Electronic Journal of Statistics
Vol. 9 (2015) 2884–2902
ISSN: 1935-7524
DOI: 10.1214/15-EJS1096

A note on nonparametric inference for species variety with Gibbs-type priors

Stefano Favaro[*]

Department of Economics and Statistics, University of Torino
Corso Unione Sovietica 218/bis, 10134, Torino, Italy
e-mail: stefano.favaro@unito.it

and

Lancelot F. James

Department of Information Systems, Business Statistics and Operations Management,
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
e-mail: lancelot@ust.hk

Abstract: A Bayesian nonparametric methodology has been recently introduced for estimating, given an initial observed sample, the species variety featured by an additional unobserved sample of size m. Although this methodology led to explicit posterior distributions under the general framework of Gibbs-type priors, there are situations of practical interest where m is required to be very large and the computational burden for evaluating these posterior distributions makes impossible their concrete implementation. In this paper we present a solution to this problem for a large class of Gibbs-type priors which encompasses the two parameter Poisson-Dirichlet prior and, among others, the normalized generalized Gamma prior. Our solution relies on the study of the large m asymptotic behaviour of the posterior distribution of the number of new species in the additional sample. In particular we introduce a simple characterization of the limiting posterior distribution in terms of a scale mixture with respect to a suitable latent random variable; this characterization, combined with the adaptive rejection sampling, leads to derive a large m approximation of any feature of interest from the exact posterior distribution. We show how to implement our results through a simulation study and the analysis of a dataset in linguistics.

MSC 2010 subject classifications: Primary 62F15, 60G57.
Keywords and phrases: Adaptive rejection sampling, Bayesian nonparametric inference, empirical linguistics, Gibbs-type priors, normalized generalized Gamma prior, species sampling asymptotics, two parameter Poisson-Dirichlet prior.

Received February 2015.

1. Introduction

Species sampling problems are associated to situations where an experimenter is sampling from a population of individuals belonging to different species with

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

unknown proportions. Given the information yielded by an initial observed sample, most of the statistical issues to be faced are related to the concept of species variety, or species richness, which can be quantified by estimating various features of an additional unobserved sample. A full range of statistical approaches, parametric and nonparametric as well as frequentist and Bayesian, have been proposed for estimating species variety. These approaches have found many applications in ecology, and their importance has grown considerably in recent years, driven by challenging applications arising from bioinformatics, genetics, statistical physics, computer science, design of experiments, linguistics, machine learning, etc. See, e.g., Bunge and Fitzpatrick [7] and Bunge et al. [8] for two comprehensive reviews, the latter with emphasis on applications to microbial diversity studies. In this paper we consider the Bayesian nonparametric approach introduced in Lijoi et al. [24] and further investigated in Favaro et al. [13] and Favaro et al. [14]. Other recent contributions to species sampling problems in the Bayesian framework are, e.g., Navarrete et al. [28], Zhang and Stern [33], Barger and Bunge [6], Bacallado et al. [3], Lee et al. [23], Airoldi et al. [1] and Guindani et al. [19].

Assuming a population with an (ideally) infinite number of species, we denote by $(X_i^*)_{i\geq 1}$ and $(q_i)_{i\geq 1}$ the species labels and the unknown species proportions, respectively. The Bayesian nonparametric approach in Lijoi et al. [24] is based on the randomization of the q_i's, whose distribution takes on the interpretation of a prior distribution over the species composition of the population. Specifically, let $\tilde{Q} = \sum_{i\geq 1} q_i \delta_{X_i^*}$ where $(q_i)_{i\geq 1}$ are nonnegative random weights such that $\sum_{i\geq 1} q_i = 1$ almost surely, and $(X_i^*)_{i\geq 1}$ are random locations independent of $(q_i)_{i\geq 1}$ and independent and identically distributed according to a nonatomic probability measure. Furthermore, let Π be the distribution of the discrete random probability measure $\tilde{Q}$ and let $\mathbf{X}_n = (X_1, \dots, X_n)$ be a sample from $\tilde{Q}$, namely

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

for any $n \geq 1$. Recall that, due to the de Finetti representation theorem, $\mathbf{X}_n$ is part of the exchangeable sequence $(X_i)_{i\geq 1}$ with directing measure Π. Under the nonparametric framework (1), and with Π belonging to the class of Gibbs-type priors by Gnedin and Pitman [17], results in Lijoi et al. [24] and Favaro et al. [14] provide explicit posterior distributions for several features of an additional unobserved sample $(X_{n+1}, \dots, X_{n+m})$, given the initial sample $\mathbf{X}_n$. Corresponding Bayesian nonparametric estimators, with respect to a squared loss function, are obtained by taking the expected value with respect to these posterior distributions. Two features of particular interest are: the number $K_{n,m}$ of new species in the additional sample, and the number $M_{n,m}(l)$ of species with frequency $1 \leq l \leq n + m$ in the enlarged sample obtained by gathering the initial sample and the additional sample.

Within species sampling problems there are several situations of practical interest where the size of the additional sample is required to be very large, and only a small portion of the population is sampled. For instance, these situations arise frequently in genomics, as witnessed in Mao and Lindsay [27] and Mao [26], and in linguistics, as witnessed in Sampson [32] and references therein. Making inference on $K_{n,m}$ and $M_{n,m}(l)$ for large m poses additional challenges to the Bayesian nonparametric approach of Lijoi et al. [24]. Indeed, while Gibbs-type priors lead to explicit expressions for the posterior distributions and moments of $K_{n,m}$ and $M_{n,m}(l)$, these expressions involve combinatorial coefficients and special functions whose evaluation for large m is cumbersome, thus preventing their concrete implementation. With regards to $K_{n,m}$, a first answer to this problem was presented by Favaro et al. [12] under the assumption of the two parameter Poisson-Dirichlet (PD) prior, which is a noteworthy example of Gibbs-type prior introduced in Pitman [29]. Their approach consisted in characterizing the limiting posterior distribution of $K_{n,m}$, for fixed n and as m tends to infinity, and devising an exact sampling algorithm for it. Monte Carlo expectation and quantiles of the limiting posterior distribution have then been used in order to approximate, for large m, the exact estimator of $K_{n,m}$ and the associated credible intervals.

In this paper we extend the methodology of Favaro et al. [12] to a large subclass of Gibbs-type priors introduced in James [22] and referred to as the Poisson-Gamma (PG) class. See also Proposition 21 in Pitman and Yor [31] for an early definition of the PG class. The PG class encompasses the two parameter PD prior and, among others, the normalized generalized Gamma (GG) prior, which is another noteworthy example of Gibbs-type prior introduced in Pitman [30] and nowadays widely used in Bayesian nonparametrics. See, e.g., James [21], Lijoi et al. [25], Argiento et al. [2], Griffin et al. [18] and Caron and Fox [9]. Within the PG class we provide a simple scale mixture characterization for the limiting posterior distributions of $K_{n,m}$ and $M_{n,m}(l)$, for fixed n and as m tends to infinity. Interestingly, under the assumption of a two parameter PD prior our result leads to an alternative limiting characterization with respect to the one obtained in Favaro et al. [12]. Such a novel characterization sheds some light on the difference between the two parameter PD prior and the normalized GG prior in the context of species variety estimation. Besides extending the main result in Favaro et al. [12], we show that under the general PG class one can still resort to an exact sampling algorithm, which relies on the adaptive rejection sampling of Gilks and Wild [15], for generating random variates from the limiting posterior distributions of $K_{n,m}$ and $M_{n,m}(l)$. Our result thus provides a practical tool for obtaining a Monte Carlo approximation, for large m, of any feature of interest from the exact posterior distributions of $K_{n,m}$ and $M_{n,m}(l)$. Under the assumption a two parameter PD prior and a normalized GG prior, we show how to implement this new tool through a simulation study and the analysis of a dataset in linguistics.

The paper is structured as follows. In Section 2 we present the scale mixture characterization of the limiting posterior distributions of $K_{n,m}$ and $M_{n,m}(l)$, for fixed n and as m tends to infinity, under the general framework of the PG class; details of such a characterization are provided for the special cases of the two parameter PD prior and the normalized GG prior. Section 3 contains an illustration of the proposed characterization through a detailed simulation study and the analysis of a real dataset in linguistics. Proofs are deferred to the Appendix.

2. Limiting posterior distributions under the PG class

Gibbs-type priors form a large class of nonparametric priors indexed by a parameter $\alpha \in (-\infty, 1)$ and a nonnegative function h. Let $\tilde{Q}_{\alpha,h}$ denote a Gibbs-type random probability measure with parameter (α, h), for $\alpha \in (0, 1)$. Among various possible definitions of $\tilde{Q}_{\alpha,h}$, a simple and intuitive one follows by combining Theorem 8 and Proposition 9 in Pitman [30]. Indeed these result provide an indirect definition of $\tilde{Q}_{\alpha,h}$ by characterizing the distribution of the exchangeable random partition induced by a sample $\mathbf{X}_n$ from $\tilde{Q}_{\alpha,h}$. Specifically, if $p_{n,k}(n_1, \ldots, n_k)$ denotes the probability of any particular partition of $\{1, \ldots, n\}$ induced by $\mathbf{X}_n$ and featuring k distinct blocks with frequencies $(n_1, \ldots, n_k)$, then

$$p_{n,k}(n_1, \ldots, n_k) = V_{n,k} \prod_{i=1}^{k} (1-\alpha)_{(n_i - 1)} \tag{2}$$

where

$$V_{n,k} = \frac{\alpha^k}{\Gamma(n - \alpha k)} \int_0^{+\infty} t^{-\alpha k} h(t) \int_0^1 (1-z)^{n-1-\alpha k} f_\alpha(zt) \mathrm{d}z \mathrm{d}t,$$

with f_α being the positive α-stable density function and $(a)_n = \prod_{i=0}^{n-1}(a+i)$ with the proviso $(a)_0 = 1$. The two parameter PD prior and the normalized GG prior are Gibbs-type priors corresponding to the choices $h(t) = \alpha t^{-\theta}\Gamma(\theta)/\Gamma(\theta/\alpha)$, for any $\theta > -\alpha$, and $h(t) = \exp\{b - b^{1/\alpha}t\}$, for any $b > 0$, respectively. Let K_n the number of distinct observations in $\mathbf{X}_n$, and by $(N_{1,n}, \ldots, N_{K_n,n})$ their frequencies. Pitman [30] showed that, as $n \to +\infty$, $n^{-\alpha}K_n \to S_{\alpha,h}$ almost surely, where $S_{\alpha,h}$ is a nonnegative random variable such that $\mathbb{P}[S_{\alpha,h} \in \mathrm{d}s] = \alpha^{-1}s^{-1/\alpha-1}h(s^{-1/\alpha})f_\alpha(s^{-1/\alpha})\mathrm{d}s$. The random variable $S_{\alpha,h}$ is referred to as the α-diversity.

Proposition 1 in Lijoi et al. [24] and Theorem 3 in Favaro et al. [14] provide explicit expressions for the posterior distributions of $K_{n,m}$ and $M_{n,m}(l)$ with respect to an initial sample $\mathbf{X}_n$ from $\tilde{Q}_{\alpha,h}$. However, as pointed out in the Introduction, for large m the computational burden for evaluating these expressions becomes overwhelming and prevents their practical use. See Favaro et al. [12] for a discussion in the context of the two parameter PD prior. In order to overcome this drawback for a flexible specification of the parameter h, we investigate the asymptotic behaviour of the posterior distributions of $K_{n,m}$ and $M_{n,m}(l)$ for the choice

$$h(t) = \int_D \mathrm{e}^{\zeta - \zeta^{1/\alpha}t} F(\mathrm{d}\zeta), \tag{3}$$

where F is any distribution function on a subset D of the positive real line. Using the terminology of James [22], the subclass of Gibbs-type priors with parameter

h of the form (3) is referred to as the PG class. If F is the Gamma distribution with shape parameter θ/α and scale parameter 1, for any $\alpha \in (0,1)$ and $\theta > 0$, then (3) reduces to the parameter h that characterizes the two parameter PD prior. Furthermore, if F is the degenerate distribution on b, for any $b > 0$, then (3) reduces to the parameter h that characterizes the normalized GG prior. Note that for $\theta \to 0$ and $b \to 0$ the two parameter PD prior coincides with the normalized GG prior.

In the next theorem we characterize the large m asymptotic behavior of the posterior distributions of $K_{n,m}$ and $M_{n,m}(l)$ under the general framework of the PG class. In order to state this result we need to introduce some additional notation. For any $a > 0$ and $b > 0$ let $G_{a,b}$ be Gamma random variable with shape parameter a and scale parameter b, and for any $c > 0$ let T_c be an exponentially tilted α-stable random variable, namely a nonnegative random variable such that

$$\mathbb{P}[T_c \in \mathrm{d}t] = \frac{\mathrm{e}^{-c^{1/\alpha}t} f_\alpha(t)\mathrm{d}t}{\mathrm{e}^{-c}}.$$

Furthermore, for any distribution function F on a subset D of the positive real line, and for any $n \geq 1$ and $k \leq n$ let U_F be a nonnegative random variable such that

$$\begin{aligned}\mathbb{P}[U_F \in \mathrm{d}u] \qquad (4)\\ = \frac{\int_{\underline{D}}^{u^\alpha} \frac{\mathrm{e}^{\zeta}\mathbb{1}_{(\underline{D}^{1/\alpha},\overline{D}^{1/\alpha})}(u)}{(u-\zeta^{1/\alpha})^{1-n}} F(\mathrm{d}\zeta) + \int_{\underline{D}}^{\overline{D}} \frac{\mathrm{e}^{\zeta}\mathbb{1}_{(\overline{D}^{1/\alpha},+\infty)}(u)}{(u-\zeta^{1/\alpha})^{1-n}} F(\mathrm{d}\zeta)}{\frac{\mathrm{e}^{u^\alpha}}{u^{k\alpha-n}} \int_D \int_{\zeta^{1/\alpha}}^{+\infty} u^{k\alpha-n}(u-\zeta^{1/\alpha})^{n-1}\mathrm{e}^{\zeta-u^\alpha}\mathrm{d}uF(\mathrm{d}\zeta)}\mathrm{d}u,\end{aligned}$$

with $\mathbb{1}$ being the indicator function, and with $\overline{D}$ and $\underline{D}$ being the infimum and the supremum, respectively, of the support D of the distribution function F. Note that for $D = \mathbb{R}^+$ one has $\mathbb{1}_{(\overline{D}^{1/\alpha},+\infty)}(u) = 0$ for all $u > 0$ and, hence, the second integral in the numerator of the density function (4) cancels. Throughout the rest of the paper, with a slight abuse of notation we denote by $X \,|\, Y$ a random variable whose distribution coincides with the conditional distribution of X given Y.

Theorem 1. *Let $\mathbf{X}_n$ be a sample from a prior in the PG class with parameter (α, F), and let $\mathbf{X}_n$ featuring $K_n = k$ and $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$. For any positive ζ, let*

$$S_{\alpha,\zeta,n,k} \overset{d}{=} T^{-\alpha}_{\zeta+G_{n/\alpha-k,1}} \left(\frac{\zeta + G_{n/\alpha-k,1}}{\zeta}\right)^{-1}. \qquad (5)$$

As $m \to +\infty$,

$$\frac{K_{n,m}}{m^\alpha} \,|\, \mathbf{X}_n \to S_{\alpha,U_F^\alpha,n,k} \qquad (6)$$

almost surely, where U_F is the random variable with density function (4). Furthermore, as $m \to +\infty$, $m^{-\alpha}M_{n,m}(l) \,|\, \mathbf{X}_n \to c_{l,\alpha}S_{\alpha,U_F^\alpha,n,k}$ almost surely, where $c_{l,\alpha} = (l!)^{-1}\alpha(1-\alpha)_{l-1}$.

The proof of Theorem 1 is deferred to the Appendix. The limiting random variable $S_{\alpha,U_F^\alpha,n,k}$ takes on the interpretation of the posterior counterpart, with respect to the initial sample $\mathbf{X}_n$, of the α-diversity introduced in Pitman [30]. In particular, observe that the distribution of the α-diversity $S_{\alpha,h}$, with h of the form (3), coincides with the distribution of $S_{\alpha,U_F^\alpha,0,0}$. Theorem 1 can be combined with a simulation algorithm for $S_{\alpha,U_F^\alpha,n,k}$ in order to obtain a Monte Carlo approximation, for large m, of any feature of interest from the exact posterior distributions of $K_{n,m}$ and $M_{n,m}(l)$. In particular we are interested in Monte Carlo expectation and quantiles of the limiting posterior distributions, which can be used to approximate exact estimators and associated credible intervals for $K_{n,m}$ and $M_{n,m}(l)$. For any positive ζ, the random variable $S_{\alpha,\zeta,n,k}$ in (5) can be easily sampled by means of the rejection sampling algorithm developed in Devroye [11] for exponentially tilted α-stable random variables. Accordingly, the problem of generating random variates from the distribution of $S_{\alpha,U_F^\alpha,n,k}$ reduces to the problem of devising a suitable sampling algorithm for the random variable U with distribution (4). We conclude this section by describing how to generate random variates from the distribution of U under the choice of F which corresponds to the two parameter PD prior and to the normalized GG prior.

Theorem 1 generalizes Proposition 2 in Favaro et al. [12] to the PG class. Let $B_{a,b}$ be a Beta random variable with parameter (a,b) and let Y_q, for any $q>-1$, be a polynomially tilted α-stable random variable, namely $\mathbb{P}[Y_q\in \mathrm{d}y]=(\Gamma(q\alpha+1)/\alpha\Gamma(q+1))y^{q-1/\alpha-1}f_\alpha(y^{-1/\alpha})\mathrm{d}y$. According to Proposition 2 in Favaro et al. [12], under the assumption of a two parameter PD prior, as $m\to+\infty$

$$\frac{K_{n,m}}{m^\alpha}\,|\,\mathbf{X}_n\to Z_{\alpha,\theta,n,k} \tag{7}$$

almost surely, where $Z_{\alpha,\theta,n,k}\overset{\mathrm{d}}{=}B_{k+\theta/\alpha,n/\alpha-k}Y_{\theta+n}^{-\alpha}$, with $B_{k+\theta/\alpha,n/\alpha-k}$ and $Y_{\theta+n}$ being independent random variables. The limit (7) is recovered as special cases of (6) by setting F to be a Gamma distribution with parameter $(\theta/\alpha,1)$. Indeed, under this assumption for F, the probability distribution displayed in (4) reduces to

$$\mathbb{P}[U\in\mathrm{d}u]=\frac{\alpha}{\Gamma(\theta/\alpha+k)}u^{\theta+k\alpha-1}\mathrm{e}^{-u^\alpha}\mathrm{d}u, \tag{8}$$

for any $u>0$, and it can be verified that $S_{\alpha,U_F^\alpha,n,k}\overset{\mathrm{d}}{=}Z_{\alpha,\theta,n,k}$ for any $\alpha\in(0,1)$ and $\theta>-\alpha$. Theorem 1 thus provides an alternative limiting characterization with respect to the one originally obtained Favaro et al. [12]. The random variable U with distribution (8) can be easily sampled by observing that the transformation U^α is distributed according to a Gamma distribution with parameter $(\theta/\alpha+k,1)$.

Besides introducing a novel characterization for the limiting posterior distribution of $K_{n,m}$ under the assumption of a two parameter PD prior, Theorem 1 provides the limiting characterization for the posterior distributions of $K_{n,m}$ and $M_{n,m}(l)$ under the assumption of the normalized GG prior. This characterization is obtained from Theorem 1 by assuming F to be the degenerate distribution on $b>0$. Under this assumption for F, the probability distribution

(4) reduces to

$$\mathbb{P}[U_F \in \mathrm{d}u] = \frac{\alpha u^{k\alpha-n}(u-b^{1/\alpha})^{n-1}\mathrm{e}^{-u^\alpha}}{\int_b^{+\infty}\mathrm{e}^{-x}x^{k-1}(1-(b/x)^{1/\alpha})^{n-1}\mathrm{d}x}\mathrm{d}u, \tag{9}$$

where

$$\int_b^{+\infty}\mathrm{e}^{-x}x^{k-1}\left(1-\left(\frac{b}{x}\right)^{1/\alpha}\right)^{n-1}\mathrm{d}x = \sum_{i=0}^{n-1}\binom{n-1}{i}(-b^{1/\alpha})^i\Gamma\left(k-\frac{i}{\alpha},b\right),$$

for any $u > b^{1/\alpha}$. In order to generate random variates from the distribution (9), it is sufficient to observe that the density function of the transformed random variable U^α is log concave. Therefore, one can easily sample U^α by means of the adaptive rejection sampling of Gilks and Wild [15]. In general, the adaptive rejection sampling can be exploited to generate random variates from (4) for any choice of F on the positive real line and for which a sampling procedure is available.

3. Illustration

For the sake of clarity we focus on the asymptotic posterior distribution of $K_{n,m}$ under a two parameter PD prior and a normalized GG prior. The same arguments described in this section apply to the asymptotic posterior distribution of $M_{n,m}(l)$, for any $l \geq 1$. In order to implement Theorem 1 with respect to these two prior assumptions, the first issue to face is the specification of the parameters (α,θ) and (α,b). Here, following Favaro et al. [12], we specify these parameters by resorting to an empirical Bayes procedure. This procedure suggests to set (α,θ) and (α,b) so to maximize (2) corresponding to the observed sample $(k,n_1,\ldots,n_k)$ and for $h(t)=\alpha t^{-\theta}\Gamma(\theta)/\Gamma(\theta/\alpha)$ and $h(t)=\exp\{b-b^{1/\alpha}t\}$. Precisely,

$$(\hat{\alpha},\hat{\theta}) = \underset{(\alpha,\theta)}{\arg\max}\left\{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_n}\prod_{i=1}^{k}(1-\alpha)_{(n_i-1)}\right\} \tag{10}$$

and

$$(\hat{\alpha},\hat{b}) = \underset{(\alpha,b)}{\arg\max}\left\{\frac{\int_b^{+\infty}\mathrm{e}^{-x}x^{k-1}(1-(b/x)^{1/\alpha})^{n-1}\mathrm{d}x}{\alpha^{1-k}\mathrm{e}^{-b}\Gamma(n)}\prod_{i=1}^{k}(1-\alpha)_{(n_i-1)}\right\}. \tag{11}$$

In this paper the maximizations (10) and (11) are obtained by using an adaptive version of the differential evolution global stochastic optimizer for non-linearly constrained real valued functions of multiple variables. We refer to Huang et al. [20] for a detailed account on this optimization method. In particular, the integral appearing in the maximization (11) is solved numerically by using the double exponential quadrature algorithm described in Bailey et al. [5], which has proved to be competitive with other numerical quadrature methods for most integrand functions.

Given $(\hat{\alpha},\hat{\theta})$ and $(\hat{\alpha},\hat{b})$, one can determine the Monte Carlo expectation $\bar{S}_{\alpha,U_F^\alpha,n,k}$ of the limiting random variable $S_{\alpha,U_F^\alpha,n,k}$ by generating random variates from its distribution as described in Section 2. Hence, given that the normalizing rate function for $K_{n,m}$ in Theorem 1 is m^α, $\bar{K}_{n,m} = m^\alpha \bar{S}_{\alpha,U_F^\alpha,n,k}$ provides a Monte Carlo approximation, for large m, of the exact estimator of $K_{n,m}$. Similarly, one can simulate appropriate quantiles from the distribution of $S_{\alpha,U_F^\alpha,n,k}$, and then determine a Monte Carlo approximation for the 95% and 99% credible intervals associated to $\bar{K}_{n,m}$. It is still an open problem to measure the accuracy of $\bar{K}_{n,m}$ with respect to m and, hence, to determine an m^* for which $\bar{K}_{n,m^*}$ provides an accurate approximation of the exact posterior expectation $\mathbb{E}[K_{n,m}\,|\,\boldsymbol{X}_n]$. As observed in Favaro et al. [12] under the assumption of two parameter PD prior, such a choice of m^* depends on the values of n and θ. In particular when θ and n are moderately large and not overwhelmingly smaller than m, Favaro et al. [12] showed that a finer normalizing rate function may be used to improve the approximation of $\mathbb{E}[K_{n,m}\,|\,\boldsymbol{X}_n]$. Such a less rough rate function, say $r_{\alpha,\theta,n}(m)$, is determined in such a way that $\mathbb{E}[K_{n,m}\,|\,\boldsymbol{X}_n] = r_{\alpha,\theta,n}(m)\mathbb{E}[S_{\alpha,U_F^\alpha,n,k}]$ and $r_{\alpha,\theta,n}(m)/m^\alpha \to 1$ as $m \to +\infty$. Unfortunately a similar approach can not be applied under the assumption of the normalized GG prior, due to intractable expressions for $\mathbb{E}[K_{n,m}\,|\,\boldsymbol{X}_n]$ and $\mathbb{E}[S_{\alpha,U_F^\alpha,n,k}]$.

We start by illustrating an application of Theorem 1 in a simulation study, and then we present an application to a real dataset arising from empirical linguistic. With regards to the simulation study, we consider two synthetic datasets, say Data-1 and Data-2, generated from a Zeta (or Zipf) distribution function with scale parameter 1.6, whose power law behavior is common in linguistics. The reader is referred to Chapter 7 in Sampson [32] and references therein for a comprehensive account on the use of the Zeta distribution in the context of empirical linguistics. Data-1 consists of 22000 observations featuring 664 species, whereas Data-2 consists of 82000 observations featuring 1564 species. We perform a cross validation study in terms of out-of-sample predictive performance. Specifically, we take a subsample without replacement of $n = 2000$ observations from Data-1 and Data-2, and we make prediction over an additional sample of size $m = 20000$ and $m = 80000$, respectively. In order to perform the cross validation study on representative subsamples, for each dataset we generated 10000 subsamples of size 2000. The empirical deciles of the generated subsamples, in terms of the observed number K_n of distinct species, are 139, 142, 145, 147, 149, 151, 153, 156, 160 for Data-1, and 137, 141, 143, 146, 148, 150, 152, 155, 158 for Data-2. We perform our analysis on subsamples belonging to a selection of these deciles.

For a subsample with $K_n = 149$ species from Data-1, results displayed in Table 1 and Table 2 show how the Monte Carlo approximate estimation $\bar{K}_{n,m}$ changes as the parameter (α,θ) and (α,b), respectively, vary. A similar analysis is performed in Table 3 and Table 4 for a subsample with $K_n = 148$ species from Data-2. The behaviour of these estimates agrees with the interpretation, with respect to the distribution of the number K_n of distinct species, of the parameter (α,θ) and (α,b) under the two parameter PD prior and the normalized GG prior,

Table 1

Data-1: initial samples of size $n = 2000$ with $K_n = 149$ species. Monte Carlo approximate estimations $\bar{K}_{n,m}$, for an additional sample of size $m = 20000$, under a two parameter PD prior with varying parameters α and θ

	$\alpha = 0.3$		$\alpha = 0.6$		$\alpha = 0.9$	
θ	$\bar{K}_{n,m}$	95% c.i.	$\bar{K}_{n,m}$	95% c.i.	$\bar{K}_{n,m}$	95% c.i.
0.5	293	$(239, 346)$	556	$(476, 647)$	1070	$(896, 1249)$
1	292	$(249, 344)$	558	$(482, 657)$	1069	$(905, 1236)$
1.5	298	$(258, 354)$	565	$(475, 666)$	1086	$(953, 1250)$
2	299	$(243, 344)$	565	$(470, 656)$	1091	$(922, 1258)$
2.5	304	$(251, 348)$	568	$(496, 655)$	1092	$(937, 1266)$
3	308	$(261, 360)$	578	$(494, 685)$	1118	$(918, 1294)$
3.5	314	$(269, 360)$	567	$(462, 650)$	1105	$(940, 1282)$
4	310	$(256, 355)$	585	$(512, 675)$	1111	$(949, 1289)$

Table 2

Data-1: initial samples of size $n = 2000$ with $K_n = 149$ species. Monte Carlo approximate estimations $\bar{K}_{n,m}$, for an additional sample of size $m = 20000$, under a normalized GG prior with varying parameters α and b

	$\alpha = 0.3$		$\alpha = 0.6$		$\alpha = 0.9$	
b	$\bar{K}_{n,m}$	95% c.i.	$\bar{K}_{n,m}$	95% c.i.	$\bar{K}_{n,m}$	95% c.i.
0.5	286	$(236, 243)$	553	$(467, 656)$	1099	$(906, 1293)$
1	287	$(237, 340)$	557	$(458, 650)$	1123	$(934, 1314)$
1.5	284	$(237, 337)$	559	$(468, 670)$	1158	$(967, 1409)$
2	288	$(243, 340)$	563	$(467, 665)$	1176	$(1005, 1376)$
2.5	286	$(244, 345)$	563	$(459, 662)$	1209	$(1049, 1431)$
3	287	$(230, 341)$	571	$(480, 681)$	1238	$(1049, 1405)$
3.5	286	$(231, 338)$	573	$(482, 675)$	1264	$(1095, 1487)$
4	287	$(244, 342)$	581	$(485, 674)$	1280	$(1116, 1482)$

respectively. On the one hand, the parameters θ and b control the location of the distribution of K_n: the bigger θ and b the larger the expected number of species tends to be. On the other hand, the parameter α controls the flatness of the distribution of K_n: the bigger α the flatter is the distribution of K_n. In particular, as discussed in De Blasi et al. [10] under the general framework of

Table 3

Data-2: initial samples of size $n = 2000$ with $K_n = 148$ species. Monte Carlo approximate estimations $\bar{K}_{n,m}$, for an additional sample of size $m = 80000$, under a two parameter PD prior with varying parameters α and θ

	$\alpha = 0.3$		$\alpha = 0.6$		$\alpha = 0.9$	
θ	$\bar{K}_{n,m}$	95% c.i.	$\bar{K}_{n,m}$	95% c.i.	$\bar{K}_{n,m}$	95% c.i.
0.5	458	$(387, 548)$	1360	$(1209, 1587)$	4072	$(3479, 4827)$
1	460	$(404, 520)$	1366	$(1144, 1536)$	4124	$(3515, 4699)$
1.5	466	$(404, 541)$	1398	$(1231, 1618)$	4128	$(3411, 4863)$
2	478	$(418, 569)$	1408	$(1202, 1636)$	4278	$(3626, 5223)$
2.5	479	$(409, 554)$	1428	$(1175, 1695)$	4338	$(3760, 4947)$
3	491	$(422, 568)$	1438	$(1253, 1732)$	4387	$(3594, 5229)$
3.5	490	$(417, 555)$	1461	$(1238, 1767)$	4447	$(3807, 5056)$
4	498	$(433, 574)$	1478	$(1288, 1703)$	4460	$(3731, 5093)$

Table 4

Data-2: initial samples of size $n = 2000$ with $K_n = 148$ species. Monte Carlo approximate estimations $\bar{K}_{n,m}$, for an additional sample of size $m = 80000$, under a normalized GG prior with varying parameters α and b

	$\alpha = 0.3$		$\alpha = 0.6$		$\alpha = 0.9$	
b	$\bar{K}_{n,m}$	95% c.i.	$\bar{K}_{n,m}$	95% c.i.	$\bar{K}_{n,m}$	95% c.i.
0.5	447	$(366, 530)$	1344	$(1144, 1607)$	4155	$(3455, 4869)$
1	449	$(372, 533)$	1364	$(1113, 1621)$	4306	$(3630, 5067)$
1.5	446	$(367, 530)$	1380	$(1170, 1618)$	4484	$(3691, 5341)$
2	453	$(381, 538)$	1397	$(1168, 1661)$	4623	$(3947, 5357)$
2.5	453	$(384, 544)$	1408	$(1155, 1671)$	4785	$(4072, 5581)$
3	456	$(364, 547)$	1441	$(1212, 1713)$	4951	$(4121, 5705)$
3.5	455	$(370, 536)$	1459	$(1254, 1737)$	5115	$(4405, 5959)$
4	458	$(389, 544)$	1483	$(1211, 1750)$	5240	$(4640, 6007)$

Gibbs-type priors, a value of α close to 1 determines a large number of species most of which with small frequencies, whereas a value of α close to 0 determines a small number of species with large frequencies. A comparison between results displayed in Table 1 and Table 2 reveals the following behaviour for the Monte Carlo approximate estimations $\bar{K}_{n,m}$: for $\alpha = 0.3$ and any $b \geq \theta$ estimates under

Table 5

Data-1: initial samples of size $n = 2000$ with K_n species and true $K_{n,m} = 664 - K_n$. Monte Carlo approximate estimations $\bar{K}_{n,m}$, with associated 95% credible intervals, for additional sample size $m = 20000$

		Two parameter PD prior			Normalized GG prior		
K_n	$K_{n,m}$	$(\hat{\alpha}, \hat{\theta})$	$\bar{K}_{n,m}$	95% c.i.	$(\hat{\alpha}, \hat{b})$	$\bar{K}_{n,m}$	95% c.i.
142	522	$(0.584, 0.940)$	517	$(438, 609)$	$(0.582, 1.714)$	520	$(440, 601)$
148	516	$(0.591, 0.943)$	550	$(471, 634)$	$(0.589, 1.701)$	550	$(466, 638)$
151	513	$(0.588, 1.000)$	551	$(467, 641)$	$(0.587, 1.801)$	551	$(468, 639)$
157	507	$(0.601, 0.862)$	581	$(494, 670)$	$(0.600, 1.542)$	582	$(495, 672)$

Table 6

Data-2: initial samples of size $n = 2000$ with K_n species and true $K_{n,m} = 1564 - K_n$ Monte Carlo approximate estimations $\bar{K}_{n,m}$ and associated 95% credible intervals, for additional sample size $m = 80000$

		Two parameter PD prior			Normalized GG prior		
K_n	$K_{n,m}$	$(\hat{\alpha}, \hat{\theta})$	$\bar{K}_{n,m}$	95% c.i.	$(\hat{\alpha}, \hat{b})$	$\bar{K}_{n,m}$	95% c.i.
139	1425	$(0.590, 0.841)$	1246	$(1046, 1460)$	$(0.588, 1.540)$	1239	$(1050, 1450)$
145	1419	$(0.612, 0.589)$	1369	$(1160, 1607)$	$(0.610, 1.097)$	1366	$(1159, 1589)$
152	1412	$(0.590, 0.960)$	1306	$(1108, 1511)$	$(0.588, 1.730)$	1301	$(1110, 1501)$
156	1408	$(0.601, 0.853)$	1395	$(1181, 1615)$	$(0.599, 1.528)$	1388	$(1183, 1607)$

the two PD prior are always greater than the corresponding estimates under the normalized GG prior, whereas for $\alpha = 0.9$ and any $b \geq \theta$ estimates under the two PD prior are always smaller than the corresponding estimates under the normalized GG prior. In other terms, for any fixed $b \geq \theta$, the aforementioned effect of the parameter α results more sharpen under the assumption of the normalized GG prior than under the assumption of the two parameter PD prior. The same behaviour emerges from the comparison between results in Table 3 and Table 4.

Table 5 and Table 6 display results related to four subsamples from Data-1 and Data-2, respectively. The maximizations approach described by Equation (10) and Equation (11) have been applied to the collection of selected subsamples, and they resulted into similar values for the parameter α and slightly different values for the parameters θ and b. Furthermore, the corresponding estimations $\bar{K}_{n,m}$ and their associated 95% credible intervals are very similar under the assumption of the two parameter PD prior and the normalized GG prior. Observe that the true values of $K_{n,m}$ is always included in the 95% credible intervals. We retain that such a behaviour is mainly determined by the fact that the two parameter PD prior, for any $\alpha \in (0,1)$ and $\theta > 0$, may be viewed as a suitable mixture of normalized GG priors. Specifically, let $\Pi_{\mathrm{PD}}(\alpha,\theta)$ denote the distribution of the two parameter PD random probability measure, let $\Pi_{\mathrm{GG}}(\alpha,b)$ denote the distribution of the normalized GG random probability measure, and let $G_{\theta/\alpha,1}$ be a Gamma random variable with parameter $(\theta/\alpha,1)$. Then, according to Proposition 21 in Pitman and Yor [31], we can write $\Pi_{\mathrm{PD}}(\alpha,\theta) = \Pi_{\mathrm{GG}}(\alpha, G_{\theta/\alpha,1})$. See also the definition of the PG class in Section 2, as well as the recent paper by James [22] for additional details. In other terms, assuming a two parameter PD prior is equivalent to assuming a normalized GG prior with an Gamma hyper prior over the parameter b. It can be verified that for large datasets the distribution of $G_{\theta/\alpha,1}\,|\,\boldsymbol{X}_n$ tends to be highly concentrated around $\hat{b}$. Therefore, the larger the sample size n and the number of species k tend to be, the more the two parameter PD prior and the normalized GG prior lead to the same inferences for $K_{n,m}$. A repeated analysis, which is not reported in the present paper, for subsamples belonging to the different deciles of Data-1 and Data-2 shows the same behaviour in terms of values for $(\hat{\alpha},\hat{\theta})$ and $(\hat{\alpha},\hat{b})$, and the corresponding Monte Carlo approximate estimations $\bar{K}_{n,m}$.

We conclude by presenting an application of Theorem 1 to a real dataset arising from a study in empirical linguistics. This is a concrete example, typically referred to as the "prosody example", which is drawn from research on speech timing reported in Bachenko and Gale [4]. We refer to the monograph by Sampson [32] for additional details. The prosody example assumes a classification of speech segments into consonants, full vowels, and reduced vowels. Species in the population are strings such as VCV, VCRCV, VCCRCRCV, and so on, using C, V and R to represent the three class of speech segments. We use the TIMIT database (`https://catalog.ldc.upenn.edu/LDC93S1`) as a sample, which consists of $n = 30902$ individual strings featuring $K_n = 309$ species. The maximizations (10) and (11) have been applied to this sample and they resulted in

$$(\hat{\alpha},\hat{\theta}) = (0.393, 3.506) \tag{12}$$

and

$$(\hat{\alpha},\hat{b}) = (0.393, 8.959), \tag{13}$$

respectively. Based on these values (12) and (13), Table 7 displays the Monte Carlo approximate estimation $\bar{K}_{n,m}$ for $m = 20n$, $m = 40n$, $m = 60n$, $m = 80n$ and $m = 100n$, as well as the associated 95% and 99% credible intervals. We observe that, similarly to the simulation study, Monte Carlo approximate estimations are very similar under the assumption of the two parameter PD prior and the normalized GG prior. This is not surprising due to the values for the parameters (α,θ) and (α,b) that we obtained by means of the maximizations (10) and (11).

Table 7

Prosody example: initial sample of size $n = 30902$ with $K_n = 309$ species. Monte Carlo approximate estimations $\bar{K}_{n,m}$ and associated 95% and 99% credible intervals, for various values of the additional sample size m

	Two parameter PD prior			Normalized GG prior		
m	$\bar{K}_{n,m}$	95% c.i.	99% c.i.	$\bar{K}_{n,m}$	95% c.i.	99% c.i.
$20n$	1032	$(919, 1151)$	$(893, 1197)$	1033	$(929, 1145)$	$(884, 1180)$
$40n$	1356	$(1206, 1512)$	$(1173, 1571)$	1356	$(1219, 1504)$	$(1150, 1535)$
$60n$	1579	$(1415, 1773)$	$(1375, 1843)$	1591	$(1430, 1764)$	$(1362, 1818)$
$80n$	1780	$(1584, 1985)$	$(1539, 2064)$	1781	$(1601, 1975)$	$(1525, 2035)$
$100n$	1943	$(1729, 2167)$	$(1681, 2253)$	1944	$(1748, 2156)$	$(1665, 2222)$

Appendix

We start by recalling the distribution of the random partition of $\{1, \ldots, n\}$ induced by a sample $\mathbf{X}_n$ from a Gibbs-type prior in the PG class. We denote by $p_{n,k}(n_1, \ldots, n_k; \alpha, F)$ the probability of any particular partition of the set $\{1, \ldots, n\}$ induced by $\mathbf{X}_n$ and featuring $K_n = k$ distinct blocks with corresponding frequencies $(N_{1,n}, \ldots, N_{n,K_n}) = (n_1, \ldots, n_k)$. By a direct application of (2) we can write

$$\begin{aligned} & p_{n,k}(n_1, \ldots, n_k; \alpha, F) \\ & \quad = \frac{\alpha^k \prod_{i=1}^{k-1} (1-\alpha)_{n_i - 1}}{\Gamma(n - \alpha k)} \\ & \qquad \times \int_0^{+\infty} t^{-\alpha k} \int_D \mathrm{e}^{\zeta - \zeta^{1/\alpha} t} F(\mathrm{d}\zeta) \int_0^1 (1-z)^{n-1-\alpha k} f_\alpha(zt) \mathrm{d}z \mathrm{d}t \\ & \quad = \frac{\alpha^k \prod_{i=1}^{k-1} (1-\alpha)_{n_i - 1}}{\Gamma(n)} \\ & \qquad \times \int_D \int_0^{+\infty} h^{n-1} (h + \zeta^{1/\alpha})^{-n + k\alpha} \mathrm{e}^{\zeta - (h + \zeta^{1/\alpha})^\alpha} F(\mathrm{d}\zeta) \mathrm{d}h. \end{aligned} \tag{14}$$

The probability (14) is obtained by: i) the change of variable $zt = y$, ii) the augmentation $t^{-n} = \Gamma(n)^{-1} \int_0^{+\infty} h^{n-1} \exp\{-th\} \mathrm{d}h$ and iii) the Lapalce transform of a positive α-stable random variable. We denote by $B_{a,b}$ a Beta random variable with parameter (a, b), and by Y_q, for any $q > -1$, a polynomially tilted α-stable random variable, namely $\mathbb{P}[Y_q \in \mathrm{d}y] = (\Gamma(q\alpha + 1)/ \alpha\Gamma(q + 1)) y^{q - 1/\alpha - 1} f_\alpha(y^{-1/\alpha}) \mathrm{d}y$.

Proof of Theorem 1. It is sufficient to prove the almost sure limit for $K_{n,m}$. The almost sure limit for $M_{n,m}(l)$ follows by combining the almost sure limit for $K_{n,m}$ with Corollary 21 in Gnedin et al. [16]. Let $\mathbb{Q}_{\alpha,F}$ be the posterior distribution, with respect to the initial sample $\mathbf{X}_n$, of a prior in the PG class. Furthermore, let $\mathbb{Q}_{\alpha,0}$ be the posterior distribution, with respect to the initial sample $\mathbf{X}_n$, of a prior in the PG class with F being the degenerate distribution in 0. Let $\mathbb{E}_{\alpha,F}$ be the expected value with respect to $\mathbb{Q}_{\alpha,F}$. If we denote by $\mathscr{F}_{n,m}$ the sigma-algebra generated by an additional sample $(X_{n+1},\dots,X_{n+m})$, then we can compute

$$\begin{aligned}
L_{n,m} &= \left.\frac{\mathrm{d}\mathbb{Q}_{\alpha,F}}{\mathrm{d}\mathbb{Q}_{\alpha,0}}\right|_{\mathscr{F}_{n,m}} \\
&= \frac{p_{n+m,K_n+K_{n,m}}(N_{1,n+m},\dots,N_{K_n+K_{n,m},n+m};\alpha,F)}{p_{n,K_n}(N_{1,n},\dots,N_{K_n,n};\alpha,F)} \\
&\quad\times \frac{p_{n,K_n}(N_{1,n},\dots,N_{K_n,n};\alpha,0)}{p_{n+m,K_n+K_{n,m}}(N_{1,n+m},\dots,N_{K_n+K_{n,m},n+m};\alpha,0)} \\
&= \frac{\int_D\int_0^{+\infty} h^{n+m-1}(h+\zeta^{1/\alpha})^{-n-m+\alpha K_n+\alpha K_{n,m}}\mathrm{e}^{\zeta-(h+\zeta^{1/\alpha})^\alpha}\mathrm{d}hF(\mathrm{d}\zeta)}{(K_n)_{K_{n,m}}\int_D\int_0^{+\infty} h^{n-1}(h+\zeta^{1/\alpha})^{-n+\alpha K_n}\mathrm{e}^{\zeta-(h+\zeta^{1/\alpha})^\alpha}\mathrm{d}hF(\mathrm{d}\zeta)}.
\end{aligned}$$

Hence $(L_{n,m},\mathscr{F}_{n,m})_{m\geq1}$ is a $\mathbb{Q}_{\alpha,0}$-martingale, and by a martingale convergence theorem $(L_{n,m})_{m\geq1}$ has a $\mathbb{Q}_{\alpha,0}$ almost sure limit as $m\to+\infty$. We denote by L_n the limiting random variable, for which $\mathbb{E}_{\alpha,0}[L_n]=1$. Let $E_{n,m}=E_1+\ldots+E_{K_n+K_{n,m}}$, where E_i is a negative exponential random variable with mean 1 and E_r is independent of E_s and $(K_n,K_{n.m})$ for any $r\neq s$. Accordingly, we can write

$$\begin{aligned}
L_{n,m} &= \frac{\Gamma(K_n)}{\int_D\int_{\zeta^{1/\alpha}}^{+\infty}(v-\zeta^{1/\alpha})^{n-1}v^{-n+\alpha K_n}\mathrm{e}^{\zeta-v^\alpha}\mathrm{d}h\mathrm{d}\zeta} \\
&\quad\times\frac{1}{\alpha}\int_D \mathrm{e}^{\zeta}\mathbb{E}\left[\mathbb{1}_{(\zeta,+\infty)}(E_{n,m})\left(1-\frac{\zeta^{1/\alpha}}{E_{n,m}^{1/\alpha}}\right)^{n+m-1}\mid\mathscr{F}_{n,m}\right]
\end{aligned}$$

and, as $m\to+\infty$,

$$\begin{aligned}
L_{n,m} &\approx \frac{\Gamma(K_n)}{\int_D\int_{\zeta^{1/\alpha}}^{+\infty}(v-\zeta^{1/\alpha})^{n-1}v^{-n+\alpha K_n}\mathrm{e}^{\zeta-v^\alpha}\mathrm{d}hF(\mathrm{d}\zeta)} \\
&\quad\times\frac{1}{\alpha}\int_D \mathrm{e}^{\zeta}\left(1-\frac{\zeta^{1/\alpha}}{(K_n+K_{n,m})^{1/\alpha}}\right)F(\mathrm{d}\zeta) \\
&\approx \frac{\Gamma(K_n)}{\alpha\int_D\int_{\zeta^{1/\alpha}}^{+\infty}(v-\zeta^{1/\alpha})^{n-1}v^{-n+\alpha K_n}\mathrm{e}^{\zeta-v^\alpha}\mathrm{d}hF(\mathrm{d}\zeta)} \\
&\quad\times\int_D\exp\left\{\zeta-m\frac{\zeta^{1/\alpha}}{K_{n,m}^{1/\alpha}}\right\}F(\mathrm{d}\zeta).
\end{aligned}$$

Furthermore, there exists a nonnegative random variable, say $S_{\alpha,0}$, such that $m^{-1}K_{n,m}^{1/\alpha}\to S_{\alpha,0}$ almost surely with respect to $\mathbb{Q}_{\alpha,0}$, as $m\to+\infty$. This observation follows directly from the fact that $L_{n,m}\to L_n$ almost surely with respect to $\mathbb{Q}_{\alpha,0}$, as $m \to +\infty$. In particular, according to Proposition 2 in Favaro et al. [12] with $\theta = 0$, the limiting random variable $S_{\alpha,0}$ has the following density function

$$
\begin{aligned}
&\mathbb{P}[S_{\alpha,0} \in \mathrm{d}s] \\
&\quad = \frac{\alpha\Gamma(n)}{\Gamma(K_n)\Gamma(n/\alpha - K_n)} s^{-\alpha-1}\mathrm{d}s \\
&\quad\quad \times \int_{s^{-\alpha}}^{+\infty} v^{n/\alpha-1-1/\alpha-1}\left(\frac{s^{-\alpha}}{v}\right)^{K_n-1}\left(1-\frac{s^{-\alpha}}{v}\right)^{n/\alpha-K_n-1} f_\alpha(v^{1/\alpha})\mathrm{d}v \\
&\quad = \frac{\alpha\Gamma(n)}{\Gamma(K_n)\Gamma(n/\alpha - K_n)} s^{-n}\mathrm{d}s \\
&\quad\quad \times \int_0^1 y^{-n/\alpha+1/\alpha+K_n-1}(1-y)^{n/\alpha-K_n-1} f_\alpha(sy^{1/\alpha})\mathrm{d}y.
\end{aligned}
$$

In particular, $S_{\alpha,0} \overset{\mathrm{d}}{=} (B_{K_n,n/\alpha-K_n} Y_{n/\alpha})^{-1/\alpha}$. Since $\mathbb{Q}_{\alpha,F}$ and $\mathbb{Q}_{\alpha,0}$ are mutually absolutely continuous, almost sure convergence holds true with respect to $\mathbb{Q}_{\alpha,F}$, as well. In order to identify the limiting distribution with respect to $\mathbb{Q}_{\alpha,F}$, it is sufficient to exploit the change of measure suggested by $\mathbb{Q}_{\alpha,F}(\cdot) = \int_{(\cdot)}(\mathrm{d}\mathbb{Q}_{\alpha,F}/\mathrm{d}\mathbb{Q}_{\alpha,0})\mathrm{d}\mathbb{Q}_{\alpha,0}$. If we denote by $S_{\alpha,F}$ the random variable with this distribution, then

$$
\begin{aligned}
&\mathbb{P}[S_{\alpha,F} \in \mathrm{d}s] \\
&\quad = \frac{\Gamma(n)}{\alpha\Gamma(n/\alpha - K_n)\int_D \int_{\zeta^{1/\alpha}}^{+\infty}(v-\zeta^{1/\alpha})^{n-1} v^{-n+\alpha K_n} \mathrm{e}^{\zeta - v^\alpha}\mathrm{d}hF(\mathrm{d}\zeta)} \\
&\quad\quad \times \int_D \mathrm{e}^{(\zeta-\zeta^{1/\alpha}s^{-1/\alpha})} s^{n/\alpha-1/\alpha-1} \\
&\quad\quad\quad \times \int_0^1 y^{-n/\alpha+1/\alpha+K_n-1}(1-y)^{n/\alpha-K_n-1} f_\alpha(s^{-1/\alpha}y^{1/\alpha})\mathrm{d}yF(\mathrm{d}\zeta)\mathrm{d}s.
\end{aligned}
\tag{15}
$$

The proof is completed by showing that (15) corresponds to the density function of $S_{\alpha,U_F^\alpha,n,K_n}$. Simple algebraic manipulations of the density function (15) lead to the representation

$$
\mathbb{P}[S_{\alpha,F} \in \mathrm{d}s] = \int_0^{+\infty} \mathbb{P}[S_{\alpha,F,u} \in \mathrm{d}s]\mathbb{P}[U^\alpha \in \mathrm{d}u] \tag{16}
$$

where

$$
\begin{aligned}
&\mathbb{P}[S_{\alpha,F,u} \in \mathrm{d}s] \\
&\quad = \frac{1}{\alpha\Gamma(n/\alpha - K_n)} s^{-1/\alpha-1}\mathrm{d}s \int_0^{+\infty} \frac{\mathrm{e}^{(u-u^{1/\alpha}s^{-1/\alpha})}}{u^{-n/\alpha+K_n}} \\
&\quad\quad \times \int_0^1 y^{-n/\alpha+1/\alpha+K_n-1}(1-y)^{n/\alpha-K_n-1} f_\alpha(s^{-1/\alpha}y^{1/\alpha})\mathrm{d}y
\end{aligned}
\tag{17}
$$

and

$$\begin{aligned}&\mathbb{P}[U \in \mathrm{d}u]\\&\quad = \frac{\int_{\underline{D}}^{u^\alpha} \frac{(u-\zeta^{1/\alpha})^{n-1}}{\mathrm{e}^{-\zeta}\mathbb{1}_{(\underline{D}^{1/\alpha},\overline{D}^{1/\alpha})}(u)}F(\mathrm{d}\zeta) + \int_{\underline{D}}^{\overline{D}} \frac{(u-\zeta^{1/\alpha})^{n-1}}{\mathrm{e}^{-\zeta}\mathbb{1}_{(\overline{D}^{1/\alpha},+\infty)}(u)}F(\mathrm{d}\zeta)}{\frac{\mathrm{e}^{u^\alpha}}{u^{k\alpha-n}}\int_D\int_{\zeta^{1/\alpha}}^{+\infty} u^{k\alpha-n}(u-\zeta^{1/\alpha})^{n-1}\mathrm{e}^{\zeta-u^\alpha}\mathrm{d}uF(\mathrm{d}\zeta)}\mathrm{d}u.\end{aligned}$$

Let $G_{a,b}$ be a Gamma random variable with shape parameter a and scale parameter b, and for any $c > 0$ let T_c be an exponentially tilted α-stable random variable, namely $\mathbb{P}[T_c \in \mathrm{d}t] = \exp\{c - c^{1/\alpha}t\}f_\alpha(t)\mathrm{d}t$. For any positive u, (17) is the density function of $T^{-\alpha}_{u+G_{n/\alpha-K_n,1}}((u + G_{n/\alpha-K_n,1})/u)^{-1}$. The proof is completed.

Alternative proof of Theorem 1. The proof exploits Proposition 9 and Proposition 13 in [30], which provide two characterizations of the random partition $(K_n, N_{1,n}, \dots, N_{K_n,n})$ induced by a sample of size n from a Gibbs-type prior. Specifically, let $\mathbf{X}_n$ be a sample of size n from a Gibbs-type prior belonging to the PG class. According to Proposition 13 in [30], as $n \to +\infty$, $n^{-\alpha}K_n \to S_{\alpha,F}$ almost surely, where $S_{\alpha,F}$ is a nonnegative random variable with density function of the form

$$\mathbb{P}[S_{\alpha,F} \in \mathrm{d}s] = \frac{1}{\alpha}s^{-1/\alpha-1}\mathrm{d}s\int_D \mathrm{e}^{\zeta-\zeta^{1/\alpha}s^{-1/\alpha}}\mathrm{d}F(\mathrm{d}\zeta)f_\alpha(s^{-1/\alpha}). \tag{18}$$

Let $T_{\alpha,F} = S^{-1/\alpha}_{\alpha,F}$ and let $p_{n,k}(n_1,\dots,n_k;t)$ be the conditional probability, given $T_{\alpha,F} = t$, of any particular partition of $\{1,\dots,n\}$ induced by the initial sample $\mathbf{X}_n$ and featuring $K_n = k$ distinct blocks with corresponding frequencies $(N_{1,n},\dots,N_{n,K_n}) = (n_1,\dots,n_k)$. According to Proposition 9 in Pitman [30], one has

$$\begin{aligned}&p_{n,k}(n_1,\dots,n_k;t)\\&\quad = \frac{\alpha^k\prod_{i=1}^k(1-\alpha)_{n_i-1}}{\Gamma(n-k\alpha)}t^{-\alpha k}\int_0^1 p^{n-\alpha k-1}\frac{f_\alpha((1-p)t)}{f_\alpha(t)}\mathrm{d}p.\end{aligned} \tag{19}$$

Of course by integrating the conditional probability (19) with respect to the distribution of $T_{\alpha,F}$ one obtains the probability (14). By combining (18) and (19), a straightforward application of Bayes theorem leads to the following density function

$$\begin{aligned}&\mathbb{P}[T_{\alpha,F} \in \mathrm{d}t \,|\, K_n = k, (N_{1,n},\dots,N_{n,K_n}) = (n_1,\dots,n_k)]\\&\quad = \frac{\Gamma(n)}{\Gamma(n/\alpha - K_n)\int_D\int_{\zeta^{1/\alpha}}^{+\infty}(v-\zeta^{1/\alpha})^{n-1}v^{-n+\alpha K_n}\mathrm{e}^{\zeta-v^\alpha}\mathrm{d}hF(\mathrm{d}\zeta)}\\&\qquad \times t^{-n}\mathrm{d}t\int_D \mathrm{e}^{(\zeta-\zeta^{1/\alpha}t)}\\&\qquad\quad \times\int_0^1 y^{-n/\alpha+1/\alpha+K_n-1}(1-y)^{n/\alpha-K_n-1}f_\alpha(ty^{1/\alpha})\mathrm{d}yF(\mathrm{d}\zeta).\end{aligned}$$

Due to the sufficiency of $(K_n, N_{1,n}, \ldots, N_{K_n,n})$ for $\mathbf{X}_n$ we can write $\mathbb{P}[T_{\alpha,F} \in \mathrm{d}t \,|\, K_n = k, (N_{1,n}, \ldots, N_{n,K_n}) = (n_1, \ldots, n_k)] = \mathbb{P}[T_{\alpha,F} \in \mathrm{d}t \,|\, \mathbf{X}_n]$. Moreover, it can be easily verified that $\mathbb{P}[T_{\alpha,F}^{-\alpha} \in \mathrm{d}s \,|\, \mathbf{X}_n] = \mathbb{P}[S_{\alpha,F} \in \mathrm{d}s]$, where $\mathbb{P}[S_{\alpha,F} \in \mathrm{d}s]$ is the density function in (15). Accordingly, $\mathbb{P}[T_{\alpha,F}^{-\alpha} \in \mathrm{d}s \,|\, \mathbf{X}_n]$ admits the representation (16). This proves the almost sure limit for $K_{n,m}$. The almost sure limit for $M_{n,m}(l)$ follows by a direct application of Corollary 21 in Gnedin et al. [16].

Acknowledgements

The authors are grateful to two Referees for their valuable remarks and suggestions that have lead to a substantial improvement of the presentation. Special thanks are due to Stefano Peluchetti (http://www.scilua.org/) for numerous useful discussions on numerical optimization in Lua. Stefano Favaro is supported by the European Research Council (ERC) through StG N-BNP 306406. Lancelot F. James is supported in part by the grant RGC-HKUST 601712 of the HKSAR.

References

[1] Airoldi E., Costa T., Leisen F., Bassetti F. and Guindani M. (2014). Generalized species sampling priors with latent beta reinforcements. *J. Amer. Statist. Assoc.*, **109**, 1466-1480. MR3293604

[2] Argiento, R., Guglielmi, A. and Pievatolo, A. (2010). Bayesian density estimation and model selection using nonparametric hierarchical mixtures. *Comput. Statist. Data Anal.*, **54**, 816-832. MR2580918

[3] Bacallado, S., Favaro, S. and Trippa, L. (2013). Bayesian nonparametric analysis of reversible Markov chains. *Ann. Statist.*, **41**, 870-896. MR3099124

[4] Bachenko, J. and Gale, W.A. (1993). A corpus-based model of interstress timing and structure. *J. Aco. Soc. Am.*, **94**, 1797.

[5] Bailey, D.H., Jeyabalan, K. and Li, X.S. (2006). A comparison of three high-precision quadrature schemes. *Experiment. Math.*, **14**, 317-329. MR2172710

[6] Barger, K. and Bunge, J. (2010). Objective Bayesian estimation of the number of species. *Bayesian Anal.*, **5**, 619-639. MR2740156

[7] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *J. Amer. Statist. Assoc.*, **88**, 364-373.

[8] Bunge, J., Willis, A. and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annu. Rev. Sta. Appl.*, **1**, 427-445.

[9] Caron, F. and Fox, E.B. (2015). Sparse graphs using exchangeable random measures. *Preprint arXiv:1401.1137*.

[10] De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I. and Ruggiero, M. (2014). Are Gibbs-type priors the most natural generaliza-

tion of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.*, **37**, 212-229.

[11] Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Trans. Model. Comp. Simul.*, **19**, 4.

[12] Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B*, **71**, 993-1008. MR2750254

[13] Favaro, S., Lijoi, A. and Prünster, I. (2012). A new estimator of the discovery probability. *Biometrics*, **68**, 1188-1196. MR3040025

[14] Favaro, S., Lijoi, A. and Prünster, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.*, **23**, 1721-1754. MR3114915

[15] Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337-348.

[16] Gnedin, S., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power law. *Probab. Surv.*, **4**, 146-171. MR2318403

[17] Gnedin, A. and Pitman J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, **138**, 5674-5685. MR2160320

[18] Griffin, J.E., Kolossiatis, M. and Steel, M.F.J. (2013). Comparing distributions by using dependent normalized random measure mixtures. *J. Roy. Statist. Soc. Ser. B*, **75**, 499-529 MR3065477

[19] Guindani, M., Sepulveda, N., Paulino, C.D. and Müller, P. (2014). A Bayesian semiparametric approach for the differential analysis of sequence data. *J. Roy. Statist. Soc. Ser. C*, **63**, 385-404. MR3238158

[20] Huang, V.L., Qin, A.K. and Suganthan, P.N. (2006). Self-adaptive differential evolution algorithm for constrained real-parameter optimization. *Proc. IEEE Congress on Evolutionary Computation, 2006.*

[21] James, L.F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *Preprint arXiv:math/0205093.*

[22] James, L.F. (2013). Stick-breaking PG(α,ζ)-generalized Gamma processes. *Preprint arXiv:1308.6570.*

[23] Lee, J., Quintana, F.A., Müller, P. and Trippa, L. (2013). Defining predictive probability functions for species sampling models. *Statist. Sci.*, **28**, 209-222. MR3112406

[24] Lijoi, A., Mena, R.H. and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, **94**, 715-740. MR2416792

[25] Lijoi, A., Mena, R.H. and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. Roy. Statist. Soc. Ser. B*, **69**, 769-786. MR2370077

[26] Mao, C.X. (2004). Prediction of the conditional probability of discovering a new class. *J. Amer. Statist. Assoc.*, **99**, 1108-1118. MR2109499

[27] Mao, C.X. and Lindsay, B.G. (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika*, **89**, 669-82. MR1929171

[28] Navarrete, C., Quintana, F. and Müller, P. (2008). Some issues on nonparametric Bayesian modeling using species sampling models. *Stat. Model.*, **8**, 3-21. MR2750628

[29] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, **102**, 145-158. MR1337249

[30] Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: a Festschrift for Terry Speed* (D.R. Goldstein, Ed.) *Lecture Notes Monograph Series* **40** 1-34. IMS, Beachwood, OH. MR2004330

[31] Pitman, J. and Yor, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, **25**, 855-900. MR1434129

[32] Sampson, G. (2001). *Empirical Linguistics.* Continuum Press, London - New York.

[33] Zhang, H. and Stern, H. (2009). Sample size calculation for finding unseen species. *Bayesian Anal.*, **4**, 763-792. MR2570088