

Collegio Carlo Alberto



Rediscovery of Good-Turing estimators
via Bayesian nonparametrics

Stefano Favaro
Bernardo Nipoti
Yee Whye Teh

No. 486
December 2016

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Rediscovery of Good-Turing estimators via Bayesian nonparametrics

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

The problem of estimating discovery probabilities has regained popularity in recent years due to its frequent appearance in challenging applications arising from ecology, biology, bioinformatics, genetics, linguistic, etc. In this paper, for the first time, a connection is established between the celebrated Good-Turing estimator and a class of Bayesian nonparametric estimators for discovery probabilities recently appeared in the literature. Specifically, under a two parameter Poisson-Dirichlet prior and for a large sample size, we show that Bayesian nonparametric estimators for discovery probabilities are approximately equal to suitably smoothed Good-Turing estimators. Such an approximation is then thoroughly discussed with a view towards smoothing techniques typically adopted for the Good-Turing estimator. As a by-product of our results, we also present a novel methodology for associating exact and asymptotic credible intervals to Bayesian nonparametric estimators for discovery probabilities. This methodology is then illustrated through the analysis of some benchmark Expressed Sequence Tags datasets.

KEYWORDS: asymptotics; Bayesian nonparametric inference; credible intervals; discovery probability; Expressed Sequence Tags analysis; Good-Toulmin estimator; Good-Turing estimator; local smoothing; model-based smoothing; species sampling problem; two parameter Poisson-Dirichlet prior.

1 Introduction

The problem of estimating discovery probabilities is associated to situations where an experimenter is sampling from a population of individuals belonging to different species with unknown proportions. Assuming an ideally infinite number of species, species labels are denoted by $(X_i^*)_{i \geq 1}$ and their respective proportions in the population by $(p_i)_{i \geq 1}$. Given an initial observed sample of size n , the estimation of discovery probabilities consists in determining the probability that the $(n + m + 1)$ -th observation coincides with a species whose frequency, within an enlarged sample of size $(n + m)$, is exactly l , for any $m \geq 0$ and $l = 0, 1, \dots, n + m$. This probability is referred to as the $(m; l)$ -discovery. In terms of the unknown species proportions p_i 's, the determination of the $(m; l)$ -discovery corresponds to estimating

$$\mathcal{D}_{n,m}(l) = \sum_{i \geq 1} p_i \mathbb{1}_{\{l\}}(N_{i,n+m}),$$

where $N_{i,n+m}$ is the frequency with which the species of type X_i^* is recorded in the enlarged sample of size $(n + m)$. Clearly $\mathcal{D}_{n,m}(0)$ is the proportion of yet unobserved distinct species or, equivalently, the probability of discovering a new species. The complement $1 - \mathcal{D}_{n,m}(0)$ is the sample coverage, namely the proportion of distinct species in the sample. The problem of estimating the $(m; l)$ -discovery has regained popularity in recent years due to its frequent appearance in challenging applications arising from ecology, biology, bioinformatics, genetics, linguistic, etc.

A noteworthy approach for estimating discovery probabilities was developed by Alan M. Turing and Irving J. Good during their collaboration at Bletchley Park in the 1940s. This approach led to a nonparametric estimator for the $(0; l)$ -discovery, which first appeared in the seminal paper by Good (1953). Specifically, let H be a statistical hypothesis that determines the species proportions $(p_i)_{i \geq 1}$. Also, let $M_{l,n}$ be the number of distinct species with frequency l in the initial sample of size n , and let m_l denote the corresponding realization in the observed sample. Then, according to Good (1953), an estimator of the $(0; l)$ -discovery is

$$\check{\mathcal{D}}_{H,n,0}(l) = (l + 1) \frac{\mathbb{E}_H(M_{l+1,n})}{n} \tag{1}$$

for any $l = 0, 1, \dots, n$, and where \mathbb{E}_H is the expected value with respect to $(p_i)_{i \geq 1}$. Any parameter of the distribution $(p_i)_{i \geq 1}$ specified by H is typically estimated from the sample. Examples of these distributions are presented in Good (1953) and, among them, we mention the Zipf laws and the Pearson type laws. In order to obtain a distribution-free estimator, Good (1953) suggested to approximate (1) by replacing $\mathbb{E}_H(M_{l+1,n})$ with the observed value m_{l+1} . Specifically,

$$\check{\mathcal{D}}_{H,n,0}(l) \approx \check{\mathcal{D}}_{n,0}(l) = (l + 1) \frac{m_{l+1}}{n}, \tag{2}$$

where the symbol \approx is used to mean ‘‘approximately equal to’’. The estimator $\check{\mathcal{D}}_{n,0}(l)$ in (2) is known as the Good-Turing estimator. The paper by Good (1953) is considered as one of the precursors of what is known as the empirical Bayes method. With this regards, it is worth remarking the close resemblance between the estimator (2) and formula (19) of Robbins (1956), and between the argument given in Good (1953) for justifying estimators (1) and (2) and the empirical Bayes method described

by Robbins (1956). See also Robbins (1964).

While one may intuitively expect that m_l is the main ingredient for estimating the $(0; l)$ -discovery, $\check{D}_{n,0}(l)$ is in fact a function of m_{l+1} . Hence, for instance, if $m_{l+1} = 0$ then one estimates as zero the $(0; l)$ -discovery, and this regardless of m_l . Moreover, $\check{D}_{n,0}(l)$ is a good approximation only if n is large and the m_l 's are smoothed in a neighborhood of m_l and m_{l+1} . These features represent a significant drawback of the Good-Turing estimator: indeed, as l becomes larger compared to n , the m_l 's become increasingly noisy and the gaps where $m_l = 0$ become larger. For this reason, as pointed out by Good (1953), any concrete application of $\check{D}_{n,0}(l)$ requires a choice of some method for smoothing the m_l 's. Smoothing techniques are needed to obtain a more regular collection of m_l 's which can be used as good proxies in (2). Smoothing is typically done by making local assumptions, for example that $\sqrt{m_l}$, as a function of l , is approximately parabolic. In general, if S characterizes a local smoothing assumption, then $S(m_l)$'s are the smoothed values and a more accurate form of (2) is obtained by setting

$$\check{D}_{n,0}(l) \approx \check{D}_{S,n,0}(l) = (l+1) \frac{S(m_{l+1})}{n}$$

for large n . Of course the best value for $S(m_{l+1})$ would be $\mathbb{E}_H(M_{l+1,n})$ where H is the true statistical hypothesis on the p_i 's. See the monograph by Sampson (2001) and references therein for a comprehensive account on local smoothing techniques. An alternative method of smoothing would be to make use of all possible H 's by assuming, for instance, H to be selected from a superpopulation with an assigned distribution. This nonparametric approach was hinted at in Good (1953) and then left as an open problem by stating: "unfortunately, this theoretically satisfying method leads to a mathematical problem that I have not solved". See Section 2 and Section 3 in Good (1953) for a thorough discussion on smoothing techniques.

Good and Toulmin (1956) provided some interesting developments to the theory originally expounded in Good (1953). In particular they introduced an estimator of the $(m; 0)$ -discovery, here denoted by $\check{D}_{H,n,m}(0)$, for any $m \geq 1$. Specifically, given an initial observed sample of size n , let us consider an additional unobserved sample of size $m \geq 1$. Furthermore, let $\lambda = n/m$. The main result in Good and Toulmin (1956), then, can be expressed as follows

$$\check{D}_{H,n,m}(0) = \frac{1}{n} \sum_{i \geq 1} (-\lambda)^{i-1} i \mathbb{E}_H(M_i)$$

and

$$\check{D}_{H,n,m}(0) \approx \check{D}_{n,m}(0) = \frac{1}{n} \sum_{i \geq 1} (-\lambda)^{i-1} i m_i. \quad (3)$$

The estimator $\check{D}_{n,m}(0)$ in (3) is known as the Good-Toulmin estimator. As observed by Good and Toulmin (1956), a more accurate estimator than (3) is obtained by replacing the m_i 's with suitably smoothed values. The papers by Good (1953) and Good and Toulmin (1956), along with many others in the field, have originated a rich and consolidated body of literature that is well surveyed in Bunge and Fitzpatrick (1993) and Colwell and Coddington (1994). More recent contributions are reviewed in Chao (2005). Among them we recall Mao and Lindsay (2002) and Mao (2004) in which alternative derivations of $\check{D}_{n,m}(0)$ are presented.

In this paper, for the first time, a connection is established between the Good-Turing estimator and a class of Bayesian nonparametric estimators for discovery probabilities recently proposed by Lijoi et al. (2007) and Favaro et al. (2012). Bayesian nonparametric inference for discovery probabilities is based on the randomization of the unknown species proportions p_i 's in $\mathcal{D}_{n,m}(l)$. This is in line with the superpopulation approach hinted at by Good (1953). Specifically, let $\tilde{P} = \sum_{i \geq 1} p_i \delta_{X_i^*}$ be a discrete random probability measure, namely $(p_i)_{i \geq 1}$ are nonnegative random weights with some distribution such that $\sum_{i \geq 1} p_i = 1$ almost surely, and $(X_i^*)_{i \geq 1}$ are random locations independent of $(p_i)_{i \geq 1}$ and independent and identically distributed according to a nonatomic probability measure ν_0 . A sample of n individuals (X_1, \dots, X_n) is taken from a population with composition directed by \tilde{P} , namely

$$\begin{aligned} X_i | \tilde{P} &\stackrel{\text{iid}}{\sim} \tilde{P} & i = 1, \dots, n \\ \tilde{P} &\sim \Pi, \end{aligned} \quad (4)$$

for any $n \geq 1$, with Π playing the role of the prior. Then, according to the de Finetti representation theorem, $(X_i)_{i \geq 1}$ is an exchangeable sequence. In this framework, and assuming Π in the class of Gibbs-type priors introduced by Gnedin and Pitman (2006), Lijoi et al. (2007) and Favaro et al. (2012) derived a Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,m}(l)$ of the $(m;l)$ -discovery, for any $m \geq 0$ and $l = 0, 1, \dots, n + m$. Further references addressing related issues are Christen and Nakamura (2003), Lijoi et al. (2008), Favaro et al. (2009), Zhang and Stern (2009), Barger and Bunge (2010) and Guindani et al. (2013). The main advantages of the Bayesian nonparametric approach, with respect to the approach in Good (1953) and Good and Toulmin (1956), consist in dispensing with the specification of a statistical hypothesis H and, in particular, with the specification of large n approximations and ad hoc smoothing techniques. With this regards the Bayesian nonparametric approach provides a powerful approach, which naturally leads to a simple and exact expression for $\hat{\mathcal{D}}_{n,m}(l)$.

Within the class of Gibbs-type priors we focus on the two parameter Poisson-Dirichlet prior introduced by Perman et al. (1992), which includes the Dirichlet prior by Ferguson (1973) as a special case. See Pitman and Yor (1997) for details. Such a prior stands out for both mathematical tractability and inferential properties and, therefore, it represents a natural candidate for applications. See De Blasi et al. (2013) for a thorough discussion. Under the model (4) with Π being a two parameter Poisson-Dirichlet prior, let (X_1, \dots, X_n) be a sample featuring $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$. We show that the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,0}(l)$ is approximately equal, for large n , to a Good-Turing estimator $\check{\mathcal{D}}_{n,0}(l)$ with suitably smoothed m_i 's. Our result thus suggests that, for a large sample size, the randomization of the species proportions p_i 's can be interpreted as a smoothing technique in the sense of Good (1953). Such a smoothing, differently from the local smoothing originally discussed in Good (1953), takes on the interpretation of a model-based smoothing. Indeed, while the local smoothing was introduced as a convenient tool for post-processing the m_i 's in order to improve the performance of $\check{\mathcal{D}}_{n,0}(l)$, the smoothing obtained by the large n approximation of $\hat{\mathcal{D}}_{n,0}(l)$ is naturally induced by the prior Π . A different choice of Π will lead to a different smoothing. We show that the smoothing induced by the two parameter Poisson-Dirichlet prior has two interesting features:

i) it is closely related to the Poisson smoothing and Zipf smoothing hinted at by Good (1953); ii) it is invariant with respect to the choice of Π in a subclass of the Gibbs-type priors including the two parameter Poisson-Dirichlet prior.

The connection between $\hat{\mathcal{D}}_{n,0}(l)$ and $\check{\mathcal{D}}_{n,0}(l)$ can be extended to $\hat{\mathcal{D}}_{n,m}(l)$. Let $(X_{n+1}, \dots, X_{n+m})$ be an additional unobserved sample thus giving rise to an enlarged sample (X_1, \dots, X_{n+m}) . Then, for a fixed n and large m , we show that $\hat{\mathcal{D}}_{n,m}(l)$ is approximately equal to $\check{\mathcal{D}}_{m,0}(l)$ with the m_l 's replaced by a smoothed version of the Bayesian nonparametric estimator $\hat{\mathcal{M}}_{l,m}^{(n)}$ of the number of distinct species with frequency l in the enlarged sample. Interestingly, such a smoothing coincides with the smoothing of the m_l 's in the large n approximation of $\hat{\mathcal{D}}_{n,0}(l)$. As a by-product of this result, we also present a novel methodology to control the uncertainty of the point estimator $\hat{\mathcal{D}}_{n,m}(l)$, for any $m \geq 1$ and $l = 0, 1, \dots, n + m$. While focusing on determining a closed form expression for the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,m}(l)$, Lijoi et al. (2007) and Favaro et al. (2012) left open the problem of associating a measure of uncertainty to $\hat{\mathcal{D}}_{n,m}(l)$. In this paper, under a two parameter Poisson-Dirichlet prior, we provide a comprehensive answer to this important problem by showing how to derive exact and asymptotic credible intervals for the estimator $\hat{\mathcal{D}}_{n,m}(l)$. We mainly focus on large m asymptotic credible intervals, since in most of the situations of practical interest the size of the additional sample is very large and the computational burden makes the evaluation of exact credible intervals almost impossible. In particular we show how to combine $\hat{\mathcal{D}}_{n,m}(l)$ with large m asymptotic credible intervals and we present an illustration dealing with some benchmark Expressed Sequence Tags (ESTs) datasets. A comparison with asymptotic confidence intervals for the Good-Toulmin estimator $\check{\mathcal{D}}_{n,m}(0)$ is also presented.

The paper is structured as follows. Section 2 provides a brief review of the Bayesian nonparametric estimators for discovery probabilities we resort to, along with some known sampling properties of the two parameter Poisson-Dirichlet prior. In Section 3 and Section 4 we present and discuss our main results, namely the connection between Good-Turing estimators and their Bayesian nonparametric counterparts under a two parameter Poisson-Dirichlet prior. In particular, in Section 4 we discuss the usefulness of our results for deriving exact and asymptotic credible intervals for the point estimator $\hat{\mathcal{D}}_{n,m}(l)$. In Section 5 we show how to implement large m asymptotic credible intervals and we present some illustrations.

2 Bayesian nonparametric estimators for discovery probabilities

As outlined in the Introduction, in this section we review Bayesian nonparametric estimators for discovery probabilities under the assumption that the p_i 's in $\mathcal{D}_{n,m}(l)$ are randomized by a two parameter Poisson-Dirichlet prior. This prior assumption consists in specifying the distribution of \tilde{P} in (4) as follows: i) for any $\sigma \in [0, 1)$ and $\theta > -\sigma$, the random probabilities p_i 's are such that $p_1 = V_1$ and $p_i = V_i \prod_{1 \leq j \leq i-1} (1 - V_j)$ where $(V_j)_{j \geq 1}$ are independent random variables with V_j distributed according to a Beta distribution with parameter $(1 - \sigma, \theta + j\sigma)$; ii) the random location X_i^* 's are independent of the p_i 's and independent and identically distributed according to a nonatomic probability measure ν_0 . The Dirichlet prior is recovered by setting $\sigma = 0$. Throughout the paper we shorten “two param-

eter Poisson-Dirichlet” by $\text{PD}(\sigma, \theta)$. Furthermore, we denote by $\tilde{P}_{\sigma, \theta}$ a $\text{PD}(\sigma, \theta)$ random probability measure and by $\mathbb{E}_{\tilde{P}_{\sigma, \theta}}$ the expected value with respect to the $\text{PD}(\sigma, \theta)$ prior.

Let (X_1, \dots, X_n) be a sample from $\tilde{P}_{\sigma, \theta}$. The discreteness of $\tilde{P}_{\sigma, \theta}$ implies a random partition of (X_1, \dots, X_n) into K_n distinct observations $X_1^*, \dots, X_{K_n}^*$ with frequencies $\mathbf{N}_n = (N_{1,n}, \dots, N_{n,K_n})$ such that $\sum_{1 \leq i \leq K_n} N_{i,n} = n$. Distinct observations identify the K_n different species being recorded and $N_{i,n}$ is the number of individuals in (X_1, \dots, X_n) of type X_i^* . Pitman (1995) first derived the joint distribution of the random variables K_n and \mathbf{N}_n , namely

$$\mathbb{P}[K_n = j, \mathbf{N}_n = (n_1, \dots, n_j)] = \frac{\prod_{i=0}^{j-1} (\theta + i\sigma)}{(\theta)_n} \frac{n!}{j!} \prod_{i=1}^j \frac{(1-\sigma)_{(n_i-1)}}{n_i!} \quad (5)$$

where $(a)_n = (a)(a+1)\cdots(a+n-1)$ stands for the n -th ascending factorial of a , with $(a)_0 \equiv 1$. In various applications, it is useful to describe the probability distribution of the random partition induced by (X_1, \dots, X_n) in terms of $\mathbf{M}_n = (M_{1,n}, \dots, M_{n,n})$ where $M_{i,n}$ stands for the number of distinct species in (X_1, \dots, X_n) with frequency i . Hence, (5) can be rewritten as

$$\mathbb{P}[\mathbf{M}_n = (m_1, \dots, m_n)] = \frac{\prod_{i=0}^{j-1} (\theta + i\sigma)}{(\theta)_n} n! \prod_{i=1}^n \frac{1}{m_i!} \left(\frac{(1-\sigma)_{(i-1)}}{i!} \right)^{m_i}, \quad (6)$$

for any (m_1, \dots, m_n) of nonnegative integer numbers such that $\sum_{1 \leq i \leq n} im_i = n$ and $\sum_{1 \leq i \leq n} m_i = j$. The distribution (6) is referred to as the Ewens-Pitman sampling formula and it includes as special case the sampling formula introduced by Ewens (1972). See Pitman (2006) for a comprehensive and stimulating account on distributions over random partitions.

The distribution (5) determines the predictive distribution characterizing $\tilde{P}_{\sigma, \theta}$. See, e.g., Pitman (1995) and Pitman (1996). Such a predictive distribution, assuming that (X_1, \dots, X_n) features $K_n = j \leq n$ distinct species with frequencies $\mathbf{N}_n = (n_1, \dots, n_j)$, has the simple expression

$$\mathbb{P}[X_{n+1} \in \cdot | X_1, \dots, X_n] = \frac{\theta + \sigma j}{\theta + n} \nu_0(\cdot) + \frac{1}{\theta + n} \sum_{i=1}^j (n_i - \sigma) \delta_{X_i^*}(\cdot), \quad (7)$$

for any $n \geq 1$. The predictive distribution (7) provides some insights into the inferential implications associated with the specification of a $\text{PD}(\sigma, \theta)$ prior. The probability of sampling a new species, namely a species not detected in (X_1, \dots, X_n) , depends on n and j . In particular, if $\sigma = 0$ this probability depends solely on n . The parameter σ tunes the probability of discovering new species: the larger σ the higher is the probability of discovering new species.

According to the definition of $\mathcal{D}_{n,0}(l)$, under a $\text{PD}(\sigma, \theta)$ prior the Bayesian nonparametric estimator of the $(0; l)$ -discovery is read directly from the predictive distribution (7). Specifically, let (X_1, \dots, X_n) be a sample from $\tilde{P}_{\sigma, \theta}$ featuring $K_n = j \leq n$ distinct species with $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$. Then, the Bayesian nonparametric estimator of the $(0; 0)$ -discovery is

$$\hat{\mathcal{D}}_{n,0}(0) = \mathbb{E}_{\tilde{P}_{\sigma, \theta}}[\mathcal{D}_{n,0}(0) | X_1, \dots, X_n] = \frac{\theta + \sigma j}{\theta + n}, \quad (8)$$

whereas the Bayesian nonparametric estimator of the $(0; l)$ -discovery, for any $l = 1, \dots, n$, is

$$\hat{\mathcal{D}}_{n,0}(l) = \mathbb{E}_{\tilde{P}_{\sigma,\theta}}[\mathcal{D}_{n,0}(l) | X_1, \dots, X_n] = (l - \sigma) \frac{m_l}{\theta + n}. \quad (9)$$

Estimators (8) and (9) provide natural Bayesian nonparametric counterparts to Good-Turing estimators $\check{\mathcal{D}}_{n,0}(0)$ and $\check{\mathcal{D}}_{n,0}(l)$, respectively, under a $\tilde{P}_{\sigma,\theta}$ prior. Note that the number j of distinct species and the sample size n are sufficient to estimate the $(0; 0)$ -discovery. If $\sigma = 0$ the sole n is sufficient to estimate the $(0; 0)$ -discovery. Differently, the number m_l of distinct species with frequency l and the sample size n are sufficient to estimate the $(0; l)$ -discovery.

In order to introduce the Bayesian nonparametric counterpart of the Good-Toulmin estimator, for any $m \geq 1$ let $(X_{n+1}, \dots, X_{n+m})$ be an additional unobserved sample from $\tilde{P}_{\sigma,\theta}$ thus giving rise to an enlarged sample (X_1, \dots, X_{n+m}) . Let the initial part (X_1, \dots, X_n) feature $K_n = j \leq n$ distinct species with $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$. Then, Lijoi et al. (2007) and Favaro et al. (2009) showed that the Bayesian nonparametric estimator of the $(m; 0)$ -discovery is

$$\hat{\mathcal{D}}_{n,m}(0) = \mathbb{E}[\mathbb{E}_{\tilde{P}_{\sigma,\theta}}[\mathcal{D}_{n,m}(0) | X_1, \dots, X_n]] = \frac{\theta + \sigma j}{\theta + n} \frac{(\theta + n + \sigma)_m}{(\theta + n + 1)_m}. \quad (10)$$

Note that, differently from (8), the inner conditional expected value in (10) is random because the additional sample $(X_{n+1}, \dots, X_{n+m})$ is assumed to be not observed, for any $m \geq 1$. Recently, Favaro et al. (2012) provided an extension of (10) by determining the Bayesian nonparametric estimator of the $(m; l)$ -discovery, for any $l = 1, \dots, n + m$. Such an estimator is

$$\begin{aligned} \hat{\mathcal{D}}_{n,m}(l) &= \mathbb{E}[\mathbb{E}_{\tilde{P}_{\sigma,\theta}}[\mathcal{D}_{n,m}(l) | X_1, \dots, X_n]] \\ &= \sum_{i=1}^l \binom{m}{l-i} m_i (i - \sigma)_{(l+1-i)} \frac{(\theta + n - i + \sigma)_{(m-l+i)}}{(\theta + n)_{(m+1)}} \\ &\quad + (1 - \sigma)_l \binom{m}{l} (\theta + \sigma j) \frac{(\theta + n + \sigma)_{(m-l)}}{(\theta + n)_{m+1}}. \end{aligned} \quad (11)$$

Estimator (10) provides a natural Bayesian nonparametric counterpart of the Good-Toulmin estimator $\check{\mathcal{D}}_{n,m}(0)$ under a $\tilde{P}_{\sigma,\theta}$ prior. Note that j and n are sufficient to estimate the $(m; 0)$ -discovery. If $\sigma = 0$ the sole n is sufficient to estimate the $(m; 0)$ -discovery. Differently, j , (m_1, \dots, m_l) and n are sufficient to estimate the $(m; l)$ -discovery. If $\sigma = 0$ the sole m_1, \dots, m_l and n are sufficient to estimate the $(m; l)$ -discovery. We point out that the estimator (11) does not have any counterpart both in the frequentist and in the Bayesian frameworks.

We conclude by presenting a generalization of the above Bayesian nonparametric estimators. Specifically, for any $m \geq 0$ and $\tau \geq 1$ let (l_1, \dots, l_τ) be a collection of distinct indexes such that $l_i = 0, 1, \dots, n + m$, for any $i = 1, \dots, \tau$. We define the $(m; l_1, \dots, l_\tau)$ -discovery as the cumulative probability $\mathcal{D}_{n,m}(l_1, \dots, l_\tau) = \sum_{1 \leq i \leq \tau} \mathcal{D}_{n,m}(l_i)$. The estimator of the $(m; l_1, \dots, l_\tau)$ -discovery is

$$\hat{\mathcal{D}}_{n,m}(l_1, \dots, l_\tau) = \sum_{i=1}^{\tau} \hat{\mathcal{D}}_{n,m}(l_i). \quad (12)$$

Such a generalization is motivated by the fact that in several applications one is interested, rather than in estimating the $(m; l)$ -discovery, in estimating the probability of discovering the so-called rare species. These are species not yet observed or species observed with a frequency below a certain threshold τ . In this case, the Bayesian nonparametric estimator of the probability of discovering a rare species coincides with the estimator (12) where $l_i = i - 1$ for any $i = 1, \dots, \tau$.

3 Good-Turing estimators via Bayesian nonparametrics

Under a $\text{PD}(\sigma, \theta)$ prior, the most notable difference between Good-Turing estimators and their Bayesian nonparametric counterparts can be traced back to the different use of the information contained in the observed sample (X_1, \dots, X_n) . The Good-Turing estimator $\check{\mathcal{D}}_{n,0}(0)$ is a function of m_1 while the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,0}(0)$ in (8) is a function of j . In general, for any $l = 1, \dots, n$, the Good-Turing estimator $\check{\mathcal{D}}_{n,0}(l)$ is a function of m_{l+1} while the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,0}(l)$ in (9) is a function of m_l . In this section we show that for a large sample size n and for any $l = 0, 1, \dots, n$, $\hat{\mathcal{D}}_{n,0}(l)$ is approximately equal to a Good-Turing estimator $\check{\mathcal{D}}_{n,0}(l)$ with suitably smoothed m_l 's.

Our result stems from the large n asymptotic behaviour of the number K_n of distinct species in a sample (X_1, \dots, X_n) from $\tilde{P}_{\sigma, \theta}$. Specifically, let us consider $\sigma \in (0, 1)$ and $\theta > -\sigma$. We denote by f_σ the density function of a positive σ -stable random variable, and we introduce a nonnegative random variable $S_{\sigma, q}$, for any real $q > -1$, with density function of the form

$$f_{S_{\sigma, q}}(y) = \frac{\Gamma(q\sigma + 1)}{\sigma\Gamma(q + 1)} y^{q-1-1/\sigma} f_\sigma(y^{-1/\sigma}). \quad (13)$$

Note that the random variable $S_{\sigma, q}^{-1/\sigma}$ is the so-called polynomially tilted positive σ -stable random variable. See Pitman (2006) for details. Pitman (1999) showed that, as $n \rightarrow +\infty$, one has

$$\frac{K_n}{n^\sigma} \xrightarrow{\text{a.s.}} S_{\sigma, \theta/\sigma}. \quad (14)$$

Then, for any $l = 1, \dots, n$, the large n behaviour of $M_{l,n}$ follows directly by combining the fluctuation (14) with Corollary 21 in Gneden et al. (2007). Specifically, as $n \rightarrow +\infty$, we obtain

$$\frac{M_{l,n}}{n^\sigma} \xrightarrow{\text{a.s.}} \frac{\sigma(1-\sigma)^{(l-1)}}{l!} S_{\sigma, \theta/\sigma}. \quad (15)$$

See Pitman (2006) for refinements of (14) and (15). For $\sigma = 0$ and $\theta > 0$ the random variables K_n and $M_{l,n}$ are characterized by a different large n asymptotic behaviour. Specifically, for any $\lambda > 0$ let P_λ be random variable distributed according to a Poisson distribution with parameter λ . Then, as $n \rightarrow +\infty$, $K_n/\log n \xrightarrow{\text{a.s.}} \theta$ and $M_{l,n} \xrightarrow{\text{a.s.}} P_{\theta/l}$. See Korwar and Hollander (1973).

Theorem 1. Let (X_1, \dots, X_n) be a sample from $\tilde{P}_{\sigma, \theta}$ featuring $K_n = j \leq n$ distinct species with $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$. Then, for large n and any $l = 0, 1, \dots, n$, one has

i) for any $\sigma \in (0, 1)$ and $\theta > -\sigma$

$$\hat{\mathcal{D}}_{n,0}(l) \approx (l+1) \frac{m_{l+1}}{n} \approx (l+1) \frac{\frac{\sigma(1-\sigma)_l j}{(l+1)!}}{n}; \quad (16)$$

ii) for $\sigma = 0$ and $\theta > 0$

$$\hat{\mathcal{D}}_{n,0}(l) \approx (l+1) \frac{m_{l+1}}{n} \approx (l+1) \frac{\theta}{n}. \quad (17)$$

Proof of Theorem 1. Approximations displayed in (16) and (17) follow by exploiting the definition of $\mathcal{D}_{n,0}(l)$ and the aforementioned fluctuations for K_n and $M_{l,n}$. Specifically, let $\sigma \in (0, 1)$ and $\theta > -\sigma$. For $l = 0$, by combining the estimator (8) with fluctuations (14) and (15) one has

$$\hat{\mathcal{D}}_{n,0}(0) \approx \frac{j\sigma}{n} \approx \frac{m_1}{n} \quad (18)$$

for large n . In general, for any $l = 1, \dots, n$, by combining the estimator (9) with fluctuations (14) and (15) one has

$$\hat{\mathcal{D}}_{n,0}(l) \approx (l-\sigma) \frac{m_l}{n} \approx (l-\sigma) \frac{\frac{\sigma(1-\sigma)_{(l-1)} j}{l!}}{n} \approx (l+1) \frac{m_{l+1}}{n} \quad (19)$$

for large n . Approximations in (16), then, are obtained by a direct application of (18) and (19). The case $\sigma = 0$ and $\theta > 0$ follows by similar arguments. Specifically, instead of exploiting fluctuations (14) and (15) we apply the fluctuation for $M_{l,n}$ under the assumption $\sigma = 0$ and $\theta > 0$. \square

Theorem 1 shows a connection between Good-Turing estimators and their Bayesian nonparametric counterparts obtained by randomizing the p_i 's in $\mathcal{D}_{n,0}(l)$ with a $\text{PD}(\sigma, \theta)$ prior. Such a connection is established by means of two large n approximations of the estimator $\hat{\mathcal{D}}_{n,0}(l)$. The first approximation in (16) and (17) shows that $\hat{\mathcal{D}}_{n,0}(l)$ is approximately equal to $\check{\mathcal{D}}_{n,0}(l)$ with unsmoothed m_l 's. The second approximation in (16) and (17), then, shows how to smooth the m_l 's appearing in the first approximation. Hence, for a large sample size n , the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,0}(l)$ is approximately equal to a smoothed Good-Turing estimator $\check{\mathcal{D}}_{n,0}(l)$, and the smoothing depends on whether $\sigma \in (0, 1)$ or $\sigma = 0$. Approximations for $\hat{\mathcal{D}}_{n,0}(l_1, \dots, l_r)$ follows from (12) by a straightforward application of Theorem 1.

Theorem 1 suggests that, for a large sample size, the randomization of the species proportions p_i 's can be interpreted as a smoothing technique in the sense of Good (1953). Such a smoothing, differently from the local smoothing discussed in Good (1953), can be interpreted as a model-based smoothing since it is completely directed by the $\text{PD}(\sigma, \theta)$ prior. For any $\sigma \in (0, 1)$ and $\theta > -\sigma$ the $\text{PD}(\sigma, \theta)$ prior induces a smoothing such that m_l is approximately equal to a proportion $\sigma(1-\sigma)_l/(l+1)!$ of the number of distinct species in the observed sample. This smoothing is related to the Poisson smoothing, hinted at by Good (1953), in which m_l is approximately equal to a proportion $e^{-\delta} \delta^l / l!$ of the number of distinct species in the observed sample. Specifically, $\sigma(1-\sigma)_l/(l+1)!$ corresponds to a Poisson smoothing under the choice of a suitable parameter δ . Indeed, according to Devroye (1993), if L_σ is a discrete random variable with distribution $\mathbb{P}[L_\sigma = l] = \sigma(1-\sigma)_l/(l+1)!$ and P_λ is a random

variable distributed according to a Poisson distribution with parameter $\lambda > 0$, then

$$L_\sigma \stackrel{d}{=} P_{G_{1,1}G_{1,1-\sigma}/G_{1,\sigma}},$$

where $G_{1,1}$, $G_{1,1-\sigma}$ and $G_{1,\sigma}$ are independent Gamma random variables with parameters $(1, 1)$, $(1, 1-\sigma)$ and $(1, \sigma)$, respectively. For this reason, we refer to the smoothing induced by a $\text{PD}(\sigma, \theta)$ prior with $\sigma \in (0, 1)$ as the Poisson-Gamma smoothing. Note that the Poisson-Gamma smoothing does not depend on the parameter $\theta > -\sigma$. This is because our approximations are obtained by combining the first order large n approximations $(\theta + j\sigma)/(\theta + n) \approx j\sigma/n$ and $1/(\theta + n) \approx 1/n$ with the large n asymptotic behaviour $M_{l,n}/K_n \xrightarrow{\text{a.s.}} \sigma(1-\sigma)_{(l-1)}/l!$. For $\sigma = 0$ and $\theta > 0$ the $\text{PD}(\sigma, \theta)$ prior, namely the Dirichlet prior, induces a smoothing such that m_l is approximately equal to $\theta/(l+1)$. This is somehow reminiscent of the Zipf smoothing discussed by Good (1953). Specifically such a smoothing is induced by assuming a Zipf distribution, or some suitable modifications of it, for the species proportions $(p_i)_{i \geq 1}$. See, e.g., Equations (53), (54) and (55) in Good (1953) for further details. See also Zipf (1949).

Approximations (16) and (17) are determined by the interplay between the large n asymptotic behaviors of K_n and $M_{l,n}$ under a $\text{PD}(\sigma, \theta)$ prior. Accordingly, a different choice of the prior will lead to a different smoothing. Hereafter we show that approximations (16) are invariant with respect to the choice of a prior in a large subclass of the Gibbs-type priors including the $\text{PD}(\sigma, \theta)$ prior. Under Gibbs-type priors Bayesian nonparametric estimators for discovery probabilities are available from Lijoi et al. (2007) and Favaro et al. (2012). Gibbs-type priors generalizes the $\text{PD}(\sigma, \theta)$ prior as follows. Let $\sigma \in (-\infty, 1)$ and let $V = (V_{n,j})_{j \leq n, n \geq 1}$ be nonnegative weights satisfying the forward recursion $V_{n,j} = V_{n+1,j+1} + (n - j\sigma)V_{n+1,j}$ with $V_{1,1} = 1$. We denote by $\tilde{P}_{\sigma,V}$ a discrete random probability measure distributed according to a Gibbs-type prior. If (X_1, \dots, X_n) is a sample from $\tilde{P}_{\sigma,V}$, then

$$\mathbb{P}[X_{n+1} \in \cdot | X_1, \dots, X_n] = \frac{V_{n+1,j+1}}{V_{n,j}} \nu_0(\cdot) + \frac{V_{n+1,j}}{V_{n,j}} \sum_{i=1}^j (n_i - \sigma) \delta_{X_i^*}(\cdot) \quad (20)$$

for any $n \geq 1$, with X_1^*, \dots, X_j^* being the $K_n = j \leq n$ distinct species in (X_1, \dots, X_n) with frequencies $\mathbf{N}_n = (n_1, \dots, n_j)$. If $V_{n,j} = \prod_{0 \leq i \leq j-1} (\theta + i\sigma)/(\theta)_{n \uparrow 1}$ then (20) reduces to the predictive distribution (7). See Pitman (2003) for an explicit expression and for characterizations of the $V_{n,j}$'s under the assumption $\sigma \in (0, 1)$. In particular, here we recall that for large n one has $V_{n+1,j+1}/V_{n,j} \approx \sigma j/n$ and $V_{n+1,j}/V_{n,j} \approx 1/n$. The parameter σ tunes the probability of discovery new species. If $\sigma \in (0, 1)$ then the larger σ the higher is the probability of generating a new species. If $\sigma \in (-\infty, 0)$ things work the other way round and each new generated species reduces the probability of generating further new species. This provides an intuition for the fact that the assumption of Gibbs-type prior with $\sigma \in (-\infty, 0)$ implies a finite number of distinct species in the population. See De Blasi et al. (2013).

As for the $\text{PD}(\sigma, \theta)$ prior, under a Gibbs-type prior the Bayesian nonparametric the estimator of the $(0; l)$ -discovery is read directly from the predictive distribution (20). Specifically, let (X_1, \dots, X_n) be a sample from $\tilde{P}_{\sigma,V}$ featuring $K_n = j \leq n$ distinct species with $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$.

Then, the Bayesian nonparametric estimator of the (0; 0)-discovery is

$$\hat{\mathcal{D}}_{n,0}(0) = \mathbb{E}_{\tilde{P}_{\sigma,V}}[\mathcal{D}_{n,0}(0) | X_1, \dots, X_n] = \frac{V_{n+1,j+1}}{V_{n,j}} \quad (21)$$

whereas the Bayesian nonparametric estimator of the (0; l)-discovery, for any $l = 1, \dots, n$, is

$$\hat{\mathcal{D}}_{n,0}(l) = \mathbb{E}_{\tilde{P}_{\sigma,V}}[\mathcal{D}_{n,0}(l) | X_1, \dots, X_n] = (l - \sigma)m_l \frac{V_{n+1,j}}{V_{n,j}}. \quad (22)$$

Estimators (21) and (22) provide natural Bayesian nonparametric counterparts to Good-Turing estimators $\check{\mathcal{D}}_{n,0}(0)$ and $\check{\mathcal{D}}_{n,0}(l)$, respectively, under a Gibbs-type prior. Note that (21) and (22) have the same sufficient statistics of the corresponding estimators (8) and (9) under a PD(σ, θ) prior. The next theorem extends Theorem 1 to the subclass of Gibbs-type priors with $\sigma \in (0, 1)$.

Theorem 2. For any $\sigma \in (0, 1)$ let (X_1, \dots, X_n) be a sample from $\tilde{P}_{\sigma,V}$ featuring $K_n = j \leq n$ distinct species with $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$. Then, for large n and any $l = 0, 1, \dots, n$, one has

$$\hat{\mathcal{D}}_{n,0}(l) \approx (l + 1) \frac{m_{l+1}}{n} \approx (l + 1) \frac{\frac{\sigma(1-\sigma)^l j}{(l+1)!}}{n}. \quad (23)$$

Proof of Theorem 2. Approximations displayed in (23) follow by exploiting the definition of $\mathcal{D}_{n,0}(l)$ and fluctuation limits for K_n and $M_{l,n}$ under a general Gibbs-type prior with $\sigma \in (0, 1)$. In particular, with regards to the fluctuation limit for K_n , Pitman (2003) showed that as $n \rightarrow +\infty$

$$\frac{K_n}{n^\sigma} \xrightarrow{\text{a.s.}} T_\sigma, \quad (24)$$

where T_σ is a nonnegative and almost surely finite random variable. Then, for any $l = 1, \dots, n$, Corollary 21 in Gnedin et al. (2007) leads to the fluctuation for $M_{l,n}$. Specifically, as $n \rightarrow +\infty$

$$\frac{M_{l,n}}{n^\sigma} \xrightarrow{\text{a.s.}} \frac{\sigma(1-\sigma)^{(l-1)}}{l!} T_\sigma. \quad (25)$$

Since $V_{n+1,j+1}/V_{n,j} \approx \sigma j/n$ and $V_{n+1,j}/V_{n,j} \approx 1/n$ then $\hat{\mathcal{D}}_{n,0}(0) \approx \sigma j/n$ and $\hat{\mathcal{D}}_{n,0}(l) = (l - \sigma)m_l/n$ for large n . Then, along lines similar to the proof of Theorem 1, approximations in (23) are obtained by suitably combining these large n approximations with fluctuations in (24) and (25). \square

For any Gibbs-type prior with $\sigma \in (0, 1)$, the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,0}(l)$ is approximately equal, for large n , to a Poisson-Gamma smoothed Good-Turing estimator $\check{\mathcal{D}}_{n,0}(l)$. In other words, the smoothing of the m_l 's in the approximating $\check{\mathcal{D}}_{n,0}(l)$ does not depend on the $V_{n,j}$'s. As observed for the PD(σ, θ) prior, this is because our approximations are obtained by combining the first order large n approximations $V_{n+1,j+1}/V_{n,j} \approx j\sigma/n$ and $V_{n+1,j}/V_{n,j} \approx 1/n$ with the large n asymptotic behaviour $M_{l,n}/K_n \xrightarrow{\text{a.s.}} \sigma(1-\sigma)^{(l-1)}/l!$. Under higher order large n approximations we expect that the $V_{n,j}$'s do affect the smoothing of the m_l 's. A first attempt of deriving these higher order approximations, although in a context different from Bayesian nonparametric inference for discovery probabilities, has been recently proposed by Ruggiero et al. (2013). There, second order approximations of $V_{n+1,j+1}/V_{n,j}$ and $V_{n+1,j}/V_{n,j}$ are obtained with respect to a suitable specification

of the $V_{n,j}$'s and $\sigma = 1/2$. These specifications correspond to a Gibbs-type prior introduced in Lijoi et al. (2005) and known as the normalized inverse-Gaussian prior. Unfortunately, apart from this special case, we do not have a comprehensive answer for the problem of extending Theorem 2 to higher order large n approximations of $V_{n+1,j+1}/V_{n,j}$ and $V_{n+1,j}/V_{n,j}$. Work on this is ongoing.

4 On Bayesian nonparametric estimators for the $(m; l)$ -discovery

For any $m \geq 1$ let $(X_{n+1}, \dots, X_{n+m})$ be an additional unobserved sample from $\tilde{P}_{\sigma, \theta}$ thus giving rise to an enlarged sample (X_1, \dots, X_{n+m}) . Also, let the initial part (X_1, \dots, X_n) feature $K_n = j \leq n$ distinct species with $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$. Under this framework, in this section we extend Theorem 1 to the $(m; l)$ -discovery, for any $m \geq 1$ and $l = 0, 1, \dots, n + m$. This extension requires the study of the large m asymptotic behaviour of the random probability

$$D_{n,m}(l) = \mathbb{E}_{\tilde{P}_{\sigma, \theta}}[\mathcal{D}_{n,m}(l) | X_1, \dots, X_n].$$

As already pointed out in Section 2, we remark that $D_{n,m}(l)$ is random because the additional sample $(X_{n+1}, \dots, X_{n+m})$ is assumed to be not observed, for any $m \geq 1$. The distribution of $D_{n,m}(l)$ takes on the interpretation of the posterior distribution, with respect to (X_1, \dots, X_n) , of the $(m; l)$ -discovery. Hereafter we derive an explicit expression for the distribution of $D_{n,m}(l)$ and, in particular, we characterize its large m asymptotic behaviour. Besides leading to an extension of Theorem 1, these distributional results provide a useful tool for associating exact and asymptotic credible intervals to the point estimator $\hat{D}_{n,m}(l) = \mathbb{E}[D_{n,m}(l)]$. With a slightly abuse of notation, throughout this section we write $X | Y$ to denote the random variable whose distribution corresponds to the conditional distribution of X given Y .

We start by considering the random probability $D_{n,m}(0)$. Let $K_m^{(n)}$ be the number of new edible t species in the additional sample. This is the number of distinct species generated by $(X_{n+1}, \dots, X_{n+m})$ and not coinciding with species already detected in (X_1, \dots, X_n) . Then, according to (8),

$$\mathbb{P}[D_{n,m}(0) = p] = \mathbb{P}\left[\frac{\theta + \sigma j + \sigma K_m^{(n)}}{\theta + n + m} = p | K_n = j\right], \quad (26)$$

where the posterior distribution $\mathbb{P}[K_m^{(n)} \in \cdot | K_n = j]$ has been first obtained in Lijoi et al. (2007) and then investigated in Favaro et al. (2009). Specifically, for any $\sigma \in (0, 1)$ and $\theta > -\sigma$, one has

$$\mathbb{P}[K_m^{(n)} = x | K_n = j] = \frac{(\frac{\theta}{\sigma} + j)_x}{(\theta + n)_m} \mathcal{C}(m, x; \sigma, -n + \sigma j) \quad (27)$$

and

$$\hat{K}_m^{(n)} = \mathbb{E}[K_m^{(n)} | K_n = j] = \frac{(\frac{\theta}{\sigma} + j)}{(\theta + n)_m} \mathcal{C}(m, 1, \sigma, -\theta - n - \sigma) \quad (28)$$

for $x = 0, 1, \dots, m$, where $\mathcal{C}(m, x; \sigma, \gamma) = (x!)^{-1} \sum_{i=0}^x (-1)^i \binom{x}{i} (-i\sigma - \gamma)_m$ denotes the noncentral

generalized factorial coefficient. The case $\sigma = 0$ and $\theta > 0$ is recovered from (27) and (28) by taking $\sigma \rightarrow 0$. Indeed $\lim_{\sigma \rightarrow 0} \sigma^{-x} \mathcal{C}(m, x; \sigma, \gamma) = |s(m, x; -\gamma)|$ where $|s(m, x; -\gamma)|$ denotes the noncentral Stirling number of the first type. Equation (28) takes on the interpretation of the Bayesian nonparametric estimator of the number of new distinct species generated by the additional sample. Note that by combining (26) with (28) one obtains the estimator (10).

Let $M_{l,m}^{(n)}$ be the number of distinct species with frequency l in the enlarged sample. This is the number of new distinct species with frequency l in $(X_{n+1}, \dots, X_{n+m})$ plus the number of distinct species with frequency l that arise by updating, via $(X_{n+1}, \dots, X_{n+m})$, the frequencies already induced by (X_1, \dots, X_n) . Then, the identity (26) can be extended in order to determine the distribution of the random probability $D_{n,m}(l)$, for any $l = 1, \dots, n + m$. According to (9),

$$\begin{aligned} & \mathbb{P}[D_{n,m}(l) = p] \\ &= \mathbb{P} \left[(l - \sigma) \frac{M_{l,m}^{(n)}}{\theta + n + m} = p \mid K_n = j, (M_{1,n}, \dots, M_{l,n}) = (m_1, \dots, m_l) \right], \end{aligned} \quad (29)$$

where the posterior distribution $\mathbb{P}[M_{l,m}^{(n)} \in \cdot \mid K_n = j, (M_{1,n}, \dots, M_{l,n}) = (m_1, \dots, m_l)]$ has been derived and investigated in Favaro et al. (2013). Specifically, for any $\sigma \in [0, 1)$ and $\theta > -\sigma$, one has

$$\begin{aligned} & \mathbb{P}[M_{l,m}^{(n)} = x \mid K_n = j, (M_{1,n}, \dots, M_{l,n}) = (m_1, \dots, m_l)] \\ &= \frac{1}{x!} \sum_{t=0}^j t! \sum_{(c_1, \dots, c_t) \in \mathcal{C}_{j,t}} \prod_{i=1}^t (n_{c_i} - \sigma)_{(l - n_{c_i})} \\ & \times \sum_{y=0}^{\lfloor \frac{m + \sum_{i=1}^t n_{c_i} - x}{l} \rfloor} \frac{(-1)^y}{y!} \binom{x+y}{t} \binom{m}{l, \dots, l, l - n_{c_1}, \dots, l - n_{c_t}, m - (x+y)l + \sum_{i=1}^t n_{c_i}} \\ & \times (\sigma(1 - \sigma)_{(l-1)})^{x+y-t} \frac{(\frac{\theta}{\sigma} + j)_{x+y-t} (\theta + (x+y)\sigma + n - \sum_{i=1}^t n_{c_i})_{m - (x+y)l + \sum_{i=1}^t n_{c_i}}}{(\theta + n)_m} \end{aligned} \quad (30)$$

and

$$\begin{aligned} \hat{\mathcal{M}}_{l,m}^{(n)} &= \mathbb{E}[M_{l,m}^{(n)} \mid K_n = j, (M_{1,n}, \dots, M_{l,n}) = (m_1, \dots, m_l)] \\ &= \sum_{i=1}^l \binom{m}{l-i} m_i (i - \sigma)_{(l-i)} \frac{(\theta + n - i + \sigma)_{(m-l+i)}}{(\theta + n)_m} \\ & \quad + (1 - \sigma)_{(l-1)} \binom{m}{l} (\theta + \sigma j) \frac{(\theta + n + \sigma)_{(m-l)}}{(\theta + n)_m}, \end{aligned} \quad (31)$$

where $\mathcal{C}_{j,t}$ denotes the set of the t -combinations (without repetitions) of the set $\{1, \dots, j\}$, and $\lfloor x \rfloor$ stands for the integer part of x . Equation (31) takes on the interpretation of the Bayesian nonparametric estimator of the number of distinct species with frequency l in the enlarged sample of size $(n + m)$. In particular, note that by combining (29) with (31) one obtains the estimator (11).

The next theorem extends Theorem 1 to the $(m; l)$ -discovery. Such a theorem exploits identities (26) and (29) together with some results in Favaro et al. (2009) and Favaro et al. (2013) on the large

m asymptotic behaviour of the random variable $K_m^{(n)} | (K_n = j)$ under $\tilde{P}_{\sigma,\theta}$. Specifically, let $\sigma \in (0, 1)$ and $\theta > -\sigma$. Let us introduce the random variable $S_{\sigma,\theta,n,j}$ defined as follows

$$S_{\sigma,\theta,n,j} \stackrel{d}{=} B_{j+\theta/\sigma,n/\sigma-j} S_{\sigma,(\theta+n)/\sigma} \quad (32)$$

with $B_{a,b}$ being a random variable distributed according to Beta distribution with parameter (a, b) , and $S_{\sigma,\theta}$ being a random variable with density function (13). Also, $B_{j+\theta/\sigma,n/\sigma-j}$ is assumed to be independent of $S_{\sigma,(\theta+n)/\sigma}$. Favaro et al. (2009) showed that, as $m \rightarrow +\infty$, one has

$$\frac{K_m^{(n)}}{m^\sigma} | (K_n = j) \xrightarrow{\text{a.s.}} S_{\sigma,\theta,n,j}. \quad (33)$$

The large m behaviour of $M_{l,m}^{(n)}$, for any $l = 1, \dots, n+m$, follows directly by combining the fluctuation (33) with Corollary 21 in Gnedin et al. (2007). Specifically, as $m \rightarrow +\infty$, we obtain

$$\frac{M_{l,m}^{(n)}}{m^\sigma} | (K_n = j, (M_{1,n}, \dots, M_{l,n}) = (m_1, \dots, m_l)) \xrightarrow{\text{a.s.}} \frac{\sigma(1-\sigma)^{(l-1)}}{l!} S_{\sigma,\theta,n,j}. \quad (34)$$

For $\sigma = 0$ and $\theta > 0$ the random variables $K_m^{(n)}$ and $M_{l,m}^{(n)} | (K_n = j, (M_{1,n}, \dots, M_{l,n}) = (m_1, \dots, m_l))$ have a different large m asymptotic behaviour. Indeed Favaro et al. (2013) showed that such a large m behaviour coincides with the large n asymptotic behaviour of the random variables K_n and $M_{l,n}$, respectively. See Favaro et al. (2009) for refinements on these fluctuations.

Theorem 3. Let (X_1, \dots, X_n) be a sample from $\tilde{P}_{\sigma,\theta}$ featuring $K_n = j \leq n$ distinct species with $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$. Then, for large m and any $l = 0, 1, \dots, n+m$, one has

i) for any $\sigma \in (0, 1)$ and $\theta > -\sigma$

$$\hat{D}_{n,m}(l) \approx (l+1) \frac{\hat{M}_{l+1,m}^{(n)}}{m} \approx (l+1) \frac{\frac{\sigma(1-\sigma)^l}{(l+1)!} \hat{K}_m^{(n)}}{m} \quad (35)$$

ii) for $\sigma = 0$ and $\theta > 0$

$$\hat{D}_{n,m}(l) \approx (l+1) \frac{\hat{M}_{l+1,m}^{(n)}}{m} \approx (l+1) \frac{\frac{\theta}{l+1}}{m} \quad (36)$$

Proof of Theorem 3. The proof is along lines similar to the proof of Theorem 1. Indeed approximations (35) and (36) follow by combining identities (26) and (29) with the aforementioned fluctuations for $K_m^{(n)} | K_n$ and $M_{l,m}^{(n)} | (K_n, (M_{1,n}, \dots, M_{l,n}))$. We start by considering the case $\sigma \in (0, 1)$ and $\theta > -\sigma$. By combining (26) and (29) with (33) and (34), respectively, one has

$$\frac{D_{n,m}(l)}{m^{\sigma-1}} \xrightarrow{\text{a.s.}} \frac{\sigma(1-\sigma)^l}{l!} S_{\sigma,\theta,n,j} \quad (37)$$

as m tends to infinity. Then, for any $p \in [0, 1]$, fluctuations (33) and (34) lead to the large m

approximations

$$\begin{aligned} & \mathbb{P}[D_{n,m}(l) = p] \\ & \approx \mathbb{P}\left[(l+1)\frac{M_{l+1,m}^{(n)}}{m} = p \mid K_n = j, (M_{1,n}, \dots, M_{l+1,n}) = (m_1, \dots, m_{l+1})\right] \end{aligned} \quad (38)$$

and

$$\mathbb{P}[D_{n,m}(l) = p] \approx \mathbb{P}\left[(l+1)\frac{\frac{\sigma(1-\sigma)^l}{(l+1)!} K_m^{(n)}}{m} = p \mid K_n = j\right], \quad (39)$$

respectively. Approximations in (35) follows by taking the expected value with respect to the right-hand side of (38) and (39), respectively. The case $\sigma = 0$ and $\theta > 0$ follows by similar arguments. Specifically, we replace (37) with $mD_{n,m}(l) \xrightarrow{\text{a.s.}} (lP_{\theta/l}\mathbb{1}_{\mathbb{N}}(l) + \theta\mathbb{1}_{\{0\}}(l))$ as m tends to infinity, and we apply the corresponding fluctuations for the $M_{l,m}^{(n)} \mid (K_n, (M_{1,n}, \dots, M_{l,n}))$'s. \square

Theorem 3 shows a connection between Good-Turing estimators and the Bayesian nonparametric estimator $\hat{D}_{n,m}(l)$ under a PD(σ, θ) prior. As in Theorem 1 we have two approximations. The first approximation shows that $\hat{D}_{n,m}(l)$ is approximately equal, for fixed n and large m , to $\check{D}_{m,0}(l)$ with the m_l 's replaced by the Bayesian nonparametric estimators $\hat{\mathcal{M}}_{l,m}^{(n)}$'s in (34). The second approximation, then, show how to smooth the $\hat{\mathcal{M}}_{l,m}^{(n)}$'s appearing in the first approximation. Interestingly, for $l = 0$ Theorem 3 shows that the Bayesian nonparametric counterpart of the Good-Toulmin estimator is approximately equal, for fixed n and large m , to a smoothed Good-Turing estimator $\check{D}_{m,0}(0)$ with m_1 replaced by the estimated number of distinct species with frequency 1 generated by the additional unobserved sample. Approximations for $\hat{D}_{n,m}(l_1, \dots, l_\tau)$ follows from (12) by a straightforward application of Theorem 3.

Interestingly, the smoothing of the $\hat{\mathcal{M}}_{l,m}^{(n)}$'s in (35) coincides with the smoothing of the m_l 's provided in (16). This is the Poisson-Gamma smoothing discussed in Section 3. However, differently from Theorem 1, such a smoothing is applied to the Bayesian nonparametric estimator $\hat{\mathcal{K}}_m^{(n)}$ in (28). Accordingly, since $\hat{\mathcal{K}}_m^{(n)} \approx m^\sigma \mathbb{E}[S_{\sigma,\theta,n,j}]$ for large m , and since $\mathbb{E}[S_{\sigma,\theta,n,j}] = (j + \theta/\sigma)\Gamma(\theta+n)/\Gamma(\theta+n+\sigma)$, we can write the following approximation

$$\hat{D}_{n,m}(l) \approx (l+1) \frac{\frac{\sigma(1-\sigma)^l}{(l+1)!} m^\sigma \left(\frac{\theta}{\sigma} + j\right) \frac{\Gamma(\theta+n)}{\Gamma(\theta+n+\sigma)}}{m}. \quad (40)$$

For $\sigma = 0$ the PD(σ, θ) prior induces a smoothing such that $\hat{\mathcal{M}}_{l,m}^{(n)}$ is approximately equal to $\theta/(l+1)$. Observe that, along lines similar to those described in Section 3, Theorem 3 can be extended to Gibbs-type priors with $\sigma \in (0, 1)$. Specifically, let (X_1, \dots, X_n) be a sample from $\tilde{P}_{\sigma,V}$ featuring $K_n = j \leq n$ distinct species with $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$. Then, for large m and for any $l = 0, 1, \dots, n+m$ the Bayesian nonparametric estimator $\hat{D}_{n,m}(l)$ can be approximated as in (35) with $\hat{\mathcal{K}}_m^{(n)}$ and $\hat{\mathcal{M}}_{l,m}^{(n)}$ replaced by the corresponding estimators under a Gibbs-type prior with $\sigma \in (0, 1)$. Explicit expressions for these estimators are available from Lijoi et al. (2007) and Favaro et al. (2013). See also Favaro et al. (2012).

We conclude this section by pointing out the usefulness of the above distributional results in

controlling the uncertainty of $\hat{D}_{n,m}(l)$, for any $m \geq 1$ and $l = 0, 1, \dots, n + m$. While deriving the closed form expressions for $\hat{D}_{n,m}(l)$ in (10) and (11), Lijoi et al. (2007), Favaro et al. (2009) and Favaro et al. (2012) did not consider the problem of associating a measure of uncertainty to $\hat{D}_{n,m}(l)$. Our distributional results for $D_{n,m}(l)$ fill this important gap by providing a concrete tool for associating credible intervals to the point estimator $\hat{D}_{n,m}(l)$. The problem of deriving credible intervals for $\hat{D}_{n,m}(l)$ reduces to the problem of evaluating the distribution of the random probability $D_{n,m}(l)$. In such a context, then, identities (26) and (29) play a crucial role. In particular by combining (26) and (29) with (27) and (30), respectively, a simulation algorithm can be implemented in order to generate values from $D_{n,m}(l)$. These values, then, are used to evaluate appropriate quantiles for obtaining exact credible intervals for $\hat{D}_{n,m}(l)$.

There are, however, situations of practical interest where j , n and m are required to be very large and the computational burden for evaluating the distribution of $D_{n,m}(l)$ becomes overwhelming. This happens, for instance, in various genomic applications where one has to deal with relevant portions of cDNA libraries which typically consist of millions of genes. In order to overcome this drawback we exploit the asymptotic behaviour of $D_{n,m}$, for fixed n and large m , and use the appropriate quantiles of the limiting random variable to obtain asymptotic credible intervals for $\hat{D}_{n,m}(l)$. According to Theorem 3, for any $\sigma \in (0, 1)$ and $\theta > -\sigma$, one has

$$\frac{D_{n,m}(l)}{m^{\sigma-1}} \xrightarrow{\text{a.s.}} \frac{\sigma(1-\sigma)^l}{l!} S_{\sigma,\theta,n,j} \quad (41)$$

and, for $\sigma = 0$ and $\theta > 0$,

$$mD_{n,m}(l) \xrightarrow{\text{a.s.}} (lP_{\theta/l} \mathbb{1}_{\mathbb{N}}(l) + \theta \mathbb{1}_{\{0\}}(l)), \quad (42)$$

as $m \rightarrow +\infty$. Hence we can resort to a simulation algorithm for generating values from the limiting random variables in (41) and (42), and then use these values to evaluate appropriate quantiles. In Section 5 we present this simulation algorithm and we show how to combine $\hat{D}_{n,m}(l)$ with asymptotic credible intervals. The next proposition provides a generalization of (41) and (42) to the random probability $D_{n,m}(l_1, \dots, l_\tau) = \sum_{1 \leq i \leq \tau} D_{n,m}(l_i)$. Such a result, then, provides a fundamental tool for obtaining asymptotic credible intervals for the estimator $\hat{D}_{n,m}(l_1, \dots, l_\tau)$.

Proposition 1. Let (X_1, \dots, X_n) be a sample from $\tilde{P}_{\sigma,\theta}$ featuring $K_n = j \leq n$ distinct species with $(M_{1,n}, \dots, M_{n,n}) = (m_1, \dots, m_n)$. Furthermore, for any $\tau \geq 1$ let (l_1, \dots, l_τ) be a collection of distinct indexes such that $l_i = 0, \dots, n + m$ for any $i = 1, \dots, \tau$. Then, one has

i) for any $\sigma \in (0, 1)$ and $\theta > -\sigma$

$$\frac{D_{n,m}(l_1, \dots, l_\tau)}{m^{\sigma-1}} \xrightarrow{\text{w.}} \left(\sum_{i=1}^{\tau} \frac{\sigma(1-\sigma)^{l_i}}{l_i!} \right) S_{\sigma,\theta,n,j}; \quad (43)$$

ii) for $\sigma = 0$ and $\theta > 0$

$$mD_{n,m}(l_1, \dots, l_\tau) \xrightarrow{\text{w.}} \sum_{i=1}^{\tau} (l_i P_{\theta/l_i} \mathbb{1}_{\mathbb{N}}(l_i) + \theta \mathbb{1}_{\{0\}}(l_i)). \quad (44)$$

Proof of Proposition 1. Fluctuations (43) and (44) arise from the study of the large m behaviour of $\mathbb{E}[\prod_{1 \leq i \leq \tau} (M_{l_i})^{r_i} | X_1, \dots, X_n]$ with $r_i \geq 0$ for any $i = 1, \dots, \tau$. Let $\sigma \in (0, 1)$ and $\theta > -\sigma$. By combining the definition of $D_{n,m}(l_1, \dots, l_\tau)$ with the multinomial theorem, we can write

$$\mathbb{E} \left[\left(\frac{D_{n,m}(l_1, \dots, l_\tau)}{m^{\sigma-1}} \right)^r \right] = \sum_{(r_1, \dots, r_\tau) \in \mathcal{P}_{r,\tau}} \binom{r}{r_1, \dots, r_\tau} \mathbb{E} \left[\prod_{i=1}^{\tau} \left(\frac{D_{n,m}(l_i)}{m^{\sigma-1}} \right)^{r_i} \right]$$

where we set $\mathcal{P}_{r,\tau} = \{(r_1, \dots, r_\tau) : r_i \geq 0 \text{ for } i = 1, \dots, \tau \text{ and } \sum_{1 \leq i \leq \tau} r_i = r\}$. According to the fluctuation (41), the last expression can be approximated, for large m , by the following quantity

$$\sum_{(r_1, \dots, r_\tau) \in \mathcal{P}_{r,\tau}} \binom{r}{r_1, \dots, r_\tau} \prod_{i=1}^{\tau} \frac{(l_i + 1)^{r_i}}{m^{\sigma r_i}} \mathbb{E} \left[\prod_{i=1}^{\tau} (M_{l_i+1,m})^{r_i} | X_1, \dots, X_n \right],$$

where an explicit expression for $\mathbb{E}[\prod_{1 \leq i \leq \tau} (M_{l_i+1,m})^{r_i} | X_1, \dots, X_n]$ is given by Corollary 5 in Cesari et al. (2012). The fluctuation (43), then, follows by a direct application of the approximation $\Gamma(x+y)/\Gamma(x) \approx x^y$ as $x \rightarrow +\infty$. The case $\sigma = 0$ and $\theta > 0$ follows by similar arguments. \square

5 Applications

Under a $\text{PD}(\sigma, \theta)$ prior, we show how to combine $\hat{D}_{n,m}(l)$ and $\hat{D}_{n,m}(l_1, \dots, l_\tau)$ with large m asymptotic credible intervals, and we present an illustration dealing with some benchmark ESTs datasets. This completes the framework set up in Lijoi et al. (2007) and Favaro et al. (2012), thus making Bayesian nonparametric estimators for discovery probabilities fully amenable of use. ESTs represent an efficient way to characterize expressed genes from an organism. The rate of gene discovery depends on the degree of redundancy of the cDNA 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 therefore of fundamental importance since it allows one to optimize the use of expensive experimental sampling techniques. See Susko and Roger (2004) for details.

We consider five cDNA libraries. The tomato-flower cDNA library. See Quackenbush et al. (2000). For this library the observed sample consists of $n = 2586$ ESTs with $j = 1825$ distinct genes whose frequencies are $m_{i,2586} = 1434, 253, 71, 33, 11, 6, 2, 3, 1, 2, 2, 1, 1, 1, 2, 1, 1$ with $i \in \{1, 2, \dots, 14\} \cup \{16, 23, 27\}$. Two cDNA libraries of the amitochondriate protist *Mastigamoeba balamuthi*: the first library is non-normalized, whereas the second library is normalized, namely it undergoes a normalization protocol which aims at making the frequencies of genes in the library more uniform so to increase the discovery rate. See Susko and Roger (2004). For the *Mastigamoeba* non-normalized the observed sample consists of $n = 715$ ESTs with $j = 460$ distinct genes whose frequencies are $m_{i,715} = 378, 33, 21, 9, 6, 1, 3, 1, 1, 1, 1, 5$ with $i \in \{1, 2, \dots, 10\} \cup \{13, 15\}$. For the the *Mastigamoeba* normalized the observed sample consists of $n = 363$ with $j = 248$ distinct genes whose frequencies are $m_{i,363} = 200, 21, 14, 4, 3, 3, 1, 0, 1, 1$ with $i \in \{1, 2, \dots, 9\} \cup \{14\}$. Two *Naegleria gruberi* cDNA libraries prepared from cells grown under different culture conditions, namely aerobic and anaerobic. See Susko and Roger (2004). For the *Naegleria gruberi* aerobic the sample consists of $n = 959$ ESTs with $j = 473$

distinct genes whose frequencies are $m_{i,959} = 346, 57, 19, 12, 9, 5, 4, 2, 4, 5, 4, 1, 1, 1, 1, 1, 1$ with $i = \{1, 2, \dots, 12\} \cup \{16, 17, 18\} \cup \{27\} \cup \{55\}$. For the *Naegleria gruberi* anaerobic the sample consists of $n = 969$ ESTs with $j = 631$ distinct genes whose frequencies are $m_{i,969} = 491, 72, 30, 9, 13, 5, 3, 1, 2, 0, 1, 0, 1$ with $i \in \{1, 2, \dots, 13\}$. Previous inferential studies of these data sets, both in the frequentist and in the Bayesian frameworks, are in, e.g., Quackenbush et al. (2000), Mao and Lindsay (2002), Mao (2004), Susko and Roger (2004), Lijoi et al. (2007), Lijoi et al. (2008), Favaro et al. (2009), Favaro et al. (2012) and Favaro et al. (2013).

In order to implement asymptotic credible intervals for $\hat{D}_{n,m}(l)$ and $\hat{D}_{n,m}(l_1, \dots, l_\tau)$, the first issue to face is represented by the specification of (σ, θ) in the prior. Here we specify σ and θ by resorting to an empirical Bayes procedure. Specifically, let (X_1, \dots, X_n) be a sample from $\tilde{P}_{\sigma, \theta}$ featuring $K_n = j \leq n$ distinct species with respective frequencies $\mathbf{N}_n = (n_1, \dots, n_j)$. Then, the empirical Bayes procedure consists in choosing the values of θ and σ that maximize the likelihood function given in (5). This corresponds to setting $(\sigma, \theta) = (\hat{\sigma}, \hat{\theta})$, where

$$(\hat{\sigma}, \hat{\theta}) = \arg \max_{(\sigma, \theta)} \left\{ \frac{\prod_{i=0}^{j-1} (\theta + i\sigma)}{(\theta)_n} \prod_{i=1}^j (1 - \sigma)^{(n_i - 1)} \right\}.$$

The estimates $(\hat{\sigma}, \hat{\theta})$ obtained for the considered libraries are shown in Table 1. Notice that, alternatively, one could specify a prior distribution of (σ, θ) . Nonetheless, posterior distributions for discovery probabilities are highly peaked and not very sensitive to such prior specification and, therefore, for the sake of simplicity, we undertook the empirical Bayes procedure.

(Table 1 about here)

5.1 Preliminaries

Fluctuations (41) and (43) provide a useful tool to approximate the distribution of the random probabilities $D_{n,m}(l)$ and $D_{n,m}(l_1, \dots, l_\tau)$. The same limiting results would clearly hold true for any scaling factor $r(m)$ such that $r(m) \approx m^{\sigma-1}$. Moreover, as it is clear by inspecting Table 2, as soon as θ and n are not overwhelmingly smaller than m , the asymptotic estimator

$$\hat{D}'_{n,m}(l) = m^{\sigma-1} \frac{\sigma(1-\sigma)^l}{l!} \mathbb{E}[S_{\sigma, \theta, n, j}] \quad (45)$$

can be far from the exact estimator $\hat{D}_{n,m}(l)$. See formulae (10) and (11) for the exact estimator and formula (40) for the asymptotic estimator. Hence, the corresponding asymptotic credible intervals could be far from the exact estimates. For this reasons we introduce the scaling $r^*(m, l) \approx m^{\sigma-1}$ such that $\hat{D}_{n,m}(l) = r^*(m, l) \frac{\sigma(1-\sigma)^l}{l!} \mathbb{E}[S_{\sigma, \theta, n, j}]$, and we define the unbiased estimator

$$\hat{D}^*_{n,m}(l) = r^*(m, l) \frac{\sigma(1-\sigma)^l}{l!} \mathbb{E}[S_{\sigma, \theta, n, j}] \quad (46)$$

Similar observations holds true for the estimator $\hat{D}_{n,m}(l_1, \dots, l_\tau)$. Note that, according to (43), the asymptotic counterpart of this estimator coincides with $\hat{D}'_{n,m}(l_1, \dots, l_\tau) = \sum_{1 \leq i \leq \tau} \hat{D}'_{n,m}(l_i)$. In

particular, we introduce the scaling $r^*(m, l_1, \dots, l_\tau) \approx m^{\sigma-1}$ such that

$$\hat{\mathcal{D}}_{n,m}(l_1, \dots, l_\tau) = r^*(m, l_1, \dots, l_\tau) \sum_{1 \leq i \leq \tau} (\sigma(1-\sigma)_{l_i}/l_i!) \mathbb{E}[S_{\sigma,\theta,n,j}]$$

and we define the unbiased estimator

$$\hat{\mathcal{D}}_{n,m}^*(l_1, \dots, l_\tau) = r^*(m, l_1, \dots, l_\tau) \left(\sum_{i=1}^{\tau} \frac{\sigma(1-\sigma)_{l_i}}{l_i!} \right) \mathbb{E}[S_{\sigma,\theta,n,j}]. \quad (47)$$

It is immediate to verify that $r^*(m, l) \approx m^{\sigma-1}$ and $r^*(m, l_1, \dots, l_\tau) \approx m^{\sigma-1}$. To keep the exposition as simple as possible we do not provide the expression for the factors $r^*(m, l) \approx m^{\sigma-1}$ and $r^*(m, l_1, \dots, l_\tau)$. See Favaro et al. (2009) for a similar approach in the context of Bayesian nonparametric inference for the number of new distinct species generated by the additional sample.

We make use of (41) and (43) to quantify the uncertainty of estimates. In particular we are interested in asymptotic credible intervals for $\hat{\mathcal{D}}_{n,m}(l)$ and $\hat{\mathcal{D}}_{n,m}(l_1, \dots, l_\tau)$. This can be readily done by evaluating appropriate quantiles of the distribution of the limiting random variable $S_{\sigma,\theta,n,j}$. For instance let s_1 and s_2 be quantiles of the distribution of $S_{\sigma,\theta,n,j}$ such that (s_1, s_2) is the 95% credible interval with respect to this distribution. Then, according to (41) and (46), the set

$$\left(r^*(m, l) \frac{\sigma(1-\sigma)_l}{l!} s_1, r^*(m, l) \frac{\sigma(1-\sigma)_l}{l!} s_2 \right), \quad (48)$$

is a 95% asymptotic credible interval for $\hat{\mathcal{D}}_{n,m}(l)$. Analogous observations hold true for the estimator $\hat{\mathcal{D}}_{n,m}(l_1, \dots, l_\tau)$. In order to determine the quantiles s_1 and s_2 , we resort to a simulation algorithm for sampling the limiting random variable $S_{\sigma,\theta,n,j}$. Note that, from the distributional identity (32), this procedure involves sampling the random variable $S_{\sigma,q}$ with density function (13).

A strategy for sampling $S_{\sigma,q}$ was proposed by Favaro et al. (2009). Let $W_{\sigma,q} = S_{\sigma,q}^{-1/\sigma}$ and we introduce a random variable U_q distributed according to a Gamma distribution with parameter $(q, 1)$. Then, conditionally on $U_q = u$, the distribution of $W_{\sigma,q}$ has density function proportional to $f_\sigma(w) \exp\{-uw\}$. Therefore, the problem of sampling from (13) boils down to the problem of sampling from an exponentially tilted stable distribution. Here we improve the sampling scheme proposed in Favaro et al. (2009) by resorting recent algorithms for sampling from an exponentially tilted positive σ -stable random variable. See, e.g., Devroye (2009) and Hofert (2011). In particular, we consider the so-called fast rejection algorithm proposed in Hofert (2011). This is an exact sampling method that exploits a clever factorization of the Laplace transform of exponentially tilted positive σ -stable random variables. Summarizing, in order to sample from $S_{\sigma,\theta,n,j}$ we devised the following algorithm:

1. sample $B_{j+\theta/\sigma, n/\sigma-j}$;
2. sample $G_{(\theta+n)/\sigma, 1}$ and set $U_{(\theta+n)/\sigma} = G_{(\theta+n)/\sigma, 1}^{1/\sigma}$;
3. given $U_{(\theta+n)/\sigma} = u$, sample $W_{\sigma, (\theta+n)/\sigma}$ from density proportional to $f_\sigma(w) \exp\{-uw\}$, by means of the fast rejection sampling, and set $S_{\sigma, (\theta+n)/\sigma} = W_{\sigma, (\theta+n)/\sigma}^{-\sigma}$;
4. set $S_{\sigma,\theta,n,j} = B_{j+\theta/\sigma, n/\sigma-j} S_{\sigma, (\theta+n)/\sigma}$.

Observe that an alternative approach to sample $S_{\sigma,q}$ would consist in using the sampling scheme for polynomially tilted positive σ -stable random variable proposed in Devroye (2009).

5.2 Illustrations

The analysis we carry out has three main focuses. First, we show that the tuning of the scaling factor that appears in (41) and (43) is of fundamental importance when the additional sample size m is not much larger than original sample size n and θ . To this end, we provide a comparison between the exact estimator $\hat{D}_{n,m}(0)$ in (10) and the corresponding asymptotic estimators $\hat{D}'_{n,m}(0)$ and $\hat{D}^*_{n,m}(0)$ in (45) and (46), respectively. We also compare the corresponding asymptotic credible intervals. Second, we compare the performance of the Good-Toulmin estimators $\check{D}_{n,m}(0)$ and its Bayesian nonparametric counterpart $\hat{D}_{n,m}(0)$. More specifically, along with point estimates for the $(m;0)$ -discovery, we focus on uncertainty quantification for the two classes of estimators, where, as for Good-Toulmin estimators, we consider confidence intervals as devised with a moment-based approach in Mao (2004). Finally, we provide asymptotic credible intervals for $\hat{D}_{n,m}(l)$ and $\hat{D}_{n,m}(l_1, \dots, l_\tau)$. These asymptotic credible intervals complete the study recently presented in Favaro et al. (2012).

Table 2 compares estimates obtained by means of the exact estimator $\hat{D}_{n,m}(0)$ and by means of the corresponding asymptotic estimators $\hat{D}'_{n,m}(0)$ and $\hat{D}^*_{n,m}(0)$. With regards to the asymptotic estimators, 95% credible intervals are also provided. It is apparent that $\hat{D}'_{n,m}(0)$ provides estimates that are not precise unless the additional sample size m is very large, namely $m = 100n$. This is in contrast with the performance of $\hat{D}^*_{n,m}(0)$. Nonetheless, for $m = 10n$, asymptotic 95% credible intervals for $\hat{D}'_{n,m}(0)$ cover the exact estimates of the $(m;0)$ -discovery.

(Table 2 about here)

Next, we compare estimates of the $(m;0)$ -discovery obtained via the Good-Toulmin estimator $\check{D}_{n,m}(0)$ and its Bayesian nonparametric counterpart $\hat{D}_{n,m}(0)$. Specifically we consider confidence intervals for $\check{D}_{n,m}(0)$, as devised in Mao (2004) via a moment-based approach and asymptotic credible intervals for $\hat{D}_{n,m}(0)$. Note that this asymptotic credible intervals are obtained as in (41). We focus our attention on the range $[0, n]$ for the size m of the additional sample: such choice reflects the fact that the moment-based estimator we are considering is known to be a good estimator for small m , namely $m \leq n$. See Mao (2004) and Tierny and Lambert (1984). Figure 1 highlights the presence of some common features characterizing the estimates obtained for all the datasets considered. First of all we notice that when m is close to 0, interestingly both methods provide similar estimates for the $(m;0)$ -discovery. On the other hand, already for small values of m , asymptotic credible intervals (black dashed lines) are narrower than the corresponding moment-based 95% confidence intervals (red dashed lines). Such difference becomes more apparent when m increases: while the Bayesian asymptotic credible intervals show a regular behavior around the corresponding point estimates, with intervals that tend to get narrow very slowly, estimates obtained with Good-Toulmin estimator and corresponding confidence intervals feature a more unpredictable trend. The latter approach can lead to estimates with very different behaviors, as m approaches n . For example, while Figure 1(a) provides fairly stable estimates even for m close to n , Figure 1(d) shows intervals that get narrow around point

estimates that rapidly increase in proximity of n . On the contrary, point estimates in Figure 1(e) are eventually decreasing and the corresponding confidence intervals become significantly large.

(Figure 1 about here)

Finally, we confine ourselves to consider the two *Naegleria* libraries and we determine asymptotic credible intervals for $\hat{\mathcal{D}}_{n,m}(l)$ and $\hat{\mathcal{D}}_{n,m}(l_1, \dots, l_\tau)$. With regards to these two EST libraries, Bayesian nonparametric inference for discovery probabilities have been recently considered in Favaro et al. (2012), where estimates for discovery probabilities and cumulative discovery probabilities are obtained. However, in Favaro et al. (2012) no measures of uncertainty are provided for these estimates. Table 3 and Table 4 summarize posterior estimates of the $(m; l)$ -discovery for $l = 0, \dots, 4$ and estimates of $(m; l_1, \dots, l_\tau)$ -discovery for $\tau = 3, 4, 5$. These estimates are endowed with asymptotic 95% credible intervals obtained by combining results (41) and (43) with the choice of the scaling factors $r^*(m, l)$ and $r^*(m, l_1, \dots, l_\tau)$. Along the same line, Figure 2 shows estimated $(m; l_1, \dots, l_\tau)$ -discovery, for $\tau = 3, 4, 5$, with asymptotic 95% credible intervals as m ranges in $[0, 3000]$, for both *Naegleria* datasets.

(Table 3 and Table 4 about here)

(Figure 2 about here)

Acknowledgements

The authors are grateful to Shui Feng and Antonio Lijoi for useful and stimulating discussions. Stefano Favaro is supported by the European Research Council (ERC) through StG N-BNP 306406.

References

- Barger, K. and Bunge, J. (2010). Objective Bayesian estimation of the number of species. *Bayesian Anal.*, **5**, 619–639.
- Bunge, J. and Fitzpatrick, M. (1993) Estimating the number of species: a review. *J. Am. Statist. Assoc.*, **88**, 364–373.
- Cesari, O., Favaro, S. and Nipoti, B. (2012) Posterior analysis of rare variants in Gibbs-type species sampling models. *Preprint*.
- Chao, A. (2005). Species estimation and applications. In *Encyclopedia of Statistical Sciences* (N. Balakrishnan, C.B. Read and B. Vidakovic, Eds), Vol. **12**, 7907-7916. Wiley, New York.
- Charalambides, C.A. (2005). *Combinatorial methods in discrete distributions*. Wiley.

- Christen, J.A. and Nakamura, M. (2003). Sequential stopping rules for species accumulation. *J. Agric. Bio. Environ. Stat.*, **8**, 184–195.
- Colwell, R.K. and Coddington, J.A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. Roy. Soc. London Ser. B*, **345**, 101–118.
- De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I. and Ruggiero, M. (2013). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* In press.
- Devroye, L. (1993). A triptych of discrete distributions related to the sable law. *Statist. Probab. Lett.*, **18**, 349–351.
- Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Trans. Model. Comput. Simul.*, **4**, 18.
- Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, **3**, 87–112.
- Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B*, **71**, 993–1008.
- Favaro, S., Lijoi, A. and Prünster, I. (2012). A new estimator of the discovery probability *Biometrics*, **68**, 1188–1196.
- Favaro, S., Lijoi, A. and Prünster, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.*, **23**, 1721–1754.
- Gnedin, A. and Pitman J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, **138**, 5674–5685.
- Gnedin, S., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power law. *Probab. Surv.*, **4**, 146–171.
- Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–64.
- Good, I.J. and Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, **43**, 45–63.
- Guindani, M., Sepulveda, N., Paulino, C.D. and Müller, P. (2013). A Bayesian semiparametric approach for the differential analysis of sequence data. *J. Roy. Statist. Soc. Ser. C*. In press.
- Hofert, M. (2011). Efficiently sampling nested Archimedean copulas. *Comput. Statist. Data Anal.*, **55**, 57–70.
- Korwar, R.M. and Hollander, M. (1973). Contribution to the theory of Dirichlet processes. *Ann. Probab.*, **1**, 705–711.

- Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc.*, **100**, 1278-1291.
- 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. (2008). A Bayesian Nonparametric approach for comparing clustering structures in EST libraries. *J. Comput. Biol.*, **15**, 1315-1327.
- Mao, C.X. and Lindsay, B.G. (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika*, **89**, 669-681.
- Mao, C.X. (2004). Prediction of the conditional probability of discovering a new class. *J. Am. Statist. Assoc.*, **99**, 1108-1118.
- Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields*, **92**, 21-39.
- Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, **102**, 145-158.
- Pitman, J. and Yor, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, **25**, 855-900.
- Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: a Festschrift for Terry Speed* (D.R. Goldstein, Ed.) *Lecture Notes Monograph Series* **40** 1-34. IMS, Beachwood, OH.
- Pitman, J. (2006). *Combinatorial Stochastic Processes*. Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875. New York: Springer.
- Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J. (2000). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research* **29**, 159-164.
- Robbins, E.H. (1956) An empirical Bayes approach to statistics. In *Proc. of the third Berkley Symp. on Math. Statist. and Probability*. University of California Press, Berkley, California.
- Robbins, E.H. (1964) The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, **35**, 1-20.
- 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
- 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.
- Tierney, L., and Lambert, D. (1984). Asymptotic efficiency of estimators of functionals of mixed distributions. *Ann. Statist.*, **12**, 1380-1387.

Zhang, H. and Stern, H. (2009). Sample size calculation for finding unseen species. *Bayesian Anal.*, 4, 763–792.

Zipf, G.K. (1949). *Human behaviour and the principle of least effort* Addison-Wesley Press.

Table 1: Estimates for the hyperparameter (σ, θ) .

Library	$\hat{\sigma}$	$\hat{\theta}$
Tomato flowers	0.612	741.0
<i>Mastigamoeba</i> non-normalized	0.770	46.0
<i>Mastigamoeba</i> normalized	0.700	57.0
<i>Naegleria</i> aerobic	0.670	46.3
<i>Naegleria</i> anaerobic	0.660	155.5

Table 2: Comparison between the exact estimator $\hat{D}_{n,m}(0)$ and corresponding asymptotic estimators $\hat{D}'_{n,m}(0)$ and $\hat{D}^*_{n,m}(0)$. For the asymptotic estimators 95% credible intervals (c.i.) are provided.

Library	m	$\hat{D}_{n,m}(0)$	rate $m^{\sigma-1}$		rate $r^*(m, 0)$	
			$\hat{D}'_{n,m}(0)$	95% c.i.	$\hat{D}^*_{n,m}(0)$	95% c.i.
Tomato flowers ($n = 2586$)	n	0.447	0.616	(0.598, 0.634)	0.447	(0.434, 0.460)
	$10n$	0.241	0.252	(0.245, 0.259)	0.240	(0.233, 0.248)
	$100n$	0.103	0.103	(0.100, 0.107)	0.103	(0.100, 0.106)
<i>Mastigamoeba</i> non-normalized ($n = 715$)	n	0.452	0.533	(0.498, 0.570)	0.452	(0.422, 0.483)
	$10n$	0.307	0.314	(0.294, 0.336)	0.307	(0.287, 0.328)
	$100n$	0.185	0.185	(0.173, 0.198)	0.185	(0.172, 0.197)
<i>Mastigamoeba</i> normalized ($n = 363$)	n	0.456	0.574	(0.528, 0.621)	0.456	(0.419, 0.493)
	$10n$	0.278	0.288	(0.264, 0.311)	0.278	(0.256, 0.301)
	$100n$	0.144	0.144	(0.133, 0.156)	0.144	(0.132, 0.156)
<i>Naegleria</i> aerobic ($n = 959$)	n	0.290	0.367	(0.342, 0.393)	0.290	(0.270, 0.310)
	$10n$	0.166	0.172	(0.160, 0.184)	0.166	(0.155, 0.178)
	$100n$	0.080	0.080	(0.075, 0.086)	0.080	(0.074, 0.086)
<i>Naegleria</i> anaerobic ($n = 969$)	n	0.412	0.534	(0.505, 0.562)	0.412	(0.389, 0.433)
	$10n$	0.236	0.244	(0.231, 0.257)	0.235	(0.223, 0.248)
	$100n$	0.111	0.112	(0.106, 0.118)	0.111	(0.106, 0.117)

Table 3: *Naegleria* aerobic library. $\hat{\mathcal{D}}_{n,m}(l)$, for $l = 0, 1, 2, 3, 4$, and $\hat{\mathcal{D}}_{n,m}(0, \dots, \tau)$, for $\tau = 3, 4, 5$, and corresponding unbiased asymptotic 95% credible intervals (c.i.).

	$m = n$		$m = 2n$		$m = 3n$	
	estimate	95% c.i.	estimate	95% c.i.	estimate	95% c.i.
$(m; 0)$ -discovery	0.290	(0.270, 0.310)	0.254	(0.237, 0.272)	0.231	(0.215, 0.248)
$(m; 1)$ -discovery	0.093	(0.087, 0.100)	0.083	(0.077, 0.088)	0.075	(0.070, 0.081)
$(m; 2)$ -discovery	0.061	(0.057, 0.066)	0.054	(0.051, 0.058)	0.050	(0.046, 0.053)
$(m; 3)$ -discovery	0.046	(0.043, 0.049)	0.041	(0.039, 0.044)	0.038	(0.036, 0.041)
$(m; 4)$ -discovery	0.036	(0.034, 0.039)	0.034	(0.031, 0.036)	0.031	(0.029, 0.034)
$(m; 0, 1, 2, 3)$ -discovery	0.490	(0.456, 0.525)	0.432	(0.402, 0.463)	0.395	(0.367, 0.423)
$(m; 0, 1, 2, 3, 4)$ -discovery	0.526	(0.490, 0.563)	0.466	(0.434, 0.499)	0.426	(0.397, 0.456)
$(m; 0, 1, 2, 3, 4, 5)$ -discovery	0.556	(0.518, 0.596)	0.494	(0.460, 0.530)	0.452	(0.421, 0.485)

Table 4: *Naegleria* anaerobic library. $\hat{\mathcal{D}}_{n,m}(l)$, for $l = 0, 1, 2, 3, 4$, and $\hat{\mathcal{D}}_{n,m}(0, \dots, \tau)$, for $\tau = 3, 4, 5$, and corresponding asymptotic 95% credible intervals (c.i.).

	$m = n$		$m = 2n$		$m = 3n$	
	estimate	95% c.i.	estimate	95% c.i.	estimate	95% c.i.
$(m; 0)$ -discovery	0.412	(0.389, 0.433)	0.362	(0.342, 0.380)	0.330	(0.311, 0.346)
$(m; 1)$ -discovery	0.129	(0.122, 0.136)	0.117	(0.110, 0.123)	0.108	(0.102, 0.113)
$(m; 2)$ -discovery	0.080	(0.076, 0.084)	0.074	(0.070, 0.078)	0.069	(0.065, 0.073)
$(m; 3)$ -discovery	0.059	(0.055, 0.062)	0.055	(0.052, 0.058)	0.052	(0.049, 0.055)
$(m; 4)$ -discovery	0.045	(0.042, 0.047)	0.044	(0.041, 0.046)	0.042	(0.040, 0.044)
$(m; 0, 1, 2, 3)$ -discovery	0.680	(0.642, 0.715)	0.608	(0.574, 0.640)	0.560	(0.528, 0.587)
$(m; 0, 1, 2, 3, 4)$ -discovery	0.725	(0.684, 0.762)	0.651	(0.615, 0.685)	0.600	(0.567, 0.631)
$(m; 0, 1, 2, 3, 4, 5)$ -discovery	0.760	(0.718, 0.799)	0.687	(0.649, 0.722)	0.635	(0.600, 0.668)

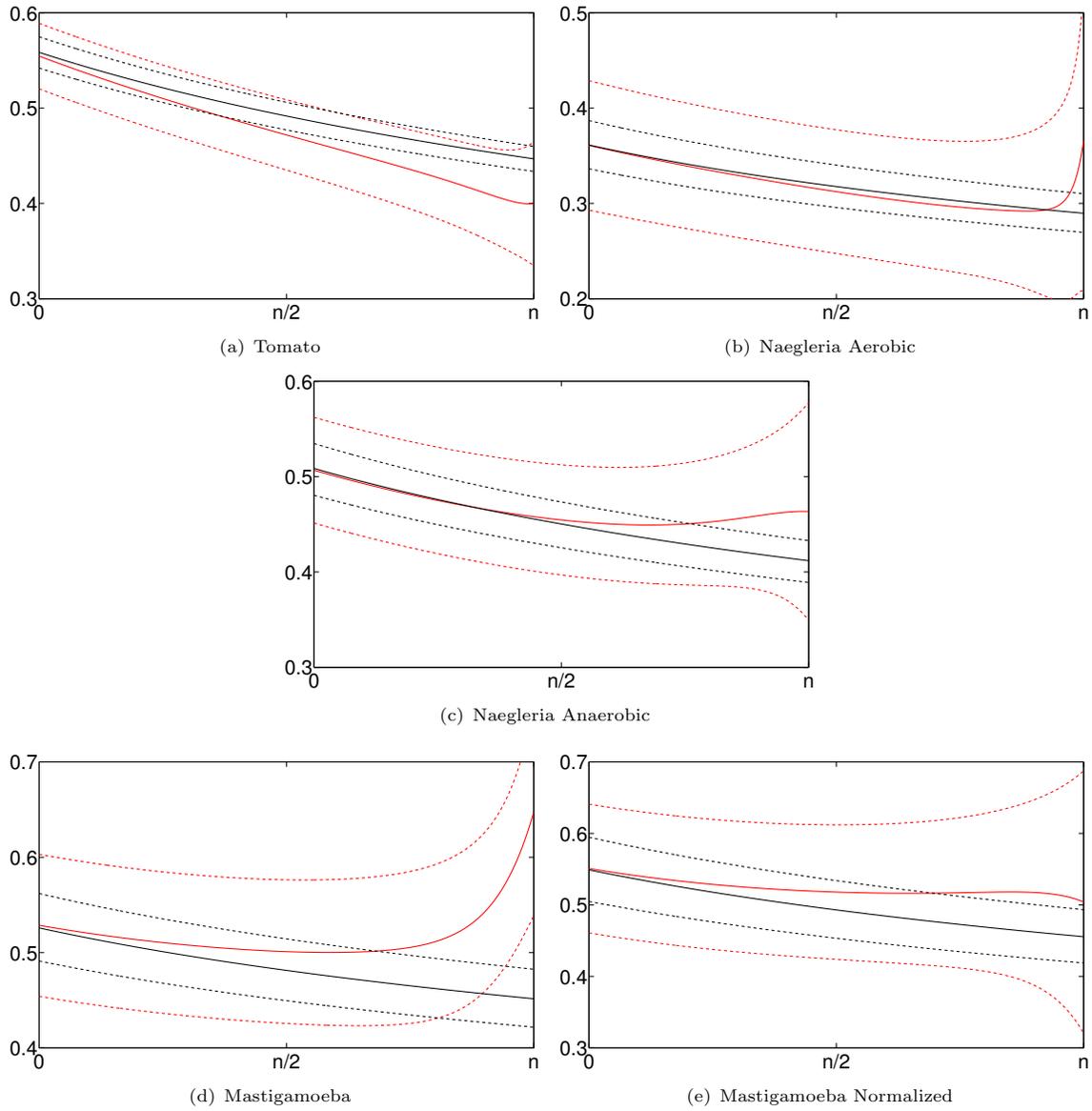


Figure 1: Comparison of Good-Toulmin estimator $\check{D}_{n,m}(0)$ (red solid curves) and Bayesian nonparametric estimator $\hat{D}_{n,m}(0)$ (black solid curves) for m ranging in $[0, n]$. The Good-Toulmin estimates are endowed with 95% confidence intervals (red dashed curves). Bayesian nonparametric estimators are endowed with asymptotic 95% credible intervals (black dashed curves).

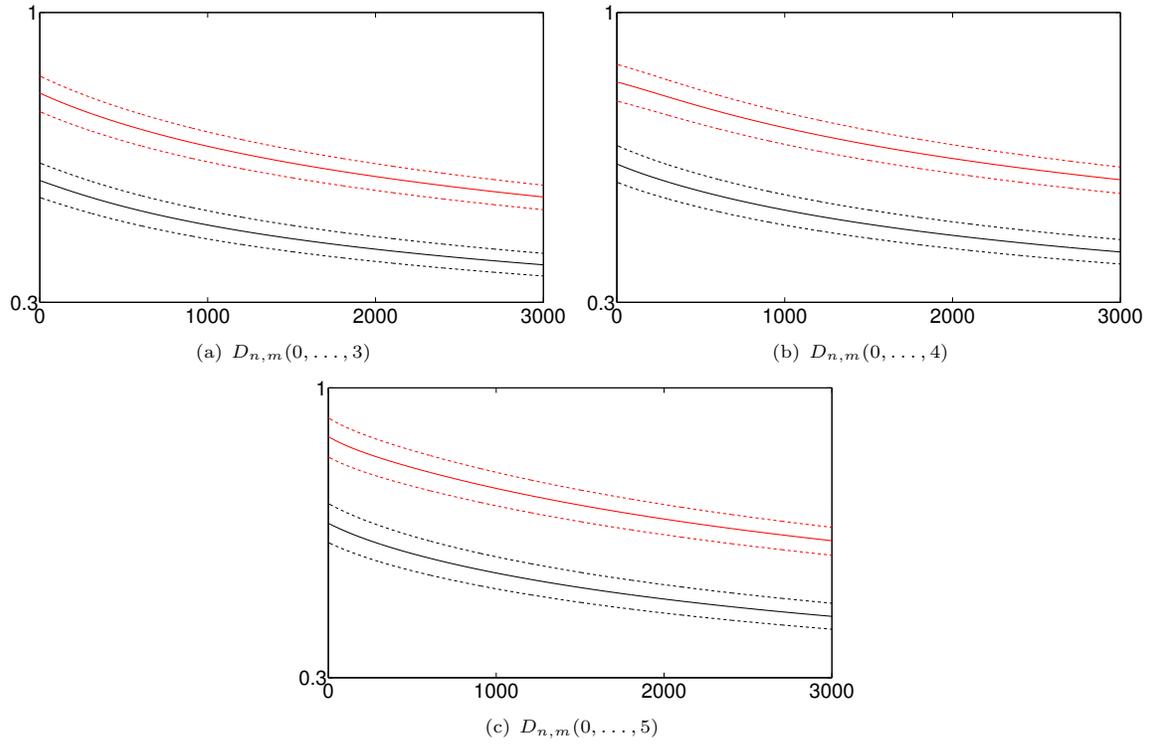


Figure 2: *Naegleria* libraries. Exact estimates $\hat{D}_{n,m}(0, \dots, \tau)$, for $\tau = 3, 4, 5$, (solid curves) together with asymptotic 95% credible intervals (dashed curves). Black curves refer to *Naegleria* aerobic, red curves to *Naegleria* anaerobic. The size m of the additional sample ranges in $[0, 3000]$.