Rediscovery of Good-Turing estimators via Bayesian nonparametrics

This is the author's manuscript

Original Citation:

Availability:
This version is available http://hdl.handle.net/2318/1591184 since 2016-09-03T11:55:58Z

Published version:
DOI:10.1111/biom.12366

Terms of use:
Open Access
Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.
SUMMARY. The problem of estimating discovery probabilities originated in the context of statistical ecology over many years it has become popular due to its frequent appearance in challenging applications arising in genetics, linguistics, designs of experiments, machine learning, etc. A full range of statistical approaches, parametric and as well as frequentist and Bayesian, has been proposed for estimating discovery probabilities. In this article, the relationships between the celebrated Good–Turing approach, which is a frequentist nonparametric approach, introduced 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 probability 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 approximable credible regions for estimating asymptotic credible regions 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 distribution

1. Introduction

Consider a population of individuals \((X_i)_{i=1}^\infty\) belonging to an (ideally) infinite number of species \((X'_i)_{i=1}^\infty\) with unknown proportions \((p_i)_{i=1}^\infty\). 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 species with frequency \(l \geq 0\) in the enlarged sample of size \(n+m\), with the additional sample being unobserved. Formally, if \(N_{i,m}\) denotes the frequency of \(X'_i\) in the enlarged sample, then

\[
D_{n,m}(l) = \sum_{i \geq l} p_i \mathbb{I}_{\{N_{i,m} \geq l\}}.
\]

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 \((0;l)\)-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, has 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 the 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 Iving 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), and first developed by I. J. Good and Toulmin (1956), an estimator of \(D_{n,m}(l)\) for any \(m \geq 1\) let us consider the additional unobserved sample \((X_{n+1}, \ldots, X_{n+m})\), and define \(Y_{n,m}\) the number of species with frequency \(l \leq 0\) in the enlarged sample, then

\[
D_{n,m}(l) = \mathbb{E}_{X_{n+1}, \ldots, X_{n+m}}[Y_{n,m}(l); X_{n+1}, \ldots, X_{n+m}].
\]

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 approximable credible regions for estimating asymptotic credible regions 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.

1.1. The Good–Turing Approach

Let \(\mathcal{H}\) be a parametric statistical hypothesis that is \(\mathcal{H}\) determines the species composition by specifying a distribution function \(F\), and with a finite number of unknown parameters \((X_1, \ldots, X_n)\) be a random sample from \(\mathcal{H}\), and let us consider the additional unobserved sample \((X_{n+1}, \ldots, X_{n+m})\), and define \(Y_{n,m}\) the number of species with frequency \(l \leq 0\) in the enlarged sample, then

\[
D_{n,m}(l) = \mathbb{E}_{X_{n+1}, \ldots, X_{n+m}}[Y_{n,m}(l); X_{n+1}, \ldots, X_{n+m}].
\]

Several exact and approximate credible regions for estimating asymptotic credible regions 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.
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 $D_{n,m}(0; \mathcal{H})$. Specifically,

$$D_{n,m}(0; \mathcal{H}) \approx D_{n,m}(0) = \frac{1}{n} \sum_{i=1}^{n} (-\gamma)^{i-1} i m_{i,n}.$$  \hfill (3)

$D_{n,m}(0)$ 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 $D_{n,m}(0)$ 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 $D_{n,0}(l)$ is that it depends on $m_{1+1,n}$, and not on $m_{1,n}$ as one would intuitively expect for an estimator of the $(0;l)$-discovery. Such a feature, combined with the irregular behavior of the $m_{i,n}$’s for large $l$, makes $D_{n,0}(l)$ a sensible approximation only if $l$ is sufficiently small with respect to $n$. Indeed for some large $l$ one might observe that $m_{1,n} > 0$ and $m_{1+1,n} = 0$, which provides the absurd estimate $D_{n,0}(l) = 0$, or that $m_{1,n} < m_{1+1,n}$ although the overall observed trend for $m_{i,n}$ is to decrease as $l$ increases. In order to overcome these drawbacks Good (1953) suggested to smooth the irregular series of $m_{i,n}$’s into a more regular series to be used as a proxy. If $m'_{i,n}$’s are the smoothed $m_{i,n}$’s with respect to a smoothing rule $\mathcal{H}$, then $D_{n,0}(l; \mathcal{H}) = (l + 1)m'_{1+1,n}/n$ is a more accurate approximation than $D_{n,0}(l)$. Common smoothing rules consider $m'_{i,n}$, as a function of $l$, to be approximately parabolic or, alternatively, $m'_{i,n}$ to be a certain proportion of the number of species in $X_n$. An alternative method assumes $\mathcal{H}$ 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 $X_n$ is drawn from a population with species composition determined by $P$, i.e.,

$$X_i | P \overset{\text{iid}}{\sim} P \quad i = 1, \ldots, n \quad P \sim \mathcal{H},$$  \hfill (4)

Under the framework (4), and with a Poisson–Dirichlet prior, Lijoi et al. (2007) and Favaro et al. (2012) derived a Bayesian nonparametric estimator of the $(l; n)$-discovery. Specifically, let $X_n$ be a superpopulation of species with correspondence $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$. From Lijoi et al. (2007), the Bayesian nonparametric estimator $D_{n,m}(0)$, with respect to a squared loss,

$$D_{n,m}(0) = \frac{\theta + \sigma k_m}{\theta + n} (\theta + n)$$

for any $m \geq 0$, where $a_n = \prod_{0 \leq i \leq n-1} \frac{1}{(a_i)_{i \leq 1}}, (a_0)_{i \leq 1} = 1$. For any $m \geq 0$, let $(X_{n+1}, \ldots, X_m)$ be an additional unobserved sample from $P_{\mathcal{H}}$. According to Theorem 2 in Favaro et al. (2012), the Bayesian nonparametric estimator of $D_{n,m}(l)$, with respect to a squared loss,

$$D_{n,m}(l) = \sum_{i=1}^{l} \left( \frac{m}{l} \right) m_{i,n}(i - \sigma_{l+1-i}) + (1 - \sigma) \left( \frac{m}{l} \right) (\theta + \sigma k_m)$$

for any $l = 1, \ldots, n + m$. According to theorems (5) and (6), the Bayesian nonparametric measure takes advantage with respect to the Good–Turing estimator (i) it leads directly to exact estimators, thus avoiding the use of large $n$ approximations; (ii) $D_{n,0}(l)$ is $m_{l,n}$ and not of $m_{i+1,n}$, thus avoiding the smoothing techniques to prevent absurd estimates of the irregular behavior of the $m_{i,n}$’s.

1.3. Contributions of the Paper and Conclusion

Let $a_n \sim b_n$ mean that $\lim_{n \rightarrow +\infty} a_n/b_n = 1$ are asymptotically equivalent as $n$ tends to infinity. In this paper we show that the Bayesian nonparametric estimator $D_{n,0}(l)$ is asymptotically equivalent to $n$ tends to infinity, to a Good–Turing estimator with suitably smoothed frequency counts. More precisely, we prove that $D_{n,0}(l) \approx D_{n,0}(l; \mathcal{H})$ as a smoothing rule such that $m_{l,n}$ is small.

$$m_{l,n} = \frac{\sigma(1 - \sigma^{l-1})}{l!} k_l$$

While smoothing techniques were introduced as an ad hoc tool for post processing, the priors improve the performance of $D_{n,0}(l)$, our
Good–Turing estimator $D_{n,0}(l)$ in which $m_{l+1,n}$ is replaced by a smoothed version, via $\mathcal{I}_{PD}$, of the Bayesian nonparametric estimator $M_{n,m}(l + 1)$ 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 $D_{n,m}(l)$, 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 $D_{n,m}(l)$, 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 $D_{n,0}(l; \mathcal{I}_{PD})$ 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 $D_{n,0}(l)$ and $\hat{D}_{n,0}(l)$. A second numerical illustration is devoted to the large $m$ asymptotic credible intervals for the Bayesian nonparametric estimator $\hat{D}_{n,m}(l)$. Using ESTs data, we compare asymptotic confidence intervals for the Good–Toulmin estimator $\hat{D}_{n,m}(0)$ with asymptotic credible intervals for its Bayesian nonparametric counterpart $D_{n,m}(0)$. 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 $\hat{D}_{n,m}(l)$. 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 in-
and

served sample. Let \( \lambda \) be the frequency count \( m_{l,n} \) by taking the proportion \( (1 - \sigma)_{-1} / l! \) of \( k_a \). Such a smoothing rule is somehow related to the Poisson smoothing \( \mathcal{F}_{\text{Poi}} \), originally introduced by Good (1953), in which the frequency count \( m_{l,n} \) is approximately equal to a proportion \( e^{-\lambda} l^{-l-1} / (l - 1)! \) of \( k_a \), for any \( \lambda \) > 0 and \( l \geq 0 \) such that \( \sum_{l \geq 0} \mathcal{D}_{0,l}(l) = 1 \). See Chapter 2 in Engen (1978) for a common example of Poisson smoothing where \( \tau = 1 \) and \( \lambda = n/k_a \). In particular \( \mathcal{F}_{\text{PD}} \) 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 \( P_u \) a discrete random variable with distribution \( \mathbb{P}[P_u = l] = e^{-\lambda} l^{-l-1} / (l - 1)! \), 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 \( \mathbb{P}[L_{\sigma} = l] = \sigma(1 - \sigma)_{-1} / l! \), then according to Devroye (1993) \( L_{\sigma} \overset{d}{=} \sum_{i=1}^n \sim \mathcal{G}_{1,1} \). 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 \( \mathcal{F}_{\text{PD}} \) 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 \( \mathcal{D}_{0,l}(l) \), 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 \( \mathcal{D}_{n,m}(l) \)

Let \( X_n \) be a sample of size \( n \) from \( P_{m,n} \) featuring \( K_n = k_a \) species with frequency counts \( (M_{1,n}, \ldots, M_{m,n}) = (m_{1,n}, \ldots, m_{m,n}) \). For any \( m \geq 1 \) let \( (X_{n+1}, \ldots, X_{n+m}) \) be an additional unobserved sample. Let \( K_{n,m}^{(m)} \) be the number of new species in \( (X_{n+1}, \ldots, X_{n+m}) \) and let \( M_{l,m}^{(m)} \) denote the number of species with frequency \( l \) in \( (X_{n+1}, \ldots, X_{n+m}) \). Since the additional sample is assumed to be not observed, let us introduce a randomized version of \( \mathcal{D}_{n+m,0}(0) \) and \( \mathcal{D}_{n+m,0}(l) \) as

\[
\mathcal{D}_{n,m}^{(m)} = \frac{\theta + \sigma K_{n,m}^{(m)}}{\theta + n + m}
\]

according to the expression (6), \( (K_n, M_{n,m}) \) consistent statistic for \( \mathcal{D}_{0,m}(l) \) and, therefore, \( \mathcal{D}_{0,m}^{(m)}(l) \) \( X_n \) takes on the interpretation of a distribution, with respect to \( X_n \), of \( \mathcal{D}_{n,m}(l) \).

By means of the identities introduced for the distribution of \( \mathcal{D}_{0,m}^{(m)}(l) \) and \( \mathcal{D}_{n,m}^{(m)}(l) \), distribution of \( K_{n,m}^{(m)} \) \( X_n \) and \( M_{l,m}^{(m)} \) \( X_n \), it has been obtained in Lijoi et al. (2007) and Favaro et al. (2012) for the web appendix for details on these distributions. In particular, Proposition 1 in Favaro et al. (2012) reads

\[
\hat{K}_{n,m} = \mathbb{E}[K_{n,m}^{(m)} | X_n] = \frac{(\theta + \sigma K_n)}{(\theta + n + m)} ((\theta + n + m) \mathbb{E}[\sigma K_n] + n)
\]

which is the Bayesian nonparametric equivalent to a squared loss function of \( K_{n,m}^{(m)} \). For \( l = 1, \ldots, n + m \), Proposition 7 in Favaro et al. (2012) reads

\[
\hat{M}_{l,m} = \mathbb{E}[M_{l,m}^{(m)} | X_n]
\]

\[
= \sum_{i=1}^{n} \left( \frac{m}{l} \right) m_{l,m}(i - \sigma)_{-1} \left( \frac{\theta + \sigma K_n}{\theta + n + m} \right)
\]

\[
+ (1 - \sigma)_{-1} \left( \frac{m}{l} \right) \left( \theta + \sigma K_n \right)
\]

which is the Bayesian nonparametric respect to a squared loss function, of \( M_{l,m}^{(m)} \). For means of (11) and (12) one obtains \( \hat{D}_{n,m}^{(m)}(l) \) \( (\theta + \sigma K_n + \sigma \hat{K}_{n,m}) / (\theta + n + m) \) and \( \hat{D}_{n,m}^{(m)}(l) \) \( (l - \sigma) \mathcal{M}_{l,m}(l) / (\theta + n + m) \), which provides a representation for the estimators of the \( \mathcal{M}_{l,m}(l) \)-discovery, respectively.

Similarly to Theorem 1, an asymptotic equivalence \( \mathcal{D}_{n,m}(l) \) and \( \mathcal{D}_{n,m,0}(l) \) relies on the interplay of large \( m \) asymptotic behaviors of the random \( M_{l,m}^{(m)} \) \( X_n \). Specifically, for any \( n \geq 1 \), by means of Proposition 2 in Favaro et al. (2009) and Gnedenko et al. (2007) one obtains the equivalence

\[
M_{l,m}^{(m)} | X_n \overset{a.s.}{\sim} \frac{\sigma(1 - \sigma)_{-1} l!}{\theta + n + m}
\]

as \( m \to +\infty \). In other terms, under a \( P_{m,n} \) \( m \) asymptotic equivalence between \( M_{l,m}^{(m)} \) incides with the large \( m \) asymptotic equivalence and \( K_n \). We refer to the web appendix for details.
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 \(\{1, \ldots, l\}\) 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

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

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 \(r\). Of course, large \(n\) and large \(m\) asymptotic equivalents for the estimator \(\hat{D}_{n,m}(l_1, \ldots, l_\tau)\) follow by a direct application of Theorem 1 and Theorem 2, respectively.

3. Credible Intervals for \(D_{n,m}(l_1, \ldots, l_\tau)\)

While deriving the estimator \(\hat{D}_{n,m}(l)\), Lijoi et al. (2007) and Favaro et al. (2012) did not consider the problem of associating a measure of uncertainty to \(\hat{D}_{n,m}(l)\). Such a problem reduces to the problem of evaluating the distribution of \(D_{n,m}(l)\mid X_n\) by combining (11) and (12) with the distributions of \(K_{n,m}\mid X_n\) and \(M_{1,n,m}\mid X_n\) recalled in the web appendix. While the distribution of \(D_{n,m}(l)\mid X_n\) 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 \(D_{n,m}(l)\mid X_n\) in order to associate asymptotic credible intervals to the estimator \(\hat{D}_{n,m}(l)\).

Let \(X_n\) be a sample from \(P_{\theta,\sigma}\) featuring \(K_n = k_n\) species \(X_1, \ldots, X_{k_n}\) with frequencies \((N_{1,a}, \ldots, N_{k_n,a}) = (n_1, \ldots, n_{k_n})\). Let \(Z_{a,k_n}^{(a)} \equiv B_{a+b/n+\theta/n}, k_n Z_{\theta+b/n}^{(a)}\) where \(B_{a,b}\) is a Beta random variable with parameter \((a, b)\) and \(Z_{\theta+b/n}^{(a)}\) has density function \(f_{Z_{\theta+b/n}^{(a)}}(z) = \Gamma(q\sigma + 1)z^{q-1-1/\sigma} f_{\sigma}(z^{-1/\sigma})/\sigma\Gamma(q + 1)\).

\[
\hat{D}_{n,m}(l) \approx (l + 1) \frac{M_{n,m}(l + 1)}{m} \approx (l + 1) \frac{(q + 1)\Gamma(K_{n,m})}{m}.
\]

(14)

The distribution \(D_{n,m}(l)\mid X_n\) is required to be

\[
D_{(l_1, \ldots, l_\tau),m}(l) = \sum_{1 \leq i \leq \tau} D_{n,m}(l_i).
\]

In the next proposition we substitute the fluctuation limit (15) to the cumulative probability \(\hat{D}_{n,m}(l_1, \ldots, l_\tau)\) of \(D_{(l_1, \ldots, l_\tau),m}\mid X_n\).

**Proposition 1.** Let \(X_n\) be a sample of size \(n\) featuring \(K_n = k_n\) species with corresponding frequencies \((M_{1,n,m}, \ldots, M_{n,m}) = (m_1, \ldots, m_{n,m})\). Then, \(X_n\)

\[
\hat{D}_{n,m}(l_1, \ldots, l_\tau) \Rightarrow \sum_{i=1}^{r} \frac{\sigma(1 - \sigma)\theta_1}{l_1!}
\]

Fluctuation limits (15) and (16) provide useful tools for approximating the distribution of \(D_{n,m}(l)\mid X_n\) and \(\hat{D}_{n,m}(l)\mid X_n\). The same fluctuation limits hold for any sequence \(\theta_1, \sigma_2\) such that, as \(m \to +\infty\), \(r(m) \simeq m^{\sigma-1}\). This suggests a scaling factor \(m^{\sigma-1}\) smaller than \(m\).

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

with \(\mathbb{E}[Z_{\sigma,\theta}^{(a)}] = (k_\theta + \theta/\sigma)\Gamma(\theta + n)/\Gamma(\theta + n + 1)\). Hence, the corresponding credible intervals could be far from the data.

Of course, the same issue appears for \(\hat{D}_{n,m}(l_1, \ldots, l_\tau)\). For this reason we consider factors \(r^*(m, l)\) and \(r^*(m, l_1, \ldots, l_\tau)\) in such a way that \(r^*(m, l_1, \ldots, l_\tau)\) is defined as follows. For any \(i = 1, \ldots, \tau\), let \(D_{n,m}(l_i)\mid X_n\) and \(\hat{D}_{n,m}(l_i)\mid X_n\) be defined in (15) and (16) and let \(\hat{D}_{n,m}(l_1, \ldots, l_\tau)\) be defined in (14). Hence, the corresponding asymptotic credible intervals for \(\hat{D}_{n,m}(l_1, \ldots, l_\tau)\) follow by a direct application of Theorem 1 and Theorem 2, respectively.

\[
\hat{D}_{n,m}(l_1, \ldots, l_\tau) = \sum_{i=1}^{\tau} \frac{\sigma(1 - \sigma)\theta_1}{l_i!}\mathbb{E}[Z_{\sigma,\theta}^{(a)}],
\]

It can be easily verified that, as \(m \to +\infty\), \(r^*(m, l_1, \ldots, l_\tau) \simeq m^{\sigma-1}\). Explicit expressions for \(r^*(m, l)\) and \(r^*(m, l_1, \ldots, l_\tau)\) are given in the appendix. The reader is referred to Favaro for more details.
Hereafter, following the approach of Lijoi et al. (2007) and according to the definition of \( Z \), the distribution of \( Z \) can be readily done by evaluating appropriate quantiles of \( B_{kn} \).

One could also specify a prior distribution on the parameter \( G \), and we introduce a Gamma random variable \( Z_{\sigma,q} \) for sampling the limiting random variable \( Z_{\sigma,q} \). This procedure leads to a \( \sigma \)-stable random variable. Summarizing, in order to obtain the smoothed values \( D_{n,l} \), the smoothed Good–Turing estimator \( \bar{D}_{n,l}(l) \), the Poisson smoothed Good–Turing estimator \( \bar{D}_{n,l}(l; \mathcal{F}_{PD}) \), and the Poisson smoothed Good–Turing estimator \( \bar{D}_{n,l}(l; \mathcal{F}_{SGT}) \) with \( \tau = 1 \) and \( \lambda = n/k \),

We draw 500 samples of size \( n = 1000 \) from a Zeta distribution \( \zeta = 1 \) and \( \sigma = 1 \), and we introduce a Gamma random variable \( U_0 \) with parameter \( (q,1) \). Specifically, let \( L_{aq} = Z_{aq} - \sigma \) and we introduce a Gamma random variable \( U_0 \) with parameter \( (q,1) \). Then, conditionally on \( U_0 = u \), the distribution of \( L_{aq} \) has density function proportional to \( f_u(x) \exp(-ux) \). Therefore, the problem of sampling from \( L_{aq} \) 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 \( Z_{aq} \), we have the following steps: (i) sample \( B_{kn} \) from a Gamma distribution \( \Gamma \) and we introduce a Gamma random variable \( U_0 \) with parameter \( (q,1) \). Given \( U_0 = u \), sample \( L_{aq} \) from density proportional to \( f_u(x) \exp(-ux) \), by means of the fast rejection sampling, and set \( Z_{aq} = L_{aq} - \sigma \); (iv) set \( \bar{D}_{n,l} = B_{kn} + i\sigma / \sigma \).

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 \( X_{kn} \) be a sample from \( P_{\theta} \) featuring \( K_n = k_n \) species with frequencies \( (n_1, n_2, \ldots, n_{k_n}) = (n_1, n_2, \ldots, n_{k_n}) \). The empirical Bayes procedure consists in choosing \( \theta \) and \( \sigma \) that maximize the distribution of \( X_{kn} \). This, under a PD(\(\sigma, \theta)\) prior, corresponds to setting \((\sigma, \theta) = (\hat{\sigma}, \hat{\theta})\), where

\[
(\hat{\sigma}, \hat{\theta}) = \arg \max_{(\sigma, \theta)} \left\{ \prod_{i=1}^{k_n} (\theta + i\sigma) / \theta \right\} \prod_{i=1}^{k_n} (1 - \sigma)_{n_i - 1} \tag{18}
\]

One could also specify a prior distribution on the parameter \((\sigma, \theta)\) and then seek a full Bayes solution. However, in terms of estimating \( \bar{D}_{n,l}(l) \), 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 we compare the performance of the empirical Bayes estimators for the \((0,l)\)-discovery with corresponding Good–Turing estimators and smoothed Good–Turing estimators, for some choices of \( l \).

We draw 500 samples of size \( n = 1000 \) from a Zeta distribution \( \zeta = 1 \) and \( \sigma = 1 \), and we introduce a Gamma random variable \( U_0 \) with parameter \( (q,1) \). Specifically, for \( i = 1, 2, \ldots, 5 \), the \( i \)-th sample is composed by 100 samples featuring observed distinct species \( k_n \) and we set \( k_n = i \). As expected, the Good–Turing estimator \( \hat{D}_{n,l}(l) \) shows that the parametric assumptions of observed distinct species \( k_n \) and \( l \) are not relevant differences.

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 \( X_{kn} \) be a sample from \( P_{\theta} \) featuring \( K_n = k_n \) species with frequencies \( (n_1, n_2, \ldots, n_{k_n}) = (n_1, n_2, \ldots, n_{k_n}) \). The empirical Bayes procedure consists in choosing \( \theta \) and \( \sigma \) that maximize the distribution of \( X_{kn} \). This, under a PD(\(\sigma, \theta)\) prior, corresponds to setting \((\sigma, \theta) = (\hat{\sigma}, \hat{\theta})\), where

\[
(\hat{\sigma}, \hat{\theta}) = \arg \max_{(\sigma, \theta)} \left\{ \prod_{i=1}^{k_n} (\theta + i\sigma) / \theta \right\} \prod_{i=1}^{k_n} (1 - \sigma)_{n_i - 1} \tag{18}
\]

One could also specify a prior distribution on the parameter \((\sigma, \theta)\) and then seek a full Bayes solution. However, in terms of estimating \( \bar{D}_{n,l}(l) \), 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
underlying the smoothing rules $\mathcal{S}_{\text{PoI}}$ and $\mathcal{S}_{\text{PD}}$ are not suitable for data generated according to a Zeta distribution.

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

We illustrate the implementation of the asymptotic credible intervals for the Bayesian nonparametric estimator $\hat{D}_{n,m}(l_1, \ldots, l_i)$ through the analysis of EST by sequencing a benchmark cDNA library, an efficient way to characterize expressed genes. The rate of gene discovery depends on the degree of redundancy of the cDNA library from which the EST are obtained. Correctly estimating the rela...
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 \( k_a = 473 \) distinct genes and \( m_{a,959} = 346, 57, 19, 12, 9, 5, 4, 2, 4, 4, 1, 1, 1, 1, 1, 1, 1 \), 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 \( k_a = 631 \) distinct genes and \( m_{a,969} = 491, 72, 30, 9, 13, 5, 3, 1, 2, 0, 1, 0, 1, 1, 1 \), for \( i = \{1, 2, \ldots, 13\} \). 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 \(\hat{D}_{a,0}(0) = 0.36\) for the Naegleria gruberi aerobic library and \(\hat{D}_{a,0}(0) = 0.51\) 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 \(D_n\), the choice of the parameter \((\sigma, \theta)\).

We now focus on the Naegleria gruberi libraries and observe that the estimates of the parameters \(D_{n,m}(l)\) provided by the exact estimator \(D_{n,m}(0)\), are 0.289, 0.165, 0.080, respectively, while the estimates provided by the asymptotic estimator \(\hat{D}_{a,0}(0)\) are 0.367, 0.171, 0.080. It is apparent that the two methods provide estimates that are close to the exact estimates and are very large. This motivates the use of \(\hat{D}_{a,m}(0)\) with a more accurate scaling factor. Similar considerations hold for the Naegleria gruberi anaerobic library. A comparative study between the asymptotic estimator \(\hat{D}_{a,m}(0)\) and \(\hat{D}_{a,m}(0)\), as well as the correspondingcredible intervals are also compared. We focus on \(m \in \{0, n\}\), the fact that \(\hat{D}_{a,m}(0)\) is known to be small for small \(m\), namely \(m \leq n\). See Mao (2004) for details.

Figure 1. Contour lines of the posterior distribution of the parameter \((\sigma, \theta)\). The cross marks denote the estimates (\(\hat{\sigma}, \hat{\theta}\)) provided by the exact estimator \(D_{n,m}(0)\), as well as the corresponding credible intervals, is known to be a good estimator for \(D_n\), for small \(m\), namely \(m \leq n\). See Mao (2004) for details. This difference is substantial when \(m\) increases. While the asymptotic intervals show a regular behavior around the point estimates, with intervals that tend to get wider, estimates obtained with the Good–Toulmin estimations hold for the Naegleria gruberi libraries. When \(m\) increases, approaches provide similar estimates for \(D_{n,m}(0)\). However, even for small values of \(m\), the asymptotic intervals are narrower than the corresponding 95% confidence intervals. This difference becomes more apparent as \(m\) increases. While the empirical Bayes approach provides a regular behavior around the point estimates, with intervals that tend to get narrower, estimates obtained with the asymptotic approach and corresponding confidence intervals show a similar behavior. The latter approach can be very different behaviors, as \(m\) approach...
Table 2

<table>
<thead>
<tr>
<th>Library</th>
<th>( \hat{D}_{n,m}(0) )</th>
<th>( \hat{D}'_{n,m}(0) )</th>
<th>( \hat{D}^*_{n,m}(0) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naegleria aerobic</td>
<td>0.289 (0.267, 0.312)</td>
<td>0.253 (0.234, 0.273)</td>
<td>0.231 (0.213, 0.249)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.409 (0.387, 0.431)</td>
<td>0.358 (0.339, 0.378)</td>
<td>0.326 (0.309, 0.344)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.093 (0.084, 0.101)</td>
<td>0.083 (0.076, 0.089)</td>
<td>0.075 (0.069, 0.081)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.130 (0.123, 0.137)</td>
<td>0.117 (0.111, 0.124)</td>
<td>0.108 (0.102, 0.114)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.061 (0.057, 0.066)</td>
<td>0.054 (0.050, 0.059)</td>
<td>0.050 (0.046, 0.054)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.080 (0.076, 0.085)</td>
<td>0.075 (0.071, 0.079)</td>
<td>0.070 (0.066, 0.074)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.046 (0.042, 0.049)</td>
<td>0.041 (0.038, 0.045)</td>
<td>0.038 (0.035, 0.043)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.059 (0.056, 0.062)</td>
<td>0.055 (0.052, 0.058)</td>
<td>0.052 (0.049, 0.056)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.036 (0.033, 0.039)</td>
<td>0.034 (0.031, 0.036)</td>
<td>0.031 (0.028, 0.034)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.045 (0.042, 0.047)</td>
<td>0.044 (0.042, 0.046)</td>
<td>0.042 (0.039, 0.044)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.490 (0.452, 0.528)</td>
<td>0.432 (0.399, 0.465)</td>
<td>0.394 (0.361, 0.427)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.679 (0.642, 0.716)</td>
<td>0.606 (0.573, 0.640)</td>
<td>0.556 (0.523, 0.589)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.526 (0.485, 0.563)</td>
<td>0.465 (0.430, 0.501)</td>
<td>0.425 (0.390, 0.459)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.724 (0.685, 0.763)</td>
<td>0.650 (0.615, 0.686)</td>
<td>0.599 (0.564, 0.634)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.556 (0.514, 0.599)</td>
<td>0.494 (0.456, 0.532)</td>
<td>0.452 (0.415, 0.490)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.760 (0.718, 0.801)</td>
<td>0.686 (0.649, 0.723)</td>
<td>0.634 (0.596, 0.672)</td>
</tr>
</tbody>
</table>

Figure 2. Comparison of Good–Toulmin estimator \( \hat{D}_{n,m}(0) \) (inner dashed curves) and Bayesian nonparametric estimator \( \hat{D}_{n,m}(0) \) (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

<table>
<thead>
<tr>
<th>Library</th>
<th>( \hat{D}_{n,m}(l) )</th>
<th>( \hat{D}_{n,m}(l,\ldots,\tau) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naegleria aerobic</td>
<td>0.289 (0.267, 0.312)</td>
<td>0.253 (0.234, 0.273)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.409 (0.387, 0.431)</td>
<td>0.358 (0.339, 0.378)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.093 (0.084, 0.101)</td>
<td>0.083 (0.076, 0.089)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.130 (0.123, 0.137)</td>
<td>0.117 (0.111, 0.124)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.061 (0.057, 0.066)</td>
<td>0.054 (0.050, 0.059)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.080 (0.076, 0.085)</td>
<td>0.075 (0.071, 0.079)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.046 (0.042, 0.049)</td>
<td>0.041 (0.038, 0.045)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.059 (0.056, 0.062)</td>
<td>0.055 (0.052, 0.058)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.036 (0.033, 0.039)</td>
<td>0.034 (0.031, 0.036)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.045 (0.042, 0.047)</td>
<td>0.044 (0.042, 0.046)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.490 (0.452, 0.528)</td>
<td>0.432 (0.399, 0.465)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.679 (0.642, 0.716)</td>
<td>0.606 (0.573, 0.640)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.526 (0.485, 0.563)</td>
<td>0.465 (0.430, 0.501)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.724 (0.685, 0.763)</td>
<td>0.650 (0.615, 0.686)</td>
</tr>
<tr>
<td>Naegleria aerobic</td>
<td>0.556 (0.514, 0.599)</td>
<td>0.494 (0.456, 0.532)</td>
</tr>
<tr>
<td>Naegleria anaerobic</td>
<td>0.760 (0.718, 0.801)</td>
<td>0.686 (0.649, 0.723)</td>
</tr>
</tbody>
</table>
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 $\hat{D}_{n,m}(l_1,\ldots,l_r)$ 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