Discovery Probabilities when Uncertainty Matters

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

[1] *Department of Decision Sciences and IGIER, Bocconi University, Milan, Italy,*
julyan.arbel@unibocconi.it
[2] *Department of Economics and Statistics, University of Torino, Italy,*
stefano.favaro@unito.it
[3] *School of Computer Science and Statistics, Trinity College, Dublin, Ireland,*
bernardo.nipoti@gmail.com
[4] *Department of Statistics, University of Oxford, United Kingdom,*
y.w.teh@stats.ox.ac.uk

Résumé. Étant donné un échantillon de taille n d'une population constituée d'espèces dont les proportions sont inconnues, on s'intéresse au tirage d'un $(n+1)$-ème individu, et plus précisément à la probabilité que cet individu coïncide avec une espèce déjà observée avec une fréquence donnée. Ces différentes probabilités sont appelées probabilités de découverte. Nous montrons qu'en spécifiant une distribution a priori de type Gibbs, on obtient naturellement des intervalles de crédibilité pour un estimateur bayésien non paramétrique des probabilités de découverte.

Mots-clés. Intervalles de crédibilité, Probabilité de découverte, Statistique bayésienne non paramétrique.

Abstract. Given a sample of size n from a population of species with unknown proportions, a common problem of practical interest consists in making inference on the probability that the $(n+1)$-th draw coincides with a species already observed with a given frequency. These probabilities are termed discovery probabilities. Under the general framework of Gibbs-type priors we show how to derive credible intervals for a Bayesian nonparametric estimator of discovery probabilities.

Keywords. Bayesian nonparametrics, Credible intervals, Discovery probabilities.

1 Introduction

The problem of estimating discovery probabilities is associated to situations where an experimenter is sampling from a population of individuals $(X_i)_{i\geq 1}$ belonging to an (ideally) infinite number of species $(Y_i)_{i\geq 1}$ with unknown proportions $(q_i)_{i\geq 1}$. Given a sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ interest lies in estimating the probability that the $(n+1)$-th draw coincides with a species with frequency l in $\mathbf{X}_n$, for any $l = 0, 1, \ldots, n$. This probability is denoted by $D_n(l)$ and commonly referred to as the l-discovery. In terms of the species proportions q_i's, one has $D_n(l) = \sum_{i\geq 1} q_i \mathbb{1}_{\{l\}}(\tilde{N}_{i,n})$, where $\tilde{N}_{i,n}$ denotes the frequency of the species Y_i in the sample. See [3] for an up-to-date review on the full range of statistical approaches, parametric and nonparametric as well as frequentist and Bayesian, for estimating the l-discovery and related quantities.

A Bayesian nonparametric approach for estimating $D_n(l)$ was proposed in [6] and [4], and it relies on the randomization of the unknown species proportions q_i's. Specifically, consider the random probability measure $Q = \sum_{i\geq 1} q_i \delta_{Y_i}$, where $(q_i)_{i\geq 1}$ are nonnegative random weights such that $\sum_{i\geq 1} q_i = 1$ almost surely, and $(Y_i)_{i\geq 1}$ are random locations independent of $(q_i)_{i\geq 1}$ and independent and identically distributed according to a nonatomic probability measures ν_0 on a space $\mathbb{X}$. Then, it is assumed that

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

for any $n \geq 1$, where $\mathscr{Q}$ takes on the interpretation of the prior distribution over the unknown species composition of the population. Under the Bayesian nonparametric model (1), the estimator of $D_n(l)$ with respect to a squared loss function, say $\hat{D}_n(l)$, arises directly from the predictive distributions characterizing the exchangeable sequence $(X_i)_{i\geq 1}$. Assuming $\mathscr{Q}$ in the large class of Gibbs-type priors introduced in [5], we consider in this paper the problem of deriving credible intervals for the estimator $\hat{D}_n(l)$.

Let $\mathbf{X}_n$ be a sample from a Gibbs-type random probability measure Q and featuring $K_n = k$ species $X_1^*, \ldots, X_{K_n}^*$ with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k,n})$, and let $A_0 := \mathbb{X} \setminus \{X_1^*, \ldots, X_{K_n}^*\}$ and $A_l := \{X_i^* : N_{i,n} = l\}$, for any $l = 1, \ldots, n$. Since $\hat{D}_n(l) = \mathbb{E}[Q(A_l) \,|\, \mathbf{X}_n]$, the problem of deriving credible intervals for $\hat{D}_n(l)$ boils down to the problem of characterizing the distribution of $Q(A_l) \,|\, \mathbf{X}_n$, where with a slight abuse of notation we denote by $A \,|\, B$ a random variable whose distribution coincides with the conditional distribution of A given B. Indeed this distribution takes on the interpretation of the posterior distribution of $D_n(l)$ with respect to $\mathbf{X}_n$. We present an explicit expression for $E_{n,r}(l) := \mathbb{E}[(Q(A_l))^r \,|\, \mathbf{X}_n]$, for any $r \geq 1$. Due to the boundedness of the support of $Q(A_l) \,|\, \mathbf{X}_n$, the sequence $(E_{n,r}(l))_{r\geq 1}$ characterizes uniquely the distribution of $Q(A_l) \,|\, \mathbf{X}_n$ and, in principle, it can be used to obtain an approximate evaluation of it. An illustration of our results is presented.

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

We start by recalling the predictive distribution characterizing a Gibbs-type prior. Let $\mathbf{X}_n$ be a sample from a Gibbs-type random probability measure Q and featuring $K_n = k$ species $X_1^*, \ldots, X_{K_n}^*$ with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k,n})$. According to the celebrated de Finetti's representation theorem, the sample $\mathbf{X}_n$ is part of an exchangeable sequence $(X_i)_{i\geq 1}$ whose distribution has been characterized in [5] as follows: for any set A in the Borel sigma-algebra of $\mathbb{X}$, one has

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

where $\sigma \in [0,1)$ and $(V_{n,k})_{k\leq n, n\geq 1}$ are nonnegative weights such that $V_{1,1} = 1$ and $V_{n,k} = (n - \sigma k)V_{n+1,k} + V_{n+1,k+1}$. The conditional probability (2) is typically referred to as the predictive distribution of Q. For any $a > 0$ and nonnegative integer n, let $(a)_n := \prod_{0\leq i\leq n-1}(a+i)$ with $(a)_0 := 1$. The two parameter Poisson–Dirichlet prior in [7] is an example of Gibbs-type prior corresponding to the choice $V_{n,k} = \prod_{0\leq i\leq k-1}(\theta + i\sigma)/(\theta)_n$, for any $\sigma \in [0,1)$ and $\theta > -\sigma$. We refer to [6] for other examples.

Let $M_{l,n}$ be the number of species with frequency l in the sample $\mathbf{X}_n$, and $m_{l,n}$ the corresponding observed value. The predictive distribution of Q plays a fundamental role in determining the Bayesian nonparametric estimator $\hat{D}_n(l)$ of $D_n(l)$, as well as the corresponding credible intervals. Indeed, recalling the definition of A_l provided in the Introduction, by a direct applications of (2) one obtains the following expressions

$$E_{n,r}(0) = \mathbb{E}[(Q(A_0))^r \,|\, \mathbf{X}_n] = \sum_{i=0}^{r}\binom{r}{i}(-1)^i\frac{V_{n+i,k}}{V_{n,k}}(n - \sigma k)_i \tag{3}$$

and

$$E_{n,r}(l) = \mathbb{E}[(Q(A_l))^r \,|\, \mathbf{X}_n] = \frac{V_{n+r,k}}{V_{n,k}}((l-\sigma)m_{l,n})_r. \tag{4}$$

We refer to Theorem 1 in [1] for details. Equations (3) and (4) take on the interpretation of the r-th moments of the posterior distribution of $D_n(0)$ and $D_n(l)$ respectively, under the assumption of a Gibbs-type prior. In particular for $r = 1$, by using the recursion for the $V_{n,k}$'s, the posterior moments (3) and (4) reduce to $V_{n+1,k+1}/V_{n,k}$ and $(l-\sigma)m_{l,n}V_{n+1,k}/V_{n,k}$, respectively, which are the Bayesian nonparametric estimators of the l-discovery.

The distribution of $Q(A_l) \,|\, \mathbf{X}_n$ is on $[0,1]$ and, therefore, it is characterized by $(E_{n,r}(l))_{r\geq 1}$. The approximation of a distribution given its moments is a longstanding problem which has been tackled by various approaches such as expansions in polynomial bases, maximum entropy methods and mixtures of distributions. For instance, the polynomial approach consists in approximating the density function of $Q(A_l) \,|\, \mathbf{X}_n$ with a linear combination of orthogonal polynomials, where the coefficients of the combination are determined by

equating $E_{n,r}(l)$ with the corresponding moments of the approximating density. The higher the degree of the polynomials, or equivalently the number of moments used, the more accurate the approximation. The approximating density function of $Q(A_l) \,|\, \mathbf{X}_n$ can then be used to obtain an approximate evaluation of the credible intervals for $\hat{D}_n(l)$. See [2] for details.

Under the assumption of the two parameter Poisson–Dirichlet prior, moments (3) and (4) lead to explicit and simple characterizations for the distributions of $Q(A_l) \,|\, \mathbf{X}_n$. We refer to [1] for another example of Gibbs-type priors leading to explicit characterizations of $Q(A_l) \,|\, \mathbf{X}_n$. In particular, for any $a, b > 0$ let $B_{a,b}$ be a random variable distributed according to a Beta distribution with parameter (a, b). By combining (3) and (4) with $V_{n,k} = \prod_{0 \le i \le k-1} (\theta + i\sigma)/(\theta)_n$, it can be easily verified that

$$Q(A_0) \,|\, \mathbf{X}_n \stackrel{\mathrm{d}}{=} B_{\theta+\sigma k, n-\sigma k} \tag{5}$$

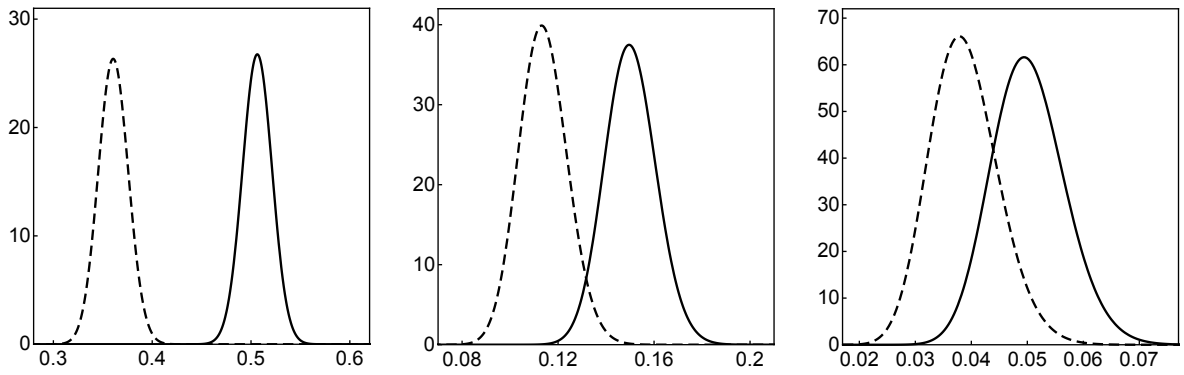
and

$$\begin{aligned} Q(A_l) \,|\, \mathbf{X}_n &\stackrel{\mathrm{d}}{=} B_{(l-\sigma)m_{l,n}, n-\sigma k-(l-\sigma)m_{l,n}} (1 - B_{\theta+\sigma k, n-\sigma k}) \\ &\stackrel{\mathrm{d}}{=} B_{(l-\sigma)m_{l,n}, \theta+n-(l-\sigma)m_{l,n}}. \end{aligned} \tag{6}$$

According to the distributional identities (5) and (6), credible intervals for the Bayesian nonparametric estimator $\hat{D}_n(l)$ can be determined by performing a numerical (Monte Carlo) evaluation of appropriate quantiles of the distribution of $Q(A_l) \,|\, \mathbf{X}_n$. Note that in the special case of the Beta distribution quantiles can be also determined explicitly as solutions of a certain class of non-linear ordinary differential equations. See [8] and references therein for a detailed account on this approach.

3 Illustration

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



(a) $D_n(0)$ (b) $D_n(1)$ (c) $D_n(5)$

Figure 1: *Naegleria gruberi* aerobic and anaerobic libraries. Posterior distributions (dashed curve for aerobic, solid for anaerobic) of discovery probabilities $D_n(l)$, for $l \in \{0, 1, 5\}$.

As for specifying the parameters σ and θ characterizing the two-parameter Poisson–Dirichlet prior, we undertake an empirical Bayes approach. In other terms we choose the values of (σ, θ) that maximize the likelihood function with respect to the sample $\mathbf{X}_n$ featuring $K_n = k_n$ and $(N_{1,n}, \ldots, N_{K_n,n}) = (n_{1,n}, \ldots, n_{k_n,n})$. For details, see [6]. This procedure leads to the estimates $\hat{\sigma} = 0.669$, $\hat{\theta} = 46.241$ for the *Naegleria gruberi* aerobic library, and $\hat{\sigma} = 0.656$, $\hat{\theta} = 155.408$ for the the *Naegleria gruberi* anaerobic library.

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

	$l = 0$		$l = 1$		$l = 5$		$l = 10$	
Aerobic	0.361	(0.331, 0.391)	0.114	(0.095, 0.134)	0.039	(0.028, 0.052)	0.046	(0.034, 0.060)
Anaerobic	0.509	(0.478, 0.537)	0.148	(0.129, 0.169)	0.050	(0.038, 0.064)	0	(0, 0)

Table 1 summarizes the estimated discovery probabilities $\hat{D}_n(l)$, for $l \in \{0, 1, 5, 10\}$, for both libraries, together with the associated 95% posterior credible intervals. Notice that the values of $\hat{D}_n(10)$ and corresponding credible interval, for the anaerobic library, reflect the fact that, since $m_{10,n} = 0$, the posterior distribution of $D_n(10)$ is degenerate at 0. Similarly, Figure 1 compares the posterior distributions of $D_n(l)$, for $l = 0, 1, 5$, corresponding to the two DNA libraries.

Acknowledgement

Stefano Favaro is also affiliated to the Collegio Carlo Alberto, Moncalieri, Italy. Stefano Favaro and Julyan Arbel are supported by the European Research Council (ERC) 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

[1] Arbel, J., Favaro, S., Nipoti, B. and Teh, Y.W.: Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. Preprint arXiv:1506.04915.

[2] Arbel, J., Lijoi, A. and Nipoti, B.: Full Bayesian inference with hazard mixture models. Comput. Statist. Data Anal. **93**, 359-372 (2016)

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

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

[5] Gnedin, A. and Pitman, J.: Exchangeable Gibbs partitions and Stirling triangles. J. Math. Sci. **138**, 5674-5685 (2006)

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

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

[8] Steinbrecher, G. and Shaw, W.T.: Quantile mechanics. European J. Appl. Math. **19**, 87-112. (2008)

[9] 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.