On Bayesian nonparametric inference for discovery probabilities

Sull'inferenza Bayesiana nonparametrica per le probabilità di scoperta

Julyan Arbel, Stefano Favaro, Bernardo Nipoti and Yee Whye Teh

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 $D_n(l)$ that the $(n+1)$-th draw coincides with a species with frequency l in the sample, for any $l \geq 0$. Under the general framework of Gibbs-type priors we show how to derive credible intervals for a Bayesian nonparametric estimator of $D_n(l)$.
Abstract Dato un campione di ampiezza n da una popolazione di specie con proporzioni ignote, un problema di interesse pratico consiste nel fare inferenza sulla probabilità $D_n(l)$ che l'osservazione $(n+1)$-esima sia una specie con frequenza l nel campione, per ogni $l \geq 0$. Per distribuzioni priori di tipo Gibbs, mostriamo come derivare intervalli di credibilità per uno stimatore Bayesiano nonparametrico di $D_n(l)$.

Key words: 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

Julyan Arbel
Collegio Carlo Alberto, Moncalieri, Italy; e-mail: julyan.arbel@gmail.com

Stefano Favaro
Department of Economics and Statistics, University of Torino; e-mail: stefano.favaro@unito.it

Bernardo Nipoti
Department of Economics and Statistics, University of Torino; e-mail: bernardo.nipoti@unito.it

Yee Whye Teh
Department of Statistics, University of Oxford; e-mail: y.w.teh@stats.ox.ac.uk

an (ideally) infinite number of species $(X_i^*)_{i\geq 1}$ 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 $D_n(l) = \sum_{i\geq 1} q_i \mathbb{1}_{\{l\}}(N_{i,n})$, where $N_{i,n}$ denotes the frequency of the species X_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 [5], and it relies on the randomization of the unknown species proportions q_i's. Specifically, consider the random probability measure $Q = \sum_{i\geq 1} q_i \delta_{X_i^*}$, where $(q_i)_{i\geq 1}$ are nonnegative random weights such that $\sum_{i\geq 1} q_i = 1$ almost surely, and $(X_i^*)_{i\geq 1}$ 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 [4], in this paper we consider 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 $Q(A_l) \,|\, \mathbf{X}_n$. 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 bounded support of $Q(A_l) \,|\, X_1,\ldots,X_n$, 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 such a distribution. 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 $X_1^*, \dots, X_{K_n}^*$ with corresponding frequencies $(N_{1,n}, \dots, N_{K_n,n}) = (n_{1,n}, \dots, n_{k,n})$. 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 [4] 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 [6] 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 [5] for other examples.

Let $M_{l,n}$ be the number of species with frequency l in the sample $\mathbf{X}_n$, and by $m_{l,n}$ 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 $D_n(l)$ under the assumption of a Gibbs-type prior. In particular for $r = 1$, by using the recursion 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 moments of the approximating density. The higher the degree of the polynomials, or equivalently the number of moments used, the more accurate the approximation. As a rule of thumb, ten moments turn out to be enough in most cases. 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 $\hat{D}_n(l)$. This is typically done by generating random variates, via rejection sampling, from the approximating distribution of $Q(A_l)\,|\,\mathbf{X}_n$. 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 an other example of Gibbs-type priors lading 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 $V_{n,k}=\prod_{0\leq i\leq k-1}(\theta+i\sigma)/(\theta)_n$, 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

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

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 [7] and references therein for a detailed account on this approach.

3 Illustration

Acknowledgements Stefano Favaro and Bernardo Nipoti are affiliated to the Collegio Carlo Alberto, Moncalieri, Italy. Stefano Favaro is 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. Gnedin, A. and Pitman, J.: Exchangeable Gibbs partitions and Stirling triangles. J. Math. Sci. **138**, 5674-5685 (2006)

5. Lijoi, A., Mena, R.H. and Prünster, I.: Bayesian nonparametric estimation of the probability of discovering new species. Biometrika **94**, 769-786 (2007)
6. Pitman, J.: Exchangeable and partially exchangeable random partitions. Probab. Theory Related Fields **102**, 145–158 (1995)
7. Steinbrecher, G. and Shaw, W.T.: Quantile mechanics. European J. Appl. Math. **19**, 87-112. (2008)