Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics

Julyan Arbel, Stefano Favaro, Bernardo Nipoti, Yee Whye Teh

Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics

Julyan Arbel
Collegio Carlo Alberto, Italy
email: julyan.arbel@carloalberto.org

Stefano Favaro
University of Torino and Collegio Carlo Alberto, Italy
email: stefano.favaro@unito.it

Bernardo Nipoti
University of Torino and Collegio Carlo Alberto, Italy
email: bernardo.nipoti@carloalberto.org

Yee Whye Teh
University of Oxford, UK
email: y.w.teh@stats.ox.ac.uk

Abstract

Given a sample of size n from a population of individual belonging to different species with unknown proportions, a popular problem of practical interest consists in making inference on the probability $D_n(l)$ that the $(n+1)$-th draw coincides with a species with frequency l in the sample, for any $l = 0, 1, \ldots, n$. This paper contributes to the methodology of Bayesian nonparametric inference for $D_n(l)$. Specifically, under the general framework of Gibbs-type priors we show how to derive credible intervals for the Bayesian nonparametric estimator of $D_n(l)$, and we investigate the large n asymptotic behaviour of such an estimator. Of particular interest are special cases of our results obtained under the assumption of the two parameter Poisson-Dirichlet prior and the normalized generalized Gamma prior, which are two of the most commonly used Gibbs-type priors. With respect to these two prior assumptions, the proposed results are illustrated through a simulation study and a benchmark Expressed Sequence Tags dataset. To the best our knowledge, this illustration provides the first comparative study between the two parameter Poisson-Dirichlet prior and the normalized generalized Gamma prior in the context of Bayesian nonparemetric inference for $D_n(l)$.

KEYWORDS: Asymptotics; Bayesian nonparametrics; credible intervals; discovery probability; Gibbs-

type priors; Good–Turing estimator; normalized generalized Gamma prior; smoothing technique; two parameter Poisson-Dirichlet prior.

1 Introduction

The problem of estimating discovery probabilities is typically associated to situations where an experimenter is sampling from a population of individuals $(X_i)_{i\geq 1}$ belonging to an (ideally) infinite number of species $(X_i^*)_{i\geq 1}$ with unknown proportions $(q_i)_{i\geq 1}$. Given an observable sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, interest lies in estimating the probability that the $(n+1)$-th draw coincides with a species with frequency l in $\boldsymbol{X}_n$, for any $l = 0, 1, \ldots, n$. This probability is denoted by $D_n(l)$ and referred to as the l-discovery. In terms of the species proportions q_i's, we can write

$$D_n(l) = \sum_{i\geq 1} q_i \mathbb{1}_{\{l\}}(N_{i,n}),$$

where $N_{i,n}$ denotes the frequency of the species X_i^* in the sample. Clearly $D_n(0)$ is the proportion of yet unobserved species or, equivalently, the probability of discovering a new species. The reader is referred to Bunge and Fitzpatrick (1993) and Bunge et al. (2014) for two comprehensive reviews on the full range of statistical approaches, parametric and nonparametric as well as frequentist and Bayesian, for estimating the l-discovery and related quantities.

The estimation of the l-discovery has found numerous applications in ecology and linguistics, and its importance has grown considerably in recent years, driven by challenging applications in bioinformatics, genetics, machine learning, design of experiments, etc. For examples, Efron and Thisted (1976) and Church and Gale (1991) discuss applications in empirical linguistics; Good (1953) and Chao and Lee (1992), among many others, discuss the probability of discovering new species of animals in a population; Mao and Lindsay (2002), Navarrete et al. (2008), Lijoi et al. (2007a) and Guindani et al. (2014) study applications in genomics and molecular biology; Zhang (2005) considers applications to network species sampling problems and data confidentiality; Caron and Fox (2015) discuss applications arising from bipartite and sparse random graphs; Rasmussen and Starr (1979) and Chao et al. (2009) investigate optimal stopping procedures in finding new species; Bubeck et al. (2013) study applications within the framework of multi-armed bandits for security analysis of electric power systems.

This paper contributes to the methodology of Bayesian nonparametric inference for $D_n(l)$. As observed in Lijoi et al. (2007), a natural Bayesian nonparametric approach for estimating $D_n(l)$ consists in randomizing the q_i's. Specifically, consider the random probability measure $Q = \sum_{i\geq 1} q_i \delta_{X_i^*}$, 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 as a nonatomic probability measures ν_0 on a space $\mathbb{X}$. Then, it is assumed that

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

for any $n \geq 1$, where $\mathscr{Q}$ is the prior distribution over the species composition. Under the Bayesian nonparametric model (1), the estimator of $D_n(l)$ with respect to a squared loss function, say $\hat{\mathcal{D}}_n(l)$, arises from the predictive distributions characterizing $(X_i)_{i\geq 1}$. Assuming Q in the large class of Gibbs-type random probability measures by Pitman (2003), in this paper we consider the problem of deriving credible intervals for $\hat{\mathcal{D}}_n(l)$, and we study the large n asymptotic behaviour of $\hat{\mathcal{D}}_n(l)$. Before introducing our esults, we briefly review some aspects of $\hat{\mathcal{D}}_n(l)$.

1.1 Preliminaries on $\hat{\mathcal{D}}_n(l)$

We start by recalling the predictive distribution characterizing a Gibbs-type prior. Specifically, let $\boldsymbol{X}_n$ be a sample from a Gibbs-type random probability measure Q and featuring $K_n = k_n$ species $X_1^*, \ldots, X_{K_n}^*$ with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k_n,n})$. According to the celebrated de Finetti's representation theorem, $\boldsymbol{X}_n$ is part of an exchangeable sequence $(X_i)_{i\geq 1}$ whose distribution has been characterized in Pitman (2003) and Gnedin and Pitman (2006) as follows: for any set A in the Borel sigma-algebra of $\mathbb{X}$,

$$\mathbb{P}[X_{n+1} \in A \,|\, \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}}\nu_0(A) + \frac{V_{n+1,k_n}}{V_{n,k_n}}\sum_{i=1}^{k_n}(n_{i,n} - \sigma)\delta_{X_i^*}(A) \tag{2}$$

where $\sigma \in (0,1)$ and $(V_{n,k_n})_{k_n \leq n, n\geq 1}$ are nonnegative weights such that $V_{1,1} = 1$ and $V_{n,k_n} = (n - \sigma k_n)V_{n+1,k_n} + V_{n+1,k_n+1}$. The conditional probability (2) is referred to as the predictive distribution of Q. Two peculiar features of Q emerge directly from (2): i) the probability that $X_{n+1} \notin \{X_1^*, \ldots, X_{K_n}^*\}$ depends only on k_n; ii) the probability that $X_{n+1} = X_i^*$ depends only on $(k_n, n_{i,n})$. See De Blasi et al. (2015) for a review on Gibbs-type priors in Bayesian nonparametrics.

Two of the most commonly used nonparametric priors are of Gibbs-type; these are the two parameter Poisson-Dirichlet (PD) prior in Pitman (1995) and Pitman and Yor (1997), and the normalized generalized Gamma (GG) prior in James (2002) and Pitman (2003). The Dirichlet process by Ferguson (1973) is a limiting special case for $\sigma \to 0$. For any $\sigma \in (0,1)$ and $\theta > -\sigma$, the predictive distribution of the two parameter PD prior is of the form (2) with

$$V_{n,k_n} = \frac{\prod_{i=0}^{k_n-1}(\theta + i\sigma)}{(\theta)_n}, \tag{3}$$

where $(a)_n := \prod_{0\leq i\leq n-1}(a+i)$ with $(a)_0 := 1$; see Pitman (1995) for details on (3). For any $\sigma \in (0,1)$ and $\tau > 0$, the predictive distribution of the normalized GG prior is of the form (2) with

$$V_{n,k_n} = \frac{\sigma^{k_n-1}\mathrm{e}^{\tau^\sigma}}{\Gamma(n)}\sum_{i=0}^{n-1}\binom{n-1}{i}(-\tau)^i\Gamma\left(k_n - \frac{i}{\sigma}; \tau^\sigma\right), \tag{4}$$

where $\Gamma(a,b) := \int_b^{+\infty} x^{a-1}\exp\{-x\}\mathrm{d}x$; see Lijoi et al. (2007b) for details on (4). According to (2), the parameter σ admits an interpretation in terms of the distribution of K_n: the larger σ the higher is the number of species and, among these, most of them have small abundances. In other terms, the larger σ the flatter is the distribution of K_n. The parameters θ and τ are location parameters, namely

the bigger they are the larger the expected number of species tends to be.

Let us denote by $M_{l,n}$ the number of species with frequency l in $\boldsymbol{X}_n$, and by $m_{l,n}$ the corresponding observed value. The predictive distribution of Q has a fundamental role in determining the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ of $D_n(l)$. Indeed, according to the definition of $D_n(l)$, the estimator $\hat{\mathcal{D}}_n(l)$ arises from (2) by suitably specifying the Borel set A. In particular, if $A_0 := \mathbb{X} \setminus \{X_1^*, \ldots, X_{K_n}^*\}$ and $A_l := \{X_i^* : N_{i,n} = l\}$, for any $l = 1, \ldots, n$, then one has

$$\hat{\mathcal{D}}_n(0) = \mathbb{P}[X_{n+1} \in A_0 \,|\, \boldsymbol{X}_n] = \mathbb{E}[Q(A_0) \,|\, \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} \tag{5}$$

and

$$\hat{\mathcal{D}}_n(l) = \mathbb{P}[X_{n+1} \in A_l \,|\, \boldsymbol{X}_n] = \mathbb{E}[Q(A_l) \,|\, \boldsymbol{X}_n] = (l - \sigma) m_{l,n} \frac{V_{n+1,k_n}}{V_{n,k_n}}. \tag{6}$$

Estimators (5) and (6) provide Bayesian counterparts to the celebrated Good–Turing estimator $\check{\mathcal{D}}_n(l) = (l+1)m_{l+1,n}/n$, for any $l = 0, 1, \ldots, n-1$, which is frequentist nonparametric estimator of $D_n(l)$ introduced in Good (1953). The most notable difference between $\hat{\mathcal{D}}_n(l)$ and $\check{\mathcal{D}}_n(l)$ consists in the use of the information in $\boldsymbol{X}_n$: $\check{\mathcal{D}}_n(l)$ is a function of $m_{l+1,n}$, and not on $(k_n, m_{l,n})$ as one would intuitively expect for an estimator of $D_n(l)$. See Favaro et al. (2012) for details.

Under the two parameter PD prior, Favaro et al. (2015) established a large n asymptotic relationship between $\hat{\mathcal{D}}_n(l)$ and $\check{\mathcal{D}}_n(l)$. Due to the irregular behaviour of the $m_{l,m}$'s, the peculiar dependency on $m_{l+1,n}$ makes $\check{\mathcal{D}}_n(l)$ a sensible estimator only if l is sufficiently small with respect to n. See, e.g., Good (1953) and Sampson (2001) for examples of absurd estimates determined by $\check{\mathcal{D}}_n(l)$. In order to overcome this drawback, Good (1953) suggested to smooth $(m_{l,n})_{l\geq 1}$ into a more regular series $(m'_{l,n})_{l\geq 1}$, where $m'_{l,n} = p_l k_n$ with $\mathscr{S} = (p_l)_{l\geq 1}$ being nonnegative weights such that $\sum_{l\geq 0}(l+1)m'_{l+1,n}/n = 1$. The resulting smoothed estimator is

$$\check{\mathcal{D}}_n(l; \mathscr{S}) = (l+1)\frac{m'_{l+1,n}}{n}.$$

See Chapter 7 in Sampson (2001) and references therein for a comprehensive account on smoothing techniques for $\check{\mathcal{D}}_n(l)$. According to Theorem 1 in Favaro et al. (2015), as n becomes large, $\hat{\mathcal{D}}_n(l)$ is asymptotically equivalent to $\check{\mathcal{D}}_n(l; \mathscr{S}_{\mathrm{PD}})$, where $\mathscr{S}_{\mathrm{PD}}$ denotes a smoothing rule such that

$$m'_{l,n} = \frac{\sigma(1-\sigma)_{l-1}}{l!} k_n. \tag{7}$$

Note that (7) is a proper smoothing rule since $\sum_{i\geq 1} \sigma(1-\sigma)_{l-1}/l! = 1$. While the smoothing approach were introduced as an ad hoc tool for post processing the irregular $m_{l,n}$'s in order to improve the performance of $\check{\mathcal{D}}_n(l)$, Theorem 1 in Favaro et al. (2015) shows that, for a large sample size n, a similar smoothing mechanism underlies the Bayesian nonparametric framework (1) with a two parameter PD prior. Interestingly, the smoothing rule $\mathscr{S}_{\mathrm{PD}}$ has been proved to be a generalization of the Poisson smoothing rule discussed in Good (1953) and Engen (1978).

1.2 Contributions of the paper and outline

While proposing a Bayesian nonparametric framework for estimating the l-discovery, Lijoi et al. (2007) did not consider the problem of associating a measure of uncertainty to $\hat{\mathcal{D}}_n(l)$. In this paper we provide an answer to this important problem. With a slight abuse of notation, throughout the paper we write $X \,|\, Y$ to denote a random variable whose distribution coincides with the conditional distribution of X given Y. Since $\hat{\mathcal{D}}_n(l) = \mathbb{E}[Q(A_l) \,|\, \boldsymbol{X}_n]$, the problem of deriving credible intervals for $\hat{\mathcal{D}}_n(l)$ boils down to the problem of characterizing the distribution of $Q(A_l) \,|\, \boldsymbol{X}_n$, for any $l = 0, 1, \ldots, n$. Indeed this distribution takes on the interpretation of the posterior distribution of $D_n(l)$ with respect to the sample $\boldsymbol{X}_n$. For any Gibbs-type priors we provide an explicit expression for $\mathcal{E}_{n,r}(l) := \mathbb{E}[(Q(A_l))^r \,|\, \boldsymbol{X}_n]$, for any $r \geq 1$. Due to the bounded support of $Q(A_l) \,|\, \boldsymbol{X}_n$, the sequence $(\mathcal{E}_{n,r}(l))_{r\geq 1}$ characterizes uniquely the distribution of $Q(A_l) \,|\, \boldsymbol{X}_n$ and, in principle, it can be used to obtain an approximate evaluation of such a distribution. In particular, under the two parameter PD prior and the normalized GG prior we present an explicit and simple characterization of the distribution of $Q(A_l) \,|\, \boldsymbol{X}_n$.

We also study the large n asymptotic behaviour of $\hat{\mathcal{D}}_n(l)$, thus extending Theorem 1 in Favaro et al. (2015) to Gibbs-type priors. Specifically, we show that, as n tends to infinity, $\hat{\mathcal{D}}_n(0)$ and $\hat{\mathcal{D}}_n(l)$ are asymptotically equivalent to $\hat{\mathcal{D}}'_n(0) = \sigma k_n/n$ and $\hat{\mathcal{D}}'_n(l) = (l-\sigma)m_{l,n}/n$, respectively. In other terms, at the order of asymptotic equivalence, any Gibbs-type prior leads to the same approximating estimator $\hat{\mathcal{D}}'_n(l)$. As a corollary we have that $\hat{\mathcal{D}}_n(l)$ is asymptotically equivalent to the smoothed Good–Turing estimator $\check{\mathcal{D}}_n(l; \mathscr{S}_{\mathrm{PD}})$, namely $\mathscr{S}_{\mathrm{PD}}$ is invariant with respect to any prior of Gibbs-type. Refinements of $\hat{\mathcal{D}}'_n(l)$ are presented for the two parameter PD prior and the normalized GG prior. A thorough study of the large n asymptotic behaviour of (2) reveals that: i) for V_{n,k_n} in (3) and (4) the estimator $\hat{\mathcal{D}}_n(l)$ admits large n asymptotic expansions whose first order truncations coincide with $\hat{\mathcal{D}}'_n(l)$; ii) second order truncations depend on $\theta > -\sigma$ and $\tau > 0$, respectively, thus providing approximating estimators which are different between the two parameter PD prior and the normalized GG prior. A discussion of these second order asymptotic refinements is presented with a view towards the problem of finding corresponding refinements of the relationship between $\hat{\mathcal{D}}_n(l)$ and $\check{\mathcal{D}}_n(l; \mathscr{S}_{\mathrm{PD}})$.

Our results are illustrated through a simulation study and the analysis of a benchmark Expressed Sequence Tags (ESTs) dataset. To the best of our knowledge, only the two parameter PD prior has been so far applied in the context of Bayesian nonparametric inference for the discovery probability. In this paper we consider both the two parameter PD prior and the normalized GG prior, thus providing their comparative study. It turns out that the two parameter PD prior leads to estimates of the l-discovery, as well as associated credible intervals, which are very close to those obtained under the assumption of the normalized GG prior. This unexpected behaviour is motivated by resorting to a representation of the two parameter PD prior in terms of a suitable mixture of normalized GG priors. Credible intervals for $\hat{\mathcal{D}}_n(l)$ are also compared with corresponding confidence intervals for the Good–Turing estimator, which were obtained in Mao (2004) and Baayen (2001). A second numerical illustration is devoted to the large n asymptotic behaviour of $\hat{\mathcal{D}}_n(l)$. In particular, by using simulated data we compare the exact estimator $\hat{\mathcal{D}}_n(l)$ with its first order and second order approximations.

In Section 2 we present some distributional results for $Q(A_l) \,|\, \boldsymbol{X}_n$; these results provide a fundamental tool for deriving credible intervals for the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$. In section 3 we investigate the large n asymptotic behaviour of $\hat{\mathcal{D}}_n(l)$, and we discuss its relationships

with smoothed Good–Turing estimators. Section 4 contains some numerical illustrations. Proofs and technical derivations are postponed to the Appendix.

2 Credible intervals for $\hat{\mathcal{D}}_n(l)$

We first recall an integral representation for the V_{n,k_n}'s characterizing the predictive distributions (2). This representation was introduced by Pitman (2003), and it leads to a useful parameterization for Gibbs-type priors. See also Gnedin and Pitman (2006) for details. For any $\sigma \in (0,1)$ let f_σ be the density function of a positive σ-stable random variable, that is $\int_0^{+\infty} \exp\{-tx\} f_\sigma(x)\mathrm{d}x = \exp\{-t^\sigma\}$ for any $t > 0$. Then, for any nonnegative function h, one has

$$V_{n,k_n} = V_{h,(n,k_n)} := \frac{\sigma^{k_n}}{\Gamma(n-\sigma k_n)} \int_0^{+\infty} h(t) t^{-\sigma k_n} \int_0^1 p^{n-1-\sigma k_n} f_\sigma((1-p)t)\mathrm{d}p\mathrm{d}t. \tag{8}$$

According to (2) and (8), a Gibbs-type prior is parameterized by (σ, h, ν_0). We denote by Q_h a Gibbs-type random probability measure with parameter (σ, h, ν_0). The expression (3) for the two parameter PD prior is recovered from (8) by setting $h(t) = p(t;\sigma,\theta) := \sigma\Gamma(\theta)t^{-\theta}/\Gamma(\theta/\sigma)$, for any $\sigma \in (0,1)$ and $\theta > -\sigma$. The expression (4) for the normalized GG prior is recovered from (8) by setting $h(t) = g(t;\sigma,\tau) := \exp\{\tau^\sigma - \tau t\}$, for any $\tau > 0$. See Section 5.4 in Pitman (2003) for details.

Besides providing a parameterization for Gibbs-type priors, the representation (8) leads to a simple numerical evaluation of $V_{h,(n,k_n)}$. Specifically, let $B_{a,b}$ be a Beta random variable with parameter (a,b) and, for any $\sigma \in (0,1)$ and $c > -1$, let $S_{\sigma,c}$ be a positive random variable with density function $f_{S_{\sigma,c}}(x) = \Gamma(c\sigma+1)x^{-c\sigma} f_\sigma(x)/\Gamma(c+1)$. $S_{\sigma,c}$ is typically referred to as the polynomially tilted σ-stable random variable. Simple algebraic manipulations of (8) lead to write

$$V_{h,(n,k_n)} = \frac{\sigma^{k_n-1}\Gamma(k_n)}{\Gamma(n)} \mathbb{E}\left[h\left(\frac{S_{\sigma,k_n}}{B_{\sigma k_n, n-\sigma k_n}}\right)\right], \tag{9}$$

with $B_{\sigma k_n, n-\sigma k_n}$ independent of S_{σ,k_n}. According to (9) a Monte Carlo evaluation of $V_{h,(n,k_n)}$ can be performed by sampling from $B_{\sigma k_n, n-\sigma k_n}$ and S_{σ,k_n}. In this respect, an efficient rejection sampling for $S_{\sigma,c}$ has been proposed by Devroye (2009). The next theorem, combined with (9), provides a practical tool for obtaining an approximate evaluation of the credible intervals for $\hat{\mathcal{D}}_n(l)$.

THEOREM 1. Let $\boldsymbol{X}_n$ be a sample from Q_h featuring $K_n = k_n$ species, labelled by $X_1^*, \ldots, X_{K_n}^*$, with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k_n,n})$. Furthermore, for any set A in the Borel sigma-algebra of $\mathbb{X}$ let $\mu_{n,k_n}(A) = \sum_{1\le i\le k_n}(n_{i,n}-\sigma)\delta_{X_i^*}(A)$. Then, for any $r \ge 1$

$$\begin{aligned}\mathbb{E}[(Q_h(A))^r \mid \boldsymbol{X}_n] = &\sum_{i=0}^{r} \frac{V_{h,(n+r,k_n+r-i)}}{V_{h,(n,k_n)}} (\nu_0(A))^{r-i} \\ &\times \sum_{0\le j_1\le\cdots\le j_i\le r-i} \prod_{q=1}^{i} (\mu_{n,k_n}(A) + j_q(1-\sigma) + q - 1).\end{aligned} \tag{10}$$

Let $\boldsymbol{M}_n := (M_{1,n}, \ldots, M_{n,n}) = (m_{1,n}, \ldots, m_{n,n})$ be the frequency counts from a sample $\boldsymbol{X}_n$ from Q_h. As pointed out in the Introduction, in order to obtain credible intervals for $\hat{\mathcal{D}}_n(l)$ we are interested

in Theorem 1 for two particular specifications of the Borel set A, namely $A_0 = \mathbb{X} \setminus \{X_1^*, \ldots, X_{K_n}^*\}$ and $A_l = \{X_i^* : N_{i,n} = l\}$, for any $l = 1, \ldots, n$. For these Borel sets, (10) reduces to

$$\mathcal{E}_{n,r}(0) = \mathbb{E}[(Q_h(A_0))^r \,|\, \boldsymbol{X}_n] = \sum_{i=0}^{r} \binom{r}{i} (-1)^i \frac{V_{h,(n+i,k_n)}}{V_{h,(n,k_n)}} (n - \sigma k_n)_i \tag{11}$$

and

$$\mathcal{E}_{n,r}(l) = \mathbb{E}[(Q_h(A_l))^r \,|\, \boldsymbol{X}_n] = \frac{V_{h,(n+r,k_n)}}{V_{h,(n,k_n)}} ((l - \sigma) m_{l,n})_r, \tag{12}$$

respectively. Equations (11) and (12) take on the interpretation of the r-th moments of the posterior distribution of $D_n(0)$ and $D_n(l)$ under the assumption of a Gibbs-type prior. In particular for $r = 1$, by using the recursion $V_{h,(n,k_n)} = (n - \sigma k_n) V_{h,(n+1,k_n)} + V_{h,(n+1,k_n+1)}$, (11) and (12) reduce to the Bayesian nonparametric estimators of $D_n(l)$ displayed in (5) and (6), respectively.

The distribution of $Q_h(A_l) \,|\, \boldsymbol{X}_n$ is on $[0, 1]$ and, therefore, it is characterized by $(\mathcal{E}_{n,r}(l))_{r \geq 1}$. The approximation of a distribution given its moments, is a longstanding problem which has been tackled by various approaches such as expansions in polynomial bases, maximum entropy methods and mixtures of distributions. For instance, the polynomial approach consists in approximating the density function of $Q_h(A_l) \,|\, \boldsymbol{X}_n$ with a linear combination of orthogonal polynomials, where the coefficients of the combination are determined by equating $\mathcal{E}_{n,r}(l)$ with the moments of the approximating density. The higher the degree of the polynomials, or equivalently the number of moments used, the more accurate the approximation. As a rule of thumb, ten moments turn out to be enough in most cases. See Provost (2005) for details. The approximating density function of $Q_h(A_l) \,|\, \boldsymbol{X}_n$ can then be used to obtain an approximate evaluation of the credible intervals for $\hat{\mathcal{D}}_n(l)$. This is typically done by generating random variates, via rejection sampling, from the approximating distribution of $Q_h(A_l) \,|\, \boldsymbol{X}_n$. See Arbel et al. (2015) for details.

Under the assumption of the two parameter PD prior and the normalized GG prior, (11) and (12) lead to explicit and simple characterizations for the distributions of $Q_p(A_l) \,|\, \boldsymbol{X}_n$ and $Q_g(A_l) \,|\, \boldsymbol{X}_n$, respectively. Before stating these results, let us introduce some additional notation. Let $G_{a,1}$ be a Gamma random variable with parameter $(a, 1)$ and, for any $\sigma \in (0, 1)$ and $b > 0$, let $R_{\sigma,b}$ be a random variable with density function $f_{R_{\sigma,b}}(x) = \exp\{b^\sigma - bx\} f_\sigma(x)$. $R_{\sigma,b}$ is typically referred to as the exponentially tilted σ-stable random variable. Finally, let us define

$$W_{a,b} = \frac{b R_{\sigma,b}}{b R_{\sigma,b} + G_{a,1}}, \tag{13}$$

where $G_{a,1}$ is independent of $R_{\sigma,b}$. Note that the random variable $W_{a,b}$ is nonnegative and defined on the set $[0, 1]$. In the next propositions we show that the distributions of $Q_p(A_l) \,|\, \boldsymbol{X}_n$ and $Q_g(A_l) \,|\, \boldsymbol{X}_n$, for any $l = 0, 1, \ldots, n$, are obtained by a suitable randomization of $W_{a,b}$ over b.

PROPOSITION 1. Let $\boldsymbol{X}_n$ be a sample from Q_p featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$. Furthermore, let Z_p be a nonnegative random variable with density function of the form

$$f_{Z_p}(x) = \frac{\sigma}{\Gamma(\theta/\sigma + k_n)} x^{\theta + \sigma k_n - 1} \mathrm{e}^{-x^\sigma} \mathbb{1}_{(0,+\infty)}(x).$$

Then,

$$Q_p(A_0) \,|\, \boldsymbol{X}_n \stackrel{\mathrm{d}}{=} W_{n-\sigma k_n, Z_p} \stackrel{\mathrm{d}}{=} B_{\theta+\sigma k_n, n-\sigma k_n}$$

and

$$Q_p(A_l) \,|\, \boldsymbol{X}_n \stackrel{\mathrm{d}}{=} B_{(l-\sigma)m_{l,n}, n-\sigma k_n-(l-\sigma)m_{l,n}}(1 - W_{n-\sigma k_n, Z_p}) \stackrel{\mathrm{d}}{=} B_{(l-\sigma)m_{l,n}, \theta+n-(l-\sigma)m_{l,n}}.$$

Proposition 2. Let $\boldsymbol{X}_n$ be a sample from Q_g featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$. Furthermore, let Z_g be a nonnegative random variable with density function of the form

$$f_{Z_g}(x) = \frac{\sigma x^{\sigma k_n - n}(x-\tau)^{n-1} \exp\{-x^\sigma\} \mathbb{1}_{(\tau,+\infty)}(x)}{\sum_{0 \le i \le n-1} \binom{n-1}{i} (-\tau)^i \Gamma(k_n - i/\sigma; \tau^\sigma)}. \tag{14}$$

Then,

$$Q_g(A_0) \,|\, \boldsymbol{X}_n \stackrel{\mathrm{d}}{=} W_{n-\sigma k_n, Z_g}$$

and

$$Q_g(A_l) \,|\, \boldsymbol{X}_n \stackrel{\mathrm{d}}{=} B_{(l-\sigma)m_{l,n}, n-\sigma k_n-(l-\sigma)m_{l,n}}(1 - W_{n-\sigma k_n, Z_g}).$$

According to Proposition 1 and Proposition 2, the random variables $Q_p(A_0) \,|\, \boldsymbol{X}_n$ and $Q_g(A_0) \,|\, \boldsymbol{X}_n$ have a common structure driven by (13). Moreover, for any $l = 1, \ldots, n$, note that $Q_p(A_l) \,|\, \boldsymbol{X}_n$ and $Q_g(A_l) \,|\, \boldsymbol{X}_n$ are obtained by taking the same random proportion $B_{(l-\sigma)m_{l,n}, n-\sigma k_n-(l-\sigma)m_{l,n}}$ of $(1 - W_{n-\sigma k_n, Z_p})$ and $(1 - W_{n-\sigma k_n, Z_g})$, respectively. Under the assumption of the two parameter PD prior and the normalized GG prior, Proposition 1 and Proposition 2 provide practical tools for deriving credible intervals for the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$, for any $l = 0, 1, \ldots, n$. This is typically done by performing a numerical evaluation of appropriate quantiles of the distribution of $Q_p(A_l) \,|\, \boldsymbol{X}_n$ and $Q_g(A_l) \,|\, \boldsymbol{X}_n$. Note that, in the special case of the Beta distribution, quantiles can be also determined explicitly as solutions of a certain class of non-linear ordinary differential equations. See Steinbrecher and Shaw (2008) and references therein for a detailed account on this approach.

In this paper we resort to a Monte Carlo evaluation of the credible intervals of $\hat{\mathcal{D}}_n(l)$; this approach requires to generate random variates from the distribution of $Q_p(A_l) \,|\, \boldsymbol{X}_n$ and $Q_g(A_l) \,|\, \boldsymbol{X}_n$. With regards to the two parameter PD prior, sampling from $Q_p(A_l) \,|\, \boldsymbol{X}_n$, for any $l = 0, 1, \ldots, n$, is straightforward. Indeed, according to Proposition 1, it requires to generate random variates from a Beta distribution. With regards to the normalized GG prior, sampling from $Q_p(A_l) \,|\, \boldsymbol{X}_n$, for any $l = 0, 1, \ldots, n$, is also straightforward. First, let us consider the problem of sampling from Z_g with density function (14). It can be easily verified that the density function of the transformed random variable Z_g^σ is log-concave and, therefore, one can sample from Z_g^σ by means of the adaptive rejection sampling of Gilks and Wild (1992). Given Z_g, the problem of sampling from $W_{n-\sigma k_n, Z_g}$ boils down to the problem of generating random variates from the distribution of the exponentially tilted σ-stable random variable R_{σ, Z_g}. This can be done by resorting to an efficient rejection sampling proposed by Devroye (2009).

3 Large sample asymptotics for $\hat{\mathcal{D}}_n(l)$

We investigate the large n asymptotic behavior of the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$, with a view towards its asymptotic relationships with smoothed Good–Turing estimators. We recall from the Introduction that, under a Gibbs-type prior, the most notable difference between the Good–Turing estimator $\check{\mathcal{D}}_n(l)$ and $\hat{\mathcal{D}}_n(l)$ can be traced back to the different use of the information contained in the sample $\boldsymbol{X}_n$. Specifically: i) $\check{\mathcal{D}}_n(0)$ is a function of $m_{1,n}$ while $\hat{\mathcal{D}}_n(0)$ is a function of k_n; ii) $\check{\mathcal{D}}_n(l)$ is a function of $m_{l+1,n}$ while $\hat{\mathcal{D}}_n(l)$ is a function of $m_{l,n}$, for any $l = 1, \ldots, n$. Let $a_n \simeq b_n$ mean that $\lim_{n\to+\infty} a_n/b_n = 1$, namely a_n and b_n are asymptotically equivalent as n tends to infinity. Hereafter we show that, as n tends to infinity, $\hat{\mathcal{D}}_n(l) \simeq \check{\mathcal{D}}_n(l; \mathscr{S}_{\text{PD}})$, where $\mathscr{S}_{\text{PD}}$ is the smoothing rule displayed in (7). Such a result thus generalizes Theorem 1 in Favaro et al. (2015) to the entire class of Gibbs-type priors.

THEOREM 2. Let $\boldsymbol{X}_n$ be a sample from Q_h featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$. Then,

$$\hat{\mathcal{D}}_n(0) = \frac{\sigma k_n}{n} + o\left(\frac{k_n}{n}\right) \tag{15}$$

and

$$\hat{\mathcal{D}}_n(l) = (l-\sigma)\frac{m_{l,n}}{n} + o\left(\frac{m_{l,n}}{n}\right). \tag{16}$$

The asymptotic equivalence between $\hat{\mathcal{D}}_n(l)$ and $\check{D}_n(l; \mathscr{S}_{\text{PD}})$ arises by combining Theorem 2 with an interesting interplay between the large n asymptotic behaviors of K_n and $M_{l,n}$. Specifically, let $A_n \overset{\text{a.s.}}{\simeq} B_n$ as $n \to +\infty$ mean that $\lim_{n\to+\infty} A_n/B_n = 1$ almost surely, namely A_n and B_n are almost surely asymptotically equivalent as n tends to infinity. By a direct application of Proposition 13 in Pitman (2003) and Corollary 21 in Gnedin et al. (2007) we can write

$$M_{l,n} \overset{\text{a.s.}}{\simeq} \frac{\sigma(1-\sigma)_{l-1}}{l!} K_n, \tag{17}$$

as $n \to +\infty$. That is, as n tends to infinity the number of species with frequency l becomes a proportion $\sigma(1-\sigma)_{l-1}/l!$ of the number of species. By suitably combining (15) and (16) with (17), we obtain

$$\hat{\mathcal{D}}_n(l) \simeq (l+1)\frac{m_{l+1,n}}{n} \simeq (l+1)\frac{\frac{\sigma(1-\sigma)_l}{(l+1)!}k_n}{n}, \tag{18}$$

for any $l = 0, 1, \ldots, n$. See the Appendix for details on (18). The first equivalence in (18) shows that, as n tends to infinity, $\hat{\mathcal{D}}_n(l)$ is asymptotically equal to the Good–Turing estimator $\check{\mathcal{D}}_n(l)$, whereas the second equivalence shows that, as n tends to infinity, the frequency counts $m_{l,n}$ in $\check{\mathcal{D}}_n(l)$ are smoothed via $\mathscr{S}_{\text{PD}}$. We refer to Section 2 in Favaro et al. (2015) for a relationship between the smoothing rule $\mathscr{S}_{\text{PD}}$ and the Poisson smoothing in Good (1953).

A peculiar feature of $\mathscr{S}_{\text{PD}}$ is that it does not depend on the function h characterizing the Gibbs-type prior. In other terms the smoothing rule $\mathscr{S}_{\text{PD}}$ is invariant with respect to the choice of any prior of Gibbs-type; for instance, the two parameter PD prior and the normalized GG prior lead to the same smoothing rule $\mathscr{S}_{\text{PD}}$. This invariance property of $\mathscr{S}_{\text{PD}}$ is clearly determined by the fact that the asymptotic equivalences in (18) arise by combining (17), which does not depend on h, with

(15) and (16), which also do not depend of h. Intuitively, smoothing rules depending on the function h, if any exists, necessarily require to combine refinements of the asymptotic expansions (15) and (16) with corresponding refinements of the asymptotic equivalence (17). Under the assumption of the two parameter PD prior and the normalized GG prior, the next propositions provide asymptotic refinements of Theorem 2.

PROPOSITION 3. Let $\boldsymbol{X}_n$ be a sample from Q_p featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$. Then,

$$\hat{\mathcal{D}}_n(0) = \frac{\sigma k_n}{n} + \frac{\theta}{n} + o\left(\frac{k_n}{n}\right)$$

and

$$\hat{\mathcal{D}}_n(l) = (l-\sigma)\frac{m_{l,n}}{n} - (l-\sigma)\frac{\theta m_{l,n}}{n^2} + o\left(\frac{m_{l,n}}{n^2}\right).$$

PROPOSITION 4. Let $\boldsymbol{X}_n$ be a sample from Q_g featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$. Then,

$$\hat{\mathcal{D}}_n(0) = \frac{\sigma k_n}{n} + \frac{\tau n k_n^{-1/\sigma}}{n} + o\left(\frac{k_n}{n}\right)$$

and

$$\hat{\mathcal{D}}_n(l) = (l-\sigma)\frac{m_{l,n}}{n} - (l-\sigma)\frac{\tau n k_n^{-1/\sigma} m_{l,n}}{n^2} + o\left(\frac{m_{l,n}}{n^2}\right).$$

According to Proposition 3 and Proposition 4, the large n asymptotic approximations in Theorem 2 can be interpreted as first order approximations, in the sense that they coincide with the one-term truncations of the asymptotic series expansions of $\hat{\mathcal{D}}_n(0)$ and $\hat{\mathcal{D}}_n(l)$, respectively. The combination of these first order approximations with (17) led to the asymptotic relationship in (18). As a direct consequence $\mathscr{S}_{\text{PD}}$ takes on the interpretation of a first order smoothing rule, namely a smoothing rule independent of the function h. In Proposition 3 and Proposition 4 we introduced second order approximations of $\hat{\mathcal{D}}_n(0)$ and $\hat{\mathcal{D}}_n(l)$ by considering a two-term truncation of the corresponding asymptotic series expansions. Note that it is sufficient to include the second term in order to introduce the dependency on $\theta > -\sigma$ and $\tau > 0$, respectively, and then obtaining approximations of $\hat{\mathcal{D}}_n(0)$ and $\hat{\mathcal{D}}_n(l)$ which are different between the two parameter PD prior and the normalized GG prior.

Despite the availability of the second order approximations in Proposition 3 and Proposition 4, it can be easily verified that their combination with corresponding second order refinements of (17) does not lead to a second order refinement of (18). Indeed such a combination still leads to the first order asymptotic equivalence displayed in (18). Specifically, if we let $A_n = O(B_n)$ as $n \to +\infty$ mean that $\limsup_{n\to+\infty} A_n/B_n < +\infty$ almost surely, then a second order refinement of (17), arising from Gnedin et al. (2007), can be expressed as follows

$$M_{l,n} = \frac{\sigma(1-\sigma)_{l-1}}{l!} K_n + O\left(\frac{K_n}{n^{\sigma/2}}\right). \tag{19}$$

However, second order terms in Propositions 3 and Proposition 4 are absorbed by $O\left(K_n/n^{\sigma/2}\right)$ in (19). Furthermore, even if a finer version of (19) was available, its combination with Propositions 3 and Proposition 4 would produce higher order terms preventing the resulting expression from being

interpreted as a Good–Turing estimator and, therefore, any smoothing rule from being elicited. In other terms, under the two parameter PD prior and the normalized GG prior, the relationship between $\hat{\mathcal{D}}_n(l)$ and $\check{\mathcal{D}}_n(l)$ only holds at the order of asymptotic equivalence.

4 Illustrations

We illustrate our results through the analysis of synthetic data and real data. Synthetic data are generated from the Zeta distribution, whose power law behavior is common in a variety of applications. See Sampson (2001) and references therein for applications of the Zeta distribution in empirical linguistics. Recall that a Zeta random variable Z is such that $\mathbb{P}[Z = z] = z^{-s}/C(s)$, for $z = \{1, 2, \ldots\}$ and $s > 1$, where $C(s) = \sum_{i \geq 1} i^{-s}$. We consider a Zeta distribution with parameter $s = 1.1$ and $s = 1.5$. For each one of these values we draw 500 samples of size $n = 1000$ from Z, we order them according to the number of observed species k_n, and we split them in 5 groups: for $i = 1, 2, \ldots, 5$, the i-th group of samples will be composed by 100 samples featuring a total number of observed species k_n that stays between the quantiles of order $(i-1)/5$ and $i/5$ of the empirical distribution of k_n. Then we pick at random one sample for each group and label it with the corresponding index i. This procedure leads to five sample for each one of the two values of the parameter s, namely $s = 1.1$ and $s = 1.5$.

With regards to the analysis of real data, we consider ESTs data generated by sequencing two *Naegleria gruberi* complementary DNA libraries; these are prepared from cells grown under different culture conditions, namely aerobic and anaerobic conditions. The rate of gene discovery depends on the degree of redundancy of the library from which such sequences are obtained. Correctly estimating the relative redundancy of such libraries, as well as other quantities such as the probability of sampling a new or a rarely observed gene, is of great importance since it allows one to optimize the use of expensive experimental sampling techniques. The *Naegleria gruberi* aerobic library consists of $n = 959$ ESTs with $k_n = 473$ distinct genes and $m_{l,959} = 346, 57, 19, 12, 9, 5, 4, 2, 4, 5, 4, 1, 1, 1, 1, 1, 1,$ for $l = \{1, 2, \ldots, 12\} \cup \{16, 17, 18\} \cup \{27\} \cup \{55\}$. The *Naegleria gruberi* anaerobic library consists of $n = 969$ ESTs with $k_n = 631$ distinct genes and $m_{l,969} = 491, 72, 30, 9, 13, 5, 3, 1, 2, 0, 1, 0, 1,$ for $l \in \{1, 2, \ldots, 13\}$. We refer to Susko and Roger (2004) for a detailed account on the *Naegleria gruberi* libraries.

We focus on the two parameter PD prior and the normalized GG prior. In order to apply our results, we need to specify (σ, θ) and (σ, τ). Although one can undertake a full Bayesian approach by specifying a prior distribution for these parameters, for the sake of simplicity here we undertake an empirical Bayes approach. In other terms we choose the values of (σ, θ) and (σ, τ) that maximize the likelihood function with respect to the sample $\boldsymbol{X}_n$ featuring $K_n = k_n$ and $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k_n,n})$. Formally, we set $(\sigma, \theta) = (\hat{\sigma}, \hat{\theta})$ and $(\sigma, \tau) = (\hat{\sigma}, \hat{\tau})$ where

$$(\hat{\sigma}, \hat{\theta}) = \arg\max_{(\sigma,\theta)} \left\{ \frac{\prod_{i=0}^{k_n - 1} (\theta + i\sigma)}{(\theta)_n} \prod_{i=1}^{k_n} (1 - \sigma)_{(n_{i,n} - 1)} \right\} \tag{20}$$

and

$$(\hat{\sigma},\hat{\tau}) = \arg\max_{(\sigma,\tau)} \left\{ \frac{\mathrm{e}^{\tau^\sigma}\sigma^{k_n-1}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau)^i \Gamma\left(k_n - \frac{i}{\sigma}; \tau^\sigma\right) \prod_{i=1}^{k_n} (1-\sigma)_{(n_{i,n}-1)} \right\}. \tag{21}$$

Under the assumption of the two parameter PD prior, Favaro et al. (2015) showed that for large datasets there are no relevant differences between the full Bayesian approach and the empirical Bayes approach. This is because the posterior distribution of the parameter (σ,θ) is highly concentrated around $(\hat{\sigma},\hat{\theta})$. It can be checked that a similar behaviour characterizes the posterior distribution of (σ,τ). We refer to Favaro et al. (2015) and Lijoi et al. (2007) for a detailed discussion of the empirical Bayes approach in relationship with the full Bayesian approach.

For each one of the proposed datasets, Table 1 reports the sample size n, the number of species k_n, and the values of $(\hat{\sigma},\hat{\theta})$ and $(\hat{\sigma},\hat{\tau})$ obtained by the maximizations (20) and (21), respectively. Note that the value of $\hat{\sigma}$ obtained under the two parameter PD prior coincides, up to a negligible error, with the value of $\hat{\sigma}$ obtained under the normalized GG prior. In general, we expect the same behaviour for any Gibbs-type prior. This is not surprising if we look at the likelihood function of a sample $\boldsymbol{X}_n$ from a Gibbs-type random probability measure Q_h, i.e.,

$$\frac{\sigma^{k_n}\prod_{i=1}^{k_n}(1-\sigma)_{(n_i-1)}}{\Gamma(n-\sigma k_n)} \int_0^{+\infty} h(t) t^{-\sigma k_n} \int_0^1 p^{n-1-\sigma k_n} f_\sigma((1-p)t)\mathrm{d}p\mathrm{d}t. \tag{22}$$

Apart from σ, any other parameter is introduced in (22) via the function h, which does not depend on the sample size n and the number of species k_n. Then, it seems reasonable to expect that for large n and k_n the maximization of (22) with respect to σ leads to a value $\hat{\sigma}$ which is very close to the value that would be obtained by maximizing (22) with $h(t)=1$. In other terms, the larger n and k_n tend to be, the more the effect of the function h on $\hat{\sigma}$ tends to vanish.

Table 1: Simulated data and *Naegleria gruberi* libraries. For each sample we report the sample size n, the number of species k_n and the maximum likelihood values $(\hat{\sigma},\hat{\theta})$ and $(\hat{\sigma},\hat{\tau})$.

				PD		GG	
	sample	n	k_n	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\tau}$
Simulated data: $s=1.1$	1	1000	642	0.914	2.086	0.913	2.517
	2	1000	650	0.905	3.812	0.905	4.924
	3	1000	656	0.910	3.236	0.910	4.060
	4	1000	663	0.916	2.597	0.916	3.156
	5	1000	688	0.920	3.438	0.920	4.225
Simulated data: $s=1.5$	1	1000	128	0.624	1.207	0.622	3.106
	2	1000	135	0.675	0.565	0.673	0.957
	3	1000	138	0.684	0.487	0.682	0.795
	4	1000	146	0.656	1.072	0.655	2.302
	5	1000	149	0.706	0.377	0.704	0.592
Naegleria	Aerobic	959	473	0.669	46.241	0.684	334.334
	Anaerobic	969	631	0.656	155.408	0.656	4151.075

4.1 Credible intervals

We apply Propositions 1 and Proposition 2 in order to endow the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ with credible intervals. With regards to the two parameter PD prior, for $l = 0$ we generate 5000 draws from the distribution of a beta random variable $B_{\hat{\theta}+\hat{\sigma}k_n, n-\hat{\sigma}k_n}$ while, for $l \geq 1$ we sample 5000 draws from the distribution of a beta random variable $B_{(l-\hat{\sigma})m_{l,n}, \hat{\theta}+n-(l-\hat{\sigma})m_{l,n}}$. In both cases, we compute the quantiles of order $\{0.025, 0.975\}$ of the empirical distribution and obtain 95% posterior credible intervals for $\hat{\mathcal{D}}_n(l)$. The procedure for the normalized GG case is only slightly more elaborate but still quite straightforward. By exploiting the adaptive rejection algorithm by Gilks and Wild (1992), we sample 5000 draws from Z_g with density function (14). In turn, we sample 5000 draws from $W_{n-\hat{\sigma}k_n, Z_g}$. We then use the quantiles of order $\{0.025, 0.975\}$ of the empirical distribution of $W_{n-\hat{\sigma}k_n, Z_g}$ to obtain 95% posterior credible intervals for $\hat{\mathcal{D}}_n(0)$. Similarly, if $l \geq 1$, we sample 5000 draws from the distribution of a beta random variable $B_{(l-\hat{\sigma})m_{l,n}, n-\hat{\sigma}k_n-(l-\hat{\sigma})m_{l,n}}$ and use the quantiles of the empirical distribution of $B_{(l-\hat{\sigma})m_{l,n}, n-\hat{\sigma}k_n-(l-\hat{\sigma})m_{l,n}}(1-W_{n-\hat{\sigma}k_n, Z_g})$ as extremes of the posterior credible interval for $\hat{\mathcal{D}}_n(l)$.

Table 2: Simulated data with $s = 1.1$. We report the true value of the probability $D_n(l)$ and the Bayesian nonparametric estimates of $D_n(l)$ with 95% credible intervals.

			Good–Turing		PD		GG	
	sample	$D_n(l)$	$\check{\mathcal{D}}_n(l)$	95%-c.i.	$\hat{\mathcal{D}}_n(l)$	95%-c.i.	$\hat{\mathcal{D}}_n(l)$	95%-c.i.
	1	0.599	0.588	(0.440, 0.736)	0.587	(0.557, 0.618)	0.588	(0.558, 0.620)
	2	0.592	0.590	(0.454, 0.726)	0.590	(0.559, 0.621)	0.591	(0.562, 0.620)
$l = 0$	3	0.600	0.599	(0.462, 0.736)	0.598	(0.568, 0.628)	0.599	(0.567, 0.630)
	4	0.605	0.609	(0.473, 0.745)	0.609	(0.579, 0.638)	0.608	(0.577, 0.638)
	5	0.599	0.634	(0.499, 0.769)	0.634	(0.603, 0.664)	0.635	(0.604, 0.663)
	1	0.050	0.044	(0.037, 0.051)	0.051	(0.038, 0.065)	0.051	(0.038, 0.065)
	2	0.052	0.054	(0.046, 0.062)	0.056	(0.043, 0.071)	0.055	(0.042, 0.070)
$l = 1$	3	0.051	0.046	(0.039, 0.053)	0.054	(0.040, 0.068)	0.053	(0.040, 0.068)
	4	0.055	0.046	(0.039, 0.053)	0.051	(0.038, 0.065)	0.051	(0.038, 0.065)
	5	0.061	0.052	(0.045, 0.059)	0.051	(0.038, 0.065)	0.050	(0.038, 0.064)
	1	0.015	0.030	(0.022, 0.038)	0.016	(0.009, 0.025)	0.016	(0.009, 0.025)
	2	0.022	0	(0, 0)	0.016	(0.009, 0.025)	0.016	(0.009, 0.025)
$l = 5$	3	0.019	0.012	(0.008, 0.016)	0.020	(0.013, 0.030)	0.021	(0.012, 0.030)
	4	0.015	0.006	(0.003, 0.009)	0.020	(0.013, 0.030)	0.021	(0.013, 0.031)
	5	0.007	0.012	(0.007, 0.017)	0.008	(0.004, 0.015)	0.008	(0.003, 0.015)
	1	0	0.011	n.a.	0	(0, 0)	0	(0, 0)
	2	0.007	0	(0, 0)	0.009	(0.004, 0.016)	0.009	(0.004, 0.016)
$l = 10$	3	0.011	0	(0, 0)	0.009	(0.004, 0.016)	0.009	(0.004, 0.016)
	4	0.011	0	(0, 0)	0.009	(0.004, 0.016)	0.009	(0.004, 0.016)
	5	0	0.011	n.a.	0	(0, 0)	0	(0, 0)

Under the two parameter PD prior and the normalized GG prior, and with respect to the synthetic data, Table 2 and Table 3 show the estimated l-discoveries, for $l = 0, 1, 5, 10$, and the corresponding 95% posterior credible intervals. It is apparent that the two parameter PD prior and the normalized GG prior lead to the same inferences for the l-discovery. We retain that such a behaviour is mainly

Table 3: Simulated data with $s = 1.5$. We report the true value of the probability $D_n(l)$ and the Bayesian nonparametric estimates of $D_n(l)$ with 95% credible intervals.

			Good–Turing		PD		GG	
	sample	$D_n(l)$	$\check{\mathcal{D}}_n(l)$	95%-c.i.	$\hat{\mathcal{D}}_n(l)$	95%-c.i.	$\hat{\mathcal{D}}_n(l)$	95%-c.i.
	1	0.099	0.080	(0.010, 0.150)	0.081	(0.065, 0.098)	0.081	(0.065, 0.098)
	2	0.103	0.092	(0.012, 0.172)	0.092	(0.075, 0.110)	0.091	(0.075, 0.110)
$l = 0$	3	0.095	0.096	(0.014, 0.178)	0.095	(0.078, 0.114)	0.095	(0.076, 0.113)
	4	0.096	0.096	(0.015, 0.177)	0.097	(0.079, 0.116)	0.097	(0.080, 0.115)
	5	0.093	0.108	(0.019, 0.197)	0.106	(0.087, 0.126)	0.105	(0.087, 0.124)
	1	0.030	0.038	(0.031, 0.045)	0.030	(0.020, 0.042)	0.030	(0.021, 0.042)
	2	0.037	0.030	(0.024, 0.036)	0.030	(0.021, 0.041)	0.030	(0.020, 0.042)
$l = 1$	3	0.034	0.034	(0.028, 0.040)	0.030	(0.021, 0.042)	0.031	(0.021, 0.042)
	4	0.029	0.040	(0.033, 0.047)	0.033	(0.023, 0.045)	0.033	(0.022, 0.044)
	5	0.040	0.026	(0.021, 0.031)	0.032	(0.022, 0.044)	0.032	(0.023, 0.043)
	1	0.013	0.012	(0.008, 0.016)	0.013	(0.007, 0.021)	0.013	(0.007, 0.021)
	2	0.011	0.006	(0.003, 0.009)	0.004	(0.001, 0.009)	0.004	(0.001, 0.009)
$l = 5$	3	0.010	0.012	(0.007, 0.017)	0.009	(0.004, 0.015)	0.009	(0.004, 0.016)
	4	0.010	0.036	(0.024, 0.048)	0.009	(0.004, 0.015)	0.009	(0.004, 0.015)
	5	0.012	0	(0, 0)	0.013	(0.007, 0.021)	0.013	(0.006, 0.021)
	1	0.019	0	(0, 0)	0.019	(0.011, 0.028)	0.019	(0.011, 0.028)
	2	0	0.011	n.a.	0	(0, 0)	0	(0,0)
$l = 10$	3	0.011	0.011	(0.006, 0.016)	0.009	(0.004, 0.016)	0.009	(0.004, 0.016)
	4	0	0	n.a.	0	(0,0)	0	(0,0)
	5	0.006	0	(0, 0)	0.009	(0.004, 0.016)	0.009	(0.004, 0.017)

determined by the fact that the two parameter PD prior, for any $\sigma \in (0,1)$ and $\theta > 0$, may be viewed as a mixture of normalized GG priors. Specifically, let $\mathscr{Q}_p(\sigma,\theta)$ be the distribution of the two parameter PD random probability measure, let $\mathscr{Q}_g(\sigma,b)$ be the distribution of the normalized GG random probability measure, and let $G_{\theta/\sigma,1}$ be a Gamma random variable with parameter $(\theta/\sigma,1)$. Then, according to Proposition 21 in Pitman and Yor (1997), we can write $\mathscr{Q}_p(\sigma,\theta) = \mathscr{Q}_g(\sigma, G^{1/\sigma}_{\theta/\sigma,1})$. 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 $\tau^{1/\sigma}$. As we pointed out before, for large datasets the distribution of $G^{1/\sigma}_{\theta/\sigma,1} \,|\, \boldsymbol{X}_n$ tends to be highly concentrated around $\hat{\tau}$. Therefore, the larger the sample size n and the number of species k_n tend to be, the more the two parameter PD prior and the normalized GG prior lead to the same inferences for the l-discovery.

Our study is completed by comparing the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ with the Good–Turing estimator $\check{\mathcal{D}}_n(l)$. As expected, Good–Turing estimates are not reliable as soon as l is not very small compared to n. See, e.g., the cases $l = 5$ and $l = 10$. Of course, as pointed out in the Introduction, these estimates may be improved by introducing a suitable smoothing rule for the frequency counts $m_{l,n}$'s. We are not aware of a non-asymptotic approach for devising confidence intervals for $\check{\mathcal{D}}_n(l)$; furthermore, we found that different procedures are used according to the choice of $l = 0$ and $l \geq 1$. We relied on the asymptotic confidence interval by Mao (2004) for $l = 0$ and on the confidence interval described by Church and Gale (1991) for $l \geq 1$. See also Baayen (2001) for details. We observe that the confidence intervals for $\check{\mathcal{D}}_n(l)$ are wider than the corresponding credible intervals for $\hat{\mathcal{D}}_n(l)$ when $l = 0$, and narrower if $l \geq 1$. Differently from the credible intervals for

$\hat{\mathcal{D}}_n(l)$, the confidence intervals for $\check{\mathcal{D}}_n(l)$ are symmetric about $\check{\mathcal{D}}_n(l)$; such a behaviour is determined by the Gaussian approximation used by Mao (2004) and Church and Gale (1991) to derive confidence intervals.

Table 4: *Naegleria gruberi* aerobic and anaerobic libraries. For each sample and for $l = 0, 1, 5, 10$, we report the Bayesian nonparametric estimates of $D_n(l)$ with 95% credible intervals.

		Good–Turing		PD		GG	
	sample	$\check{\mathcal{D}}_n(l)$	95%-c.i.	$\hat{\mathcal{D}}_n(l)$	95%-c.i.	$\hat{\mathcal{D}}_n(l)$	95%-c.i.
$l = 0$	Aerobic	0.361	(0.293, 0.429)	0.361	(0.331, 0.391)	0.361	(0.332, 0.389)
	Anaerobic	0.507	(0.451, 0.562)	0.509	(0.478, 0.537)	0.507	(0.480, 0.532)
$l = 1$	Aerobic	0.119	(0.107, 0.131)	0.114	(0.095, 0.134)	0.110	(0.092, 0.131)
	Anaerobic	0.149	(0.135, 0.162)	0.148	(0.129, 0.169)	0.150	(0.131, 0.172)
$l = 5$	Aerobic	0.031	(0.024, 0.038)	0.039	(0.028, 0.052)	0.039	(0.028, 0.053)
	Anaerobic	0.031	(0.024, 0.038)	0.050	(0.038, 0.064)	0.050	(0.038, 0.064)
$l = 10$	Aerobic	0.046	(0.037, 0.055)	0.046	(0.034, 0.060)	0.047	(0.034, 0.061)
	Anaerobic	0.011	n.a.	0	(0, 0)	0	(0, 0)

4.2 Large sample approximations

We conclude our illustration by analyzing the accuracy of the large n approximations of $\hat{\mathcal{D}}_n(l)$ introduced in Theorem 2, Proposition 3 and Proposition 4. To this end we consider the simulated datasets described above. Under the assumption of the two parameter PD prior and the normalized GG prior, and for $l = 0, 1, 5, 10$, we compare the true discovery probabilities $D_n(l)$ with the Bayesian nonparametric estimates of $D_n(l)$ and with their first order and second order approximations. Note that Theorem 2 shows that the first order approximation of $\hat{\mathcal{D}}_n(l)$ is invariant within the whole class of Gibbs-type priors and involves only the parameter σ. As displayed in Table 1, the empirical Bayes estimates for the parameter σ can be slightly different under the two parameter PD and the normalized GG prior. Nonetheless, given that this difference is almost negligible, in this illustration we considered only the first oder approximation of $\hat{\mathcal{D}}_n(l)$ with the parameter $\sigma = \hat{\sigma}$ set as indicated in (20).

Results of this comparative study are reported in Table 5. We also include, as an overall measure of the performance of the exact and approximate estimators, the sum of squared errors (SSE), defined, for a generic estimator $\hat{D}_n(l)$ of the l-discovery, as $\text{SSE}(\hat{D}_n(l)) = \sum_{0 \leq l \leq n} (\hat{D}_n(l) - d_n(l))^2$, with $d_n(l)$ being the true value of $D_n(l)$. It is interesting to notice that, for all the considered samples, there are not substantial differences between the SSEs of the exact Bayesian nonparametric estimates and the SSEs of the first and second order approximate Bayesian nonparametric estimates. Arguably, given the sample size of the datasets we are considering, the first order approximation is already pretty accurate and, thus, the approximation error does not contribute significantly to increase the SSE. Finally, as expected, the order of magnitude of the SSE referring to the not-smoothed Good–Turing estimator is much larger than the one corresponding to the Bayesian nonparametric estimators.

Table 5: Simulated data with $s = 1.1$. We report the true value of the probability $D_n(l)$, the Good–Turing estimates of $D_n(l)$ and the exact and approximate Bayesian nonparametric estimates of $D_n(l)$.

		$s = 1.1$				
		1	2	3	4	5
$l = 0$	$D_n(l)$	0.599	0.592	0.600	0.605	0.599
	$\check{\mathcal{D}}_n(l)$	0.588	0.590	0.599	0.609	0.634
	$\hat{\mathcal{D}}_n(l)$ under PD	0.587	0.590	0.598	0.609	0.634
	$\hat{\mathcal{D}}_n(l)$ under GG	0.588	0. 591	0.599	0.608	0.635
	1st ord.	0.587	0.588	0.597	0.608	0.633
	2nd ord. PD	0.589	0.592	0.600	0.610	0.6366
	2nd ord. GG	0.589	0.592	0.600	0.610	0.636
$l = 1$	$D_n(l)$	0.050	0.052	0.051	0.055	0.061
	$\check{\mathcal{D}}_n(l)$	0.044	0.054	0.046	0.046	0.052
	$\hat{\mathcal{D}}_n(l)$ under PD	0.051	0.056	0.054	0.051	0.051
	$\hat{\mathcal{D}}_n(l)$ under GG	0.051	0.055	0.053	0.051	0.050
	1st ord.	0.051	0.056	0.054	0.051	0.051
	2nd ord. PD	0.051	0.056	0.054	0.051	0.051
	2nd ord. GG	0.051	0.056	0.054	0.051	0.0512
$l = 5$	$D_n(l)$	0.015	0.022	0.019	0.015	0.007
	$\check{\mathcal{D}}_n(l)$	0.030	0	0.012	0.006	0.012
	$\hat{\mathcal{D}}_n(l)$ under PD	0.016	0.016	0.020	0.020	0.008
	$\hat{\mathcal{D}}_n(l)$ under GG	0.016	0.016	0.021	0.021	0.008
	1st ord.	0.016	0.016	0.020	0.020	0.008
	2nd ord. PD	0.016	0.016	0.020	0.020	0.008
	2nd ord. GG	0.016	0.016	0.020	0.020	0.008
$l = 10$	$D_n(l)$	0	0.007	0.011	0.011	0
	$\check{\mathcal{D}}_n(l)$	0.011	0	0	0	0.011
	$\hat{\mathcal{D}}_n(l)$ under PD	0	0.009	0.009	0.009	0
	$\hat{\mathcal{D}}_n(l)$ under GG	0	0.009	0.009	0.009	0
	1st ord.	0	0.009	0.009	0.009	0
	2nd ord. PD	0	0.009	0.009	0.009	0
	2nd ord. GG	0	0.009	0.009	0.009	0
$10^4 \times \text{SSE}(\check{\mathcal{D}}_n)$		289.266	275.881	256.886	254.416	255.655
$10^4 \times \text{SSE}(\hat{\mathcal{D}}_n)$ under PD		3.534	2.057	1.137	4.883	15.437
$10^4 \times \text{SSE}(\hat{\mathcal{D}}_n)$ under GG		3.399	2.080	1.149	4.852	15.045
$10^4 \times \text{SSE}(\hat{\mathcal{D}}_n)$ 1st ord.		3.780	2.142	1.180	4.776	14.456
$10^4 \times \text{SSE}(\hat{\mathcal{D}}_n)$ 2st ord. PD		3.275	2.011	1.128	5.041	17.007
$10^4 \times \text{SSE}(\hat{\mathcal{D}}_n)$ 2st ord. GG		3.279	2.014	1.130	5.035	16.984

A Appendix

This appendix contains: i) the proofs of Theorems 1, Proposition 1, Proposition 2, Theorem 2, Proposition 3 and Proposition 4; ii) details on the derivation of the asymptotic equivalence between $\hat{\mathcal{D}}_n(l)$ and $\check{\mathcal{D}}_n(l;\mathscr{S}_{\mathrm{PD}})$.

Let $\boldsymbol{X}_n = (X_1,\dots,X_n)$ be a sample from a Gibbs-type RPM Q_h. Recall that, due to the discreteness of Q_h, the sample $\boldsymbol{X}_n$ features $K_n = k_n$ species, labelled by $X_1^*,\dots,X_{K_n}^*$, with corresponding frequencies $(N_{1,n},\dots,N_{K_n,n}) = (n_{1,n},\dots,n_{k_n,n})$. Furthermore, let $M_{l,n} = m_{l,n}$ be the number of species with frequency l, namely $M_{l,n} = \sum_{1\le i\le K_n} \mathbb{1}_{\{N_{i,n}=l\}}$ such that $\sum_{1\le i\le n} M_{i,n} = K_n$ and $\sum_{1\le i\le n} iM_{i,n} = n$. For any $\sigma\in(0,1)$ let f_σ be the density function of a positive σ-stable random variable. According to Proposition 13 in Pitman (2003), as $n\to+\infty$

$$\frac{K_n}{n^\sigma} \xrightarrow{\text{a.s.}} S_{\sigma,h} \tag{23}$$

and

$$\frac{M_{l,n}}{n^\sigma} \xrightarrow{\text{a.s.}} \frac{\sigma(1-\sigma)_{l-1}}{l!} S_{\sigma,h}, \tag{24}$$

where $S_{\sigma,h}$ is a random variable with density function $f_{S_{\sigma,h}}(s) = \sigma^{-1}s^{-1/\sigma-1}h(s^{-1/\sigma})f_\sigma(s^{-1/\sigma})$. Note that by the fluctuation limits displayed in (23) and (24), as n tends to infinity the number of species with frequency l in a sample of size n from Q_h becomes, almost surely, a proportion $\sigma(1-\sigma)_{l-1}/l!$ of the total number of species in the sample. All the random variables introduced in this Appendix are meant to be assigned on a common probability space $(\Omega,\mathscr{F},\mathbb{P})$.

A.1 Proofs

Proof of Theorem 1. We proceed by induction. Note that the result holds for $r=1$, and obviously for any sample size $n\ge 1$. Let us assume that it holds for a given $r\ge 1$, and also for any sample size $n\ge 1$. Then, the $(r+1)$-th moment of $Q_h(A)\,|\,\boldsymbol{X}_n$ can be written as follows

$$\begin{aligned}
&\mathbb{E}[Q_h^r(A)\,|\,\boldsymbol{X}_n]\\
&\quad= \int_A\cdots\int_A \mathbb{P}[X_{n+r+1}\in A\,|\,\boldsymbol{X}_n, X_{n+1}=x_{n+1},\dots,X_{n+r}=x_{n+r}]\\
&\qquad\times \mathbb{P}[X_{n+r}\in \mathrm{d}x_{n+r}\,|\,\boldsymbol{X}_n, X_{n+1}=x_{n+1},\dots,X_{n+r-1}=x_{n+r-1}]\\
&\qquad\quad\times\cdots\times \mathbb{P}[X_{n+2}\in \mathrm{d}x_{n+2}\,|\,\boldsymbol{X}_n, X_{n+1}=x_{n+1}]\mathbb{P}[X_{n+1}\in \mathrm{d}x_{n+1}\,|\,\boldsymbol{X}_n]\\
&\quad= \int_A \mathbb{E}[Q_h^r(A)\,|\,\boldsymbol{X}_n, X_{n+1}=x_{n+1}]\\
&\qquad\times\left(\frac{V_{h,(n+1,k_n+1)}}{V_{h,(n,k_n)}}\nu_0(\mathrm{d}x_{n+1}) + \frac{V_{h,(n+1,k_n)}}{V_{h,(n,k_n)}}\sum_{i=1}^{k_n}(n_i-\sigma)\delta_{X_i^*}(\mathrm{d}x_{n+1})\right).
\end{aligned}$$

Further, by the assumption on the r-th moment and by dividing A into $(A\setminus\boldsymbol{X}_n)\cup(A\cap\boldsymbol{X}_n)$, one obtains

$$\mathbb{E}[Q_h^{r+1}(A)\,|\,\boldsymbol{X}_n]$$

$$= \sum_{i=0}^{r} \frac{V_{n+r+1,k_n+r+1-i}}{V_{h,(n,k_n)}} [\nu_0(A)]^{r+1-i} R_{r,i}(\mu_{n,k_n}(A) + 1 - \sigma)$$
$$+ \sum_{i=1}^{r+1} \frac{V_{n+r+1,k_n+r+1-i}}{V_{h,(n,k_n)}} [\nu_0(A)]^{r+1-i} \mu_{n,k_n}(A) R_{r,i-1}(\mu_{n,k_n}(A) + 1),$$

where we defined $R_{r,i}(\mu) := \sum_{0 \le j_1 \le \cdots \le j_i \le r-i} \prod_{1 \le l \le i} (\mu + j_l(1-\sigma) + l - 1)$. The proof is completed by noting that, by means of simple algebraic manipulations, $R_{r+1,i}(\mu) = R_{r,i}(\mu + 1 - \sigma) + \mu R_{r,i-1}(\mu + 1)$. Note that when $\nu_0(A) = 0$ and $i = r$, the convention $\nu_0(A)^{r-i} = 0^0 = 1$ is adopted. □

Proof of Proposition 1. Let us consider the Borel sets $A_0 := \mathbb{X} \setminus \{X_1^*, \ldots, X_{K_n}^*\}$ and $A_l := \{X_i^* : N_{i,n} = l\}$, for any $l = 1, \ldots, n$. The two parameter PD prior is a Gibbs-type prior with $h(t) = p(t; \sigma, \theta) := \sigma\Gamma(\theta) t^{-\theta} / \Gamma(\theta/\sigma)$, for any $\sigma \in (0,1)$ and $\theta > -\sigma$. Therefore one has $V_{n,k_n} = V_{p,(n,k_n)} = [(\theta)_n]^{-1} \prod_{0 \le i \le k_n - 1} (\theta + i\sigma)$. By a direct application of Theorem 1 we can write

$$\begin{aligned}
\mathbb{E}[Q_h^r(A_0) \mid \boldsymbol{X}_n] &= \sum_{i=0}^{r} \binom{r}{i} (-1)^i \frac{(\theta)_n}{(\theta)_{n+i}} (n - \sigma k_n)_i \\
&= (\theta)_n \frac{(\theta + \sigma k_n)_r}{(\theta)_n (\theta + n)_r} \\
&= \frac{(\theta + \sigma k_n)_r}{(\theta + \sigma k_n + n - \sigma k_n)_r},
\end{aligned}$$

which is r-th moment of a Beta random variable with parameter $(\theta + \sigma k, n - \sigma k)$. Let us define the random variable $Y = Z_p R_{\sigma, Z_p}$. Then, it can be easily verified that Y has density function

$$\begin{aligned}
f_Y(y) &= \int_0^\infty \frac{1}{z} f_{R_{\sigma,z}}(y/z) f_{Z_p}(z) \mathrm{d}z \\
&= \frac{\sigma}{\Gamma(\theta/\sigma + k_n)} \int_0^\infty \mathrm{e}^{z^\sigma - y - z^\sigma} z^{\theta + \sigma k_n - 2} f_\sigma(y/z) \mathrm{d}z \\
&= \frac{\sigma}{\Gamma(\theta/\sigma + k_n)} y^{\theta + \sigma k_n - 1} \mathrm{e}^{-y} \int_0^\infty u^{-(\theta + \sigma k_n)} f_\sigma(u) \mathrm{d}u
\end{aligned}$$

where, by Equation 60 in Pitman (2003), $\int_0^\infty u^{-(\theta+\sigma k_n)} f_\sigma(u) \mathrm{d}u = \Gamma(\theta/\sigma + k_n) / \sigma\Gamma(\theta + \sigma k_n)$. Hence Y is a Gamma random variable with parameter $(\theta + \sigma k_n, 1)$. Accordingly, we have $W_{n-\sigma k_n, Z_p} \stackrel{\mathrm{d}}{=} B_{\theta+\sigma k_n, n-\sigma k_n}$. Similarly, by a direct application of Theorem 1, for any $l > 1$ we can write

$$\begin{aligned}
\mathbb{E}[Q_h^r(A_l) \mid \boldsymbol{X}_n] &= \frac{(\theta)_n}{(\theta)_{n+r}} ((l - \sigma) m_{l,n})_r \\
&= \frac{((l - \sigma) m_{l,n})_r}{((l - \sigma) m_{l,n})_r + \theta + n - (l - \sigma) m_{l,n}},
\end{aligned}$$

which is the r-th moment of a Beta random variable with parameter $((l-\sigma)m_{l,n}, \theta + n - (l-\sigma)m_{l,n})$. Finally, the decomposition $B_{(l-\sigma)m_{l,n}, \theta+n-(l-\sigma)m_{l,n}} \stackrel{\mathrm{d}}{=} B_{(l-\sigma)m_{l,n}, n-\sigma k_n-(l-\sigma)m_{l,n}} (1 - W_{n-\sigma k_n, Z_p})$ follows from a characterization of Beta random variables in Theorem 1 in Jambunathan (1954). It can be also easily verified by using the moments of Beta random variables. □

Proof of Proposition 2. Let us consider the Borel sets $A_0 := \mathbb{X} \setminus \{X_1^*, \ldots, X_{K_n}^*\}$ and $A_l :=$

$\{X_i^* : N_{i,n} = l\}$, for any $l = 1, \ldots, n$. The two parameter PD prior is a Gibbs-type prior with $h(t) = g(t; \sigma, \tau) := \exp\{\tau^\sigma - \tau t\}$, for any $\tau > 0$. By a direct application of Theorem 1 we can write

$$\begin{aligned} &\mathbb{E}[Q_g^r(A_0) \,|\, \boldsymbol{X}_n] \\ &\quad = \frac{\sigma\Gamma(n)}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k_n)} \int_0^1 w^r (1-w)^{n-1-\sigma k_n} \int_0^{+\infty} t^{-\sigma k_n} \mathrm{e}^{-\tau t} f_\sigma(wt) \mathrm{d}t\mathrm{d}w, \end{aligned} \tag{25}$$

where

$$\begin{aligned} C_{\sigma,\tau,n,k_n} &:= \frac{\sigma\Gamma(n)}{\Gamma(n-\sigma k_n)} \int_0^{+\infty} t^{-\sigma k_n} \mathrm{e}^{-\tau t} \int_0^1 (1-w)^{n-1-\sigma k_n} f_\sigma(wt) \mathrm{d}w\mathrm{d}t \\ &= \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau)^i \Gamma(k - i/\sigma; \tau^\sigma). \end{aligned}$$

Hereafter we show that (25) coincides with the r-th moment of the random variable $W_{n-\sigma k_n, Z_g}$. Given $Z_g = z$ it is easy to find that the distribution of $W_{n-\sigma k_n, z}$ has the following density function

$$f_{W_{n-\sigma k_n, z}}(w) = \frac{\exp\{z^\sigma\}}{z\Gamma(n-k_n\sigma)} (1-w)^{n-k_n\sigma-1} \int_0^{+\infty} u^{n-k_n\sigma} \mathrm{e}^{-u} f_\sigma\left(\frac{uw}{z}\right) \mathrm{d}u.$$

By randomizing over z with respect to the distribution of Z_g provides the distribution of $W_{n-\sigma k_n, Z_g}$. Specifically,

$$\begin{aligned} f_{W_{n-\sigma k_n, Z_g}}(w) &= \frac{\sigma}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k_n)} (1-w)^{n-\sigma k_n-1} \\ &\quad \times \int_\tau^\infty z^{-n+\sigma k_n-1} (z-\tau)^{n-1} \int_0^\infty u^{n-\sigma k_n} \mathrm{e}^{-u} f_\sigma\left(\frac{uw}{z}\right) \mathrm{d}u\mathrm{d}z \\ &= \frac{\sigma}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k)} (1-w)^{n-\sigma k_n-1} \\ &\quad \times \int_\tau^\infty (z-\tau)^{n-1} \int_0^\infty t^{n-\sigma k_n} \mathrm{e}^{-tz} f_\sigma(wt) \,\mathrm{d}t\mathrm{d}z \\ &= \frac{\sigma\Gamma(n)}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k_n)} (1-w)^{n-\sigma k_n-1} \int_0^\infty t^{-\sigma k_n} \mathrm{e}^{-\tau t} f_\sigma(wt) \,\mathrm{d}t. \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathbb{E}[W_{n-\sigma k_n, Z_g}^r] \\ &\quad = \frac{\sigma\Gamma(n)}{C_{\sigma,\tau,n,k_n}\Gamma(n-\sigma k_n)} \int_0^1 w^r (1-w)^{n-\sigma k_n-1} \int_0^\infty t^{-\sigma k_n} \mathrm{e}^{-\tau t} f_\sigma(wt) \,\mathrm{d}t\mathrm{d}w \end{aligned}$$

which coincides with (25). We complete the proof by determining the distribution of the random variable $Q_g(A_l) \,|\, \boldsymbol{X}_n$, for any $l > 1$. Again, by a direct application of Theorem 1 we can write

$$\begin{aligned} &\mathbb{E}[Q_g^r(A_l) \,|\, \boldsymbol{X}_n] \\ &= ((l-\sigma)m_{l,n})_r \frac{\frac{\sigma^{k_n}}{\Gamma(n-\sigma k_n+r)}}{\frac{\sigma^{k_n}}{\Gamma(n-\sigma k_n)}} \frac{\int_0^{+\infty} t^{-\sigma k_n} \exp\{-\tau t\} \int_0^1 (1-z)^{n+r-1-\sigma k_n} f_\sigma(zt) \mathrm{d}t\mathrm{d}z}{\int_0^{+\infty} t^{-\sigma k_n} \exp\{-\tau t\} \int_0^1 (1-z)^{n-1-\sigma k_n} f_\sigma(zt) \mathrm{d}t\mathrm{d}z} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\Gamma(n-\sigma k_n)}{\Gamma\left((l-\sigma)m_{l,n}\right)\Gamma(\sum_{1\leq i\neq l\leq n} i m_{i,n} - \sigma\sum_{1\leq i\neq l\leq n} m_{i,n})} \\
&\quad\times \int_0^1 x^{(l-\sigma)m_{l,n}+r-1}(1-x)^{\sum_{1\leq i\neq l\leq n} i m_{i,n}-\sigma\sum_{1\leq i\neq l\leq n} m_{i,n}-1} \\
&\qquad\times \frac{\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 (1-z)^{n+r-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z}{\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 (1-z)^{n-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z}\mathrm{d}x \\
&= \frac{\Gamma(n-\sigma k_n)}{\Gamma\left((l-\sigma)m_{l,n}\right)\Gamma(\sum_{1\leq i\neq l\leq n} i m_{i,n} - \sigma\sum_{1\leq i\neq l\leq n} m_{i,n})} \\
&\quad\times \int_0^1 x^{(l-\sigma)m_{l,n}-1}(1-x)^{\sum_{1\leq i\neq l\leq n} i m_{i,n}-\sigma\sum_{1\leq i\neq l\leq n} m_{i,n}-1} \\
&\qquad\times \frac{\frac{\sigma\Gamma(n)}{\Gamma(n-\sigma k_n)}\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 x^r(1-z)^r(1-z)^{n-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z}{\frac{\sigma^{k_n}}{\Gamma(n-\sigma k_n)}\int_0^{+\infty} t^{-\sigma k_n}\exp\{-\tau t\}\int_0^1 (1-z)^{n-1-\sigma k_n} f_\sigma(zt)\mathrm{d}t\mathrm{d}z}\mathrm{d}x,
\end{aligned}
$$

which is the r-th moment of the scale mixture $B_{(l-\sigma)m_{l,n},n-\sigma k_n-(l-\sigma)m_{l,n}}(1-W_{n-\sigma k_n,Z_g})$, where $W_{n-\sigma k_n,Z_g}$ is the random variable characterized above, and where the Beta random variable $B_{(l-\sigma)m_{l,n},n-\sigma k_n-(l-\sigma)m_{l,n}}$ is independent of the random variable $(1-W_{n-\sigma k_n,Z_g})$. The proof is completed. □

Proof of Theorem 2. According to the fluctuation limit (23) there exists a nonnegative and finite random variable $S_{\sigma,h}$ such that $n^{-\sigma}K_n \xrightarrow{\text{a.s.}} S_{\sigma,h}$ as $n\to+\infty$. Let $\Omega_0 := \{\omega\in\Omega : \lim_{n\to+\infty} n^{-\sigma}K_n(w) = S_{\sigma,h}(\omega)\}$. Furthermore, let us define $g_{0,h}(n,k_n) = V_{h,(n+1,k_n+1)}/V_{h,(n,k_n)}$, where $V_{h,(n,k_n)} = \sigma^{k_n-1}\Gamma(k_n)\mathbb{E}[h(S_{\sigma,k_n}/B_{\sigma k_n,n-\sigma k_n})]/\Gamma(n)$. Then we can write the following expression

$$
g_{0,h}(n,k_n) = \frac{\sigma k_n}{n}\frac{\mathbb{E}\left[h\left(\frac{S_{\sigma,k_n+1}}{B_{\sigma k_n+1,n+1-\sigma(k_n+1)}}\right)\right]}{\mathbb{E}\left[h\left(\frac{S_{\sigma,k_n}}{B_{\sigma k_n,n-\sigma k_n}}\right)\right]}. \tag{26}
$$

We have to show that the ratio of the expectations in (26) converges to 1 as $n\to+\infty$. For this, it is sufficient to show that, as $n\to+\infty$, the random variable $T_{\sigma,n,k_n} = S_{\sigma,k_n}/B_{\sigma k_n,n-\sigma k_n}$ converges almost surely to a random variable $T_{\sigma,h}$. This is shown by computing the moment of order r of T_{σ,n,k_n}, i.e.,

$$
\mathbb{E}(T^r_{\sigma,n,k_n}) = \frac{\Gamma(n)}{\Gamma(n-r)}\frac{\Gamma(k_n-r/\sigma)}{\Gamma(k_n)} \simeq \frac{n^r}{k_n^{r/\sigma}}.
$$

For any $\omega\in\Omega_0$ the ratio $n/K_n^{1/\sigma}(\omega) = n/k_n^{1/\sigma}$ converges to $S^{-1/\sigma}_{\sigma,h}(\omega) = T_{\sigma,h}(\omega) = t$. Accordingly, $n^r/k_n^{r/\sigma}$ converges to $\mathbb{E}[T^r_\sigma(\omega)] = t^r$ for any $\omega\in\Omega_0$. Since $\mathbb{P}[\Omega_0]=1$, the almost sure limit, as n tends to infinity, of the random variable T_{σ,n,K_n} is identified with the nonnegative random variable $T_{\sigma,h}$, which has density function $f_{T_{\sigma,h}}(t) = h(t)f_\sigma(t)$. The proof is completed.

Proof of Proposition 3. Let $h(t) = p(t;\sigma,\theta) := \sigma\Gamma(\theta)t^{-\theta}/\Gamma(\theta/\sigma)$, for any $\sigma\in(0,1)$ and $\theta > -\sigma$. Furthermore, let us define $g_{0,p}(n,k_n) = V_{p,(n+1,k_n+1)}/V_{p,(n,k_n)}$ and $g_{1,p}(n,k_n) = 1 - V_{p,(n+1,k_n+1)}/V_{p,(n,k_n)}$, so that we have $g_0(n,k_n) = (\theta+\sigma k_n)/(\theta+n)$ and $g_1(n,k_n) = 1/(\theta+n)$. Then,

$$
g_{0,p}(n,k_n) = \frac{\sigma k_n}{n} + \frac{\theta}{n} + o\left(\frac{1}{n}\right) \tag{27}
$$

and

$$g_{1,p}(n,k_n) = \frac{1}{n} - \frac{\theta}{n^2} + o\left(\frac{1}{n^2}\right) \tag{28}$$

follow by a direct application of the Taylor series expansion to $g_0(n,k_n)$ and $g_1(n,k_n)$, respectively, and then truncating the series at the second order. The proof is completed by combining (27) and (28) with the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ under a two parameter PD prior. □

Proof of Proposition 4. The proof is along lines similar to the proof of Proposition 3.2. in Ruggiero et al. (2015), which, however, considers a different parameterization for the normalized GG prior. Let $h(t) = g(t;\sigma,\tau) := \exp\{\tau^\sigma - \tau t\}$, for any $\sigma \in (0,1)$ and $\tau > 0$, and let $g_{0,g}(n,k_n) = V_{g,(n+1,k_n+1)}/V_{g,(n,k_n)}$ and $g_{1,p}(n,k_n) = 1 - V_{g,(n+1,k_n+1)}/V_{g,(n,k_n)}$, where we have

$$V_{g,(n,k_n)} = \frac{\sigma^{k_n}\exp\{\tau^\sigma\}}{\Gamma(n)} \int_0^{+\infty} x^{n-1}(\tau+x)^{-n+\sigma k_n} \mathrm{e}^{-(\tau+x)^\sigma}\,\mathrm{d}x.$$

Note that, by using the triangular relation characterizing the nonnegative weight $V_{g,(n,k_n)}$, we can write

$$g_{0,g}(n,k_n) = \frac{V_{g,(n,k_n)} - (n-\sigma k_n)V_{g,(n+1,k_n)}}{V_{g,(n,k_n)}} = 1 - \left(1 - \frac{\sigma k_n}{n}\right) w(n,k_n),$$

where

$$w(n,k_n) = \frac{\int_0^\infty x^n \exp\{-[(\tau+x)^\sigma - \tau^\sigma]\}(\tau+x)^{\sigma k_n - n - 1}\,\mathrm{d}x}{\int_0^\infty x^{n-1}\exp\{-[(\tau+x)^\sigma - \tau^\sigma]\}(\tau+x)^{\sigma k_n - n}\,\mathrm{d}x}.$$

Let us denote by $f(x)$ the integrand function of the denominator of $1 - w(n,k_n)$, and let $f_N(x) = \tau f(x)/(\tau+x)$. That is, $f_N(x)$ is the denominator of $1 - w(n,k_n)$. Therefore we can write

$$1 - w(n,k_n) = \frac{\int_0^\infty \tau f(x)/(\tau+x)\,\mathrm{d}x}{\int_0^\infty f(x)\,\mathrm{d}x}.$$

Since $f(x)$ is unimodal, by means of the Laplace approximation method it can be approximated with a Gaussian kernel with mean $x^* = \arg\max_{x>0} x^{n-1}\exp\{-[(\tau+x)^\sigma - \tau^\sigma]\}(\tau+x)^{\sigma k_n - n}$ and with variance $-[(\log\circ f)''(x^*)]^{-1}$. The same holds for $f_N(x)$. Then, we obtain the approximation

$$1 - w(n,k_n) \simeq \frac{f_N(x_N^*)C(x_N^*, -[(\log\circ f_N)''(x_N^*)]^{-1})}{f(x_D^*)C(x_D^*, -[(\log\circ f)''(x_D^*)]^{-1})},$$

where x_N^* and x_D^* denote the modes of f_N and f, respectively, and where $C(x,y)$ denotes the normalizing constant of a Gaussian kernel with mean x and variance y. Specifically, this yields to

$$1 - w(n,k_n) \simeq \frac{f_N(x_N^*)}{f(x_D^*)} \left(\frac{(\log\circ f_N)''(x_N^*)}{(\log\circ f)''(x_D^*)}\right)^{-1/2}. \tag{29}$$

The mode x_D^* is the only positive real root of the function $G(x) = \sigma x(\tau+x)^\sigma - (n-1)\tau - (\sigma k_n - 1)x$. A study of G shows that x_D^* is bounded by below by a positive constant times $n^{1/(1+\sigma)}$, which implies

that the terms involving τ are negligible in the following renormalization of $G(x_D^*)$

$$\sigma \frac{x_D^*}{n}\left(\frac{\tau}{n}+\frac{x_D^*}{n}\right)^{\sigma}-\frac{n-1}{n^{\sigma+1}}\tau-\frac{\sigma k_n-1}{n^{\sigma}}\frac{x_D^*}{n}.$$

The same calculation holds for x_N^*. According to the fluctuation limit (23) there exists a nonnegative and finite random variable $S_{\sigma,g}$ such that $n^{-\sigma}K_n \xrightarrow{\text{a.s.}} S_{\sigma,g}$ as $n \to +\infty$. Let $\Omega_0 := \{\omega \in \Omega : \lim_{n\to+\infty} n^{-\sigma}K_n(w) = S_{\sigma,h}(\omega)\}$, and let $S_{\sigma,g}(\omega) = s_\sigma$ for any $\omega \in \Omega_0$. Then, we have

$$\frac{x_N^*}{n} \simeq \frac{x_D^*}{n} \simeq s_\sigma^{1/\sigma}. \tag{30}$$

In order to make use of (29), we also need an asymptotic equivalence for $x_D^* - x_N^*$. Note that $G(x_D^*) = 0$ and $G(x_N^*) = -x_N^*$ allow us to resort to a first order Taylor bound on G at x_N^* and shows that $x_D^* - x_N^*$ has a lower bound equivalent to $s_\sigma^{(1-\sigma)/\sigma} n^{1-\sigma}/\sigma^2$. The same argument applied to $G(x) + x$ at x_D^* provides an upper bound with the same asymptotic equivalence, thus

$$\frac{x_D^* - x_N^*}{n^{1-\sigma}} \simeq \frac{s_\sigma^{(1-\sigma)/\sigma}}{\sigma^2}. \tag{31}$$

By studying f and f_N, as well as the second derivative of their logarithm, together with asymptotic equivalences (30) and (31), we can write $f(x_D^*) \simeq f(x_N^*)$ and $(\log\circ f)''(x_D^*) \simeq (\log\circ f)''(x_N^*) \simeq (\log\circ f_N)''(x_N^*)$. Hence, from (29) one obtains $1 - w(n,k_n) \simeq \tau/(\tau + x_N^*) \simeq \tau s_\sigma^{-1/\sigma}/n$, which leads to

$$\begin{aligned} g_{0,g}(n,k_n) &= 1 - \left(1-\frac{\sigma k_n}{n}\right)\left(1 - \tau s_\sigma^{-1/\sigma}\frac{1}{n} + o\left(\frac{1}{n}\right)\right), \\ &= \frac{\sigma k_n}{n} + \tau s_\sigma^{-1/\sigma}\frac{1}{n} + o\left(\frac{1}{n}\right), \end{aligned} \tag{32}$$

and

$$\begin{aligned} g_{1,g}(n,k_n) &= \frac{1-g_{0,g}(n,k_n)}{n-\sigma k_n} = \frac{1}{n}\left(1 - \frac{\tau s_\sigma^{-1/\sigma}/n + o\left(\frac{1}{n}\right)}{1-\frac{\sigma k}{n}}\right), \\ &= \frac{1}{n}\left(1 - \frac{\tau s_\sigma^{-1/\sigma}}{n} + o\left(\frac{1}{n}\right)\right). \end{aligned} \tag{33}$$

Expressions (32) and (33) provide second order approximations of $g_{0,g}(n,k_n)$ and $g_{1,g}(n,k_n)$, respectively. Recall that for any ω in Ω_0 we have $n^{-\sigma}k_n \simeq s_\sigma$, namely we can replace s_σ with $n^{-\sigma}k_n$. This is because of the fluctuation limit displayed in (23). The proof is completed by combining (32) and (33) with the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ under a normalized GG prior. □

A.2 Details on the derivation of $\hat{\mathcal{D}}_n(l) \simeq \check{\mathcal{D}}_n(l;\mathscr{S}_{\mathbf{PD}})$

Let us define $c_{\sigma,l} = \sigma(1-\sigma)_{l-1}/l!$ and recall that $\hat{\mathcal{D}}_n(0) = V_{n+1,k_n+1}/V_{n,k_n}$ and $\hat{\mathcal{D}}_n(l) = (l-\sigma)m_{l,n}V_{n+1,k_n}/V_{n,k_n}$. The relationship between the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ and the smoothed Good-Turing estimator $\check{\mathcal{D}}_n(l;\mathscr{S}_{\text{PD}})$ follows by combining Theorem 2 with the fluctuation

limits (23) and (24). For any $\omega \in \Omega$, a version of the predictive distributions of $Q_{\sigma,h}$ is

$$\frac{V_{n+1,K_n(\omega)+1}}{V_{n,K_n(\omega)}}\nu_0(\cdot) + \frac{V_{n+1,K_n(\omega)}}{V_{n,K_n(\omega)}} \sum_{i=1}^{K_n(\omega)} (N_{i,n}(\omega) - \sigma)\delta_{X_i^*(\omega)}(\cdot).$$

According to (23) and (24), $\lim_{n\to+\infty} c_{\sigma,l} M_{l,n}/K_n = 1$ almost surely. See Lemma 3.11 in Pitman (2006) for additional details. By Theorem 2 we have $V_{n+1,K_n+1}/V_{n,K_n} \overset{\text{a.s.}}{\simeq} \sigma K_n/n$, and $M_{1,n} \overset{\text{a.s.}}{\simeq} \sigma K_n$, as $n \to +\infty$. Then, a version of the Bayesian nonparametric estimator of the 0-discovery coincides with

$$\begin{aligned} \frac{V_{n+1,K_n(\omega)+1}}{V_{n,K_n(\omega)}} &\simeq \frac{\sigma K_n(\omega)}{n} \\ &\simeq \frac{M_{1,n}(\omega)}{n}, \end{aligned} \tag{34}$$

as $n \to +\infty$. By Theorem 2 we have $V_{n+1,K_n}/V_{n,K_n} \overset{\text{a.s.}}{\simeq} 1/n$, and $M_{l,n} \overset{\text{a.s.}}{\simeq} c_{\sigma,l} K_n$, as $n \to +\infty$. Accordingly, a version of the Bayesian nonparametric estimator of the l-discovery coincides with

$$\begin{aligned} (l-\sigma) M_{l,n}(\omega) \frac{V_{n+1,K_n(\omega)}}{V_{n,K_n(\omega)}} &\simeq (l-\sigma)\frac{M_{l,n}(\omega)}{n} \\ &\simeq c_{\sigma,l}(l-\sigma)\frac{K_n(\omega)}{n} \\ &\simeq (l+1)\frac{M_{l+1,n}(\omega)}{n}, \end{aligned} \tag{35}$$

as $n \to +\infty$. Let $\Omega_0 := \{\omega \in \Omega : \lim_{n\to+\infty} n^{-\sigma} K_n(w) = Z_{\sigma,\theta/\sigma}(\omega), \lim_{n\to+\infty} n^{-\sigma} M_{l,n}(\omega) = c_{\sigma,l} Z_{\sigma,\theta/\sigma}(\omega)\}$. From (23) and (24) we have $\mathbb{P}[\Omega_0] = 1$. Fix $\omega \in \Omega_0$ and denote by $k_n = K_n(\omega)$ and $m_{l,n} = M_{l,n}(\omega)$ the number of species generated and the number of species with frequency l generated by the sample $\boldsymbol{X}_n(\omega)$. Accordingly, $\hat{\mathcal{D}}_n(l) \simeq \check{\mathcal{D}}_n(l; \mathscr{S}_{\text{PD}})$ follows from (34) and (35).

Acknowledgments

The authors are grateful to Alexander Gnedin for suggesting the asymptotic relationship (19). Julyan Arbel and Stefano Favaro are supported by the European Research Council through StG N-BNP 306406. Yee Whye Teh is supported by the European Research Council through the European Unions Seventh Framework Programme (FP7/2007-2013) ERC grant agreement 617411.

References

Arbel, J., Lijoi, A. and Nipoti, B. (2015). Full Bayesian inference with hazard mixture models. *Comput. Statist. Data Anal.*, to appear.

Baayen, R. H. (2001). *Word frequency distributions*. Springer Science and Business Media.

Bubeck, S., Ernst, D. and Garivier, A. (2013). Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality. *J. Mach. Learn. Res.*, **14**, 601–623.

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

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

Caron, F. and Fox, E.B. (2015). Sparse graphs with exchangeable random measures. *Preprint ArXiv:1401.1137*.

Chao, A. and Lee, S. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.*, **87**, 210–217.

Chao, A., Colwell, R.K., Lin, C.W. and Gotelli, N.J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, **90**, 1125–1133.

Church, K.W. and Gale, W.A. (1991). A comparison of the enhanced Good–Turing and delated estimation methods for estimating probabilities of english bigrams. *Comput. Speech Lang.*, **5**, 19–54.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I. and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.*, **37**, 212–229.

Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Trans. Model. Comput. Simul.*, **4**: 18.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63**, 435–447.

Engen, S. (1978). *Stochastic abundance models.* Chapman and Hall.

Favaro, S., Lijoi, A. and Prünster, I. (2012). A new estimator of the discovery probability. *Biometrics*, **68**, 1188–1196.

Favaro, S., Nipoti, B. and Teh, Y.W. (2015). Rediscovery Good–Turing estimators via Bayesian nonparametrics. *Biometrics*, to appear.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.

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

Gnedin, A., 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.

Gnedin, A. and Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, **138**, 5674–5685.

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–64.

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.

Jambunathan, M.V. (1954). Some Properties of Beta and Gamma Distributions. *Ann. Math. Statist.*, **25**, 401–405.

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

Lijoi, A., Mena, R.H. and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, **94**, 769-786.

Lijoi, A., Mena, R.H. and Prünster, I. (2007a). A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinformatics*, **8**: 339.

Lijoi, A., Mena, R.H. and Prünster, I. (2007b). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. Roy. Statist. Soc. Ser. B*, **69**, 769–786.

Mao, C. X. (2004). Predicting the conditional probability of discovering a new class. *J. Amer. Statist. Assoc.*, **99**, 1108–1118.

Mao, C.X. and Lindsay, B.G. (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika*, **89**, 669-681.

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

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, **102**, 145–158.

Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: A Festschrift for Terry Speed*, Goldstein, D.R. Eds. Lecture notes monograph series, **40**, Institute of Mathematical Statistics.

Pitman, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875. New York: Springer.

Pitman, J. and Yor, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, **27**, 1870–1902.

Provost, S.B. (2005). Moment-based density approximants. *Math. J.*, **9**, 727–756.

Rasmussen, S.L. and Starr, N (1979). Optimal and adaptive stopping in the search for new species. *J. Amer. Statist. Assoc.*, **74**, 661–667.

Ruggiero, M., Walker, S.G. and Favaro, S. (2013). Alpha-diversity processes and normalized inverse Gaussian diffusions. *Ann. Appl. Probab.*, **23**, 386–425.

Sampson, G. (2001). *Empirical linguistics.* Continuum, London - New York.

Steinbrecher, G. and Shaw, W.T. (2008). Quantile mechanics. *European J. Appl. Math.*, **19**, 87–112.

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

Zhang, C.H. (2005). Estimation of sums of random variables: examples and information bounds. *Ann. Statist.*, **33**, 2022–2041.