**Rediscovery of Good-Turing estimators via Bayesian nonparametrics**

Biometrics 72, 136–145
March 2016
DOI: 10.1111/biom.12366

# Rediscovery of Good–Turing estimators via Bayesian Nonparametrics

**Stefano Favaro,[1,*] Bernardo Nipoti,[1] and Yee Whye Teh[2]**

[1]Department of Economics and Statistics, University of Torino and Collegio Carlo Alberto, Torino, Italy
[2]Department of Statistics, University of Oxford, Oxford, United Kingdom
[*]*email:* stefano.favaro@unito.it

Summary. The problem of estimating discovery probabilities originated in the context of statistical ecology, and in recent years it has become popular due to its frequent appearance in challenging applications arising in genetics, bioinformatics, linguistics, designs of experiments, machine learning, etc. A full range of statistical approaches, parametric and nonparametric as well as frequentist and Bayesian, has been proposed for estimating discovery probabilities. In this article, we investigate the relationships between the celebrated Good–Turing approach, which is a frequentist nonparametric approach developed in the 1940s, and a Bayesian nonparametric approach recently introduced in the literature. Specifically, under the assumption of a two parameter Poisson-Dirichlet prior, we show that Bayesian nonparametric estimators of discovery probabilities are asymptotically equivalent, for a large sample size, to suitably smoothed Good–Turing estimators. As a by-product of this result, we introduce and investigate a methodology for deriving exact and asymptotic credible intervals to be associated with the Bayesian nonparametric estimators of discovery probabilities. The proposed methodology is illustrated through a comprehensive simulation study and the analysis of Expressed Sequence Tags data generated by sequencing a benchmark complementary DNA library.

Key words: Asymptotic equivalence; Bayesian nonparametrics; Credible intervals; Discovery probability; Expressed Sequence Tags; Good–Toulmin estimator; Good–Turing estimator; Smoothing technique; Two parameter Poisson-Dirichlet prior.

## 1. Introduction

Consider a population of individuals $(X_i)_{i\geq 1}$ belonging to an (ideally) infinite number of species $(X_i^*)_{i\geq 1}$ with unknown proportions $(p_i)_{i\geq 1}$. Given an initial observed sample of size $n$, a quantity of practical interest is the probability $D_{n,m}(l)$ of observing at the $(n+m+1)$-th drawn a species with frequency $l \geq 0$ in the enlarged sample of size $n+m$, with the additional sample being unobserved. Formally, if $N_{i,n+m}$ denotes the frequency of $X_i^*$ in the enlarged sample, then

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

Clearly $D_{n,m}(0)$ corresponds to the proportion of yet unobserved species or, equivalently, the probability of discovering a new species at the $(n+m+1)$-th drawn. The random probability (1) is typically referred to as the $(m;l)$-discovery. While the $(0;l)$-discovery is of interest for estimating the probability of discovering new species or rare species, the $(m;l)$-discovery is typically of interest in decision problems regarding the size of the additional sample to collect.

A full range of statistical approaches, parametric and nonparametric as well as frequentist and Bayesian, have been proposed for estimating $D_{n,m}(l)$. These approaches have originally found applications in ecology, and their importance has grown considerably in recent years, driven by challenging applications arising in genetics, bioinformatics, linguistics, designs of experiments, machine learning, etc. See Bunge and Fitzpatrick (1993) and Bunge et al. (2014) for comprehensive reviews. In this article we investigate the relationships between two approaches for estimating $D_{n,m}(l)$: (i) the frequentist nonparametric approach which appeared in the seminal paper by Good (1953), and first developed by Alan M. Turing and Irving J. Good during their collaboration at Bletchley Park in the 1940s; (ii) the Bayesian nonparametric approach recently introduced by Lijoi et al. (2007) and Favaro et al. (2012). In order to state our main contributions, we briefly review the relevant aspects of these two nonparametric approaches.

### 1.1. *The Good–Turing Approach*

Let $\mathscr{H}$ be a parametric statistical hypothesis on the $p_i$'s, that is $\mathscr{H}$ determines the species composition of the population by specifying a distribution function over species and with a finite number of unknown parameters. Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ be a random sample from $\mathscr{H}$, and let us denote by $M_{l,n}$ the number of species with frequency $l$ in $\boldsymbol{X}_n$. According to Good (1953), an estimator of $D_{n,0}(l)$ is $\check{\mathcal{D}}_{n,0}(l;\mathscr{H}) = (l+1)\mathbb{E}_{\mathscr{H}}[M_{l+1,n+1}]/(n+1)$, where $\mathbb{E}_{\mathscr{H}}$ denotes the expected value with respect to the distribution function specified by $\mathscr{H}$. For any $m \geq 1$ let us consider the additional unobserved sample $(X_{n+1}, \ldots, X_{n+m})$, and define $\gamma = m/n$. According to Good and Toulmin (1956), an estimator of $D_{n,m}(0)$ is $\check{\mathcal{D}}_{n,m}(0;\mathscr{H}) = \sum_{i\geq 1}(-\gamma)^{i-1} i \mathbb{E}_{\mathscr{H}}[M_{i,n+m}]/n$. Note that, in principle, $\mathbb{E}_{\mathscr{H}}[M_{l+1,n+1}]$ and $\mathbb{E}_{\mathscr{H}}[M_{i,n+m}]$ do not depend on the initial observed sample, unless the parameters characterizing $\mathscr{H}$ are estimated using such a sample. Several examples of $\mathscr{H}$ are thoroughly discussed in Good (1953) and, among them, we mention the Zipf-type distributions and the discretized Pearson distributions.

In order to dispense with the specification of the parametric statistical hypothesis $\mathscr{H}$, Good (1953) proposed a large $n$ approximation of $\check{\mathcal{D}}_{n,0}(l;\mathscr{H})$ by replacing $\mathbb{E}_{\mathscr{H}}[M_{l+1,n+1}]/(n+1)$ with $m_{l+1,n}/n$, where $m_{l,n}$ denotes the number of species with frequency $l$ in the observed sample. In particular, if $x_n \backsimeq y_n$ means that $x_n$ is approximately equal to $y_n$ for large $n$, then we can write

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

The large $n$ approximate estimator (2) is known as the Good–Turing estimator. A similar large $n$ approximation was proposed in Good and Toulmin (1956) for $\check{\mathcal{D}}_{n,m}(0;\mathscr{H})$. Specifically,

$$\check{\mathcal{D}}_{n,m}(0;\mathscr{H}) \backsimeq \check{\mathcal{D}}_{n,m}(0) = \frac{1}{n}\sum_{i\geq 1}(-\gamma)^{i-1} i m_{i,n}. \tag{3}$$

$\check{\mathcal{D}}_{n,m}(0)$ is known as the Good–Toulmin estimator for the $(m;0)$-discovery. As observed by Good and Toulmin (1956), due to the alternating sign of the series which appears in the estimator (3), if $\gamma$ is large then $\check{\mathcal{D}}_{n,m}(0)$ can yield inadmissible estimates. This instability arises even for values of $m$ moderately larger than $n$, typically $m$ greater than $n$ is enough for it to appear.

A peculiar feature of $\check{\mathcal{D}}_{n,0}(l)$ is that it depends on $m_{l+1,n}$, and not on $m_{l,n}$ as one would intuitively expect for an estimator of the $(0;l)$-discovery. Such a feature, combined with the irregular behavior of the $m_{l,n}$'s for large $l$, makes $\check{\mathcal{D}}_{n,0}(l)$ a sensible approximation only if $l$ is sufficiently small with respect to $n$. Indeed for some large $l$ one might observe that $m_{l,n} > 0$ and $m_{l+1,n} = 0$, which provides the absurd estimate $\check{\mathcal{D}}_{n,0}(l) = 0$, or that $m_{l,n} < m_{l+1,n}$ although the overall observed trend for $m_{l,n}$ is to decrease as $l$ increases. In order to overcome these drawbacks Good (1953) suggested to smooth the irregular series of $m_{l,n}$'s into a more regular series to be used as a proxy. If $m'_{l,n}$'s are the smoothed $m_{l,n}$'s with respect to a smoothing rule $\mathscr{S}$, then $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}) = (l+1)m'_{l+1,n}/n$ is a more accurate approximation than $\check{\mathcal{D}}_{n,0}(l)$. Common smoothing rules consider $m'_{l,n}$, as a function of $l$, to be approximately parabolic or, alternatively, $m'_{l,n}$ to be a certain proportion of the number of species in $\boldsymbol{X}_n$. An alternative method assumes $\mathscr{H}$ to be selected from a superpopulation with an assigned distribution. This flexible method was hinted at in Good (1953) and then left as an open problem.

### 1.2. *The Bayesian Nonparametric Approach*

The approach in Lijoi et al. (2007) and Favaro et al. (2012) is based on the randomization of $p_i$'s. This is somehow reminiscent of the superpopulation smoothing hinted at by Good (1953). Specifically, let $P = \sum_{i\geq 1} p_i \delta_{X_i^*}$ be a discrete random probability measure, namely $(p_i)_{i\geq 1}$ are nonnegative random weights 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 as a nonatomic distribution. The sample $\boldsymbol{X}_n$ is drawn from a population with species composition determined by $P$, i.e.,

$$X_i | P \overset{\text{iid}}{\sim} P \quad i = 1,\ldots,n \quad P \sim \mathscr{P}, \tag{4}$$

for any $n \geq 1$, where $\mathscr{P}$ is a prior distribution over the species composition. Within the large class of priors considered in Lijoi et al. (2007) and Favaro et al. (2012), we focus on the two parameter Poisson-Dirichlet prior by Pitman (1995). Such a choice corresponds to set $p_1 = V_1$ and $p_i = V_i \prod_{1\leq j\leq i-1}(1-V_j)$ where the $V_j$'s are independent Beta random variables with parameter $(1-\sigma, \theta+j\sigma)$, for any $\sigma \in (0,1)$ and $\theta > -\sigma$. We shorten "two parameter Poisson-Dirichlet" respect to the GoodTuring approach: by $\text{PD}(\sigma,\theta)$, and we denote by $P_{\sigma,\theta}$ a random probability measure distributed as $\text{PD}(\sigma,\theta)$ prior.

Under the framework (4), and with $\mathscr{P}$ being the $\text{PD}(\sigma,\theta)$ prior, Lijoi et al. (2007) and Favaro et al. (2012) derived a Bayesian nonparametric estimator of the $(m;l)$-discovery. Specifically, let $\boldsymbol{X}_n$ be a sample from $P_{\sigma,\theta}$ featuring $K_n = k_n$ species with corresponding frequency counts $(M_{1,n},\ldots,M_{n,n}) = (m_{1,n},\ldots,m_{n,n})$. From Proposition 2 in Lijoi et al. (2007), the Bayesian nonparametric estimator of $D_{n,m}(0)$, with respect to a squared loss function, is

$$\hat{\mathcal{D}}_{n,m}(0) = \frac{\theta+\sigma k_n}{\theta+n}\frac{(\theta+n+\sigma)_m}{(\theta+n+1)_m}, \tag{5}$$

for any $m \geq 0$, where $(a)_n = \prod_{0\leq i\leq n-1}(a+i)$ with the proviso $(a)_0 \equiv 1$. For any $m \geq 0$, let $(X_{n+1},\ldots,X_{n+m})$ be the additional unobserved sample from $P_{\sigma,\theta}$. According to Theorem 2 in Favaro et al. (2012), the Bayesian nonparametric estimator of $D_{n,m}(l)$, with respect to a squared loss function, is

$$\begin{aligned}\hat{\mathcal{D}}_{n,m}(l) = &\sum_{i=1}^{l}\binom{m}{l-i} m_{i,n}(i-\sigma)_{l+1-i}\frac{(\theta+n-i+\sigma)_{m-l+i}}{(\theta+n)_{m+1}} \\ &+ (1-\sigma)_l\binom{m}{l}(\theta+\sigma k_n)\frac{(\theta+n+\sigma)_{m-l}}{(\theta+n)_{m+1}},\end{aligned} \tag{6}$$

for any $l = 1,\ldots,n+m$. According to the results displayed in (5) and (6), the Bayesian nonparametric approach has two notable advantages with respect to the Good–Turing approach: (i) it leads directly to exact estimators, thus avoiding the use of large $n$ approximations; (ii) $\hat{\mathcal{D}}_{n,0}(l)$ is a function of $k_n$ and $m_{l,n}$, and not of $m_{l+1,n}$, thus avoiding the use of ad-hoc smoothing techniques to prevent absurd estimates determined by the irregular behavior of the $m_{l,n}$'s.

### 1.3. *Contributions of the Paper and Outline*

Let $a_n \simeq b_n$ mean that $\lim_{n\to+\infty} a_n/b_n = 1$, namely $a_n$ and $b_n$ are asymptotically equivalent as $n$ tends to infinity. In this article we show that the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,0}(l)$ is asymptotically equivalent, as the sample size $n$ tends to infinity, to a Good–Turing estimator with suitably smoothed frequency counts. More precisely, for any $\sigma \in (0,1)$ we show that $\hat{\mathcal{D}}_{n,0}(l) \simeq \check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\text{PD}})$ as $n \to +\infty$, where $\mathscr{S}_{\text{PD}}$ is a smoothing rule such that $m_{l,n}$ is smoothed by

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

While smoothing techniques were introduced in Good (1953) as an ad hoc tool for post processing the $m_l$'s in order to improve the performance of $\check{\mathcal{D}}_{n,0}(l)$, our result shows that, for

a large sample size, a similar smoothing mechanism underlies the Bayesian framework (4) with a PD$(\sigma,\theta)$ prior. We show that $\mathscr{S}_{\rm PD}$ is related to the Poisson smoothing introduced in Good (1953), and we discuss a natural generalization of $\mathscr{S}_{\rm PD}$ which leads to an interesting open problem.

Besides introducing an asymptotic relationship between $\hat{\mathcal{D}}_{n,0}(l)$ and $\check{\mathcal{D}}_{n,0}(l)$, we extend such a relationship to the $(m;l)$-discovery. Specifically, for any fixed $n$ and as $m$ tends to infinity, we show that $\hat{\mathcal{D}}_{n,m}(l)$ is asymptotically equivalent to a Good–Turing estimator $\check{\mathcal{D}}_{m,0}(l)$ in which $m_{l+1,m}$ is replaced by a smoothed version, via $\mathscr{S}_{\rm PD}$, of the Bayesian nonparametric estimator $\hat{\mathcal{M}}_{n,m}(l+1)$ of the number of species with frequency $l$ in the enlarged sample. As a by-product of this result, we introduce a methodology for deriving large $m$ asymptotic credible intervals for $\hat{\mathcal{D}}_{n,m}(l)$, thus completing the study in Lijoi et al. (2007) and Favaro et al. (2012). While the PD$(\sigma,\theta)$ prior leads to an explicit expression for the posterior distribution of $\boldsymbol{D}_{n,m}(l)$, this expression involve combinatorial coefficients whose evaluation for large $m$ is cumbersome, thus preventing its implementation for determining exact credible intervals. Our methodology thus provides a fundamental tool in many situations of practical interest, arising especially in genomics, where $m$ is required to be very large and only a small portion of the population is sampled.

Our results are illustrated through a simulation study and the analysis of Expressed Sequence Tags (ESTs) data generated by sequencing a benchmark complementary DNA (cDNA) library. By means of a simulation study, we compare $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\rm PD})$ with smoothed Good–Turing estimators obtained by using the Poisson smoothing and a smoothing technique in Sampson (2001). Simulated data are generated from the Zeta distribution, whose power law behavior is common in numerous applications. In order to detect the effects of different smoothing techniques, we compare the smoothed Good–Turing estimators with $\check{\mathcal{D}}_{n,0}(l)$ and $\hat{\mathcal{D}}_{n,0}(l)$. A second numerical illustration is devoted to the large $m$ asymptotic credible intervals for the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,m}(l)$. Using ESTs data, we compare asymptotic confidence intervals for the Good–Toulmin estimator $\check{\mathcal{D}}_{n,m}(0)$ with asymptotic credible intervals for its Bayesian nonparametric counterpart $\hat{\mathcal{D}}_{n,m}(0)$. This study completes the numerical illustration presented in Favaro et al. (2009) and Favaro et al. (2012) on the same ESTs data.

In Section 2, we present and discuss the asymptotic equivalence between the Good–Turing approach and the Bayesian nonparametric approach under the assumption of the PD$(\sigma,\theta)$ prior. As a by-product of this asymptotic analysis, in Section 3 we introduce a methodology for associating large $m$ asymptotic credible intervals to $\hat{\mathcal{D}}_{n,m}(l)$. Section 4 contains numerical illustrations. Proofs of our results, as well as related additional materials, and the Matlab code for computing asymptotic credible intervals are available as web-based supplementary materials.

## 2. Good Turing Estimators via Bayesian Nonparametrics

Under a PD$(\sigma,\theta)$ prior, the most notable difference between the Good–Turing estimator and its Bayesian nonparametric counterpart can be traced back to the different use of the information contained in the observed sample. As pointed out in the Introduction, $\check{\mathcal{D}}_{n,0}(0)$ is a function of $m_{1,n}$ while $\hat{\mathcal{D}}_{n,0}(0)$ in (5) is a function of $k_n$. Furthermore, for any $l=1,\ldots,n$, $\check{\mathcal{D}}_{n,0}(l)$ is a function of $m_{l+1,n}$ while $\hat{\mathcal{D}}_{n,0}(l)$ in (6) is a function of $m_{l,n}$. In this section we show that, as $n$ tends to infinity, $\hat{\mathcal{D}}_{n,0}(l)$ is asymptotically equivalent to the smoothed Good–Turing estimator $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\rm PD})$, where $\mathscr{S}_{\rm PD}$ is the smoothing rule displayed in (7). A similar asymptotic equivalence, for fixed $n$ and as $m$ tends to infinity, holds between the estimators $\hat{\mathcal{D}}_{n,m}(l)$ and $\check{\mathcal{D}}_{m,0}(l)$. With a slight abuse of notation, throughout this section we write $X\,|\,Y$ to denote a random variable whose distribution coincides with the conditional distribution of $X$ given $Y$.

### 2.1. *Large n Asymptotic Equivalences for* $\hat{\mathcal{D}}_{n,0}(l)$

We start by recalling the predictive distribution characterizing $P_{\sigma,\theta}$. Let $\boldsymbol{X}_n$ be a sample of size $n$ featuring $K_n=k_n$ species $X_1^*,\ldots,X_{K_n}^*$ with frequencies $(N_{1,n},\ldots,N_{K_n,n})=(n_{1,n},\ldots,n_{k_n,n})$. According to the de Finetti's representation theorem, $\boldsymbol{X}_n$ is part of an exchangeable sequence $(X_i)_{i\geq 1}$ whose distribution has been characterized by Pitman (1995) as follows

$$\mathbb{P}[X_{n+1}\in\cdot\,|\,\boldsymbol{X}_n]=\frac{\theta+\sigma k_n}{\theta+n}\nu_0(\cdot)+\frac{1}{\theta+n}\sum_{i=1}^{k_n}(n_{i,n}-\sigma)\delta_{X_i^*}(\cdot), \tag{8}$$

with $\nu_0$ being a nonatomic probability measure. The conditional probability (8) is referred to as the predictive distribution of $P_{\sigma,\theta}$. Note that $\hat{\mathcal{D}}_{n,0}(l)$ can be read from (8), indeed from (5) and (6) one has $\hat{\mathcal{D}}_{n,0}(0)=(\theta+\sigma k_n)/(\theta+n)$ and $\hat{\mathcal{D}}_{n,0}(l)=(l-\sigma)m_{l,n}/(\theta+n)$, respectively. See Pitman (1995) for details on (8), and on the joint distribution of $K_n$ and $(N_{1,n},\ldots,N_{K_n,n})$ induced by (8).

The asymptotic equivalence between $\hat{\mathcal{D}}_{n,0}(l)$ and $\check{D}_{n,0}(l;\mathscr{S}_{\rm PD})$ relies on an interesting interplay between the large $n$ asymptotic behaviors of $K_n$ and $M_{l,n}$ under a PD$(\sigma,\theta)$ prior. Specifically, let $A_n\overset{\text{a.s.}}{\simeq}B_n$ as $n\to+\infty$ mean that $\lim_{n\to+\infty}A_n/B_n=1$ almost surely, namely $A_n$ and $B_n$ are almost surely asymptotically equivalent as $n$ tends to infinity. By a direct application of Theorem 3.8 and Lemma 3.11 in Pitman (2006), one obtains the asymptotic equivalence

$$M_{l,n}\overset{\text{a.s.}}{\simeq}\frac{\sigma(1-\sigma)_{l-1}}{l!}K_n \tag{9}$$

as $n\to+\infty$. In other terms, under a PD$(\sigma,\theta)$ prior, as the sample size $n$ tends to infinity the number of species with frequency $l$ becomes a proportion $\sigma(1-\sigma)_{l-1}/l!$ of the total number of species. We refer to the web appendix for additional details on (9). The next theorem combines (8) and (9) in order to establish the asymptotic equivalence between $\hat{\mathcal{D}}_{n,0}(l)$ and $\check{D}_{n,0}(l;\mathscr{S}_{\rm PD})$.

THEOREM 1. *Let $\boldsymbol{X}_n$ be a sample of size $n$ from $P_{\sigma,\theta}$ featuring $K_n=k_n$ species with corresponding frequency counts*

$(M_{1,n},\dots,M_{n,n})=(m_{1,n},\dots,m_{n,n})$. *Then, as* $n\to+\infty$, *one has*

$$\hat{\mathcal{D}}_{n,0}(l)\simeq(l+1)\frac{m_{l+1,n}}{n}\simeq(l+1)\frac{\frac{\sigma(1-\sigma)_l}{(l+1)!}k_n}{n}. \tag{10}$$

The smoothing rule $\mathscr{S}_{\mathrm{PD}}$ clearly arises from the large $n$ asymptotic equivalence displayed in (9); indeed $\mathscr{S}_{\mathrm{PD}}$ smooths the frequency count $m_{l,n}$ by taking the proportion $\sigma(1-\sigma)_{l-1}/l!$ of $k_n$. Such a smoothing rule is somehow related to the Poisson smoothing $\mathscr{S}_{\mathrm{Poi}}$, originally introduced by Good (1953), in which the frequency count $m_{l,n}$ is approximately equal to a proportion $\mathrm{e}^{-\lambda}\lambda^{\tau+l-1}/(\tau+l-1)!$ of $k_n$, for any $\lambda>0$ and $\tau\geq 0$ such that $\sum_{l\geq 0}\check{D}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})=1$. See Chapter 2 in Engen (1978) for a common example of Poisson smoothing where $\tau=1$ and $\lambda=n/k_n$. In particular $\mathscr{S}_{\mathrm{PD}}$ is related to the Poisson smoothing corresponding to the choice $\tau=0$ and to a suitable randomization of the parameter $\lambda$. Specifically, let us denote by $P_\lambda$ a discrete random variable with distribution $\mathbb{P}[P_\lambda=l]=\mathrm{e}^{-\lambda}\lambda^{l-1}/(l-1)!$, that is the Poisson smoothing with $\tau=0$ and $\lambda>0$. If $G_{a,b}$ is Gamma random variable with parameter $(a,b)$ and $L_\sigma$ is a discrete random variable with distribution $\mathbb{P}[L_\sigma=l]=\sigma(1-\sigma)_{l-1}/l!$, then according to Devroye (1993) $L_\sigma\overset{\mathrm{d}}{=}1+P_{G_{1,1}G_{1,1-\sigma}/G_{1,\sigma}}$ where $G_{1,1}$, $G_{1,1-\sigma}$ and $G_{1,\sigma}$ are mutually independent.

A peculiar feature of the smoothing rule $\mathscr{S}_{\mathrm{PD}}$ is that it depends only on $\sigma\in(0,1)$. This is because $\mathscr{S}_{\mathrm{PD}}$ is obtained by suitably combining (9), which does not depend of the parameter $\theta$, with other two large $n$ asymptotic equivalences independent of $\theta$, namely: (i) $\hat{\mathcal{D}}_{n,0}(0)\simeq\sigma k_n/n$ and ii) $\hat{\mathcal{D}}_{n,0}(l)\simeq(l-\sigma)m_{l,n}/n$. We conjecture that these asymptotic equivalences, as well as (9), hold for a more general class of priors considered in Lijoi et al. (2007) and Favaro et al. (2012). This is the class of Gibbs-type priors introduced by Pitman (2003) and including two of the most commonly used nonparametric priors, i.e., the PD$(\sigma,\theta)$ prior and the normalized generalized Gamma prior. See De Blasi et al. (2015) for details. In other terms, our conjecture is that Theorem 1 holds for any Gibbs-type prior, that is the smoothing rule $\mathscr{S}_{\mathrm{PD}}$ is invariant with respect to the choice of any prior in the Gibbs class. Intuitively, different smoothing rules for different Gibbs-type priors, if they exist, necessarily require to investigate the high-order large $n$ asymptotic behavior of $\hat{\mathcal{D}}_{n,0}(l)$, and then combine it with a corresponding refinement of the asymptotic equivalence in (9). Work on this is ongoing.

### 2.2. *Large m Asymptotic Equivalences for* $\hat{\mathcal{D}}_{n,m}(l)$

Let $\boldsymbol{X}_n$ be a sample of size $n$ from $P_{\sigma,\theta}$ featuring $K_n=k_n$ species with frequency counts $(M_{1,n},\dots,M_{n,n})=(m_{1,n},\dots,m_{n,n})$. For any $m\geq 1$ let $(X_{n+1},\dots,X_{n+m})$ be an additional unobserved sample. Let $K_m^{(n)}$ be the number of new species in $(X_{n+1},\dots,X_{n+m})$ and let $M_{l,m}^{(n)}$ denote the number of species with frequency $l$ in $(X_1,\dots,X_{n+m})$. Since the additional sample is assumed to be not observed, let us introduce a randomized version of $\hat{\mathcal{D}}_{n+m,0}(0)$ and $\hat{D}_{n+m,0}(l)$ as

$$D_{0,m}^{(n)}=\frac{\theta+\sigma k_n+\sigma K_m^{(n)}}{\theta+n+m} \tag{11}$$

and

$$D_{l,m}^{(n)}=(l-\sigma)\frac{M_{l,m}^{(n)}}{\theta+n+m}, \tag{12}$$

respectively. According to the expression (5), $K_n$ is a sufficient statistics for $\hat{\mathcal{D}}_{n,m}(0)$ and, therefore, the distribution of $D_{0,m}^{(n)}\mid\boldsymbol{X}_n$ takes on the interpretation of the posterior distribution, with respect to $\boldsymbol{X}_n$, of the $(m;0)$-discovery. Similarly, according to the expression (6), $(K_n,M_{1,n},\dots,M_{l,n})$ is a sufficient statistic for $\hat{\mathcal{D}}_{n,m}(l)$ and, therefore, the distribution of $D_{n,m}^{(n)}(l)\mid\boldsymbol{X}_n$ takes on the interpretation of the posterior distribution, with respect to $\boldsymbol{X}_n$, of the $(m;l)$-discovery.

By means of the identities introduced in (11) and (12), the distribution of $D_{0,m}^{(n)}\mid\boldsymbol{X}_n$ and $D_{n,m}^{(n)}(l)\mid\boldsymbol{X}_n$ follows from the distribution of $K_m^{(n)}\mid\boldsymbol{X}_n$ and $M_{l,m}^{(n)}\mid\boldsymbol{X}_n$, respectively, which have been obtained in Lijoi et al. (2007) and Favaro et al. (2013). See the web appendix for details on these distributions. In particular, Proposition 1 in Favaro et al. (2009) showed that

$$\hat{\mathcal{K}}_{n,m}=\mathbb{E}[K_m^{(n)}\mid\boldsymbol{X}_n]=\frac{(\theta/\sigma+k_n)}{(\theta+n)_m}\left((\theta+n+\sigma)_m-(\theta+n)_m\right),$$

which is the Bayesian nonparametric estimator, with respect to a squared loss function, of $K_m^{(n)}$. Furthermore, for any $l=1,\dots,n+m$, Proposition 7 in Favaro et al. (2013) showed that

$$\begin{aligned}\hat{\mathcal{M}}_{n,m}(l)&=\mathbb{E}[M_{l,m}^{(n)}\mid\boldsymbol{X}_n]\\&=\sum_{i=1}^{l}\binom{m}{l-i}m_{i,n}(i-\sigma)_{l-i}\frac{(\theta+n-i+\sigma)_{m-l+i}}{(\theta+n)_m}\\&\quad+(1-\sigma)_{l-1}\binom{m}{l}(\theta+\sigma k_n)\frac{(\theta+n+\sigma)_{m-l}}{(\theta+n)_m},\end{aligned}$$

which is the Bayesian nonparametric estimator, with respect to a squared loss function, of $M_{l,m}^{(n)}$. Note that, by means of (11) and (12) one obtains $\hat{\mathcal{D}}_{n,m}(0)=\mathbb{E}[D_{0,m}^{(n)}\mid\boldsymbol{X}_n]=(\theta+\sigma k_n+\sigma\hat{\mathcal{K}}_{n,m})/(\theta+n+m)$ and $\hat{\mathcal{D}}_{n,m}(l)=\mathbb{E}[D_{l,m}^{(n)}\mid\boldsymbol{X}_n]=(l-\sigma)\hat{\mathcal{M}}_{n,m}(l)/(\theta+n+m)$, which provides an alternative representation for the estimators of the $(m;0)$-discovery and $(m;l)$-discovery, respectively.

Similarly to Theorem 1, an asymptotic equivalence between $\hat{\mathcal{D}}_{n,m}(l)$ and $\hat{\mathcal{D}}_{m,0}(l)$ relies on the interplay between the large $m$ asymptotic behaviors of the random variables $K_m^{(n)}\mid\boldsymbol{X}_n$ and $M_{l,m}^{(n)}\mid\boldsymbol{X}_n$. Specifically, for any $n\geq 1$, by a direct application of Proposition 2 in Favaro et al. (2009) and Corollary 21 in Gnedin et al. (2007) one obtains the following asymptotic equivalence

$$M_{l,m}^{(n)}\mid\boldsymbol{X}_n\overset{\text{a.s.}}{\simeq}\frac{\sigma(1-\sigma)_{l-1}}{l!}K_m^{(n)}\mid\boldsymbol{X}_n \tag{13}$$

as $m\to+\infty$. In other terms, under a PD$(\sigma,\theta)$ prior, the large $m$ asymptotic equivalence between $M_{l,m}^{(n)}\mid\boldsymbol{X}_n$ and $K_m^{(n)}\mid\boldsymbol{X}_n$ coincides with the large $n$ asymptotic equivalence between $M_{l,n}$ and $K_n$. We refer to the web appendix for additional details

on (13). The next theorem combines (11), (12), and (13) in order to establish an asymptotic equivalence between $\hat{\mathcal{D}}_{n,m}(l)$ and $\check{\mathcal{D}}_{m,0}(l)$.

THEOREM 2. *Let $\boldsymbol{X}_n$ be a sample of size $n$ from $P_{\sigma,\theta}$ featuring $K_n = k_n$ species with corresponding frequency counts $(M_{1,n}, \ldots, M_{n,n}) = (m_{1,n}, \ldots, m_{n,n})$. Then, as $m \to +\infty$, one has*

$$\hat{\mathcal{D}}_{n,m}(l) \simeq (l+1)\frac{\hat{\mathcal{M}}_{n,m}(l+1)}{m} \simeq (l+1)\frac{\frac{\sigma(1-\sigma)_l}{(l+1)!}\hat{\mathcal{K}}_{n,m}}{m}. \tag{14}$$

Besides discovery probabilities one is also interested in cumulative discovery probabilities, which are generalizations of the $(m;l)$-discovery defined as follows. For any $\tau \geq 1$, let $\{l_1, \ldots, l_\tau\}$ be a collection of distinct indexes such that $l_i \in \{0, 1, \ldots, n+m\}$ for any $i = 1, \ldots, \tau$. We define the $(m; l_1, \ldots, l_\tau)$-discovery as the cumulative discovery probability $D_{n,m}(l_1, \ldots, l_\tau) = \sum_{1 \leq i \leq \tau} D_{n,m}(l_i)$. Hence, the Bayesian nonparametric estimator of $(m; l_1, \ldots, l_\tau)$-discovery is

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

Such a generalization of the $(m;l)$-discovery is mainly motivated by several applications of practical interest in which one aims at estimating the probability of discovering the so-called rare species. Specifically, these are species not yet observed or observed with a frequency smaller than a certain threshold $\tau$. Of course, large $n$ and large $m$ asymptotic equivalences for the estimator $\hat{\mathcal{D}}_{n,m}(l_1, \ldots, l_\tau)$ follow by a direct application of Theorem 1 and Theorem 2, respectively.

## 3. Credible Intervals for $\hat{\mathcal{D}}_{n,m}(l_1, \ldots, l_\tau)$

While deriving the estimator $\hat{\mathcal{D}}_{n,m}(l)$, Lijoi et al. (2007) and Favaro et al. (2012) did not consider the problem of associating a measure of uncertainty to $\hat{\mathcal{D}}_{n,m}(l)$. Such a problem reduces to the problem of evaluating the distribution of $D_{l,m}^{(n)} \mid \boldsymbol{X}_n$ by combining (11) and (12) with the distributions of $K_m^{(n)} \mid \boldsymbol{X}_n$ and $M_{l,m}^{(n)} \mid \boldsymbol{X}_n$ recalled in the web appendix. While the distribution of $D_{l,m}^{(n)} \mid \boldsymbol{X}_n$ is explicit, in many situations of practical interest the additional sample size $m$ is required to be very large and the computational burden for evaluating this posterior distribution 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 this section, we show how to exploit the large $m$ asymptotic behavior of $D_{l,m}^{(n)} \mid \boldsymbol{X}_n$ in order to associate asymptotic credible intervals to the estimator $\hat{\mathcal{D}}_{n,m}(l)$.

Let $\boldsymbol{X}_n$ be a sample from $P_{\sigma,\theta}$ featuring $K_n = k_n$ species $X_1^*, \ldots, X_{K_n}^*$ with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k_n,n})$. Let $Z_{\sigma,\theta,k_n}^{(n)} \stackrel{\mathrm{d}}{=} B_{k_n+\theta/\sigma, n/\sigma-k_n} Z_{\sigma,(\theta+n)/\sigma}$ where $B_{a,b}$ is a Beta random variable with parameter $(a,b)$ and $Z_{\sigma,q}$ has density function $f_{Z_{\sigma,q}}(z) = \Gamma(q\sigma+1) z^{q-1-1/\sigma} f_\sigma(z^{-1/\sigma})/\sigma\Gamma(q+1)$, with $f_\sigma$ being the positive $\sigma$-stable density. By combining (11) and (12) with Proposition 2 in Favaro et al. (2009) and Corollary 21 in Gnedin et al. (2007), as $m \to +\infty$,

$$\frac{D_{l,m}^{(n)}}{m^{\sigma-1}} \mid \boldsymbol{X}_n \xrightarrow{\text{a.s.}} \frac{\sigma(1-\sigma)_l}{l!} Z_{\sigma,\theta,k_n}^{(n)}. \tag{15}$$

For any $\tau \geq 1$ and $\{l_1, \ldots, l_\tau\}$ such that $l_i \in \{0, 1, \ldots, n+m\}$ for any $i = 1, \ldots, \tau$, let us introduce the random variable $D_{(l_1,\ldots,l_\tau),m}^{(n)} = \sum_{1 \leq i \leq \tau} D_{l_i,m}^{(n)}$. The distribution of $D_{(l_1,\ldots,l_\tau),m}^{(n)} \mid \boldsymbol{X}_n$ takes on the interpretation of the posterior distribution of the $(m; l_1, \ldots, l_\tau)$-discovery. In the next proposition we generalize the fluctuation limit (15) to the cumulative random probability $D_{(l_1,\ldots,l_\tau),m}^{(n)} \mid \boldsymbol{X}_n$.

PROPOSITION 1. *Let $\boldsymbol{X}_n$ be a sample of size $n$ from $P_{\sigma,\theta}$ featuring $K_n = k_n$ species with corresponding frequency counts $(M_{1,n}, \ldots, M_{n,n}) = (m_{1,n}, \ldots, m_{n,n})$. Then, as $m \to +\infty$, one has*

$$\frac{D_{(l_1,\ldots,l_\tau),m}^{(n)}}{m^{\sigma-1}} \mid \boldsymbol{X}_n \xrightarrow{w} \left(\sum_{i=1}^{\tau} \frac{\sigma(1-\sigma)_{l_i}}{l_i!}\right) Z_{\sigma,\theta,k_n}^{(n)}. \tag{16}$$

Fluctuation limits (15) and (16) provide useful tools for approximating the distribution of $D_{l,m}^{(n)} \mid \boldsymbol{X}_n$ and $D_{(l_1,\ldots,l_\tau),m}^{(n)} \mid \boldsymbol{X}_n$. The same fluctuation limits hold for any scaling factor $r(m)$ such that, as $m \to +\infty$, $r(m) \simeq m^{\sigma-1}$. This allows us to introduce a scaling factor finer than $m^{\sigma-1}$. Indeed it can be easily verified that, as soon as $\theta$ and $n$ are not overwhelmingly smaller than $m$,

$$\hat{\mathcal{D}}'_{n,m}(l) = m^{\sigma-1}\frac{\sigma(1-\sigma)_l}{l!}\mathbb{E}[Z_{\sigma,\theta,k_n}^{(n)}],$$

with $\mathbb{E}[Z_{\sigma,\theta,k_n}^{(n)}] = (k_n + \theta/\sigma)\Gamma(\theta+n)/\Gamma(\theta+n+\sigma)$, can be far from $\hat{\mathcal{D}}_{n,m}(l)$. Hence, the corresponding asymptotic credible intervals could be far from the exact estimates. Of course, the same issue appears for the estimator $\hat{\mathcal{D}}_{n,m}(l_1, \ldots, l_\tau)$. For this reason we consider the scaling factors $r^*(m,l)$ and $r^*(m, l_1, \ldots, l_\tau)$ in such a way that $\hat{\mathcal{D}}_{n,m}(l) = r^*(m,l)(\sigma(1-\sigma)_l/l!)\mathbb{E}[Z_{\sigma,\theta,k_n}^{(n)}]$ and $\hat{\mathcal{D}}_{n,m}(l_1, \ldots, l_\tau) = r^*(m, l_1, \ldots, l_\tau)\sum_{1 \leq i \leq \tau}(\sigma(1-\sigma)_{l_i}/l_i!)\mathbb{E}[Z_{\sigma,\theta,k_n}^{(n)}]$, and we define

$$\hat{\mathcal{D}}_{n,m}^*(l) = r^*(m,l)\frac{\sigma(1-\sigma)_l}{l!}\mathbb{E}[Z_{\sigma,\theta,n,k_n}] \tag{17}$$

and

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

It can be easily verified that, as $m \to +\infty$, $r^*(m,l) \simeq m^{\sigma-1}$ and $r^*(m, l_1, \ldots, l_\tau) \simeq m^{\sigma-1}$. Explicit expressions of the scaling factors $r^*(m,l)$ and $r^*(m, l_1, \ldots, l_\tau)$ are provided in web appendix. The reader is referred to Favaro et al. (2009) for

a similar approach in the context of Bayesian nonparametric inference for the number of new species generated by the additional sample.

We make use of (15) and (16) for deriving large $m$ asymptotic credible intervals for $\hat{\mathcal{D}}_{n,m}(l)$ and $\hat{\mathcal{D}}_{n,m}(l_1,\ldots,l_\tau)$. This can be readily done by evaluating appropriate quantiles of the distribution of $Z^{(n)}_{\sigma,\theta,k_n}$. For instance, let $s_1$ and $s_2$ be quantiles of the distribution of $Z^{(n)}_{\sigma,\theta,k_n}$ such that $(s_1,s_2)$ is the 95% credible interval with respect to this distribution. Then, $(r^*(m,l)\sigma(1-\sigma)_l s_1/l!, r^*(m,l)\sigma(1-\sigma)_l s_2/l!)$ is a 95% asymptotic credible interval for $\hat{\mathcal{D}}_{n,m}(l)$. Analogous observations hold true for the estimator $\hat{\mathcal{D}}_{n,m}(l_1,\ldots,l_\tau)$. In order to determine the quantiles $s_1$ and $s_2$, we resort to a simulation algorithm for sampling the limiting random variable $Z^{(n)}_{\sigma,\theta,k_n}$. Note that, according to the definition of $Z^{(n)}_{\sigma,\theta,k_n}$, this procedure involves sampling from the random variable $Z_{\sigma,q}$ with density function $f_{Z_{\sigma,q}}(z)=\Gamma(q\sigma+1)z^{q-1-1/\sigma}f_\sigma(z^{-1/\sigma})/\sigma\Gamma(q+1)$.

A strategy for sampling $Z_{\sigma,q}$ was proposed by Favaro et al. (2009). Specifically, let $L_{\sigma,q}=Z_{\sigma,q}^{-1/\sigma}$ and we introduce a Gamma random variable $U_q$ with parameter $(q,1)$. Then, conditionally on $U_q=u$, the distribution of $L_{\sigma,q}$ has density function proportional to $f_\sigma(x)\exp\{-ux\}$. Therefore, the problem of sampling from $Z_{\sigma,q}$ 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 to the fast rejection algorithm recently proposed in Hofert (2011) for sampling from an exponentially tilted positive $\sigma$-stable random variable. Summarizing, in order to generate random variates from the distribution of $Z^{(n)}_{\sigma,\theta,k_n}$, we have the following steps: (i) sample $B_{k_n+\theta/\sigma,n/\sigma-k_n}$; (ii) sample $G_{(\theta+n)/\sigma,1}$ and set $U_{(\theta+n)/\sigma}=G^{1/\sigma}_{(\theta+n)/\sigma,1}$; (iii) given $U_{(\theta+n)/\sigma}=u$, sample $L_{\sigma,(\theta+n)/\sigma}$ from density proportional to $f_\sigma(x)\exp\{-ux\}$, by means of the fast rejection sampling, and set $Z_{\sigma,(\theta+n)/\sigma}=L^{-\sigma}_{\sigma,(\theta+n)/\sigma}$; (iv) set $Z^{(n)}_{\sigma,\theta,k_n}=B_{k_n+\theta/\sigma,n/\sigma-k_n}Z_{\sigma,(\theta+n)/\sigma}$.

## 4. Illustrations

In order to implement our results, the first issue to be faced is the specification of the parameter $(\sigma,\theta)$ in the PD$(\sigma,\theta)$ prior. Hereafter, following the approach of Lijoi et al. (2007) and Favaro et al. (2012), we resort to an empirical Bayes procedure. Specifically let $\boldsymbol{X}_n$ be a sample from $P_{\sigma,\theta}$ featuring $K_n=k_n$ species with frequencies $(N_{1,n},\ldots,N_{K_n,n})=(n_{1,n},\ldots,n_{k_n,n})$. The empirical Bayes procedure consists in choosing $\theta$ and $\sigma$ that maximize the distribution of $\boldsymbol{X}_n$. This, under a PD$(\sigma,\theta)$ prior, corresponds to setting $(\sigma,\theta)=(\hat\sigma,\hat\theta)$, where

$$(\hat\sigma,\hat\theta)=\arg\max_{(\sigma,\theta)}\left\{\frac{\prod_{i=0}^{k_n-1}(\theta+i\sigma)}{(\theta)_n}\prod_{i=1}^{k_n}(1-\sigma)_{n_{i,n}-1}\right\}. \quad (18)$$

One could also specify a prior distribution on the parameter $(\sigma,\theta)$ and then seek a full Bayesian inference. However, in terms of estimating $D_{n,m}(l)$, there are no relevant differences between this fully Bayes approach and the empirical Bayes approach, given the posterior distribution of $(\sigma,\theta)$ is highly concentrated; this is typically the case of large datasets since the parameter $(\sigma,\theta)$ directly describe the distribution of the observables. See Section 4.2 for a more detailed discussion on these aspects. In the sequel, in order to keep the exposition as simple as possible, we consider the specification of $(\sigma,\theta)$ via the empirical Bayes procedure (18).

### 4.1. *A Comparative Study for $\hat{\mathcal{D}}_{n,0}(l)$, $\check{\mathcal{D}}_{n,0}(l)$, and $\check{\mathcal{D}}_{n,0}(l;\mathscr{S})$*

We compare the performance of the Bayesian nonparametric estimators for the $(0;l)$-discovery with respect to the corresponding Good–Turing estimators and smoothed Good–Turing estimators, for some choices of the smoothing rule. We draw 500 samples of size $n=1000$ from a Zeta distribution with scale parameter $s=1.5$. Recall that a Zeta random variable $Z$ is such that $\mathbb{P}[Z=z]=z^{-s}/C(s)$ where $C(s)=\sum_{i\geq1}i^{-s}$, for $s>1$. Next we order the samples according to the number of observed distinct species $k_n$ and we split them in 5 groups. Specifically, for $i=1,2,\ldots,5$, the $i$-th group of samples will be composed by 100 samples featuring a total number of observed distinct species $k_n$ that stays between the quantiles of order $(i-1)/5$ and $i/5$ of the empirical distribution of $k_n$. We therefore pick at random one sample for each group and label it with the corresponding index $i$. This procedure leads to a total number of 5 samples of 1000 observations with different species compositions.

We use these simulated datasets for comparing estimators for the $(0;l)$-discovery with the true value of $D_{n,0}(l)$, for $l=0,1,5,10,20,30$. Specifically, we consider the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,0}(l)$, the Good–Turing estimator $\check{\mathcal{D}}_{n,0}(l)$, the smoothed Good–Turing estimator $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{PD}})$, and the Poisson smoothed Good–Turing estimator $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})$ with $\tau=1$ and $\lambda=n/k_n$. Finally, we also consider the so-called Simple Good–Turing estimator, denoted by $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{SGT}})$, which is a popular smoothed Good–Turing estimator discussed in Chapter 7 of Sampson (2001). Specifically, in the Simple Good–Turing estimator the smoothing rule $\mathscr{S}_{\mathrm{SGT}}$ consists in first computing, for large $l$, some values $z_{l,n}$ that take into account both the positive frequency counts $m_{l,n}$ and the surrounding zero values, and then in resorting to a line of best fit for the pairs $(\log_{10}(l),\log_{10}(z_{l,n}))$ in order to obtain the smoothed values $m'_{l,n}$.

Table 1 summarizes the result of our comparative study. As an overall measure for the performance of the estimators, we use the sum of squared error (SSE) defined, for a generic estimator $\hat D(l)$ of the $(0,l)$-discovery, as $\mathrm{SSE}(\hat D(l))=\sum_{0\leq l\leq n}(\hat D(l)-d_{n,0}(l))^2$, with $d_{n,0}(l)$ being the true value of $D_{n,0}(l)$. By looking at the SSE in Table 1 it is apparent that $\hat{\mathcal{D}}_{n,0}(l)$ and $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{SGT}})$ are much more accurate than the others. As expected, the Good–Turing estimator $\check{\mathcal{D}}_{n,0}(l)$ has a good performance only for small values of $l$, while inconsistencies arise for large frequencies thus explaining the amplitude of the resulting SSE. For instance, since sample $i=3$ features one species that has frequency $l=20$ and no species with frequency $l=21$, the Good–Turing estimator $\check{\mathcal{D}}_{n,0}(20)$ gives 0 while, clearly, there is positive probability to observe the species appeared 20 times in the sample. Finally, $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{PD}})$ yields a smaller SSE than $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})$. However, the poor accuracy of $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{PD}})$ and $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})$, compared to $\hat{\mathcal{D}}_{n,0}(l)$ and $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{SGT}})$, shows that the parametric assumptions

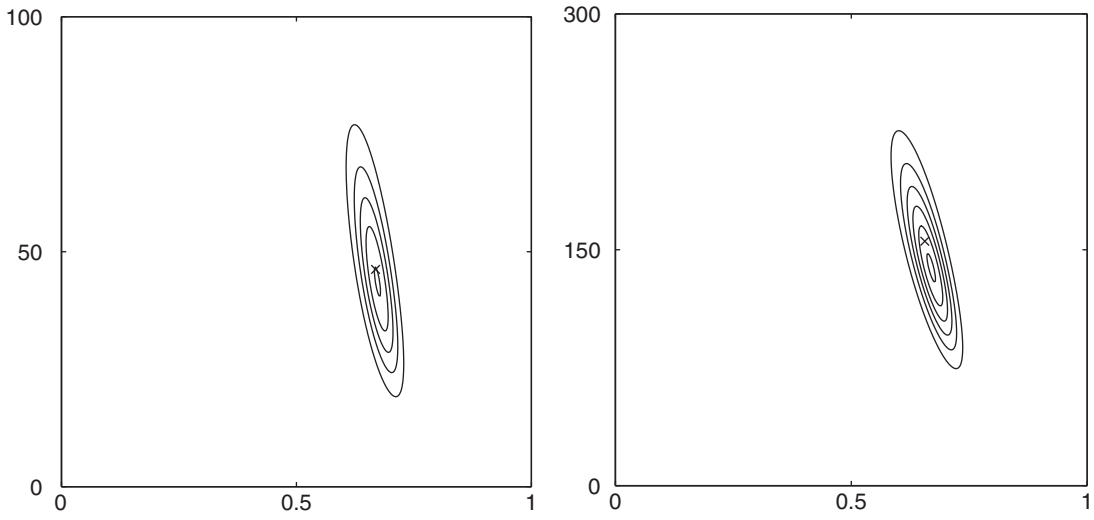
**Table 1**
*Simulated data from a Zeta distribution. Comparison between the true* $(0;l)$*-discovery* $D_{n,0}(l)$ *with the estimate obtained by* $\hat{\mathcal{D}}_{n,0}(l)$, $\check{\mathcal{D}}_{n,0}(l)$, $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{Poi})$, $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{PD})$ *and* $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{SGT})$.

| | Sample | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| | $k_n$ | 136 | 139 | 141 | 146 | 155 |
| | $\hat{\sigma}$ | 0.6319 | 0.6710 | 0.7107 | 0.6926 | 0.6885 |
| | $\hat{\theta}$ | 1.2716 | 0.6815 | 0.2334 | 0.5000 | 0.7025 |
| $l=0$ | $D_{n,0}(l)$ | 0.0984 | 0.0997 | 0.0931 | 0.0924 | 0.0927 |
| | $\hat{\mathcal{D}}_{n,0}(l)$ | 0.0871 | 0.0939 | 0.1004 | 0.1016 | 0.1073 |
| | $\check{\mathcal{D}}_{n,0}(l)$ | 0.0870 | 0.0950 | 0.1040 | 0.1040 | 0.1080 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})$ | 0.0006 | 0.0008 | 0.0008 | 0.0011 | 0.0016 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{PD}})$ | 0.0859 | 0.0933 | 0.1002 | 0.1011 | 0.1067 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{SGT}})$ | 0.0870 | 0.0950 | 0.1040 | 0.1040 | 0.1080 |
| $l=1$ | $D_{n,0}(l)$ | 0.0273 | 0.0272 | 0.0478 | 0.0365 | 0.0331 |
| | $\hat{\mathcal{D}}_{n,0}(l)$ | 0.0320 | 0.0312 | 0.0301 | 0.0319 | 0.0336 |
| | $\check{\mathcal{D}}_{n,0}(l)$ | 0.0320 | 0.0220 | 0.0160 | 0.0240 | 0.0300 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})$ | 0.0047 | 0.0054 | 0.0059 | 0.0073 | 0.0102 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{PD}})$ | 0.0316 | 0.0307 | 0.0290 | 0.0311 | 0.0332 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{SGT}})$ | 0.0319 | 0.0221 | 0.0161 | 0.0240 | 0.0300 |
| $l=5$ | $D_{n,0}(l)$ | 0.0060 | 0.0238 | 0.0132 | 0.0154 | 0.0046 |
| | $\hat{\mathcal{D}}_{n,0}(l)$ | 0.0044 | 0.0173 | 0.0086 | 0.0215 | 0.0043 |
| | $\check{\mathcal{D}}_{n,0}(l)$ | 0.0240 | 0.0180 | 0.0120 | 0.0180 | 0.0120 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})$ | 0.1148 | 0.1206 | 0.1243 | 0.1332 | 0.1470 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{PD}})$ | 0.0126 | 0.0114 | 0.0101 | 0.0111 | 0.0120 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{SGT}})$ | 0.0044 | 0.0176 | 0.0089 | 0.0219 | 0.0044 |
| $l=10$ | $D_{n,0}(l)$ | 0.0105 | 0 | 0.0105 | 0.0092 | 0.0202 |
| | $\hat{\mathcal{D}}_{n,0}(l)$ | 0.0094 | 0 | 0.0093 | 0.0093 | 0.0186 |
| | $\check{\mathcal{D}}_{n,0}(l)$ | 0 | 0 | 0.0220 | 0.0110 | 0.0110 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})$ | 0.0816 | 0.0769 | 0.0738 | 0.0664 | 0.0543 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{PD}})$ | 0.0082 | 0.0072 | 0.0062 | 0.0070 | 0.0075 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{SGT}})$ | 0.0093 | 0 | 0.0094 | 0.0093 | 0.0186 |
| $l=20$ | $D_{n,0}(l)$ | 0 | 0.0142 | 0.0169 | 0 | 0 |
| | $\hat{\mathcal{D}}_{n,0}(l)$ | 0 | 0.0193 | 0.0193 | 0 | 0 |
| | $\check{\mathcal{D}}_{n,0}(l)$ | 0 | 0 | 0 | 0 | 0 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})$ | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{PD}})$ | 0.0053 | 0.0046 | 0.0038 | 0.0043 | 0.0047 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{SGT}})$ | 0 | 0.0194 | 0.0195 | 0 | 0 |
| $l=30$ | $D_{n,0}(l)$ | 0.0260 | 0 | 0 | 0 | 0 |
| | $\hat{\mathcal{D}}_{n,0}(l)$ | 0.0293 | 0 | 0 | 0 | 0 |
| | $\check{\mathcal{D}}_{n,0}(l)$ | 0 | 0 | 0 | 0 | 0.0310 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{Poi}})$ | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{PD}})$ | 0.0041 | 0.0035 | 0.0029 | 0.0033 | 0.0036 |
| | $\check{\mathcal{D}}_{n,0}(l;\mathscr{S}_{\mathrm{SGT}})$ | 0.0292 | 0 | 0 | 0 | 0 |
| | $\mathrm{MSE}(\hat{\mathcal{D}}_{n,0})$ | 0.0006 | 0.0016 | 0.0007 | 0.0007 | 0.0006 |
| | $\mathrm{MSE}(\check{\mathcal{D}}_{n,0})$ | 0.3475 | 0.3773 | 0.3460 | 0.3575 | 0.3530 |
| | $\mathrm{MSE}(\check{\mathcal{D}}_{n,0}(\mathscr{S}_{\mathrm{Poi}}))$ | 0.2657 | 0.2723 | 0.2765 | 0.2769 | 0.2745 |
| | $\mathrm{MSE}(\check{\mathcal{D}}_{n,0}(\mathscr{S}_{\mathrm{PD}}))$ | 0.1748 | 0.1748 | 0.1753 | 0.1746 | 0.1747 |
| | $\mathrm{MSE}(\check{\mathcal{D}}_{n,0}(\mathscr{S}_{\mathrm{SGT}}))$ | 0.0007 | 0.0018 | 0.0014 | 0.0008 | 0.0007 |

underlying the smoothing rules $\mathscr{S}_{\mathrm{Poi}}$ and $\mathscr{S}_{\mathrm{PD}}$ are not suitable for data generated according to a Zeta distribution.

### 4.2. *Credible Intervals for* $\hat{\mathcal{D}}_{n,m}(l_1,\ldots,l_\tau)$

We illustrate the implementation of the asymptotic credible intervals for the Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,m}(l_1,\ldots,l_\tau)$ through the analysis of ESTs data generated by sequencing a benchmark cDNA library. 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




(a) Naegleria Aerobic (b) Naegleria Anaerobic

**Figure 1.** Contour lines of the posterior distribution of the parameter $(\sigma, \theta)$. The cross marks denote the estimates $(\hat{\sigma}, \hat{\theta})$ obtained by means of the empirical Bayes procedure (18).

such libraries, as well as other quantities such as the probability of sampling a new or a rarely observed gene, is of fundamental importance since it allows one to optimize the use of expensive experimental sampling techniques. Hereafter, we consider the *Naegleria gruberi* cDNA libraries prepared from cells grown under different culture conditions, namely aerobic and anaerobic. See Susko and Roger (2004) for additional details.

The *Naegleria gruberi* aerobic library consists of $n = 959$ ESTs with $k_n = 473$ distinct genes and $m_{i,959} = 346, 57, 19, 12, 9, 5, 4, 2, 4, 5, 4, 1, 1, 1, 1, 1, 1,$ for $i = \{1, 2, \ldots, 12\} \cup \{16, 17, 18\} \cup \{27\} \cup \{55\}$. The *Naegleria gruberi* anaerobic library consists of $n = 969$ ESTs with $k_n = 631$ distinct genes and $m_{i,969} = 491, 72, 30, 9, 13, 5, 3, 1, 2, 0, 1, 0, 1,$ for $i \in \{1, 2, \ldots, 13\}$. A fully Bayesian approach involves the specification of a prior distribution for the parameter $(\sigma, \theta)$. Let us consider independent priors for $\sigma$ and $\theta$, namely a Uniform distribution on $(0, 1)$ for $\sigma$ and a Gamma distribution with shape parameter 1 and scale parameter 100, for $\theta$. Figure 1 shows the contour lines of the posterior distribution of $(\sigma, \theta)$; note that these posterior distributions are rather concentrated on a small range of values for $\sigma$. The empirical Bayes approach (18) lead to the following estimates for $(\sigma, \theta)$: $(\hat{\sigma}, \hat{\theta}) = (0.669, 46.241)$ for the *Naegleria gruberi* aerobic library and $(\hat{\sigma}, \hat{\theta}) = (0.656, 155.408)$ for the *Naegleria gruberi* anaerobic library. These values are very close to the mode of the corresponding posterior distributions. See the cross marks in Figure 1. As a matter of fact, the fully Bayesian approach and the empirical Bayes approach lead to very similar estimates for $D_{n,m}(l)$. For instance, by adopting both the empirical Bayes approach and the fully Bayesian approach we get $\hat{\mathcal{D}}_{n,0}(0) = 0.36$ for the *Naegleria gruberi* aerobic library and $\hat{\mathcal{D}}_{n,0}(0) = 0.51$ for the *Naegleria gruberi* anaerobic library. This observation supports our choice of undertaking the empirical Bayes approach (18). The reader is referred to the web appendix for a sensitivity analysis of the asymptotic credible intervals for $\hat{\mathcal{D}}_{n,m}(0)$, with respect to the choice of the parameter $(\sigma, \theta)$.

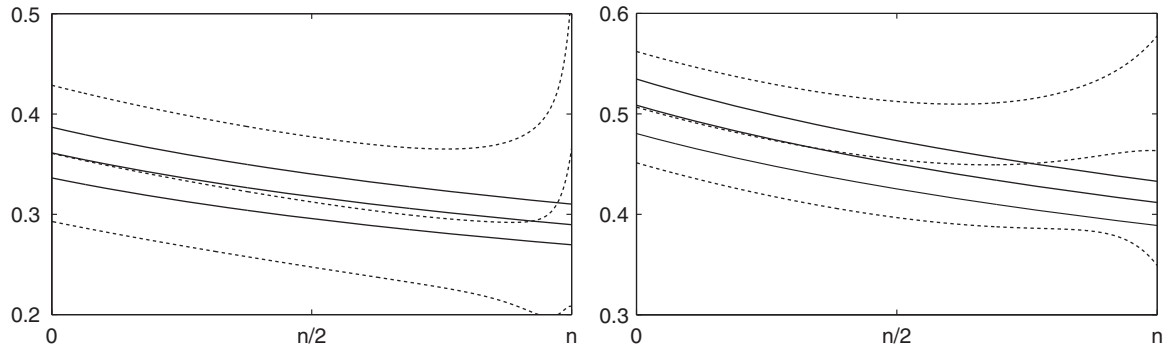
We now focus on the *Naegleria gruberi* aerobic library, and observe that the estimates of the $(m; l)$-discovery provided by the exact estimator $\hat{\mathcal{D}}_{n,m}(0)$, for $m = n, 10n, 100n$, are $0.289, 0.165, 0.080$, respectively, while the corresponding estimates provided by the asymptotic estimator $\hat{\mathcal{D}}'_{n,m}(0)$ gives $0.367, 0.171, 0.080$. It is apparent that $\hat{\mathcal{D}}'_{n,m}(0)$ provides estimates that are close to the exact estimates only when $m$ is very large. This motivates the use of asymptotic estimator $\hat{\mathcal{D}}^*_{n,m}(0)$ with a more accurate scaling factor. Similar considerations hold for the *Naegleria gruberi* anaerobic library. This comparative study between the asymptotic estimators $\hat{\mathcal{D}}'_{n,m}(0)$ and $\hat{\mathcal{D}}^*_{n,m}(0)$, as well as the corresponding credible intervals, is presented in Table 2.

The estimator $\hat{\mathcal{D}}_{n,m}(0)$ is compared with the Good–Toulmin estimator $\check{\mathcal{D}}_{n,m}(0)$. Confidence intervals for $\check{\mathcal{D}}_{n,m}(0)$, which have been devised in Mao (2004) via a moment-based approach, and asymptotic credible intervals for $\hat{\mathcal{D}}_{n,m}(0)$ are also compared. We focus on $m \in [0, n]$: such choice reflects the fact that $\check{\mathcal{D}}_{n,m}(0)$ is known to be a good estimator for small $m$, namely $m \leq n$. See Mao (2004) for details. Figure 2 highlights common features for the estimates obtained for the *Naegleria gruberi* libraries. When $m$ is close to 0 both the approaches provide similar estimates for the $(m; 0)$-discovery. However, even for small values of $m$, asymptotic credible intervals are narrower than the corresponding moment-based 95% confidence intervals. This difference becomes more substantial when $m$ increases. While the asymptotic credible intervals show a regular behavior around the corresponding point estimates, with intervals that tend to get narrow very slowly, estimates obtained with the Good–Toulmin estimator and corresponding confidence intervals feature a more irregular behavior. The latter approach can lead to estimates with very different behaviors, as $m$ approaches $n$.

**Table 2**

Naegleria *aerobic and* Naegleria *anaerobic libraries. Comparison between* $\hat{\mathcal{D}}_{n,m}(0)$ *and the corresponding asymptotic estimators* $\hat{\mathcal{D}}'_{n,m}(0)$ *and* $\hat{\mathcal{D}}^*_{n,m}(0)$*. For the asymptotic estimators* 95% *credible intervals (c.i.) are provided.*

| | | | rate $m^{\sigma-1}$ | | rate $r^*(m,0)$ | |
|---|---|---|---|---|---|---|
| Library | $m$ | $\hat{\mathcal{D}}_{n,m}(0)$ | $\hat{\mathcal{D}}'_{n,m}(0)$ | 95% c.i. | $\hat{\mathcal{D}}^*_{n,m}(0)$ | 95% c.i. |
| *Naegleria* Aerobic | $n$ | 0.289 | 0.367 | (0.339, 0.395) | 0.289 | (0.267, 0.312) |
| ($n = 959$) | $10n$ | 0.165 | 0.171 | (0.158, 0.184) | 0.165 | (0.153, 0.178) |
| | $100n$ | 0.080 | 0.080 | (0.074, 0.086) | 0.080 | (0.073, 0.086) |
| *Naegleria* Anaerobic | $n$ | 0.409 | 0.533 | (0.505, 0.561) | 0.409 | (0.387, 0.431) |
| ($n = 969$) | $10n$ | 0.232 | 0.241 | (0.229, 0.254) | 0.232 | (0.220, 0.245) |
| | $100n$ | 0.109 | 0.109 | (0.103, 0.115) | 0.109 | (0.103, 0.115) |




(a) Naegleria Aerobic (b) Naegleria Anaerobic

**Figure 2.** Comparison of Good–Toulmin estimator $\check{\mathcal{D}}_{n,m}(0)$ (inner dashed curves) and Bayesian nonparametric estimator $\hat{\mathcal{D}}_{n,m}(0)$ (inner solid curves) for $m$ ranging in $[0, n]$. The Good–Toulmin estimates are endowed with 95% confidence intervals (outer dashed curves). Bayesian nonparametric estimators are endowed with asymptotic 95% credible intervals (outer solid curves).

**Table 3**

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

| | | $m = n$ | | $m = 2n$ | | $m = 3n$ | |
|---|---|---|---|---|---|---|---|
| | Library | estimate | 95% c.i. | estimate | 95% c.i. | estimate | 95% c.i. |
| $(m;0)$-discovery | aerobic | 0.289 | (0.267, 0.312) | 0.253 | (0.234, 0.273) | 0.231 | (0.213, 0.249) |
| | anaerobic | 0.409 | (0.387, 0.431) | 0.358 | (0.339, 0.378) | 0.326 | (0.309, 0.344) |
| $(m;1)$-discovery | aerobic | 0.093 | (0.084, 0.101) | 0.083 | (0.076, 0.089) | 0.075 | (0.070, 0.081) |
| | anaerobic | 0.130 | (0.123, 0.137) | 0.117 | (0.111, 0.124) | 0.108 | (0.102, 0.114) |
| $(m;2)$-discovery | aerobic | 0.061 | (0.057, 0.066) | 0.054 | (0.050, 0.059) | 0.050 | (0.046, 0.054) |
| | anaerobic | 0.080 | (0.076, 0.085) | 0.075 | (0.071, 0.079) | 0.070 | (0.066, 0.074) |
| $(m;3)$-discovery | aerobic | 0.046 | (0.042, 0.049) | 0.041 | (0.038, 0.045) | 0.038 | (0.035, 0.041) |
| | anaerobic | 0.059 | (0.056, 0.062) | 0.055 | (0.052, 0.058) | 0.052 | (0.050, 0.055) |
| $(m;4)$-discovery | aerobic | 0.036 | (0.033, 0.039) | 0.034 | (0.031, 0.036) | 0.031 | (0.029, 0.034) |
| | anaerobic | 0.045 | (0.042, 0.047) | 0.044 | (0.042, 0.046) | 0.042 | (0.040, 0.044) |
| $(m;0,1,2,3)$-discovery | aerobic | 0.490 | (0.452, 0.528) | 0.432 | (0.399, 0.465) | 0.394 | (0.364, 0.425) |
| | anaerobic | 0.679 | (0.642, 0.716) | 0.606 | (0.573, 0.640) | 0.556 | (0.526, 0.587) |
| $(m;0,1,2,3,4)$-discovery | aerobic | 0.526 | (0.485, 0.563) | 0.465 | (0.430, 0.501) | 0.425 | (0.393, 0.459) |
| | anaerobic | 0.724 | (0.685, 0.763) | 0.650 | (0.615, 0.686) | 0.599 | (0.566, 0.631) |
| $(m;0,1,2,3,4,5)$-discovery | aerobic | 0.556 | (0.514, 0.599) | 0.494 | (0.456, 0.532) | 0.452 | (0.418, 0.487) |
| | anaerobic | 0.760 | (0.718, 0.801) | 0.686 | (0.649, 0.723) | 0.634 | (0.599, 0.668) |

We conclude this section by determining the asymptotic credible intervals for the point estimators $\hat{\mathcal{D}}_{n,m}(l)$ and $\hat{\mathcal{D}}_{n,m}(l_1,\ldots,l_\tau)$, for some choices of $l$, $\tau$ and $\{l_1,\ldots,l_\tau\}$. With regards to the *Naegleria gruberi* libraries, Bayesian nonparametric inference for discovery probabilities have been recently considered in Favaro et al. (2009) and Favaro et al. (2012), where estimates for discovery probabilities and cumulative discovery probabilities are obtained. However, in Favaro et al. (2009) and Favaro et al. (2012) no measures of uncertainty are provided for these estimates. In Table 3 we summarize estimates of the $(m;l)$-discovery for $l=0,\ldots,4$ and of the $(m;l_1,\ldots,l_\tau)$-discovery for $\tau=3,4,5$. These estimates are endowed with asymptotic 95% credible intervals obtained by combining asymptotic results displayed in (15) and (16) with the choice of the scaling factors $r^*(m,l)$ and $r^*(m,l_1,\ldots,l_\tau)$, respectively. Table 3 thus complete the illustrations presented in Favaro et al. (2009) and Favaro et al. (2012).

## 5. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2, 3 and 4 are available with this paper at the *Biometrics* website on Wiley Online Library. The Matlab code for computing the asymptotic credible intervals for $\hat{\mathcal{D}}_{n,m}(l_1,\ldots,l_\tau)$ is also available at the Biometrics website on Wiley Online Library.

## Acknowledgements

The authors are grateful to an Associate Editor and an anonymous referee for their constructive comments and suggestions. Stefano Favaro is supported by the European Research Council through StG N-BNP 306406. Yee Whye Teh is supported by the European Research Council through the European Unions Seventh Framework Programme (FP7/2007-2013) ERC grant agreement 617411.

## References

Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association* **88**, 364–373.

Bunge, J., Willis, A., and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application* **1**, 427–445.

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 Transactions on Pattern Analysis and Machine Intelligence* **37**, 212–229.

Devroye, L. (1993). A triptych of discrete distributions related to the sable law. *Statistics & Probability Letters* **18**, 349–351.

Engen, S. (1978). *Stochastic Abundance Models.* London: Chapman and Hall.

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. *Journal of the Royal Statistical Society, Series 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. *Annals of Applied Probability* **23**, 1721–1754.

Gnedin, S., Hansen, B., and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power law. *Probability Surveys* **4**, 146–171.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264.

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.

Hofert, M. (2011). Efficiently sampling nested Archimedean copulas. *Computational Statistics & Data Analysis* **55**, 57–70.

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

Mao, C. X. (2004). Prediction of the conditional probability of discovering a new class. *Journal of the American Statistical Association* **99**, 1108–1118.

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

Pitman, J. (2003). Poisson-Kingman partitions. *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.

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.

*Received October* 2014. *Revised May* 2015.
*Accepted June* 2015.