Statistica Sinica **27** (2017), 839-858
doi:http://dx.doi.org/10.5705/ss.202015.0250

BAYESIAN NONPARAMETRIC INFERENCE FOR DISCOVERY PROBABILITIES: CREDIBLE INTERVALS AND LARGE SAMPLE ASYMPTOTICS

Julyan Arbel[1,2], Stefano Favaro[1,3], Bernardo Nipoti[4] and Yee Whye Teh[5]

[1]*Collegio Carlo Alberto,* [2]*Bocconi University,* [3]*University of Torino,*
[4]*Trinity College Dublin and* [5]*University of Oxford*

Abstract: Given a sample of size n from a population of individuals belonging to different species with unknown proportions, a 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 a Bayesian nonparametric estimation 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 specification of the two parameter Poisson–Dirichlet prior and the normalized generalized Gamma prior. With respect for these prior specifications, the proposed results are illustrated through a simulation study and a benchmark Expressed Sequence Tags dataset. To the best our knowledge, this 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 $D_n(l)$.

Key words and phrases: Asymptotics, Bayesian nonparametrics, credible intervals, discovery probability, Gibbs-type priors, Good–Turing estimator, normalized generalized Gamma prior, smoothing technique, two-parameter Poisson–Dirichlet.

1. Introduction

The problem of estimating discovery probabilities arises when an experimenter is sampling from a population of individuals $(X_i)_{i\geq 1}$ belonging to an (ideally) infinite number of species $(Y_i)_{i\geq 1}$ 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, while *discovery probabilities* is used to address this class of probabilities. In terms of the species proportions q_i's, we can write

$$D_n(l) = \sum_{i\geq 1} q_i \mathbb{1}_{\{l\}}(\tilde{N}_{i,n}), \tag{1.1}$$

where $\tilde{N}_{i,n}$ denotes the frequency of the species Y_i in the sample. Here $D_n(0)$ 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, Willis and Walsh (2014) for 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 term discovery probability is also used in the literature to refer to a more general class of probabilities that originate when considering an additional unobserved sample of size $m \geq 0$. For instance, in this framework and conditionally on $\boldsymbol{X}_n$, Lijoi, Mena and Prünster (2007) consider the problem of estimating the probability that X_{n+m+1} is new, while Favaro, Lijoi and Prünster (2012) focus on the so-called m-step l-discovery, the probability that X_{n+m+1} coincides with a species that has been observed with frequency l in the enlarged sample of size $n+m$. According to this terminology, the discovery probability $D_n(l)$ introduced in (1.1) is the 0-step l-discovery.

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, Quintana and Müller (2008), Lijoi, Mena and Prünster (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, Ernst and Garivier (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, Mena and Prünster (2007) for the discovery probability of new species (0-discovery $D_n(0)$), a natural Bayesian nonparametric approach for estimating $D_n(l)$ consists in randomizing the q_i's. Specifically, consider the random probability measure $Q = \sum_{i\geq 1} q_i \delta_{Y_i}$, where $(q_i)_{i\geq 1}$ are nonnegative random weights such that $\sum_{i\geq 1} q_i = 1$ almost surely, and $(Y_i)_{i\geq 1}$ are random locations independent of $(q_i)_{i\geq 1}$ and independent and identically distributed as a nonatomic probability measure ν_0 on a space $\mathbb{X}$. Then, it is assumed that

$$\begin{aligned} X_i \,|\, Q &\overset{\text{i.i.d.}}{\sim} Q, \qquad i = 1, \ldots, n \\ Q &\sim \mathscr{Q}, \end{aligned} \tag{1.2}$$

for any $n \geq 1$, where $\mathscr{Q}$ is the prior distribution over the species composition. Under the Bayesian nonparametric model (1.2), the estimator of $D_n(l)$ with respect to a squared loss function, say $\hat{D}_n(l)$, arises from the predictive distributions characterizing $(X_i)_{i\geq 1}$. Specifying Q in the large class of Gibbs-type random probability measures by Pitman (2003), we consider the problem of deriving credible intervals for $\hat{D}_n(l)$, and study the large n asymptotic behaviour of $\hat{D}_n(l)$. Before introducing our results, we review some aspects of $\hat{D}_n(l)$.

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

Let $\boldsymbol{X}_n$ be a sample from a Gibbs-type random probability measure Q, featuring $K_n = k_n$ species $X_1^*, \ldots, X_{K_n}^*$, the unique values of $\boldsymbol{X}_n$ recorded in order of appearance, with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k_n,n})$. Here for every $i = 1, 2, \ldots, k_n$, there exists a non-negative integer ξ_i such that $X_i^* = Y_{\xi_i}$ and $N_{i,n} = \tilde{N}_{\xi_i,n}$, where $(Y_n)_{n\geq 1}$ is the sequence of random atoms in the definition of Q. Let $\sigma \in (0,1)$ and $(V_{n,k})_{k\leq n, n\geq 1}$ be a triangular array of nonnegative weights such that $V_{1,1} = 1$ and $V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}$. According to 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 $\mathbb{X}$,

$$\mathbb{P}[X_{n+1} \in A \,|\, \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} \nu_0(A) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{i=1}^{k_n} (n_{i,n} - \sigma) \delta_{X_i^*}(A). \tag{1.3}$$

The conditional probability (1.3) is referred to as the predictive distribution of Q. Two peculiar features of Q emerge directly from (1.3): the probability that $X_{n+1} \notin \{X_1^*, \ldots, X_{K_n}^*\}$ depends only on k_n; 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; the two-parameter Poisson–Dirichlet (PD) prior in Pitman (1995) and Pitman and Yor (1997); the normalized generalized Gamma (GG) prior in Pitman (2003) and Lijoi, Mena and Prünster (2007b) (see also Prünster (2002),James (2002),Lijoi and Prünster (2003), and Regazzini, Lijoi and Prünster (2003) for early appearance of normalized GG). The Dirichlet process of Ferguson (1973) can be recovered from both priors by letting $\sigma \to 0$. For any $\sigma \in (0,1)$, $\theta > -\sigma$ and $\tau > 0$, the predictive distributions of the two-parameter PD and the normalized GG priors are of the form (1.3) where V_{n,k_n}, respectively, are

$$\frac{\prod_{i=0}^{k_n-1}(\theta + i\sigma)}{(\theta)_n} \quad \text{and} \quad \frac{\sigma^{k_n-1}\mathrm{e}^{\tau^\sigma}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau)^i \Gamma\left(k_n - \frac{i}{\sigma}; \tau^\sigma\right), \tag{1.4}$$

where $(a)_n := \prod_{0\leq i\leq n-1}(a+i)$ with $(a)_0 := 1$, and $\Gamma(a,b) := \int_b^{+\infty} x^{a-1}\exp\{-x\}\mathrm{d}x$. See Pitman (1995); Lijoi, Mena and Prünster (2007b) for details on (1.4). According to (1.3), 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 the σ the flatter is the distribution of K_n. The parameters θ and τ are location parameters, the bigger they are the larger the expected number of species tends to be.

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. An estimator $\hat{\mathcal{D}}_n(l)$ arises from (1.3) 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

$$\hat{\mathcal{D}}_n(0) = \mathbb{P}[X_{n+1}\in A_0 \mid \boldsymbol{X}_n] = \mathbb{E}[Q(A_0)\mid \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}}, \tag{1.5}$$

$$\hat{\mathcal{D}}_n(l) = \mathbb{P}[X_{n+1}\in A_l \mid \boldsymbol{X}_n] = \mathbb{E}[Q(A_l)\mid \boldsymbol{X}_n] = (l-\sigma)m_{l,n}\frac{V_{n+1,k_n}}{V_{n,k_n}}. \tag{1.6}$$

Estimators (1.5) and (1.6) 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 a 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 of $(k_n, m_{l,n})$ as one would intuitively expect for an estimator of $D_n(l)$. See Favaro, Lijoi and Prünster (2012) for details.

Under the two-parameter PD prior, Favaro, Nipoti and Teh (2016) 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 for instance 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 smoothing $(m_{l,n})_{l\geq 1}$ to 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, Nipoti and Teh (2016), 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

$$m'_{l,n} = \frac{\sigma(1-\sigma)_{l-1}}{l!}k_n. \tag{1.7}$$

While the smoothing approach was 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, Nipoti and Teh (2016) shows that, for a large sample size n, a similar smoothing mechanism underlies the Bayesian nonparametric framework (1.2) 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

The problem of associating a measure of uncertainty to Bayesian nonparametric estimators for discovery probabilities was first addressed in Lijoi, Mena and Prünster (2007) where estimates of the probability of observing a new species are endowed with highest posterior density intervals. Favaro, Nipoti and Teh (2016) derive asymptotic posterior credible intervals covering also the case of species already observed with a given frequency. These contributions ultimately rely on the presence of an additional unobserved sample. While the approach of Lijoi, Mena and Prünster (2007) cannot be used to associate a measure of uncertainty to $\hat{\mathcal{D}}_n(0)$, where such additional sample is not considered, the approach of Favaro, Nipoti and Teh (2016) could be taken to derive approximate credible intervals for $\hat{\mathcal{D}}_n(l)$, $l = 0, 1, \ldots, n$. Nonetheless, due to the asymptotic nature of the approach, the resulting credible intervals are likely to perform poorly for moderate sample size n by underestimating the uncertainty associated to the estimators. They then leave essentially unaddressed the issue of quantifying the uncertainty associated to the estimators $\hat{\mathcal{D}}_n(l)$, for $l = 0, 1, \ldots, n$. In this paper we provide an answer to this problem. With a slight abuse of notation, throughout the paper we write $X \mid Y$ to denote a random variable whose distribution coincides with the conditional distribution of X given Y. Since $\hat{\mathcal{D}}_n(l) = \mathbb{E}[Q(A_l) \mid \boldsymbol{X}_n]$, the problem of deriving credible intervals for $\hat{\mathcal{D}}_n(l)$ boils down to the problem of characterizing the distribution of $Q(A_l) \mid \boldsymbol{X}_n$, 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 $\mathcal{E}_{n,r}(l) := \mathbb{E}[(Q(A_l))^r \mid \boldsymbol{X}_n]$, for any $r \geq 1$. Due to the bounded support of $Q(A_l) \mid \boldsymbol{X}_n$, the sequence $(\mathcal{E}_{n,r}(l))_{r \geq 1}$ characterizes uniquely the distribution of $Q(A_l) \mid \boldsymbol{X}_n$ 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 $Q(A_l) \mid \boldsymbol{X}_n$.

We also study the large n asymptotic behaviour of $\hat{\mathcal{D}}_n(l)$, thus extending Theorem 1 in Favaro, Nipoti and Teh (2016) 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 obtain 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 Gibbs-type prior. 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 (1.3) reveals that for V_{n,k_n} in (1.4) the estimator $\hat{\mathcal{D}}_n(l)$ admits large n asymptotic expansions whose first order truncations coincide with $\hat{\mathcal{D}}'_n(l)$, and that second order truncations depend on $\theta > -\sigma$ and $\tau > 0$, respectively, thus providing approximating estimators that differ. 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 $\check{\mathcal{D}}_n(l;\mathscr{S}_{\mathrm{PD}})$.

The estimators $\hat{\mathcal{D}}_n(l)$ depend on the values assigned to the involved parameters (see e.g. the sensitivity analysis in (Favaro, Nipoti and Teh, 2016) for the two-parameter PD case) that therefore must be suitably estimated, e.g. via an empirical Bayes approach. Taking into account the method used to estimate the parameters characterizing the underlying Gibbs-type prior would then make the analysis of the asymptotic behaviour of $\hat{\mathcal{D}}_n(l)$ more thorough, but we consider the parameters as fixed. We want to stick to the original Bayesian nonparametric framework for the estimation of discovery probabilities, as set forth in Lijoi, Mena and Prünster (2007), and we believe that this best serves the purpose of comparing the asymptotic behaviour of the two classes of estimators, highlighting the effect of the parameters in both.

Our results are illustrated in a simulation study and in the analysis of a benchmark dataset of Expressed Sequence Tags (ESTs), which are short cDNA sub-sequences highly relevant for gene identification in organisms Lijoi, Mena and Prünster (2007a). 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. We consider the two-parameter PD prior and the normalized GG prior. It turns out that the two-parameter PD prior leads to estimates of the l-discovery, as well as associated credible intervals, that are close to those obtained under the normalized GG prior specification. This surfaces due 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 as obtained by Mao (2004) and Baayen (2001). A second numerical illustration is devoted to the large n asymptotic behaviour of $\hat{\mathcal{D}}_n(l)$, 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 $Q(A_l) \mid \boldsymbol{X}_n$; 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 relationship with smoothed Good–Turing estimators. Section 4 contains some numerical illustrations. Proofs and technical derivations are available the supplementary material.

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

An integral representation for the V_{n,k_n}'s characterizing the predictive distributions (1.3) was introduced by Pitman (2003), and 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, $\int_0^{+\infty} \exp\{-tx\} f_\sigma(x)\mathrm{d}x = \exp\{-t^\sigma\}$ for any $t > 0$. Then, for some nonnegative function h, one has

$$V_{n,k_n} = V_{h,(n,k_n)} := \frac{\sigma^{k_n}}{\Gamma(n-\sigma k_n)} \int_0^{+\infty} h(t) t^{-\sigma k_n} \int_0^1 p^{n-1-\sigma k_n} f_\sigma((1-p)t)\mathrm{d}p\mathrm{d}t. \tag{2.1}$$

According to (1.3) and (2.1), a Gibbs-type prior is parameterized by (σ, h, ν_0); we denote by Q_h this Gibbs-type random probability measure. The expression (1.4) for the two-parameter PD prior is recovered from (2.1) 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 (1.4) for the normalized GG prior is recovered from (2.1) 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 (2.1) 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 (2.1) lead to

$$V_{h,(n,k_n)} = \frac{\sigma^{k_n-1}\Gamma(k_n)}{\Gamma(n)} \mathbb{E}\left[h\left(\frac{S_{\sigma,k_n}}{B_{\sigma k_n, n-\sigma k_n}}\right)\right], \tag{2.2}$$

with $B_{\sigma k_n, n-\sigma k_n}$ independent of S_{σ,k_n}. According to (2.2) 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 (2.2), 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 generated from Q_h according to* (1.2) *and featuring $K_n = k_n$ species, labelled by $X_1^*, \dots, X_{K_n}^*$, with corresponding frequencies $(N_{1,n}, \dots, N_{K_n,n}) = (n_{1,n}, \dots, n_{k_n,n})$. 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 $r \ge 1$, the rth moment $\mathbb{E}[(Q_h(A))^r \mid \boldsymbol{X}_n]$ coincides with*

$$\sum_{i=0}^{r} \frac{V_{h,(n+r,k_n+i)}}{V_{h,(n,k_n)}} (\nu_0(A))^i \sum_{0\le j_1\le\cdots\le j_i\le i} \prod_{q=0}^{r-i-1} (\mu_{n,k_n}(A) + j_q(1-\sigma) + q). \tag{2.3}$$

Let $\boldsymbol{M}_n := (M_{1,n}, \dots, M_{n,n}) = (m_{1,n}, \dots, m_{n,n})$ be the frequency counts from a sample $\boldsymbol{X}_n$ from Q_h. In order to obtain credible intervals for $\hat{\mathcal{D}}_n(l)$ we take two specifications of the Borel set A: $A_0 = \mathbb{X} \setminus \{X_1^*, \dots, X_{K_n}^*\}$ and $A_l = \{X_i^* : N_{i,n} = l\}$, for any $l = 1, \dots, n$. With them, (2.3) *reduces to*

$$\mathcal{E}_{n,r}(0) = \mathbb{E}[(Q_h(A_0))^r \mid \boldsymbol{X}_n] = \sum_{i=0}^{r} \binom{r}{i} (-1)^i \frac{V_{h,(n+i,k_n)}}{V_{h,(n,k_n)}} (n - \sigma k_n)_i, \tag{2.4}$$

$$\mathcal{E}_{n,r}(l) = \mathbb{E}[(Q_h(A_l))^r \mid \boldsymbol{X}_n] = \frac{V_{h,(n+r,k_n)}}{V_{h,(n,k_n)}} ((l-\sigma)m_{l,n})_r, \tag{2.5}$$

respectively. Equations (2.4) *and* (2.5) *take on the interpretation of the r-th moments of the posterior distribution of $D_n(0)$ and $D_n(l)$ under the specification of a Gibbs-type prior. In particular for $r = 1$, by using the recursion $V_{h,(n,k_n)} = (n - \sigma k_n)V_{h,(n+1,k_n)} + V_{h,(n+1,k_n+1)}$,* (2.4) *and* (2.5) *reduce to the Bayesian nonparametric estimators of $D_n(l)$ displayed resp. in* (1.5) *and* (1.6).

The distribution of $Q_h(A_l) \mid \boldsymbol{X}_n$ is on $[0, 1]$ and, therefore, it is characterized by $(\mathcal{E}_{n,r}(l))_{r\ge1}$. The approximation of a distribution given its moments is a longstanding problem which has been tackled by such approaches as expansions in polynomial bases, maximum entropy methods, and mixtures of distributions. For instance, the polynomial approach consists in approximating the density function of $Q_h(A_l) \mid \boldsymbol{X}_n$ 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 $Q_h(A_l) \mid \boldsymbol{X}_n$ 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 $Q_h(A_l) \mid \boldsymbol{X}_n$. See Arbel, Lijoi and Nipoti (2016) for details.

Under the specification of the two-parameter PD prior and the normalized GG prior, (2.4) and (2.5) lead to explicit and simple characterizations for the

distributions of $Q_p(A_l) \mid \boldsymbol{X}_n$ and $Q_g(A_l) \mid \boldsymbol{X}_n$, respectively. 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, define $W_{a,b} = bR_{\sigma,b}/(bR_{\sigma,b} + G_{a,1})$, where $G_{a,1}$ is independent of $R_{\sigma,b}$. The random variable $W_{a,b}$ is nonnegative and with values on the set $[0,1]$.

Proposition 1. *Let $\boldsymbol{X}_n$ be a sample generated from Q_p according to* (1.2) *and featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$. Let Z_p be a nonnegative random variable with density function of the form*

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

Then, $Q_p(A_0) \mid \boldsymbol{X}_n \stackrel{d}{=} W_{n-\sigma k_n, Z_p} \stackrel{d}{=} B_{\theta+\sigma k_n, n-\sigma k_n}$ *and*
$Q_p(A_l) \mid \boldsymbol{X}_n \stackrel{d}{=} B_{(l-\sigma)m_{l,n}, n-\sigma k_n-(l-\sigma)m_{l,n}}(1-W_{n-\sigma k_n, Z_p}) \stackrel{d}{=} B_{(l-\sigma)m_{l,n}, \theta+n-(l-\sigma)m_{l,n}}$.

Proposition 2. *Let $\boldsymbol{X}_n$ be a sample generated from Q_g according to* (1.2) *and featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$. Let Z_g be a nonnegative random variable with density function of the form*

$$f_{Z_g}(x) = \frac{\sigma x^{\sigma k_n - n}(x-\tau)^{n-1} \exp\{-x^\sigma\} \mathbb{1}_{(\tau,+\infty)}(x)}{\sum_{0 \le i \le n-1} \binom{n-1}{i} (-\tau)^i \Gamma(k_n - i/\sigma; \tau^\sigma)}. \tag{2.6}$$

Then, $Q_g(A_0) \mid \boldsymbol{X}_n \stackrel{d}{=} W_{n-\sigma k_n, Z_g}$ *and*
$Q_g(A_l) \mid \boldsymbol{X}_n \stackrel{d}{=} B_{(l-\sigma)m_{l,n}, n-\sigma k_n-(l-\sigma)m_{l,n}}(1 - W_{n-\sigma k_n, Z_g})$.

According to Propositions 1 and 2, the random variables $Q_p(A_0) \mid \boldsymbol{X}_n$ and $Q_g(A_0) \mid \boldsymbol{X}_n$ have a common structure driven by the W random variable. Moreover, for any $l = 1, \ldots, n$, $Q_p(A_l) \mid \boldsymbol{X}_n$ and $Q_g(A_l) \mid \boldsymbol{X}_n$ 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 specification of the two-parameter PD prior and the normalized GG prior, Propositions 1 and 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 $Q_p(A_l) \mid \boldsymbol{X}_n$ and $Q_g(A_l) \mid \boldsymbol{X}_n$. 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.

To obtain credible intervals for $\hat{\mathcal{D}}_n(l)$, we generate random variates from $Q_p(A_l) \mid \boldsymbol{X}_n$ and $Q_g(A_l) \mid \boldsymbol{X}_n$. With the two-parameter PD prior, sampling from

$Q_p(A_l) \mid \boldsymbol{X}_n$ for any $l = 0, 1, \ldots, n$ is straightforward, requiring generation of random variates from a Beta distribution. With the normalized GG prior, sampling from $Q_p(A_l) \mid \boldsymbol{X}_n$ for any $l = 0, 1, \ldots, n$ is also straightforward. As the density function of the transformed random variable Z_g^σ is log-concave, 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 the 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 estimator $\hat{\mathcal{D}}_n(l)$, with a view towards its asymptotic relationships with smoothed Good–Turing estimators. 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 to the different use of the information contained in the sample $\boldsymbol{X}_n$. Thus $\check{\mathcal{D}}_n(0)$ is a function of $m_{1,n}$ while $\hat{\mathcal{D}}_n(0)$ is a function of k_n, and $\check{\mathcal{D}}_n(l)$ 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 $\lim_{n \to +\infty} a_n / b_n = 1$. We show that, as n tends to infinity, $\hat{\mathcal{D}}_n(l) \simeq \check{\mathcal{D}}_n(l; \mathscr{S}_{\mathrm{PD}})$, where $\mathscr{S}_{\mathrm{PD}}$ is the smoothing rule displayed in (1.7). Such a result thus generalizes Theorem 1 in Favaro, Nipoti and Teh (2016) to the entire class of Gibbs-type priors. The asymptotic results of this section hold almost surely, but the probabilistic formalization of this idea is postponed to the proofs in the supplementary material.

Theorem 2. *For almost every sample $\boldsymbol{X}_n$ generated from Q_h according to* (1.2) *and featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$, we have*

$$\hat{\mathcal{D}}_n(0) = \frac{\sigma k_n}{n} + o\left(\frac{k_n}{n}\right), \tag{3.1}$$

$$\hat{\mathcal{D}}_n(l) = (l - \sigma)\frac{m_{l,n}}{n} + o\left(\frac{m_{l,n}}{n}\right). \tag{3.2}$$

By a direct application of Proposition 13 in Pitman (2003) and Corollary 21 in Gnedin, Hansen and Pitman (2007) we can write that, for almost every sample $\boldsymbol{X}_n$ from Q_p, featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$,

$$m_{l,n} \simeq \frac{\sigma(1 - \sigma)_{l-1}}{l!} k_n, \tag{3.3}$$

as $n \to +\infty$. By suitably combining (3.1) and (3.2) with (3.3), we obtain

$$\hat{\mathcal{D}}_n(l) \simeq (l + 1)\frac{m_{l+1,n}}{n} \simeq (l + 1)\frac{[\sigma(1 - \sigma)_l / (l + 1)!]k_n}{n}, \tag{3.4}$$

for any $l = 0, 1, \ldots, n$. See the supplementary material for details on (3.4). The first equivalence in (3.4) 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, $\mathscr{S}_{\mathrm{PD}}$ is a smoothing rule for the frequency counts $m_{l,n}$ in $\check{\mathcal{D}}_n(l)$. We refer to Section 2 in Favaro, Nipoti and Teh (2016) for a relationship between the smoothing rule $\mathscr{S}_{\mathrm{PD}}$ and the Poisson smoothing in Good (1953).

A peculiar feature of $\mathscr{S}_{\mathrm{PD}}$ is that it does not depend on the function h characterizing the Gibbs-type prior. Thus, for instance, $\mathscr{S}_{\mathrm{PD}}$ is a smoothing rule for both the two-parameter PD prior and the normalized GG prior. This invariance property of $\mathscr{S}_{\mathrm{PD}}$ is clearly determined by the fact that the asymptotic equivalences in (3.4) arise by combining (3.3), which does not depend on h, with (3.1) and (3.2), which also do not depend of h. It is worth noticing that, unlike the smoothing rule $\mathscr{S}_{\mathrm{PD}}$, the corresponding smoothed estimator $\check{\mathcal{D}}(l; \mathscr{S}_{\mathrm{PD}})$ does depend on h through k_n. Indeed, according to model (1.2), Q is the data generating process and therefore the choice of a specific Gibbs-type prior $\mathcal{Q}$ or, in other terms, the specification of h, affects the distribution of K_n. Intuitively, smoothing rules depending on the function h, if any exists, necessarily require to combine refinements of the asymptotic expansions (3.1) and (3.2) with corresponding refinements of the asymptotic equivalence (3.3). Under the specification of the two-parameter PD prior and the normalized GG prior, the next propositions provide asymptotic refinements of Theorem 2.

Proposition 3. *For almost every sample $\boldsymbol{X}_n$ generated from Q_p according to (1.2) and featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$, we have*

$$\hat{\mathcal{D}}_n(0) = \frac{\sigma k_n}{n} + \frac{\theta}{n} + o\left(\frac{1}{n}\right), \quad \hat{\mathcal{D}}_n(l) = (l - \sigma)\frac{m_{l,n}}{n}\left(1 - \frac{\theta}{n}\right) + o\left(\frac{m_{l,n}}{n^2}\right).$$

Proposition 4. *For almost every sample $\boldsymbol{X}_n$ generated from Q_g according to (1.2) and featuring $K_n = k_n$ species with $\boldsymbol{M}_n = (m_{1,n}, \ldots, m_{n,n})$, we have*

$$\hat{\mathcal{D}}_n(0) = \frac{\sigma k_n}{n} + \tau k_n^{-1/\sigma} + o\left(\frac{1}{n}\right), \quad \hat{\mathcal{D}}_n(l) = (l - \sigma)\frac{m_{l,n}}{n}\left(1 - \tau k_n^{-1/\sigma}\right) + o\left(\frac{m_{l,n}}{n^2}\right).$$

In Propositions 3 and 4, we introduce 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. Here it is sufficient to include the second term in order to introduce the dependency on $\theta > -\sigma$ and $\tau > 0$, respectively, and then the approximations of $\hat{\mathcal{D}}_n(0)$ and $\hat{\mathcal{D}}_n(l)$ differ between the two-parameter PD prior and the normalized GG prior.

The second order approximations in Propositions 3 and 4, in combination with corresponding second order refinements of (3.3), do not lead to a second

order refinement of (3.4). A second order refinement of (3.3), arising from Gnedin, Hansen and Pitman (2007), can be expressed as

$$M_{l,n} = \frac{\sigma(1-\sigma)_{l-1}}{l!} K_n + O\left(\frac{K_n}{n^{\sigma/2}}\right), \tag{3.5}$$

but second order terms in Propositions 3 and 4 are absorbed by $O\left(K_n/n^{\sigma/2}\right)$ in (3.5). Furthermore, even if a finer version of (3.5) was available, its combination with Propositions 3 and 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 and the normalized GG priors, the relationship between $\hat{\mathcal{D}}_n(l)$ and $\check{\mathcal{D}}_n(l)$ only holds at the order of asymptotic equivalence. Theorem 2 and Proposition 4, as to the normalized GG prior, provide useful approximations that might dramatically fasten up the evaluation of $\hat{\mathcal{D}}_n(l)$, for $l = 0, 1, \ldots, n$, when n is large, by avoiding the Monte Carlo evaluation of the V_{n,k_n}'s appearing in (1.5) and (1.6).

4. Illustrations

We illustrate our results with simulations and analysis of data. Data were 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. One has $\mathbb{P}[Z = z] = z^{-s}/C(s)$, for $z = \{1, 2, \ldots\}$ and $s > 1$, where $C(s) = \sum_{i \geq 1} i^{-s}$. We took $s = 1.1$ (case $s = 1.5$, typically leading to samples with a smaller number of distinct values, is presented in the supplementary material). We drew 500 samples of size n =1,000 from Z, ordered them according to the number of observed species k_n, and split them into 5 groups: for $i = 1, 2, \ldots, 5$, the i-th group of samples was composed of 100 samples featuring a total number of observed species k_n between the quantiles of order $(i-1)/5$ and $i/5$ of the empirical distribution of k_n. Then we chose at random one sample for each group and labeled it with the corresponding index i, leading to five samples (see Table 1).

We also considered ESTs data generated by sequencing two *Naegleria gruberi* complementary DNA libraries; these were prepared from cells grown under different culture conditions, 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 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 $l \in \{1, 2, \ldots, 13\}$ (see Table 1). We refer to Susko and Roger (2004) for a detailed account on the *Naegleria gruberi* libraries.

We focused on the two-parameter PD prior and the normalized GG prior. We choose the values of (σ, θ) and (σ, τ) by an empirical Bayes approach, as those that maximized the likelihood function with respect to the sample $\boldsymbol{X}_n$ featuring $K_n = k_n$ and $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k_n,n})$,

$$(\hat{\sigma}, \hat{\theta}) = \underset{(\sigma,\theta)}{\operatorname{argmax}} \left\{ \frac{\prod_{i=0}^{k_n-1}(\theta + i\sigma)}{(\theta)_n} \prod_{i=1}^{k_n} (1-\sigma)_{(n_{i,n}-1)} \right\}, \tag{4.1}$$

$$(\hat{\sigma}, \hat{\tau}) = \underset{(\sigma,\tau)}{\operatorname{argmax}} \left\{ \frac{e^{\tau^\sigma} \sigma^{k_n-1}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau)^i \Gamma\left(k_n - \frac{i}{\sigma}; \tau^\sigma\right) \prod_{i=1}^{k_n} (1-\sigma)_{(n_{i,n}-1)} \right\}. \tag{4.2}$$

As first observed by Favaro et al. (2009), under the specification of the two-parameter PD prior and for a relatively large observed sample, there is a high concentration of the posterior distribution of the parameter (σ, θ) around $(\hat{\sigma}, \hat{\theta})$. It can be checked that, under the specification of a normalized GG prior, a similar behaviour characterizes the posterior distribution of (σ, τ).

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 (4.1) and (4.2), respectively. Here 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 in light of the likelihood function of a sample $\boldsymbol{X}_n$ from a Gibbs-type random probability measure Q_h,

$$\frac{\sigma^{k_n} \prod_{i=1}^{k_n} (1-\sigma)_{(n_i-1)}}{\Gamma(n - \sigma k_n)} \int_0^{+\infty} h(t) t^{-\sigma k_n} \int_0^1 p^{n-1-\sigma k_n} f_\sigma((1-p)t) \mathrm{d}p \mathrm{d}t. \tag{4.3}$$

Apart from σ, any other parameter is introduced in (4.3) via the function h, which does not depend on the sample size n and the number of species k_n. Then, for large n and k_n the maximization of (4.3) with respect to σ should lead to a value $\hat{\sigma}$ very close to the value that would be obtained by maximizing (4.3) with $h(t) = 1$.

4.1. Credible intervals

We applied Propositions 1 and 2 in order to provide credible intervals for the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$. For the two-parameter PD prior,

Table 1. Simulated data and *Naegleria gruberi* libraries. For each sample we report the sample size n, number of species k_n and 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	1	1,000	642	0.914	2.086	0.913	2.517
	2	1,000	650	0.905	3.812	0.905	4.924
	3	1,000	656	0.910	3.236	0.910	4.060
	4	1,000	663	0.916	2.597	0.916	3.156
	5	1,000	688	0.920	3.438	0.920	4.225
Naegleria	Aerobic	959	473	0.669	46.241	0.684	334.334
	Anaerobic	969	631	0.656	155.408	0.656	4,151.075

for $l = 0$ we generated 5,000 draws from the beta $B_{\hat{\theta}+\hat{\sigma}k_n, n-\hat{\sigma}k_n}$ while, for $l \geq 1$ we sampled 5,000 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 computed the quantiles of order $\{0.025, 0.975\}$ of the empirical distribution and obtained 95% posterior credible intervals for $\hat{\mathcal{D}}_n(l)$. The procedure for the normalized GG case was only slightly more elaborate. By exploiting the adaptive rejection algorithm of Gilks and Wild (1992), we sampled 5,000 draws from Z_g with density function (2.6). In turn, we sampled 5,000 draws from $W_{n-\hat{\sigma}k_n, Z_g}$. We then used 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 sampled 5,000 draws from the beta $B_{(l-\hat{\sigma})m_{l,n}, n-\hat{\sigma}k_n-(l-\hat{\sigma})m_{l,n}}$ and used 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 $\hat{\mathcal{D}}_n(l)$. Under the two-parameter PD prior and the normalized GG prior, and with respect to these data, the top panel of Table 2 shows 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. Such a behaviour is mainly determined by the fact that the two-parameter PD prior, for any $\sigma \in (0, 1)$ and $\theta > 0$, can be viewed as a mixture of normalized GG priors. Specifically, let $\mathscr{Q}_p(\sigma, \theta)$ and $\mathscr{Q}_g(\sigma, b)$ be the distributions of the corresponding random probability measures, 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), $\mathscr{Q}_p(\sigma, \theta) = \mathscr{Q}_g(\sigma, G_{\theta/\sigma,1}^{1/\sigma})$, and specifying a two-parameter PD prior is equivalent to specifying a normalized GG prior with an Gamma hyper prior over the parameter $\tau^{1/\sigma}$. Table 2 allows us to compare the performance of the Bayesian nonparametric estimator $\hat{\mathcal{D}}_n(l)$ and 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 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 $\check{\mathcal{D}}_n(l)$, and found that different procedures are used according to the choice of $l = 0$ and $l \geq 1$. We relied on Mao (2004) for $l = 0$ and on 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 to derive confidence intervals.

4.2. Large sample approximations

We analyzed the accuracy of the large n approximations of $\hat{\mathcal{D}}_n(l)$ introduced in Theorem 2, Propositions 3 and 4. We first compared the precision of exact and approximated estimators, while a second analysis compared the behavior of first and second order approximations for varying sample sizes. For the simulated data, the specification of the two-parameter PD prior and the normalized GG prior, and for $l = 0, 1, 5, 10$, we compared the true discovery probabilities $D_n(l)$ with the Bayesian nonparametric estimates of $D_n(l)$ and with their corresponding first and second order approximations. From Table 1, the empirical Bayes estimates for σ can be slightly different under the two-parameter PD and the normalized GG priors. We considered only the first order approximation of $\hat{\mathcal{D}}_n(l)$ with the parameter $\sigma = \hat{\sigma}$ set as indicated in (4.1).

Results of this comparative study are reported in Table 3. 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 $\mathrm{SSE}(\hat{D}_n) = \sum_{0 \leq l \leq n} (\hat{D}_n(l) - d_n(l))^2$, with $d_n(l)$ being the true value of $D_n(l)$. 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. The first order approximation is already pretty accurate and, thus, the approximation error does not contribute significantly to increase the SSE. 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.

We considered simulated data with sample sizes $n = 10^2, 10^3, 10^4, 10^5$. For every n, we drew ten samples from a Zeta distribution with parameter $s = 1.1$. We focused on the two-parameter PD prior, and for each sample we determined

Table 2. Simulated data (top panel) and *Naegleria gruberi* aerobic and anaerobic libraries (bottom panel). We report the true value of the probability $D_n(l)$ (available for simulated data only) and the Bayesian nonparametric estimates of $D_n(l)$ with 95% credible intervals for $l = 0, 1, 5, 10$.

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

$(\hat{\sigma}, \hat{\theta})$ by means of the empirical Bayes procedure described in (4.1). We then evaluated, for every $l = 0, 1, \ldots, n+1$, the exact estimator $\hat{\mathcal{D}}_n(l)$ as well as its first and second order approximations. To compare the relative accuracy of the first and second order approximations $\hat{\mathcal{D}}_n^{(1)}(l)$ and $\hat{\mathcal{D}}_n^{(2)}(l)$ of the same estimator $\hat{\mathcal{D}}_n(l)$ we introduce the ratio $r_{1,2,n}$ of the sum of squared errors $\sum_{0 \le l \le n} (\hat{\mathcal{D}}_n^{(i)}(l) -$

Table 3. Simulated data. 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 $D_n(l)$.

l	Sample	1	2	3	4	5
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
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
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
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

$\hat{\mathcal{D}}_n(l))^2$ for $i = 1$ over $i = 2$. We computed the coefficient $r_{1,2,n}$ for all the samples and, for each n, the average ratio $\bar{r}_{1,2,n}$. We found the increasing values $\bar{r}_{1,2,n} = 0.163, 0.493, 1.082, 2.239$ for sizes $n = 10^2, 10^3, 10^4, 10^5$ (see Figure S1 in the supplementary material). While for small n a first order approximation turns out to be more accurate, for large values of n ($n \geq 10^4$ in our illustration), as expected, the second order approximation is more precise.

Supplementary Material

Supplementary material, available online, contains the proofs of Theorems 1, Proposition 1, Proposition 2, Theorem 2, Proposition 3 and Proposition 4, details on the derivation of the asymptotic equivalence between $\hat{\mathcal{D}}_n(l)$ and $\check{\mathcal{D}}_n(l; \mathscr{S}_{\mathrm{PD}})$, as well as additional illustrations with simulated data.

Acknowledgements

The authors are grateful to two anonymous referees for valuable comments and suggestions, and to Alexander Gnedin for suggesting the asymptotic relationship (3.5). Special thanks are due to Stefano Peluchetti (`http://www.scilua.org/`) for numerous useful discussions on numerical optimization in Lua. BN would like to thank Collegio Carlo Alberto where he was working when part of the research presented here was carried out. JA and SF are supported by the European Research Council through StG N-BNP 306406. YWT 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. (2016). Full Bayesian inference with hazard mixture models. *Comput. Statist. Data Anal.* **93**, 359-372.

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., Mena, R. H. and Prünster, I. (2009). Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B* **71**, 993-1008.

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. (2016). Rediscovery Good-Turing estimators via Bayesian nonparametrics. *Biometrics* **72**, 136-145.

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. (2006). 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.

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 non-parametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69**, 769-786.

Lijoi, A. and Prünster, I. (2003). On a normalized random measure with independent increments relevant to Bayesian nonparametric inference. *Proceedings of the 13th European Young Statisticians Meeting Bernoulli Society*, 123-134.

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* (Edited by D. R. Goldstein). Lecture notes monograph series. **40**, Institute of Mathematical Statistics.

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.

Prünster, I. (2002). Random probability measures derived from increasing additive processes and their application to Bayesian statistics. Ph.D. Thesis, University of Pavia.

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

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

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.

Collegio Carlo Alberto, via Real Collegio 30, 10024 Moncalieri, Italy.

Department of Decision Sciences, BIDSA and IGIER, Bocconi University, Milan, Italy.

E-mail: julyan.arbel@unibocconi.it

Collegio Carlo Alberto, via Real Collegio 30, 10024 Moncalieri, Italy.

Department of Economics and Statistics, University of Torino, Corso Unione Sovietica 218, 10134, Turin, Italy.

E-mail: stefano.favaro@unito.it

School of Computer Science and Statistics, Trinity College Dublin, College Green, Dublin 2, Ireland. Ireland.

E-mail: nipotib@tcd.ie

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, U.K.

E-mail: y.w.teh@stats.ox.ac.uk

(Received July 2015; accepted April 2016)