18) are as follows

$$V_{n,k} = \frac{|\sigma|^{k-1} \prod_{i=1}^{k-1}(m-i)}{(m|\sigma|+1)_{n-1}} \mathbb{1}_{\{1,\ldots,\min(n,m)\}}(k) \tag{19}$$

where $\mathbb{1}_A$ denotes the indicator function of set A. From (19) it is then easy to see (cfr. [62]) that the PY family with $\sigma < 0$ and $\theta = m|\sigma|$ corresponds to a population composed by m different species with proportions distributed according to a symmetric Dirichlet distribution with density function

$$f_m(p_1, \ldots, p_{m-1}) = \frac{\Gamma(m|\sigma|)}{\Gamma^m(|\sigma|)} \prod_{i=1}^{m-1} p_i^{|\sigma|-1} \left(1 - p_1 - \cdots p_{m-1}\right)^{|\sigma|-1}$$

for any $(p_1, \ldots, p_{m-1})$ such that $p_i \geq 0$ for any i and $\sum_{i=1}^{m-1} p_i \leq 1$. Such a model arises, in the Population Genetics literature, as the stationary law of a Wright–Fisher model.

The PY family with parameters $(\sigma, m|\sigma|)$ and $\sigma < 0$ is not only a distinguished special case of Gibbs–type prior with $\sigma < 0$ but actually is its basic building block. In fact, any Gibbs–type random probability measure with $\sigma < 0$ is obtained by specifying a prior π for the total number of species m in (19) and coincides with a species sampling model having a random (finite) number of species. Crucially, by [28], the reverse implication holds true as well: any Gibbs–type prior with $\sigma < 0$ is a mixture of PY processes with parameters $(\sigma, m|\sigma|)$, the mixing measure being a probability measure on the positive integers. Therefore, one can equivalently describe Gibbs–type priors with $\sigma < 0$ in terms of a mixture model as

$$\begin{aligned} (\tilde{p}_1, \ldots, \tilde{p}_{\tilde{m}-1}) \,|\, \tilde{m} &\sim f_{\tilde{m}} \\ \tilde{m} &\sim \pi. \end{aligned} \tag{20}$$

Interesting special cases arise by particular specifications of π. For instance, if

$$\pi(m) = \frac{\gamma(1-\gamma)_{m-1}}{m!} \tag{21}$$

for $m = 1, 2, \ldots$ with $\gamma \in (0, 1)$, one obtains the model introduced by Gnedin [27], which in the case of $\sigma = -1$ admits a completely explicit expression of the weights, namely

$$V_{n,k} = \frac{(k-1)!(1-\gamma)_{k-1}(\gamma)_{n-k}}{(n-1)!(1+\gamma)_{n-1}}. \tag{22}$$

The peculiar feature of such a model, which makes it of great use in applications, is that the heavy-tailedness of (21) implies a model with finite random number of species whose expected value is infinite. Other interesting models are obtained by specifying the mixing distribution as a Poisson distribution restricted to the positive integers with parameter $\lambda > 0$, i.e.

$$\pi(m) = \frac{\mathrm{e}^{-\lambda}}{1 - \mathrm{e}^{-\lambda}} \frac{\lambda^m}{m!} \tag{23}$$

for $m = 1, 2, \ldots$, or as a geometric mixing distribution

$$\pi(m) = (1 - \eta)\eta^{m-1} \tag{24}$$

for some $\eta \in (0, 1)$ and $m = 1, 2, \ldots$. These will be further discussed in Section 5.2.

Another important sub–class of Gibbs–type priors is the normalized generalized Gamma (NGG) process which corresponds to

$$V_{n,k} = \frac{\mathrm{e}^{\beta}\, \sigma^{k-1}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i\, \beta^{i/\sigma}\, \Gamma\left(k - \frac{i}{\sigma};\, \beta\right), \tag{25}$$

where $\sigma \in (0, 1)$, $\beta > 0$ and $\Gamma(x, a) = \int_x^{\infty} s^{a-1}\, \mathrm{e}^{-s}\, \mathrm{d}s$ is the incomplete gamma function. Also the NGG process contains several interesting special cases: if $\sigma \to 0$ one obtains the Dirichlet process, whereas $\sigma = 1/2$ yields the normalized inverse Gaussian process (N–IG) of [40], which stands out for the availability of a closed form expression of its finite–dimensional distributions. Furthermore, if $\beta = 0$, the normalized σ-stable process is also recovered from the NGG. See [67, 35, 43]. The name attributed to this particular Gibbs–type prior is motivated by the fact that it can be defined by normalizing a generalized gamma completely random measure introduced in [3] and it therefore also belongs to the class of normalized random measures with independent increments (NRMI) introduced in [67]. Interestingly, as shown in [48], it turns out to be the only random probability measure belonging to both classes, NRMIs and Gibbs–type priors. All other NRMIs, such as for instance the generalized Dirichlet process in [41, 20], are not of Gibbs–type.

In addition to specific examples described so far and still for the case of $\sigma > 0$, one might wonder whether starting from the prediction rules (15) it is possible to identify the class of random probability measures generating them. The answer is affirmative and, as shown in [28], they coincide with the so–called σ–stable Poisson–Kingman models, which are obtained by means of a particular transformation of σ–stable completely random measures. The technical background needed for precisely defining such models goes beyond the scope of this review and we refer the interested reader to [61, 28]. For our purposes it is enough to note that the derivation of posterior quantities in this setting represents a challenging issue, which has not found a satisfactory solution to date.

So far we have provided various motivations, of theoretical and practical relevance, for the use of Gibbs–type priors and the sub–classes discussed in this section provide a glimpse of the nice and simple structure they share. Nonetheless, we still need to shed some light on another distributional aspect which is important for assessing their suitability for nonparametric inference, namely their support. As mentioned in the Section 1, a large topological support is a desirable property for a nonparametric prior since the essence of being nonparametric can be

associated to the fact of assigning prior positive probability to as many "candidate models" as possible. When considering the weak topological support, which is the most natural in this framework, it can be shown (see [8]) that "genuinely nonparametric" Gibbs–type priors comply with this requirement and have full weak support: in other terms, any weak neighborhood of any distribution in $\mathbf{P}_{\mathbb{X}}$ will have a priori positive probability. Here by "genuinely nonparametric" we mean Gibbs–type priors whose realizations are discrete distributions for which the number of support points is not bounded. This essentially boils down to considering Gibbs–type priors either with $\sigma \geq 0$ or with $\sigma < 0$ and unbounded support of the prior π on the number of components in (20). Such priors can be shown to possess the full weak support property, i.e. their topological support coincides with the space of probability measures whose support is included in the support of the prior guess P^*. In particular, if the support of P^* coincides with $\mathbb{X}$, the support of Q is the whole space $\mathbf{P}_{\mathbb{X}}$.

3 Hierarchical mixture models based on Gibbs–type priors

As outlined in the Introduction an important application of discrete random probability measures and, then, of Gibbs–type priors occurs within hierarchical mixture models of the type (12): this corresponds to assuming exchangeable data $(Y_i)_{i\geq 1}$ from a random density defined by

$$\tilde{f}(y) = \int_{\mathbb{X}} f(y \mid x)\tilde{p}(\mathrm{d}x). \tag{26}$$

In particular, when $\tilde{p}$ follows a discrete prior Q, a key ingredient for prior and posterior inferences is the corresponding EPPF. Indeed, given a set of observables $Y_1, \ldots, Y_n$ modeled according to the above random density, the clustering structure among the latent variables $X_1, \ldots, X_n$ drives both the posterior distribution on the number of components and the posterior density estimation. In particular,

$$\mathrm{P}(K_n = k \mid Y_1, \ldots, Y_n) \propto \sum_{\mathsf{p}_n \in \mathcal{P}^k_{[n]}} p^{(n)}_k(n_1, \ldots, n_k) \prod_{j=1}^{k} \int_{\mathbb{X}} \prod_{i \in \mathcal{C}_j} f(y_i \mid x_j) P^*(\mathrm{d}x_j)$$

where $\mathcal{P}^k_{[n]}$ is the set of all partitions p_n of the n latent variables into k disjoint clusters and $\mathcal{C}_j$ identifies the indices of those latent variables x_i that belong to the j–th cluster in the partition $\mathsf{p}_n \in \mathcal{P}^k_{[n]}$. Therefore, the choice of Q or, equivalently, of the corresponding EPPF, is crucial for nonparametric Bayesian inferences in this framework and it can be further appreciated through some numerical illustrations we are going to provide later on in this section.

An appealing feature of Gibbs–type priors is their ability to control the prior mass allocated to different partitions through the reinforcement mechanism induced by the parameter σ and

described in Section 2. This can be appreciated by looking at the induced (prior) distribution on the number K_n of clusters. First note that the determination of the distribution of K_n follows from a marginalization of (13) and leads to

$$\mathrm{P}(K_n = k) = \frac{V_{n,k}}{\sigma^k}\,\mathscr{C}(n,k;\sigma) \tag{27}$$

with

$$\mathscr{C}(n,k;\sigma) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(-i\sigma)_n$$

denoting a generalized factorial coefficient. See [5] for details on $\mathscr{C}(n,k;\sigma)$. Substituting expressions (18) and (25) in (27) leads to the prior distributions on the number of different components for the PY and the NGG processes, respectively. Letting $\sigma \to 0$ in either of the resulting expressions one obtains the corresponding distribution for the Dirichlet process case

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

with $s(n,k)$ denoting the Stirling number of the first type. See [5].

A graphical display of these distributions is best suited to highlight their differences. To this end, fix $n = 50$ and consider the corresponding distributions of the number of components in the three above cases. For the Dirichlet process it is well–known that the total mass parameter θ controls the location of the distribution of K_{50}: larger values of θ lead to a right-shift of the distribution implying an (a priori) larger number of components essentially affecting its dispersion. In both the PY process and NGG cases the role of controlling the location is played by θ and β, respectively. Hence, it is interesting to look at the additional parameter σ. Figure 1 concerns the NGG process and displays the distribution of K_{50} for a fixed value of β and with σ ranging between 0.2 and 0.8. Note that in Figures 1, 2 and 3 the probability masses are connected by straight lines only for visual simplification. From Figure 1 it is evident that the addition of σ allows to control the flatness, or the variability, of the distribution of K_{50} thus yielding a higher degree of flexibility for the model. A similar behavior appears in the PY process. Hence, replacing the Dirichlet process with a Gibbs–type prior characterized by a value of σ in $(0,1)$ allows for a better control of the informativeness of the prior number of groups, since a larger σ flattens the prior. To better visualize this fact, it is useful to consider a simple comparative example. In addition to $n = 50$, suppose that the prior expected number of clusters is 25. This implies that a reasonable criterion for eliciting the parameters of a nonparametric prior is to fix them in a way such that $\mathrm{E}(K_{50}) = 25$. We compare five different models: Dirichlet process with $\theta = 19.233$, PY processes with $(\sigma,\theta) = (0.25, 12.2157)$ and $(\sigma,\theta) = (0.73001, 1)$, and NGG processes with $(\sigma,\beta) = (0.25, 48.4185)$ and $(\sigma,\beta) = (0.7353, 1)$, where all reported parameters

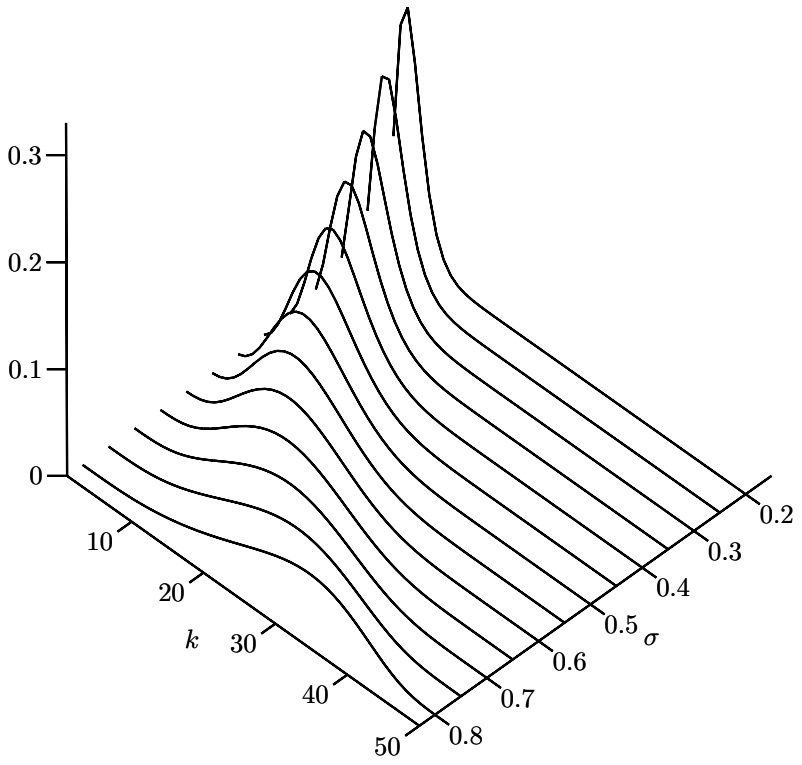

Figure 1: Prior distributions on the number of groups corresponding to the NGG process with $n = 50, \beta = 1$ and $\sigma = 0.2, 0.3, \ldots, 0.7$ and $\sigma = 0.8$.

are chosen so that $E(K_{50}) = 25$. The corresponding distributions of K_{50} are displayed in Figure 2. Clearly, by increasing the value of σ one obtains a less informative distribution on K_{50}: when moving from $\sigma = 0$ to $\sigma \approx 0.73$ the distribution of K_{50} becomes flatter, exhibiting a larger variability. The Dirichlet process, instead, implies a highly peaked distribution of K_{50}, which in terms of prior specification implies the need for a reliable prior information on the number of clusters, which is often unavailable. Furthermore, the PY and NGG processes have a similar behavior with the latter producing slightly lighter tails.

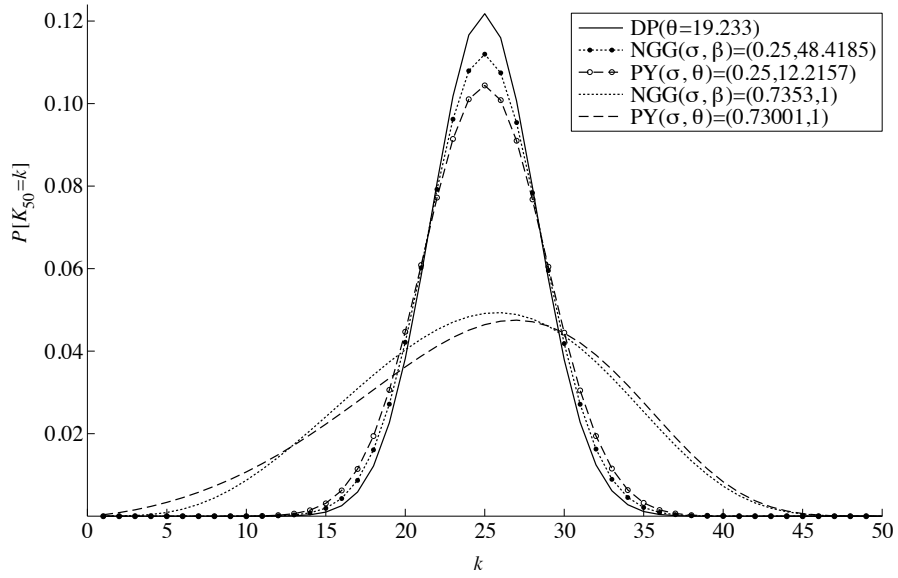

Figure 2: Prior distributions on the number of clusters corresponding to the Dirichlet (DP), the Pitman–Yor (PY) and the normalized generalized gamma (NGG) processes. The values of the parameters are set in such a way that $\mathrm{E}(K_{50}) = 25$.

Let us now take a further step and compare the above five processes in a toy example to have a closer look at the implication of such prior specifications on posterior inferences on the clustering structure of the data. To this end, assume that $n = 50$ observations are drawn from a uniform mixture of two well-separated Gaussian distributions, $\mathsf{N}(1, 0.2)$ and $\mathsf{N}(10, 0.2)$. From a classification perspective these data clearly identify two groups. We model them with the following nonparametric mixture model with standard specification

$$\begin{aligned} (Y_i \mid m_i, v_i) &\overset{\text{ind}}{\sim} \mathsf{N}(m_i, v_i), & i = 1, \ldots, n \\ (m_i, v_i \mid \tilde{p}) &\overset{\text{iid}}{\sim} \tilde{p} & i = 1, \ldots, n \\ \tilde{p} &\sim Q \end{aligned}$$

with Q corresponding to the five processes above and $P^*(\mathrm{d}m, \mathrm{d}v) = \mathsf{N}(m \mid \mu, \tau v^{-1})\mathsf{Ga}(v \mid 2, 1)\,\mathrm{d}m\,\mathrm{d}v$, where $\mathsf{N}(\cdot \mid a, b)$ denotes the Gaussian density with mean a and variance $b > 0$ and $\mathsf{Ga}(\cdot|c, d)$ is the density corresponding to a Gamma distribution with mean c/d. A further hierarchy is assumed for μ and τ, i.e. $\mu \sim \mathsf{N}(0, 0.001)$ and $\tau^{-1} \sim \mathsf{Ga}(1, 100)$. In this setup the parameter specification for the five processes (chosen so that $E(K_{50}) = 25$) corresponds to a prior opinion on K_{50} remarkably far from the true number of components in the mixture density that has generated the data. Given such a wrong prior specification one then wonders whether the models possess enough flexibility to shift a posteriori towards the correct number of components, namely 2. The results are based on 100000 iterations after 5000 of burn in adopting a standard marginal MCMC algorithm with acceleration step. See [12, 51] for further details on this algorithm.

Figure 3 depicts the posterior distribution on the number of mixture components. The most important thing to note is that a larger σ leads to better posterior estimates. Both the PY and NGG processes with $\sigma = 0.73$, have been able to shift most of the mass towards a very low number of components with the PY process exhibiting a slightly better performance. See also Table 1 for a display of the numerical values of posterior probabilities associated to the possible values of K_{50}. This shows how a stronger reinforcement mechanism, implying a flatter distribution of K_n, allows to recover more effectively the correct number of components. In contrast, the Dirichlet process is stuck around 10 components, since the high peakedness of its prior on K_n prevents it from overruling completely the wrong prior information.

Finally, it is important to point out that the above considerations concerning the advantages of the additional parameter σ hold beyond the present toy example since they represent structural properties of the models, which are by now well–understood thanks to several analytical results and computational analyses. See, e.g., [43]. As far as the estimates of the density $\tilde{f}$ in (26) are concerned, these are displayed in in Figure 4. Even if the considerable heterogeneity in the posterior inferences on the number of components is not reflected by density estimates, one can still appreciate a slightly better performance of the NGG and PY processes with $\sigma = 0.73001$ since they show a closer adherence to the depicted true density.

4 Prediction in species sampling problems

As already mentioned in Section 1, Gibbs–type priors are a powerful tool for addressing prediction and estimation in species sampling problems when observations are recorded from a population composed of individuals belonging to different types or species. This situation occurs in many applied research areas, including genetics, biology, ecology, economics and lin-

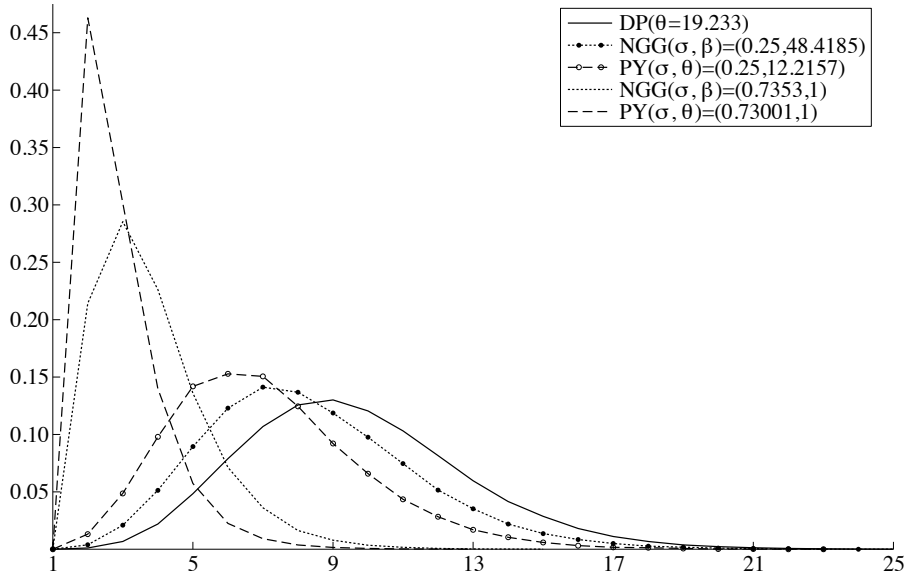

Figure 3: Posterior distributions on the number of components corresponding to mixtures of the Dirichlet (DP), the Pitman–Yor (PY) and the normalized generalized gamma (NGG) processes with $n = 50$ and parameters set so that $\mathrm{E}(K_{50}) = 25$.

k	DP(19.233)	NGG(0.25,48.4185)	PY(0.25,12.216)	NGG(0.7353,1)	PY(0.73001,1)
1	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0011	0.0039	0.0132	0.2143	**0.4630**
3	0.0068	0.0209	0.0487	**0.2854**	0.3015
4	0.0220	0.0514	0.0979	0.2263	0.1399
5	0.0484	0.0894	0.1419	0.1360	0.0573
6	0.0789	0.1229	**0.1528**	0.0713	0.0225
7	0.1069	**0.1412**	0.1506	0.0361	0.0092
8	0.1257	0.1368	0.1245	0.0163	0.0037
9	**0.1301**	0.1187	0.0921	0.0079	0.0016
10	0.1205	0.0976	0.0659	0.0035	0.0007
11	0.1031	0.0746	0.0435	0.0017	0.0003
12	0.0816	0.0516	0.0283	0.0007	0.0002
13	0.0597	0.0353	0.0170	0.0003	0.0001
≥ 14	0.1151	0.0556	0.0237	0.0004	0.0001

Table 1: Posterior distributions on the number of components arising from mixtures of the Dirichlet process (DP), the normalized generalized Gamma (NGG) process and the Pitman-Yor (PY) process centered such that the prior expected value of the number of components is 25 with the sample size $n = 50$.

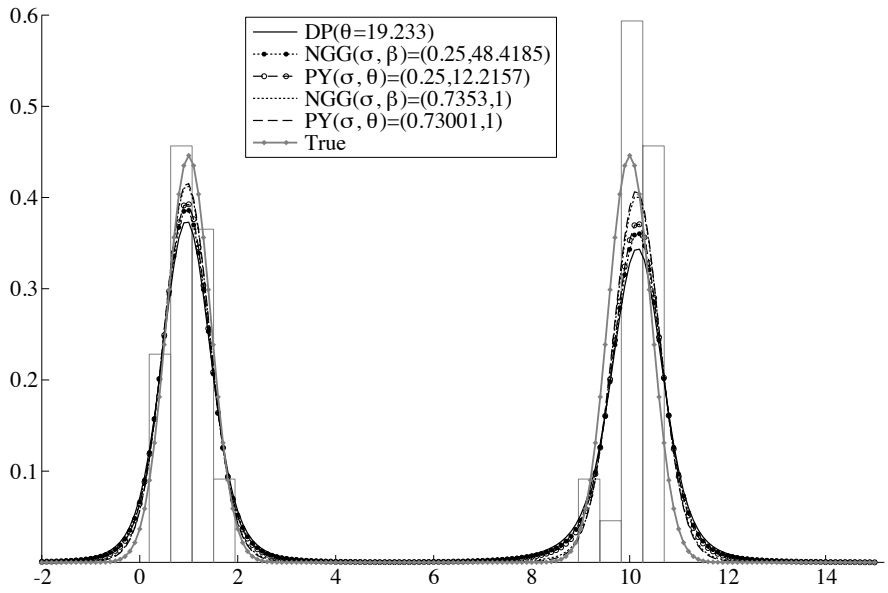

Figure 4: Density estimates corresponding to the 5 mixture models that have been considered.

guistics. Hence, in this section we will think of the observations X_i in (1) as species labels. The sample data $X_1, \ldots, X_n$ one can rely on for inferential purposes yield the following pieces of information: the number K_n of distinct species in the sample; the observed species labels $X_1^*, \ldots, X_{K_n}^*$; the frequencies $\boldsymbol{N}_n = (N_{1,n}, \ldots, N_{K_n,n})$ associated to each of the observed species. Note that the last quantity can be alternatively reformulated in terms of the frequency counts $\boldsymbol{M}_n = (M_{1,n}, \ldots, M_{n,n})$, where $M_{i,n}$ is the number of species that have appeared with frequency i in the observed sample. It is obvious that these vectors must satisfy the following constraints:

$$\sum_{i=1}^{K_n} N_{i,n} = n, \qquad \sum_{i=1}^{n} M_{i,n} = K_n, \qquad \sum_{i=1}^{n} i M_{i,n} = n.$$

In such problems species labels are typically not of interest, and the data can be efficiently summarized by either $\boldsymbol{N}_n$ or $\boldsymbol{M}_n$, namely the partition they form. Since the EPPF (5) can also be seen as the partition distribution induced by a sample, it is natural to resort to the class of priors which have the most general yet tractable partition distribution. This naturally leads to work with Gibbs–type priors which are characterized by the product–form EPPF (13).

In this framework a novel Bayesian nonparametric methodology for deriving estimators of quantities related to an additional unobserved sample $X_{n+1}, \ldots, X_{n+m}$ from $\tilde{p}$, conditional on $X_1, \ldots, X_n$, has been proposed in [42] and [47]. An important applied problem is the estimation of the so-called overall *species variety*, which can be measured by estimating the number

$K_m^{(n)} = K_{n+m} - K_n$ of "new" distinct species that will be observed in the additional sample. A generalization has been recently derived in [18] and it corresponds to the estimator of the so-called *rare species variety*

$$\widehat{M}_m^{(n)}(\tau) = \sum_{i=1}^{\tau} \widehat{M}_{i,m}^{(n)} = \sum_{i=1}^{\tau} \mathrm{E}(M_{i,n+m} \,|\, X_1, \ldots, X_n) \tag{28}$$

namely the number of distinct species with frequency less than or equal to a specific threshold of abundance τ that will be detected in the additional sample of size m. Note that both the estimator of $K_m^{(n)}$, denoted by $\widehat{K}_m^{(n)}$, and $\widehat{M}_m^{(n)}(\tau)$ can be thought of as *global measures* of overall and rare species variety respectively, since they are referred to the whole additional sample of size m. On the other hand, one may also need the corresponding *local measures*, which can be quantified in terms of the *discovery probability* at step $(n+m+1)$ of the sampling process. Bayesian estimators of the latter have been determined in [19]. More specifically, if $\Delta_{i,n+m}$ is the set including species labels that appear with frequency $i \geq 0$ in the enlarged sample $X_1, \ldots, X_{n+m}$, one is interested in estimating

$$U_{n+m,i} = \mathrm{P}(X_{n+m+1} \in \Delta_{i,n+m} \,|\, X_1, \ldots, X_n). \tag{29}$$

An estimator will be obtained by averaging over all possible realizations of the unobserved additional sample $X_{n+1}, \ldots, X_{n+m}$, conditional on the basic sample $X_1, \ldots, X_n$. Here $U_{n+m,0}$ stands for the probability of sampling a new species at step $(n+m+1)$, whereas $\sum_{i=0}^{\tau} U_{n+m,i}$ for the probability of sampling either a species not yet observed or one with frequency less than τ. Such local estimates are relevant, for example, in determining the size m of the additional sample $X_{n+1}, \ldots, X_{n+m}$: a possible criterion consists in fixing m equal to the maximum possible value for which the estimated discovery probability of new or rare species is above a certain threshold probability.

If the population is composed by a large number of unknown species (genes, agents, categories etc.) and the basic sample $X_1, \ldots, X_n$ displays only a small fraction of the species present in the population, Gibbs–type priors with $\sigma \in [0, 1)$ are particularly suited. An effective and popular example is offered by the analysis of Expressed Sequence Tags [44] or Serial Analysis of Gene Expression [31] data. Indeed, in these experiments either complementary DNA (cDNA) libraries or messenger RNA (mRNA) populations are considered and typical goals consist in identifying the genes they are composed of, the relative frequencies of such genes and also in comparing libraries/populations in terms of diversity. Due to time and cost constraints only a small portion of the whole library or population is typically sequenced and prediction is required to assess the overall characteristics. A similar experimental framework takes place in biological applications such as, for example, in the analysis of T-cell identification problems (see [71]). In this case

one can characterize the immunological status of an organism by estimating the number of distinct clonotypes in a T-cell repertoire and the clonal size distribution, which is the frequency of clonotypes with a certain clonal size. In contrast, if the population has a limited number of species, a common situation in Ecology, Gibbs–type models with $\sigma < 0$ are more appropriate [15]. In what follows, for brevity we will deal with the case $\sigma \in [0, 1)$. This implies that, when specializing the results to the PY process, one also has $\theta > -\sigma$. However, it is to be noted that most of the displayed findings carry over to the case of $\sigma < 0$.

On this topic there exists a well–estalibished frequentist literature. The most relevant contributions typically draw inspiration from papers by I.J. Good ([29]) and I.J. Good and G.H. Toulmin ([30]). See, e.g., [53] and [54]. For example, the popular Turing estimator for the discovery probability (displayed in [29] and credited to A. Turing) is

$$\check{U}_{n,i} = (i+1)\frac{M_{i+1,n}}{n}. \tag{30}$$

For $i = 0$ it provides an estimator for the probability that the $(n+1)$–th observation is new. Equivalently, $1 - \check{U}_{n,0}$ provides an estimator of the sample coverage, namely the proportion of species observed in the sample, which is an important quantity in many applied frameworks. Moreover, estimates of $K_m^{(n)}$ and of the discovery probability $U_{n+m,0}$ for any $m \geq 1$ have been established in [30] and shall be henceforth termed Good–Toulmin estimators. They coincide with

$$\check{U}_{n+m,0} = n^{-1} \sum_{i=1}^{\infty} (-\lambda)^{i-1}\, i\, M_{i,n}, \qquad \check{K}_m^{(n)} = \sum_{i=1}^{\infty} (-1)^{i-1} \lambda^i\, M_{i,n} \tag{31}$$

where $\lambda = m/n$. Due to the alternating sign of the sums, when λ is large they can yield inadmissible numerical values. This instability arises even for values of m moderately large with respect to n, typically m greater than n is enough for it to appear. An illustration is provided in Section 4.1. On the other hand, we are not aware of frequentist estimators of the discovery probabilities $U_{n+m,i}$ when both m and i are positive integers.

4.1 Bayesian inference on overall species variety

Based on the EPPF (13), an explicit expression for the distribution of the number of "new" distinct species observed in the additional sample, $K_m^{(n)}$, conditional on the information provided by $X_1, \ldots, X_n$, has been determined in [42] and is given by

$$\mathrm{P}(K_m^{(n)} = j \,|\, X_1, \ldots, X_n) = \frac{V_{n+m,k+j}}{V_{n,k}} \frac{\mathscr{C}(m, j; \sigma, -n + k\sigma)}{\sigma^j} \tag{32}$$

where $X_1, \ldots, X_n$ is partitioned into $K_n = k$ clusters with respective frequencies $n_1, \ldots, n_k$ and $\mathscr{C}(n, k; \sigma, \gamma)$ is the non–central generalized factorial coefficient

$$\mathscr{C}(m, j; \sigma, -n + k\sigma) = (j!)^{-1} \sum_{r=0}^{j} (-1)^r \binom{j}{r} (n - \sigma(r + k))_m.$$

See [5]. An important implication of (32) is the sufficiency of K_n for predicting the number of "new" distinct species. The expression in (32) serves then as a basis for determining the Bayesian nonparametric estimator, with respect to a squared loss function, of the overall species variety as

$$\widehat{K}_m^{(n)} = \mathrm{E}(K_m^{(n)} \,|\, K_n = k,\ \boldsymbol{N}_n = \boldsymbol{n}) \tag{33}$$

with $\boldsymbol{n} = (n_1, \ldots, n_k)$. This can be seen as a Bayesian counterpart of the Good–Toulmin estimator.

When $\tilde{p}$ is the PY process, with parameter (σ, θ), (32) becomes

$$\mathrm{P}(K_m^{(n)} = j \,|\, K_n = k, \boldsymbol{N}_n = \boldsymbol{n}) = \frac{(\theta/\sigma + k)_j}{(\theta + n)_m} \mathscr{C}(m, j; \sigma, -n + k\sigma). \tag{34}$$

As shown in [16], the estimator for $K_m^{(n)}$ in (33) then reduces to

$$\widehat{K}_m^{(n)} = \left(k + \frac{\theta}{\sigma}\right) \left(\frac{(\theta + n + \sigma)_m}{(\theta + n)_m} - 1\right). \tag{35}$$

The main advantage of (35), and of other estimators devised in [16] for measuring the overall species variety, is that they are explicit and can be exactly evaluated even when the size m of the additional sample is large compared to the size of the basic sample n. This happens, for instance, in genomic applications where one has to deal with relevant portions of cDNA libraries consisting of millions of genes.

In several applied contexts it is useful to accompany point estimates such as (35) with the corresponding credible intervals. These can be easily derived from the conditional distribution (34). However, if the sample sizes are very large the computation of the non-central generalized factorial coefficient may become cumbersome. To circumvent such a problem one could resort to asymptotic credible intervals. This motivates, also from a practical point of view, the asymptotic analysis of $K_m^{(n)}$, conditional on K_n, for a fixed n and as $m \to \infty$, provided in [16]. Let f_σ stand for the density function of a positive σ-stable random variable, and let U_q, for any $q \geq 0$, be a positive random variable characterized by the following density function

$$f_{U_q}(u) = \frac{\Gamma(q\sigma + 1)}{\sigma\Gamma(q + 1)} u^{q - 1 - 1/\sigma} f_\sigma\left(u^{-1/\sigma}\right).$$

Moreover, set $B_{a,b}$ as a beta random variable with parameters (a, b). Conditional on the information provided by $X_1, \ldots, X_n$, one has

$$\frac{K_m^{(n)}}{m^\sigma} \xrightarrow{\text{a.s.}} Z_{n,k}, \tag{36}$$

as $m \to \infty$, where $Z_{n,k} \overset{\text{d}}{=} B_{k+\theta/\sigma,n/\sigma-k} U_{(\theta+n)/\sigma}$ with $B_{k+\theta/\sigma,n/\sigma-k}$ and $U_{(\theta+n)/\sigma}$ being independent. A similar asymptotic result has been obtained also for the NGG process in [17]. Note that the σ-diversity discussed in (16) can be recovered from (36) by setting $n = k = 0$. Turning back to the practical uses of (36) for the determination of credible asymptotic intervals for $K_m^{(n)}$, it is apparent that one still needs to derive the quantiles of $Z_{n,k}$. From an analytical point of view this is a challenging task, which nonetheless can be avoided by resorting to straightforward computational algorithms that allow to sample from the limit random variable $Z_{n,k}$, and thus to approximate the quantiles. See [16, 9] for details.

4.2 Bayesian inference on rare species variety

The problem of deriving estimators for the rare species variety has been recently considered in [17] and [19]. One of such estimators is represented by the number of distinct species with frequencies less than or equal to a specified threshold of abundance τ, for any $\tau \leq n + m$, that are generated by the additional sample, as displayed in (28). The determination of $\widehat{M}_{i,m}^{(n)}$, under a square loss function, is eased by resorting to the decomposition

$$\widehat{M}_{i,m}^{(n)} = \widehat{N}_{i,m}^{(n)} + \widehat{O}_{i,m}^{(n)}$$

where $\widehat{N}_{i,m}^{(n)}$ is the estimator of the number of "new" distinct species with frequency i not detected in $X_1, \ldots, X_n$ and $\widehat{O}_{i,m}^{(n)}$ is the estimator of the number of "old" distinct species (i.e. included in $X_1, \ldots, X_n$) that appear with frequency i in the enlarged sample. This implies that $\widehat{M}_m^{(n)}(\tau)$ in (28) arises as the sum of two well-defined quantities: (i) the estimator of the number of "new" distinct species with frequencies less than or equal to $\tau \leq m$ and generated by the additional sample, i.e. $\widehat{N}_m^{(n)}(\tau) = \sum_{i=1}^{\tau} \widehat{N}_{i,m}^{(n)}$; (ii) the estimator of the number of "old" distinct species with frequencies less than or equal to $\tau \leq n + m$ and generated by updating the frequencies of the partition induced by the basic sample with the additional sample, i.e. $\widehat{O}_m^{(n)}(\tau) := \sum_{i=1}^{\tau} \widehat{O}_{i,m}^{(n)}$. It is apparent that if $\tau = m$ one obtains $\widehat{N}_m^{(n)}(\tau) = \widehat{K}_m^{(n)}$. In this respect, the concept of rare species variety can be interpreted as a generalization of the concept of overall species variety.

A result in [17] gives explicit expressions of the moments, of any order, of both the number of "new" species with frequency i in $X_{n+1}, \ldots, X_{n+m}$ and of the number of "old" species with frequency i in the enlarged sample $X_1, \ldots, X_{n+m}$. From these one deduces $\widehat{N}_{i,m}^{(n)}$ and $\widehat{O}_{i,m}^{(n)}$ thus

obtaining an estimator of rare species variety. It can be seen that

$$\widehat{O}_{i,m}^{(n)} = \sum_{t=1}^{i} \binom{m}{i-t} M_{t,n} \, (t-\sigma)_{i-t} \sum_{j=0}^{m} \frac{V_{n+m,k+j}}{V_{n,k}} \frac{\mathscr{C}\left(m-(i-t), j; \sigma, -n+t+(k-1)\sigma\right)}{\sigma^j}. \quad (37)$$

From (37), it is clear that $(K_n, M_{1,n}, \ldots, M_{\tau,n})$ is a sufficient statistic for predicting the number of "old" distinct species with frequency less than or equal to τ. Moreover,

$$\widehat{N}_{i,m}^{(n)} = \binom{m}{i} (1-\sigma)_{i-1} \sum_{j=0}^{i} \frac{V_{n+m,k+j+1}}{V_{n,k}} \frac{\mathscr{C}(m-i, j; \sigma, -n+k\sigma)}{\sigma^j} \quad (38)$$

and K_n is sufficient for predicting the number of "new" distinct species with frequency less than or equal to τ. Finally, $\widehat{M}_{i,m}^{(n)}$ can be derived as the sum of the estimators in (37) and (38) and, then, $\widehat{M}_{m}^{(n)}(\tau)$ from (28).

If we focus on the special case where the Gibbs–type prior is the PY process, then the expressions (37) and (38) considerably simplify and reduce to

$$\widehat{O}_{i,m}^{(n)} = \sum_{t=1}^{i} \binom{m}{i-t} M_{t,n} \, (t-\sigma)_{i-t} \, \frac{(\theta+n-t+\sigma)_{m-(i-t)}}{(\theta+n)_m}$$

$$\widehat{N}_{i,m}^{(n)} = \binom{m}{i} (1-\sigma)_{i-1} (\theta+k\sigma) \, \frac{(\theta+n+\sigma)_{m-i}}{(\theta+n)_m}.$$

It is worth noting that the determination of estimators of the rare species variety poses a major technical hurdle that does not occur when estimating the overall species variety. Indeed, one has to consider all possible modifications, induced by the observations in the additional sample, on the frequencies of the species detected in the basic sample.

In the special PY process case, one can establish the asymptotic behavior of rare species variety as $m \to \infty$. This is somehow in the spirit of (36) in the context of overall species variety. In this case, as shown in [17], one has for any $i \geq 1$

$$\frac{M_{i,n+m} \,|\, X_1, \ldots, X_n}{m^\sigma} \xrightarrow{d} \frac{\sigma(1-\sigma)_{i-1}}{i!} \, Z_{n,k},$$

as $m \to \infty$, where $Z_{n,k}$ is the limit random variable introduced in (36) and $\xrightarrow{d}$ stands for convergence in distribution. This implies that K_n is asymptotically sufficient for predicting the number of distinct species with frequency i that are generated after observing the additional sample, conditional on the information provided by the random partition of the basic sample.

Rare species variety can be further assessed locally in terms of discovery probabilities $U_{n+m,i}$ as defined in (29). This leads to the proposal of Bayesian nonparametric counterparts to the Turing and the Good–Toulmin estimators that are recalled in (30) and in (31), respectively. If

one assumes a square loss function, then an estimator of $U_{n,i}$ is

$$\widehat{U}_{n,i} = \frac{V_{n+1,k}}{V_{n,k}} \, (i - \sigma) \, M_{i,n} \tag{39}$$

for any $i \leq n$, while the discovery probability of a new species, i.e. $i = 0$, can be easily deduced from the predictive distribution (15) and is given by $\widehat{U}_{n,0} = V_{n+1,k+1}/V_{n,k}$. Note that, unlike the Turing estimator, $\widehat{U}_{n,i}$ depends on $M_{i,n}$ which seems to be more coherent with what intuition would suggest. If we now let $m \geq 1$ and $j \leq n + m$, an estimator of the discovery probability turns out to be[1]

$$\begin{aligned} \widehat{U}_{n+m,i} = \sum_{l=1}^{i} M_{l,n} (l - \sigma)_{i+1-l} \binom{m}{i-l} Q_{m,i}^{(n,k)}(l, 0, l - \sigma) \\ + \sigma (1 - \sigma)_i \binom{m}{i} Q_{m,i}^{(n,k)}(1, 1, 0) \end{aligned} \tag{40}$$

where

$$Q_{m,i}^{(n,k)}(\alpha, \beta, \gamma) = \sum_{r=\beta}^{m-i+\alpha} \frac{V_{n+m+1,k+r}}{V_{n,k}} \frac{\mathscr{C}(m - i + \alpha - \beta, r - \beta; \sigma, -n + k\sigma + \gamma)}{\sigma^r}.$$

When $i = 0$ and $m \geq 1$, this yields the following Bayesian analog of the Good–Toulmin estimator for the probability of discovering a new species

$$\widehat{U}_{n+m,0} = \sum_{i=0}^{m} \frac{V_{n+m+1,k+i+1}}{V_{n,k}} \, \frac{\mathscr{C}(m, i; \sigma, -n + k\sigma)}{\sigma^i} \tag{41}$$

whereas there is no frequentist counterpart to (40) when both m and k are positive integers. From these closed form expressions one can deduce a further measure of rare species variety as $\widehat{U}_{n+m}(\tau) = \sum_{i=0}^{\tau} \widehat{U}_{n+m,i}$.

If one adopts a specification of the $V_{n,k}$'s yielding a PY process, nice and simple forms of the estimators of the discovery probabilities and of rare species variety are obtained. For example, the analog of the Turing estimator (30) reduces to

$$\widehat{U}_{n,i} = \frac{i - \sigma}{\theta + n} \, M_{i,n}$$

and the Bayesian counterpart (41) of the Good–Toulmin estimator coincides with

$$\widehat{U}_{n+m,0} = \frac{\theta + k\sigma}{\theta + n} \, \frac{(\theta + n + \sigma)_m}{(\theta + n + 1)_m}. \tag{42}$$

[1]The estimators in (40) and (43) slightly differ from those in [19], since the latter contain a minor inaccuracy that we have corrected here.

Library	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Naegleria Aerobic	346	57	19	12	9	5	4	2	4	5	4
Naegleria Anaerobic	491	72	30	9	13	5	3	1	2	0	1
Library	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*27*	*55*	j	n
Naegleria Aerobic	1	0	0	0	1	1	1	1	1	473	959
Naegleria Anaerobic	0	1	3	0	0	0	0	0	0	631	969

Table 2: ESTs from two *Naegleria gruberi* libraries. Reported data include: frequency counts M_i, for different values of i, total number of distinct genes j and sample size n. Source: Susko and Roger (2004).

Finally, the probability that the $(n+m+1)$–th observation coincides with a species detected j times in the enlarged sample $X_1, \ldots, X_{n+m}$ is

$$\widehat{U}_{n+m,i} = \sum_{l=1}^{i} M_{l,n}(l-\sigma)_{i+1-l}\binom{m}{i-l}\frac{(\theta+n-l+\sigma)_{m-i+l}}{(\theta+n)_{m+1}} + (1-\sigma)_i\binom{m}{i}\frac{(\theta+k\sigma)(\theta+n+\sigma)_{m-i}}{(\theta+n)_{m+1}}. \tag{43}$$

To briefly illustrate the behavior of the Bayesian nonparametric estimator based on the PY process and compare it with the Good–Toulmin estimator let us consider genomic data, which consists of Expressed Sequence Tags (EST) obtained from *Naegleria gruberi* cDNA libraries. *Naegleria gruberi* is a widespread free–living soil and freshwater *amoeboflagellate* widely studied in the biological literature. The two considered datasets are sequenced from two cDNA libraries prepared from cells grown under different culture conditions, aerobic and anaerobic, and have been previously analyzed in [72, 44]. The sequenced data, which will constitute the basic samples, are reported in Table 4.2.

If one is interested in the probability of discovering a new gene at the $(n+m+1)$–th step of the sequencing process, one has two options: the Good–Toulmin estimator $\check{U}_{n+m,0}$ reported in (31) or the estimator $\hat{U}_{n+m,0}$ in (42) which is based on the PY process. To complete the specification of the latter let us mention that the parameters (σ, θ) are fixed according to an empirical Bayes specification, which yields $(0.66, 155.5)$. The results are displayed in Figure 5. It is clear that the Good–Toulmin estimator exhibits an erratic behavior for values of the additional sample relatively larger than that of the basic sample n. This phenomenon is avoided by the Bayesian nonparametric estimator since it relies on a well–defined probabilistic model in which all quantities are modeled jointly and coherently. For sizes of m for which the Good–Toulmin estimator works well, the estimators essentially coincide. Note that, in terms of the specific application, the anearobic library exhibits the clearly higher genetic diversity. Furthermore, as already mentioned, one can use such estimates to fix the size of the additional sample m as the

maximum integer for which the discovery probability lies above the desired threshold, which is typically determined also on the basis of cost considerations.

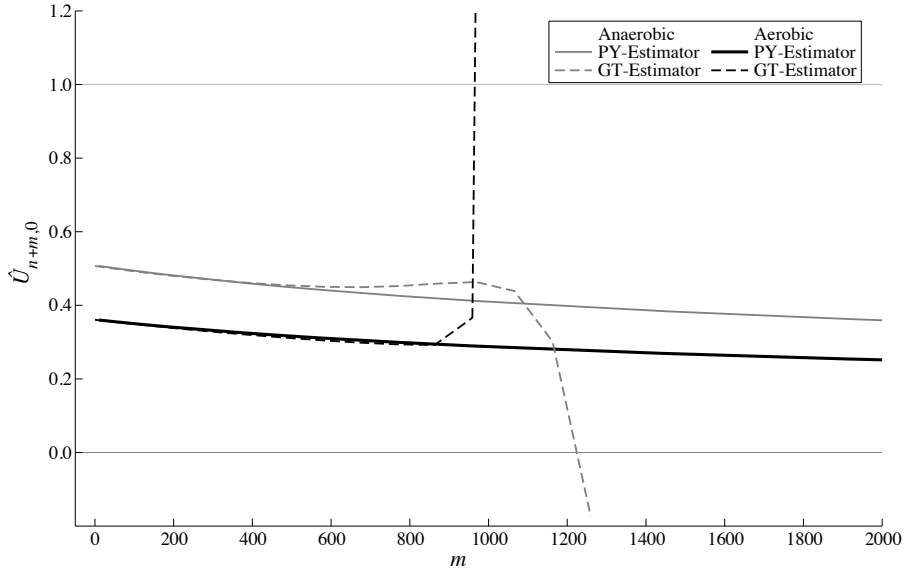

Figure 5: EST data from *Naegleria gruberi* aerobic and anaerobic cDNA libraries with basic sample $n \cong 950$: Good–Toulmin (GT) and Pitman–Yor process (PY) estimators of the probability of discovering a new gene at the $(n+m+1)$–th sampling step for $m = 1, \ldots, 2000$.

5 Frequentist asymptotics

During the last two decades frequentist consistency has gained a major role in Bayesian nonparametrics and is generally accepted as a key validation criterion for the use of a nonparametric prior in applied problems. See [25] for a recent review on the subject. The idea that underlies the study of consistency consists in assuming that the data $(X_n)_{n\geq 1}$ are iid from some "true" $P_0 \in \mathbf{P}_X$ and in verifying whether the posterior distribution $Q(\cdot\,|X_1, \ldots, X_n)$ accumulates in any suitably defined neighborhood of P_0. Therefore, while the posterior is derived based on an assumption of exchangeability of the data as described in (1), the frequentist asymptotic evaluation postulates plain independence of the data generating process. This explains why such an approach has also been termed "what if" approach by P. Diaconis. See [10]. Here we shall discuss consistency for Gibbs–type priors. In this respect, note that frequentist asymptotics of Bayesian procedures is different from the kind of asymptotics discussed in Section 4 which pre-

serves a Bayesian flavor since it aims at achieving a large sample approximation of the posterior without modifying the dependence assumption among the data.

Let us start by fixing some notation and introducing some useful concepts. First the data X_i's are assumed to be iid from some "true" P_0 or, in other terms, the distribution of the sequence of observations $(X_n)_{n\geq 1}$ is the infinite product measure $P_0^\infty = P_0 \times P_0 \times \cdots$. If A_ϵ denotes a neighborhood of P_0 of radius ϵ, the posterior is said to be consistent at P_0 if $Q(A_\epsilon|X_1,\ldots,X_n) \to 1$ almost surely with respect to P_0^∞, as $n \to \infty$ and for any $\epsilon > 0$. In the case of Gibbs–type priors, the natural choice for A_ϵ is represented by weak neighborhoods. Clearly, consistency can be achieved only at P_0 whose weak neighborhoods have a priori positive probability. In this respect, the full support property of Gibbs–type priors recalled in Section 2.2 is important since it ensures that consistency can potentially be achieved at any given P_0. Furthermore, note that the full support property represents a desirable property not only when studying consistency in the case where Gibbs–type priors are used to model directly the data, but also in the context of mixture models as in (12). Indeed, together with some other features of Gibbs–type priors, it allows to extend known consistency results for Dirichlet process mixture models in a straightforward way and the condition for it to hold will be essentially the same. See [26, 46].

As explained in some detail below, recent results suggest that Gibbs–type priors are always consistent with respect to (w.r.t.) discrete P_0's. On the other hand, when they are used to model data coming from diffuse distributions, inconsistency may arise. Possible inconsistency at diffuse P_0 should not, however, be interpreted as a serious issue: what really matters is the data generating mechanism the nonparametric prior is designed for so that consistency must hold w.r.t. choices of P_0 that are compatible with such a mechanism. Since Gibbs–type priors are discrete random probability measures, one should be primarily interested in investigating consistency w.r.t. discrete P_0's. Indeed, Gibbs–type priors, and discrete nonparametric priors in general, are designed to model discrete distributions and they should under no circumstance be used to model data coming from diffuse distributions. In the latter case they should be exploited within hierarchical mixtures.

5.1 General results

The strategy for showing consistency consists in first identifying the weak limit of the posterior, say P' in $\mathbf{P}_{\mathbb{X}}$, which will be some function of P_0, and then checking whether $P' = P_0$ so that consistency is achieved. The candidate weak limit P' is identified by investigating the asymptotic behavior of the predictive distribution (15) (i.e. the posterior expected value), which in explicit cases allows to guess P' quite easily. Then one has to show that the posterior variance of $\tilde{p}$

in (3) converges to 0, a.s.-P_0^∞, which suffices to establish that the posterior concentrates in a weak-neighborhood of the predictive distribution. See [34, 8] for details. Now, let $X_1, \ldots, X_n$ denote a sample with κ_n distinct values with corresponding frequencies $n_1, \ldots, n_{\kappa_n}$. Even if κ_n denotes the same quantity identified as K_n in previous sections, we shall use a different symbol to emphasize the fact that here the asymptotic behavior of the number of observed distinct species κ_n is dictated by P_0 from which the iid sequence is sampled and not by a Gibbs–type prior directing an exchangeable sequence according to (1). Different choices of P_0 clearly yield different (almost sure) limiting behaviors for κ_n. On the one hand, if P_0 is discrete with N point masses, for any $N \in \mathbb{N} \cup \{\infty\}$, then $P_0^\infty(\lim_n \kappa_n = N) = 1$ and $P_0^\infty(\lim_n n^{-1}\kappa_n = 0) = 1$ even if $N = \infty$. On the other hand, if P_0 is diffuse, $P_0^\infty(\kappa_n = n) = 1$ for any $n \geq 1$. Henceforth we shall focus on these two cases and adopt the shorter notation $\kappa_n \ll_{a.s.} n$ and $\kappa_n \sim_{a.s.} n$, which stand for $\kappa_n/n \to 0$ and $\kappa_n/n \to 1$ a.s.-P_0^∞, respectively. It turns out that a key quantity for studying the asymptotics of the predictive distribution is given by the probability (4) of discovering a new observation at the $(n+1)$–th sampling step, which is given by $V_{n+1,\kappa_n+1}/V_{n,\kappa_n}$ in the case of Gibbs–type priors. Considering a Gibbs–type prior with base measure P^* having support $\mathbb{X}$ and assuming that

$$\frac{V_{n+1,\kappa_n+1}}{V_{n,\kappa_n}} \to \alpha \quad \text{a.s.-}P_0^\infty \tag{H}$$

as $n \to \infty$ for some constant $\alpha \in [0, 1]$, in [8] it is shown that

$$Q(A'_\epsilon | X_1, \ldots, X_n) \;\to\; 1 \qquad\qquad \text{a.s.-}P_0^\infty$$

as $n \to \infty$ and for any $\epsilon > 0$ where A'_ϵ is a weak neighborhood of P'. Moreover, one has

$$P' = \alpha P^*(\cdot) + (1 - \alpha) P_0(\cdot). \tag{44}$$

Some comments regarding the above convergence result are in order. As for the condition (H), it is worth noting that it holds true for all Gibbs–type priors for which an explicit expression of the V_{n,κ_n}'s is known, regardless as to whether P_0 is discrete or diffuse. It therefore represents only a mild regularity condition. Moreover, the posterior distribution converges to a point mass at (44), a linear combination of the prior guess P^* and the "true" distribution P_0. Hence, weak consistency is guaranteed if $\alpha = 0$ (and in the trivial case $P^* = P_0$ to be excluded henceforth) and it is sufficient to check whether the probability of discovering a new value converges to 0, a.s.-P_0^∞. Also, one can assess the departure from consistency by looking at the size of α: the larger α, the heavier the limiting mass assigned to the prior guess P^*. One can even think of a case of "total inconsistency", i.e. $\alpha = 1$, the worst case scenario where the posterior tends to concentrate around the prior guess P^* and no learning at all takes place.

To better visualize the above convergence result it is useful to look at special cases of the PY process with $\sigma \in [0, 1)$ and $\theta > -\sigma$, for which such convergence had already been established in [34]. From the form of their predictive distributions (11), one can immediately conjecture the following result: when P_0 is discrete ($\kappa_n \ll_{a.s.} n$) we have $\alpha = 0$, implying consistency; when P_0 is diffuse ($\kappa_n \sim_{a.s.} n$), we have $\alpha = \sigma$, hence inconsistency, unless $\sigma = 0$, which corresponds to the Dirichlet case. See also [36]. An analogous result has been established for the NGG process together with some results concerning the case of Gibbs–type priors with $\sigma > 0$ in [36].

Focusing now on Gibbs–type priors with $\sigma < 0$ allows to highlight the occurrence of interesting phenomena. Recall from Section 2.2 that these priors coincide with mixtures of PY processes with parameters $\{(\sigma, m|\sigma|) : m = 1, 2, \dots, \}$ and they can be represented in hierarchical form as (20). It turns out that, according to the nature of the "true" distribution P_0, a sufficient condition can be stated in terms of the tail behavior of the mixing distribution π in (20). More precisely, for Gibbs–type priors with parameter $\sigma < 0$ and prior guess P^* whose support coincides with $\mathbb{X}$, in [8] consistency is shown to hold

(i) at any discrete P_0 if for sufficiently large m

$$\frac{\pi(m+1)}{\pi(m)} \leq 1; \tag{T1}$$

(ii) at any diffuse P_0 if for sufficiently large m and for some $M < \infty$

$$\frac{\pi(m+1)}{\pi(m)} \leq \frac{M}{m}. \tag{T2}$$

Condition (T1) is an extremely mild assumption on the regularity of the tail of the mixing π: it requires $x \mapsto \pi(x)$ to be ultimately decreasing, a condition met by the commonly used probability measures on $\mathbb{N}$. Hence one can conclude that Gibbs–type priors with parameter $\sigma < 0$ are essentially always consistent when P_0 is discrete. On the other hand, condition (T2) requires the tail of π to be sufficiently light, so when P_0 is diffuse one needs to closely investigate the tail behavior of π.

5.2 Illustrations

In light of the results stated above one is naturally led to wonder what happens when (T2) is not satisfied. To this end we consider three different Gibbs–type priors presented in Section 2.2 with $\sigma = -1$: each prior is characterized by a specific choice of the mixing distribution π. We focus on the case of diffuse P_0, which leads to some interesting conclusions. In the case of discrete P_0 it is straightforward to show that (T1) holds, hence ensuring consistency.

The first prior we consider, introduced in [27], is characterized by the heavy-tailed mixing distribution (21), which does not admit a finite expected value. Since $\pi(m+1)/\pi(m) = (m-\gamma)/(m+1)$ cannot be eventually bounded by M/m for some constant M, condition (T2) does not hold true. Given the V_{n,κ_n}'s admit the simple closed form expression (22), the weights of the prediction rule simplify to

$$\frac{V_{n+1,\kappa_n+1}}{V_{n,\kappa_n}} = \frac{\kappa_n(\kappa_n-\gamma)}{n(\gamma+n)}.$$

It is easy to see that, if P_0 is diffuse, implying $\kappa_n \sim_{a.s.} n$, condition (H) holds true with $\alpha = 1$ and the weak limit coincides with the prior guess P^*, whatever the "true" distribution of the data P_0. This means we are in the case of "total" inconsistency.

The second example has a Poisson mixing distribution (23) on the positive integers. Such a π has light tails and condition (T2) is satisfied since $\pi(m+1)/\pi(m) = \lambda/(m+1)$. Therefore, by (T2), the posterior is consistent when P_0 is diffuse.

The last sub-family of Gibbs–type priors with $\sigma = -1$ is identified by a geometric mixing distribution (24). Note that $\pi(m+1)/\pi(m) = \eta$ so that condition (T2) does not hold true. It turns out that, with P_0 diffuse and $\kappa_n \sim_{a.s.} n$, one obtains

$$\frac{V_{n+1,\kappa_n+1}}{V_{n,\kappa_n}} \to \alpha = \frac{2-\eta-2\sqrt{1-\eta}}{\eta} \in [0,1]. \tag{45}$$

See [8] for details. The limit α in (45) can be any point in $[0,1]$ according to the value of η and therefore we can obtain the whole spectrum of weak limits (44) ranging from consistency ($\alpha = 0$) to "total" inconsistency ($\alpha = 1$). In particular, α is increasing in η, so the larger η, the heavier the limiting mass assigned to the prior guess. Small values of η identify a situation similar to the second example since they yield a light-tailed π. Conversely, large values of η are more in line with the first example giving rise to heavy-tailed π. Finally, it is worth remarking that a minimal deviation from condition (T2) already produces inconsistent behaviors, even extreme ones, showing that (T2) is close to being necessary.

6 Dependent processes for Gibbs–type priors

In this section we briefly discuss possible extensions of the previous results to a dynamic setting. In particular, here we refer to time-indexed random objects, with some specification of the temporal transition mechanism, whose stationary, or at least marginal, states coincide in distribution with some random probability measure of Gibbs-type. In this respect, it is important to distinguish between two different research areas on time-dependent random probability measures, both related to Bayesian nonparametric priors. The main difference between these

two approaches, outlined below, lies in the fact that the former is mostly driven by inferential purposes, while the second is more concerned with the analytical properties of the constructed objects. If on one hand the first is closer to the interest of the Bayesian community, on the other it is our opinion that the two approaches have a strong potential of reciprocally benefitting from one another.

The first area, concerned with so-called dependent processes, is at present an extremely active front in Bayesian Nonparametrics. Besides the pioneering contributions in [6], the modern approaches to the problem can be traced back to [52]. Generally speaking, the aim is to investigate generalizations of the Dirichlet process (or other random measures) to frameworks which allow for types of dependence less restrictive than exchangeability. These include for example dependence on time or, more generally, on covariates. See, for example, [1] for some up-to-date references. Most contributions in this direction exploit the representation (3) and dependence is quite easily induced via the weights and/or the atoms. Moreover, this allows to exploit simulation techniques such as the slice sampler ([75], [7]) and the retrospective sampler [56]. The combination of these two main factors leads then to efficient inferential procedures in such non exchangeable frameworks.

The second research area has its roots in Applied Probability and is concerned with stochastic population dynamics, but is also closely related to Bayesian nonparametric modeling. The main idea underlying the constructions in this framework is that of approximating the dynamics of a large population with a diffusion process, where the process dimension depends on the number of species the population is allowed to have. When the species can be of infinitely-many types, this gives rise to infinite-dimensional or measure-valued diffusions. In some cases the individual reproduction mechanisms yield populations whose frequencies have marginal or stationary states such as the one- and two-parameter Poisson-Dirichlet distribution ([13],[58]), the Dirichlet process ([14]), the normalized-inverse Gaussian distribution ([70]). From a Bayesian perspective these clearly represent dependent priors. At least in the authors' opinion, such an approach represents a highly promising research line for the definition of dependent processes, since the possibility of studying their analytical properties also yields a deeper understanding of their behavior. Other reasons of interest for the Bayesian community include the use of Pólya urn schemes for constructing some of these dependent random probability measures ([69],[64]; see also [4]), and the investigation of the so-called σ-diversity processes (in the notation of Section 2). These constitute a dynamic counterpart of (16), and make explicit the dynamics and distributional properties concerning the evolution of the clustering structure within the population, as a consequence of the specific modeling dynamics at hand. See [70] and [68].

7 Concluding remarks

An intense research activity, started after the introduction of the Dirichlet process, has produced a vast literature concerning classes of random probability measures whose laws can be used as nonparametric priors. In current research the choice among these classes is often dictated by taste (one's "favorite prior"), mathematical tractability or a blend of the two. For instance, neutral to the right priors ([11]) are typically used in survival analysis contexts since they are conjugate also w.r.t. right censored observations. However, there is no conceptual reason to prefer a conjugate prior over a non–conjugate one and it all boils down to mathematical convenience since it allows to evaluate posterior inferences of interest. With Gibbs–type priors things go the opposite way: one makes a precise assumption on the learning mechanism according to which the prediction of a new value depends on the sample size n and on the number of distinct values observed so far $K_n = k$ but not on their frequencies $n_1, \ldots, n_k$, and only afterwards investigates the implications of such an assumption. This is very much in the spirit of de Finetti himself who constantly emphasizes in his works the importance of formulating assumptions on empirically "observable" rather than on "unobservable" quantities. In this respect Gibbs–type priors can be seen somehow as counterparts to characterizations of parametric families in terms of exchangeability and some other characteristic of the observables. Consider, for instance, Freedman's characterization [21] of exchangeable and rotational invariant sequences as mixtures of Gaussians: it is the request of rotational invariance on the observables that justifies the use of Gaussian distributions. In a nonparametric context, an analogous type of result (see [66, 50]) legitimates the use of the Dirichlet process: by assuming exchangeability and a prediction rule given by a linear combination of the prior guess and the empirical measure one automatically obtains the Dirichlet process.

Turning back to the Gibbs–case, once the assumption on the learning mechanism is made, one realizes that the high degree of mathematical tractability is nothing but an implication and not a motivation. This then allows to work out a wealth of results concerning the behavior of Gibbs–type priors. Importantly, one is not anymore constrained to a logarithmic increase of K_n as in the Dirichlet case and the whole spectrum going from a finite K_n to an almost linearly increasing K_n is available. This, in turn, produces a significantly more flexible prior on the number of components in mixture models. Furthermore, distributional properties and (often) closed form expressions for estimators of the quantities of statistical interest can be derived. An appealing feature is also represented by the fact that such quantities retain an intuitive flavor by directly relating to the key learning assumption. For instance, in the context of species sampling, one coherently has that K_n is a sufficient statistic for predictions concerning "new" values. In contrast, if predictions are required for both "new" values and already observed values

with frequency less than or equal to τ, the sufficient statistics becomes $(K_n, M_{1,n}, \ldots, M_{\tau,n})$, which includes species with frequency not larger than τ. Although more subtle, this is also in accordance with the key learning assumption and the implied reinforcement mechanism, described in the paper. Moreover, given the sound assumption on the learning scheme and its persuasive implications, it seems natural to use the Gibbs–framework also as basis for the definition of dependent processes.

Summing up, with this review we hope to have provided an affirmative and convincing answer to the question posed in the title of the paper. And we are confident that the future will see more statistical problems laid out in the well grounded general Gibbs–type framework. This would bring a solid foundation to the story and obviously would not prevent to use one's favorite Gibbs–type prior (e.g. the PY process) in the concrete application or even, if dropping the dependence on K_n is legitimated by the problem at issue, returning to the "safe" Dirichlet world.

Acknowledgment

The first three and last two authors are supported by the European Research Council (ERC) through StG "N-BNP" 306406. The fourth author is supported by CONACYT, project no. 131179.

Appendix

Proof of Proposition 1

Recall that for any species sampling model the probability of generating a new value is of the form (9). For it not to depend on $(n_1, \ldots, n_k)$ for any $n \geq 1$ and $k \leq n$, $p_k^{(n)}$ necessarily has to be of product form. [28] have shown that an EPPF associated to a infinite exchangeable random partition is of product form if and only if it is given by (13) or, equivalently, if $\tilde{p}$ is of Gibbs–type. Therefore (9) depends only on n and k if and only if it is of Gibbs–type. This proves the categorization of species sampling models $\tilde{p}$ in classes (ii) and (iii).

We are now left with showing that the Dirichlet process is the only species sampling model for which (9) neither depends on the frequencies nor on k. Given the above, this amounts to showing that the subclass (i) of the family of Gibbs–type priors (ii) contains only the Dirichlet process.

First we show by a contradiction argument that for (9) not to depend on k it must necessarily be $\sigma = 0$. Then we conclude that the only Gibbs–type prior with $\sigma = 0$ for which (9) does not depend on k is the Dirichlet process. Confining ourselves to the Gibbs–type case (since in all other cases (9) even depends on the frequencies), one has

$$\mathbb{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n) = \frac{V_{n+1,k+1}}{V_{n,k}} = 1 - (n - \sigma\, k)\frac{V_{n+1,k}}{V_{n,k}}$$

and we assume it does not depend on k. This amounts to requiring that $(n-\sigma\, k)V_{n+1,k}(V_{n,k})^{-1}$ does not depend on k, namely

$$\frac{V_{n+1,k}}{V_{n,k}} = \frac{c_n}{(n - \sigma\, k)}, \tag{46}$$

for some c_n not depending on k and, by using (14),

$$\frac{V_{n+1,k+1}}{V_{n,k}} = (1 - c_n). \tag{47}$$

The combination of (46) and (47) implies

$$\mathbb{P}(X_{n+1} \in A \mid X_1, \ldots, X_n) = (1 - c_n)P_0(A) + c_n \sum_{j=1}^{k} \frac{(n_j - \sigma)}{(n - \sigma\, k)}\delta_{X_j^*}(A). \tag{48}$$

However, this is a prediction rule corresponding to an infinite exchangeable sequence if and only if $\sigma = 0$. To see this, note that in view of [24, Proposition 3.2] infinite exchangeability requires (48) to satisfy

$$\mathbb{P}(X_{n+1} \in A,\, X_{n+2} \in B \mid X_1, \ldots, X_n) = \mathbb{P}(X_{n+1} \in B,\, X_{n+2} \in A \mid X_1, \ldots, X_n) \tag{49}$$

for any $n \geq 1$ and A, B in $\mathscr{X}$. Consider, now, two sets A and B such that $A \cap B = \varnothing$, $A \cap \{X_1, \ldots, X_n\} = \varnothing$ and $B \cap \{X_1, \ldots, X_n\} \neq \varnothing$. Hence, the left–hand side of (49) coincides with

$$c_n P_0(A) \left\{ c_{n+1} P_0(B) + (1 - c_{n+1}) \frac{1}{n + 1 - (k+1)\sigma} \sum_{j=1}^{k} (n_j - \sigma)\, \delta_{X_j^*}(B) \right\}$$

whereas the right–hand side of (49) coincides with

$$c_{n+1} P_0(A) \left\{ c_n P_0(B) + (1 - c_n) \frac{1}{n - k\sigma} \sum_{j=1}^{k} (n_j - \sigma)\, \delta_{X_j^*}(B) \right\}$$

and the two are equal if and only if, for any $k = 1, \ldots, n$, one has

$$\frac{c_{n+1}(1 - c_n)}{n - k\sigma} = \frac{c_n(1 - c_{n+1})}{n + 1 - (k+1)\sigma}.$$

Assuming $n \geq 2$, with $k = 1$ the above condition becomes

$$\frac{c_{n+1}(1 - c_n)}{n - \sigma} = \frac{c_n(1 - c_{n+1})}{n + 1 - 2\sigma} \tag{50}$$

and with $k = 2$, one has

$$\frac{c_{n+1}(1 - c_n)}{n - 2\sigma} = \frac{c_n(1 - c_{n+1})}{n + 1 - 3\sigma}. \tag{51}$$

Taking the ratios of the terms in (50) and those in (51) yields

$$\frac{n - 2\sigma}{n - \sigma} = \frac{n + 1 - 3\sigma}{n + 1 - 2\sigma}$$

and this holds true if and only if $\sigma^2 = \sigma$. This therefore contradicts the assumption that $\mathbb{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n)$ does not depend on k for $\sigma \neq 0$. A different proof can also be derived by using the recursion (14) iterated over two prediction steps.

Finally, recall that [28, Theorem 13], showed that Gibbs–type priors with $\sigma = 0$ correspond to the Dirichlet process or the Dirichlet process mixture over its total mass parameter. On the other hand, when $\sigma = 0$, (48) characterizes the Dirichlet process (see [50], [66]). Hence, Dirichlet process mixture over the total mass cannot belong to class (i), i.e. $\mathbb{P}(X_{n+1} = \text{“new”} \mid X_1, \ldots, X_n)$ must depend also on k. The proof is, then, complete. □

References

[1] A.F. Barrientos, A. Jara, F.A. Quintana, “On the support of MacEachern’s dependent Dirichlet processes and extensions”, *Bayes. Anal.*, vol.7, pp. 277–310, 2012.

[2] D. Barry and J.A. Hartigan, “A Bayesian analysis for change point problems”, *J. Amer. Statist. Assoc.*, vol.88, pp. 309-319, 1993.

[3] A. Brix, “Generalized gamma measures and shot-noise Cox processes”, *Adv. Appl. Probab.*, vol.31, pp. 929-953, 1999.

[4] F. Caron, M. Davy and A. Doucet, “Generalized Pólya urn for time-varying Dirichlet process mixtures”, *Proc. 23rd Conf. on Uncertainty in Artificial Intelligence*, Vancouver, 2007.

[5] C.A. Charalambides, *Combinatorial methods in discrete distributions*, New York: Wiley, 2005.

[6] D.M. Cifarelli and E. Regazzini, "Nonparametric statistical problems under partial exchangeability: the use of associative means" (Original title: "Problemi statistici non parametrici in condizioni di scambiabilità parziale: impiego di medie associative"), *Quaderni dell'Istituto di Matematica Finanziaria, Univ. of Torino*, vol.3(12), 1978.

[7] P. Damien, J.C. Wakefield and S.G. Walker, "Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables", *J. Roy. Statist. Soc. Ser. B*, vol.61, pp. 331–344, 1999.

[8] P. De Blasi, A. Lijoi and I. Prünster, "An asymptotic analysis of a class of discrete nonparametric priors", *Statistica Sinica*, vol. 23, pp. 1299-1322, 2013.

[9] L. Devroye, "Random variate generation for exponentially and polynomially tilted stable distributions", *ACM Trans. Model. Comp. Simul.*, vol.19, article no.18, 2009.

[10] P. Diaconis and D. Friedman, "On the consistency of Bayes estimates", *Ann. Statist.*, vol.14, pp. 1-26, 1986.

[11] K. Doksum, "Tailfree and neutral random probabilities and their posterior distributions", *Ann. Probab.*, vol. 2, pp. 183-201, 1974.

[12] M.D. Escobar and M. West, "Bayesian density estimation and inference using mixtures", *J. Amer. Statist. Assoc.*, vol.90, pp. 577-588, 1995.

[13] S.N. Ethier and T.G. Kurtz, "The infinitely-many-neutral-alleles diffusion model", *Adv. Appl. Probab.*, vol.13, pp. 429-452, 1981.

[14] S.N. Ethier and T.G. Kurtz, "Markov processes: characterization and convergence", Wiley, New York, 1986.

[15] S. Favaro, A. Lijoi, R.H. Mena and I. Prünster, " Bayesian nonparametric estimators derived from Gibbs-type priors with finitely many types", *In preparation.*

[16] S. Favaro, A. Lijoi, R.H. Mena and I. Prünster, "Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior", *J. Roy. Statist. Soc. Ser. B*, vol.71, pp. 993-1008, 2009.

[17] S. Favaro, A. Lijoi and I. Prünster, "Asymptotics for a Bayesian nonparametric estimator of species variety", *Bernoulli*, vol.18, pp. 1267-1283, 2012.

[18] S. Favaro, A. Lijoi and I. Prünster, "Conditional formulae for Gibbs-type exchangeable random partitions", *Ann. Appl. Probab.*, vol.23, pp. 1721-1754, 2013.

[19] S. Favaro, A. Lijoi and I. Prünster, "A new estimator of the discovery probability", *Biometrics*, vol.68, pp. 1188-1196, 2013.

[20] S. Favaro, I. Prünster and S.G. Walker, "On a class of random probability measures with general predictive structure", *Scand. J. Stat.*, vol.38, pp. 359-376, 2011.

[21] D.A. Freedman, "Invariants under mixing which generalize de Finettis theorem", *Ann. Math. Statist.*, vol. 33, pp. 916923, 1962.

[22] T.S. Ferguson, "A Bayesian analysis of some nonparametric problems", *Ann. Statist.*, vol.1, pp. 209-230, 1973.

[23] T.S. Ferguson, "Prior distributions on spaces of probability measures", *Ann. Statist.*, vol.2, pp. 615-629, 1974.

[24] S. Fortini, L. Ladelli and E. Regazzini, "Exchangeability, predictive distributions and parametric models", *Sankhya Ser. A*, vol.62, pp. 86-109, 2000.

[25] S. Ghosal, "The Dirichlet process, related priors, and posterior asymptotics", in *Bayesian Nonparametrics*, Hjort, N., Holmes, C., Müller, P. and Walker, S. Eds. Cambridge: Cambridge Univ. Press, 2010, pp. 35-79.

[26] S. Ghosal, J.K. Ghosh and R.V. Ramamoorthi, "Posterior consistency of Dirichlet mixtures in density estimation", *Ann. Statist.*, vol.27, pp. 143-158, 1999.

[27] A. Gnedin, "A species sampling model with finitely many types", *Elect. Comm. Probab.*, vol.15, pp. 79-88, 2010.

[28] A. Gnedin and J. Pitman, "Exchangeable Gibbs partitions and Stirling triangles", *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, vol.325, pp. 83-102, 2005.

[29] I.J. Good, "The population frequencies of species and the estimation of population parameters", *Biometrika*, vol.40, pp. 237-264, 1953.

[30] I.J. Good and G.H. Toulmin, "The number of new species, and the increase in population coverage, when a sample is increased", *Biometrika*, vol.43, pp. 45-63, 1956.

[31] M. Guindani and P. Müller, "A Bayesian Semiparametric model for the analysis of SAGE Data", Tech. Rep., 2010.

[32] J.A. Hartigan, "Partition models", *Comm. Statist. Theory Methods*, vol.19, pp. 2745-2756, 1990.

[33] H. Ishwaran and L.F. James, "Gibbs sampling methods for stick-breaking priors", *J. Amer. Stat. Ass.*, vol.96, pp. 161-173, 2001.

[34] L.F. James, "Large sample asymptotics for the two parameter Poisson Dirichlet process", in *Pushing the Limits of Contemporary Statistics*, Clarke B. and Ghosal S. Eds. Hayward: IMS, 2008, pp. 187-199.

[35] L.F. James, A. Lijoi and I. Prünster, "Posterior analysis for normalized random measure with independent increments", *Scand. J. Statist.*, vol.36, pp. 76-97, 2009.

[36] G.H. Jang, J. Lee and S. Lee, "Posterior consistency of species sampling priors", *Statist. Sinica*, vol.20, pp. 581-593, 2010.

[37] A. Jara, E. Lesaffre, M. De Iorio and F.A. Quintana, "Bayesian semiparametric inference for multivariate doubly-interval-censored data", *Ann. Appl. Statist.*, vol. 4, pp. 2126–2149, 2010.

[38] J.F.C. Kingman, "Random discrete distributions", *J. Roy. Statist. Soc. Ser. B*, vol.37, pp. 1-22, 1975.

[39] R.M. Korwar and M. Hollander, "Contribution to the theory of Dirichlet processes", *Ann. Probab.*, vol.1, pp. 705-711, 1973.

[40] A. Lijoi, R.H. Mena and I. Prünster, "Hierarchical mixture modelling with normalized inverse-Gaussian priors", *J. Amer. Stat. Assoc.*, vol.100, pp. 1278-1291, 2005.

[41] A. Lijoi, R.H. Mena and I. Prünster, "Bayesian nonparametric analysis for a generalized Dirichlet process prior" *Stat. Inference Stoch. Process.*, vol. 8, pp. 283-309, 2005

[42] A. Lijoi, R.H. Mena and I. Prünster, "Bayesian nonparametric estimation of the probability of discovering a new species", *Biometrika*, vol.94, pp. 769-786, 2007.

[43] A. Lijoi, R.H. Mena and I. Prünster, "Controlling the reinforcement in Bayesian non-parametric mixture models", *J. R. Statist. Soc. Ser. B*, vol.69, pp. 715-740, 2007.

[44] A. Lijoi, R.H. Mena and I. Prünster, I., "A Bayesian nonparametric method for prediction in EST analysis", *BMC Bioinformatics*, vol.8, article no. 339, 2007.

[45] A. Lijoi and I. Prünster, "Models beyond the Dirichlet process", in *Bayesian Nonparametrics*, N.L. Hjort, C.C. Holmes, P. Müller and S.G. Walker Eds. Cambridge: Cambridge University Press, pp. 80-136, 2010.

[46] A. Lijoi, I. Prünster and S.G. Walker, "On consistency of nonparametric normal mixtures for Bayesian density estimation", *J. Amer. Statist. Assoc.*, vol.100, pp. 1292-1296, 2005.

[47] A. Lijoi, I. Prünster and S.G. Walker, "Bayesian nonparametric estimators derived from conditional Gibbs structures", *Ann. Appl. Probab.*, vol.18, 1519-1547, 2008.

[48] A. Lijoi, I. Prünster and S.G. Walker, "Investigating nonparametric priors with Gibbs structure", *Statist. Sinica*, vol.18, pp. 1653-1668, 2008.

[49] A.Y. Lo, "On a class of Bayesian nonparametric estimates. I. Density estimates", *Ann. Statist.*, vol. 12, pp. 351-357, 1984.

[50] A.Y. Lo, "A characterization of the Dirichlet process". *Statist. Probab. Lett.*, vol. 12, pp. 185-187, 1991.

[51] S.N. MacEachern, "Estimating normal means with a conjugate style Dirichlet process prior", *Commun. Statist. Simulation Comp.*, vol 23, 727–741, 1994.

[52] S.N. MacEachern, "Dependent Nonparametric Processes". In *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statist. Assoc., Alexandria, VA, 1999.

[53] C.X. Mao and B.G. Lindsay, "A Poisson model for the coverage problem with a genomic application", *Biometrika*, vol.89, pp. 669–681, 2002.

[54] C.X. Mao, "Prediction of the conditional probability of discovering a new class", *J. Amer. Statist. Assoc.*, vol.99, pp. 1108–1118, 2004.

[55] C. Navarrete, F.A. Quintana and P. Müller, "Some issues on nonparametric Bayesian modeling using species sampling models", *Statist. Modell.*, vol.41, pp. 3–21.

[56] O. Papaspiliopoulos and G.O. Roberts, "Retrospective MCMC for dirichlet process hierarchical models", *Biometrika*, vol.95, pp. 169–186, 2008.

[57] M. Perman, J. Pitman and M. Yor, "Size–biased sampling of Poisson point processes and excursions", *Probab. Theory Related Fields*, vol. 92, pp. 21–39, 1992.

[58] L. Petrov, "Two-parameter family of diffusion processes in the Kingman simplex", *Funct. Anal. Appl.*, vol.43, pp. 279–296, 2009.

[59] J. Pitman, "Exchangeable and partially exchangeable random partitions", *Probab. Theory and Relat. Fields*, vol.102, pp. 145–158, 1995.

[60] J. Pitman, "Some developments of the Blackwell-MacQueen urn scheme", in *Statistics, Probability and Game Theory. Papers in honor of David Blackwell*, T.S. Ferguson, L.S. Shapley and J.B. MacQueen Eds., Lecture Notes, Monograph Series, vol.30, Hayward CA: IMS, 1996, pp. 245–267.

[61] J. Pitman, "Poisson-Kingman partitions", in *Statistics and Science: A Festschrift for Terry Speed*, D.R. Goldstein, Ed., Institute of Mathematical Statistics Lecture Notes-Monograph Series, vol.40, Beachwood OH: IMS, 2003, pp. 1–34.

[62] J. Pitman, *Combinatorial Stochastic Processes*, Ecole d'Eté de Probabilités de Saint-Flour XXXII, Lecture Notes in Math., vol.1875, Berlin: Springer, 2006.

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

[64] I. Prünster and M. Ruggiero, "A Bayesian nonparametric approach to modeling market share dynamics", *Bernoulli*, vol.19, pp. 64–92, 2013.

[65] F.A. Quintana and P.L. Iglesias, "Bayesian clustering and product partition models", *J. R. Stat. Soc. Ser. B*, vol.65, pp. 557-574, 2003.

[66] E. Regazzini, "Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità", *Giornale dell' Istituto Italiano degli Attuari*, vol. 41, pp. 77-89, 1978.

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

[68] M. Ruggiero, "Species dynamics in the two-parameter Poisson-Dirichlet diffusion model", *J. Appl. Probab.*, 2014, in press.

[69] M. Ruggiero and S.G. Walker, "Countable representation for infinite-dimensional diffusions derived from the two-parameter Poisson-Dirichlet process", *Electron. Comm. Probab.*, vol.14, pp. 501-517, 2009.

[70] M. Ruggiero, S.G. Walker and S. Favaro, "Alpha-diversity processes and normalized inverse-Gaussian diffusions", *Ann. Appl. Probab.*, vol.23, pp. 386-425, 2013.

[71] M. Guindani, N. Sepúlveda, C.D.M. Paulino and P. Müller, "A Bayesian semiparametric model for the analysis of sequence counts data" , *J. Roy. Statist. Soc. C*, 2013, doi: 10.1111/rssc.12041.

[72] E. Susko and A.J. Roger, " Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys", *Bioinformatics*, vol. 20, pp. 2279–2287, 2004.

[73] Y.W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes", in *Proceedings of Coling/ACL*, 2006, pp. 985-992.

[74] Y.W. Teh and M.I. Jordan, "Hierarchical Bayesian nonparametric models with applications", in *Bayesian Nonparametrics*, Hjort, N., Holmes, C., Müller, P. and Walker, S. Eds. Cambridge: Cambridge Univ. Press, 2010, pp. 158-207.

[75] S.G. Walker, "Sampling the Dirichlet mixture model with slices", *Comm. Statist. Sim. Comput.*, vol.36, pp. 45–54, 2007.

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