On a class of smoothed Good–Turing estimators

Su una classe di stimatori di Good–Turing lisciati

Stefano Favaro, Bernardo Nipoti and Yee Whye Teh

Abstract 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. A numerical illustration is presented to compare the performance between Bayesian nonparametric estimators with corresponding smoothed Good–Turing estimators.

Abstract *Nell'ipotesi di una distribuzione a priori Poisson-Dirichlet a due parametri, mostriamo che gli stimatori Bayesiani nonparametrici per probabilità di scoperta sono asintoticamente equivalenti, per un campione grande, a stimatori di Good–Turing opportunamente lisciati. Un'illustrazione è presentata per confrontare gli stimatori Bayesiani nonparametrici con i corrispondenti stimatori di Good–Turing lisciati.*

Key words: Bayesian nonparametrics, discovery probability, smoothed Good–Turing estimator

1 Introduction

The problem of estimating discovery probabilities is typically 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 $(X_i^*)_{i\geq 1}$ with unknown proportions $(q_i)_{i\geq 1}$. Given a sample $(X_1,\ldots,X_n)$, which is assumed to be observed, interest

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

lies in estimating the probability that the $(n+1)$-th draw coincides with a species with frequency l in $(X_1,\dots,X_n)$, for any $l=0,1,\dots,n$. 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. Clearly $D_n(0)$ is the proportion of yet unobserved species or, equivalently, the probability of discovering a new species. See [1] 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.

An early approach for estimating the l-discovery was developed by Alan M. Turing and Irving J. Good during their collaboration at Bletchley Park in the 1940s. This approach first appeared in [5]. Specifically, let $\mathscr{H}$ be a parametric statistical hypothesis on the q_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 $(X_1,\dots,X_n)$ be a random sample from $\mathscr{H}$, and let us denote by $M_{l,n}$ the number of species with frequency l in X_n. According to [5], an estimator of $D_n(l)$ is

$$\check{\mathscr{D}}_n(l;\mathscr{H})=(l+1)\frac{\mathbb{E}_{\mathscr{H}}[M_{l+1,n+1}]}{(n+1)}.$$

where $\mathbb{E}_{\mathscr{H}}$ is the expected value with respect to the distribution $\mathscr{H}$. In order to dispense with the specification of the parametric statistical hypothesis $\mathscr{H}$, [5] proposed to replace $\mathbb{E}_{\mathscr{H}}[M_{l+1,n+1}]/(n+1)$ with $m_{l+1,n}/n$, where $m_{l,n}$ denotes the number of species with frequency l in the observed sample. The resulting nonparametric estimator is

$$\check{\mathscr{D}}_n(l)=(l+1)\frac{m_{l+1,n}}{n},$$

which is typically referred to as the Good–Turing estimator. Note that $\check{\mathscr{D}}_n(l;\mathscr{H})$ does not depend on $(X_1,\dots,X_n)$, unless the parameters characterizing $\mathscr{H}$ are estimated using such a sample; several examples of statistical hypothesis $\mathscr{H}$ are thoroughly discussed in [5] and, among them, we mention the Zipf-type distributions and the discretized Pearson distributions. Differently from $\check{\mathscr{D}}_n(l;\mathscr{H})$, the Good–Turing estimator $\check{\mathscr{D}}_n(l)$ depends directly on $(X_1,\dots,X_n)$ through the frequency count $m_{l+1,n}$. That is, $M_{l+1,n}$ is a sufficient statistic for estimating the l-discovery via the Good–Turing approach.

A peculiar feature of $\check{\mathscr{D}}_n(l)$ 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 l-discovery. Such a feature, combined with the irregular behaviour of the $m_{l,n}$'s for large l, makes $\check{\mathscr{D}}_n(l)$ 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 $\check{\mathscr{D}}_n(l)=0$, 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 [5] 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

$\check{\mathscr{D}}_n(l;\mathscr{S}) = (l+1)m'_{l+1,n}/n$ is a more accurate approximation than $\check{\mathscr{D}}_n(l)$. 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 total number of species in the sample.

2 A smoothed $\check{\mathscr{D}}_n(l)$ via Bayesian nonparametrics

A Bayesian nonparametric approach for estimating the l-discovery was proposed in [6] and [3]. Specifically, let $Q = \sum_{i\geq 1} q_i \delta_{X_i^*}$ be a random probability measure, namely $(q_i)_{i\geq 1}$ 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 as a nonatomic distribution. Then, the sample $(X_1,\ldots,X_n)$ is assumed to be drawn from Q, namely $X_1,\ldots,X_n \,|\, Q$ are independent and identically distributed as Q, and Q is distributed according to some distribution $\mathscr{Q}$. In particular, $\mathscr{Q}$ takes on the interpretation of a prior distribution on the species composition. A common choice for $\mathscr{Q}$ is the two parameter Poisson-Dirichlet prior in [7]. Specifically, 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" by PD(σ,θ), and we denote by $Q_{\sigma,\theta}$ a random probability measure distributed as PD(σ,θ) prior.

Under a PD(σ,θ) prior, [3] introduced a Bayesian nonparametric estimator $\hat{\mathscr{D}}_n(l)$, with respect to a squared loss function, of $D_n(l)$. This estimator is obtained by a straightforward application of the predictive distribution characterizing $P_{\sigma,\theta}$, namely the conditional distribution of X_{n+1} given $(X_1,\ldots,X_n)$, for any $n \geq 1$. See [7] for details. Specifically, let $(X_1,\ldots,X_n)$ be a sample from $Q_{\sigma,\theta}$ 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,

$$\hat{\mathscr{D}}_n(0) = \frac{\theta + k_n\sigma}{\theta + n} \tag{1}$$

and

$$\hat{\mathscr{D}}_n(l) = (l-\sigma)\frac{m_{l,n}}{\theta + n}, \tag{2}$$

for any $l = 1,\ldots,n$. The estimators (1) and (2) provide Bayesian nonparametric counterpart of the Good–Turing estimator $\check{\mathscr{D}}_n(l)$. Note that 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. In particular, i) $\check{\mathscr{D}}_n(0)$ is a function of $m_{1,n}$ while $\hat{\mathscr{D}}_n(0)$ is a function of k_n; ii) $\check{\mathscr{D}}_n(l)$ is a function of $m_{l+1,n}$ while $\hat{\mathscr{D}}_n(l)$ is a function of $m_{l,n}$, for any $l = 1,\ldots,n$.

Let $a_n \simeq b_n$ as $n \to +\infty$ mean that $\lim_{n\to+\infty} a_n/b_n = 1$, namely a_n and b_n are asymptotically equivalent. When one either of a_n and b_n is a random quantity, the notation $a_n \overset{\text{a.s}}{\simeq} b_n$ means that the asymptotic relation holds with probability one. By an application of Theorem 3.8 and Lemma 3.11 in [8], one obtains $M_{l,n} \overset{\text{a.s.}}{\simeq} \sigma(1-\sigma)_{l-1}K_n/l!$ as $n \to +\infty$. That is, under a PD(σ,θ) 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. Such a result, suitably combined with (1) and (2), leads to

$$\hat{\mathscr{D}}_n(l) \simeq (l+1)\frac{m_{l+1,n}}{n} \simeq (l+1)\frac{\frac{\sigma\prod_{i=0}^{l-1}(1-\sigma+i)}{(l+1)!}k_n}{n}, \tag{3}$$

$n \to +\infty$. See [4] for details on (3). In other terms, as $n \to +\infty$, the Bayesian nonparametric estimator $\hat{\mathscr{D}}_n(l)$ is asymptotically equivalent to a smoothed Good–Turing estimator, say $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{PD}})$, where $\mathscr{S}_{\mathrm{PD}}$ is a smoothing rule such that $m_{l+1,n}$ is smoothed by $\sigma\prod_{i=0}^{l-1}(1-\sigma+i)k_n/(l+1)!$. The smoothing rule $\mathscr{S}_{\mathrm{PD}}$ clearly arises from the large n asymptotic interplay between K_n and $M_{l,n}$, under the assumption of the PD(σ,θ) prior. As a consequence, $\mathscr{S}_{\mathrm{PD}}$ does not depend on the parameter $\theta > -\sigma$.

The smoothing rule $\mathscr{S}_{\mathrm{PD}}$ is somehow related to the Poisson smoothing $\mathscr{S}_{\mathrm{Poi}}$, originally introduced by [5], in which $m_{l,n}$ 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 $\sum_{l\geq 0}\check{D}_n(l;\mathscr{S}_{\mathrm{Poi}}) = 1$. See Chapter 2 in [2] for an 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 λ. Specifically, let us denote by P_λ 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_σ is a discrete random variable with distribution $\mathbb{P}[L_\sigma = l] = \sigma(1-\sigma)_{l-1}/l!$, then it can be easily verified that L_σ is equal in distribution to $1+P_{G_{1,1}G_{1,1-\sigma}/G_{1,\sigma}}$ where $G_{1,1}$, $G_{1,1-\sigma}$ and $G_{1,\sigma}$ are mutually independent. We refer to [4] for a discussion of the smoothing rule $\mathscr{S}_{\mathrm{PD}}$ under the assumption $\sigma \to 0$.

3 Illustration

We compare the performance of the Bayesian nonparametric estimators for the 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 $C(s) = \sum_{i\geq 1} i^{-s}$, 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 $i = 1,2,\dots,5$, 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, each one with a different species composition.

We use these simulated datasets for comparing estimators for the l-discovery with the true value of $D_n(l)$, for $l=0$, $l=5$ and $l=30$. Specifically, we consider: i) the Bayesian nonparametric estimator $\hat{\mathscr{D}}_n(l)$, for which the parameter (σ,θ) are chosen by the empirical Bayes procedure described in [3]; ii) the Good–Turing estimator $\check{\mathscr{D}}_n(l)$; iii) the smoothed Good–Turing estimator $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{PD}})$; iii) the Poisson smoothed Good–Turing estimator $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{Poi}})$ with $\tau=1$ and $\lambda=n/k_n$. We also consider the so-called Simple Good–Turing estimator, denoted by $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{SGT}})$, which is a popular smoothed Good–Turing estimator discussed in Chapter 7 of [9]. 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 $\big(\log_{10}(l),\log_{10}(z_{l,n})\big)$ in order to obtain the smoothed values $m'_{l,n}$.

Table 1 about here

Table 1 summarizes the result of our comparative study. As an overall measure for the performance of the estimators, we use the mean squared error (MSE) defined, for a generic estimator $\hat{D}(l)$ of the l-discovery, as $\mathrm{MSE}(\hat{D})=\sum_{0\leq l\leq n}(\hat{D}(l)-d_n(l))^2$, with $d_n(l)$ being the true value of $D_n(l)$. By looking at the MSE in Table 1 it is apparent that $\hat{\mathscr{D}}_n(l)$ and $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{SGT}})$ are much more accurate than the others. As expected, the Good–Turing estimator $\check{\mathscr{D}}_n(l)$ has a good performance only for small values of l, while inconsistencies arise for large frequencies thus explaining the amplitude of the resulting MSE. 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 $\check{\mathscr{D}}_n(20)$ gives 0 while, clearly, there is positive probability to observe the species appeared 20 times in the sample. Finally, $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{PD}})$ yields a smaller MSE than $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{Poi}})$.

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. Bunge, J., Willis, A. and Walsh, F.: Estimating the number of species in microbial diversity studies. Annu. Rev. Sta. Appl. **1**, 427–445 (2014)
2. Engen, S.: Stochastic abundance models. Chapman and Hall (1978)
3. Favaro, S., Lijoi, A. and Prünster, I.: A new estimator of the discovery probability. Biometrics **68**, 1188–1196 (2012)
4. Favaro, S., Nipoti, B. and Teh,Y.W.: Rediscovery Good–Turing estimators via Bayesian nonparametrics. Preprint arXiv:1401.0303 (2014)

5. Good, I.J.: The population frequencies of species and the estimation of population parameters. Biometrika **40**, 237–264 (1953)
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. Pitman, J.: Combinatorial Stochastic Processes. Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics. Springer - New York (2006)
9. Sampson, G.: Empirical linguistics. Continuum, London - New York (2001)

Table 1: Simulated data from a Zeta distribution. Some comparison between the true l-discovery $D_n(l)$ with respect to the estimates obtained by $\hat{\mathscr{D}}_n(l)$, $\check{\mathscr{D}}_n(l)$, $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{Poi}})$, $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{PD}})$ and $\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{SGT}})$.

	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
$l=0$	$d_n(l)$	0.0984	0.0997	0.0931	0.0924	0.0927
	$\hat{\mathscr{D}}_n(l)$	0.0871	0.0939	0.1004	0.1016	0.1073
	$\check{\mathscr{D}}_n(l)$	0.0870	0.0950	0.1040	0.1040	0.1080
	$\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{Poi}})$	0.0006	0.0008	0.0008	0.0011	0.0016
	$\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{PD}})$	0.0859	0.0933	0.1002	0.1011	0.1067
	$\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{SGT}})$	0.0870	0.0950	0.1040	0.1040	0.1080
$l=5$	$d_n(l)$	0.0060	0.0238	0.0132	0.0154	0.0046
	$\hat{\mathscr{D}}_n(l)$	0.0044	0.0173	0.0086	0.0215	0.0043
	$\check{\mathscr{D}}_n(l)$	0.0240	0.0180	0.0120	0.0180	0.0120
	$\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{Poi}})$	0.1148	0.1206	0.1243	0.1332	0.1470
	$\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{PD}})$	0.0126	0.0114	0.0101	0.0111	0.0120
	$\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{SGT}})$	0.0044	0.0176	0.0089	0.0219	0.0044
$l=20$	$d_n(l)$	0	0.0142	0.0169	0	0
	$\hat{\mathscr{D}}_n(l)$	0	0.0193	0.0193	0	0
	$\check{\mathscr{D}}_n(l)$	0	0	0	0	0
	$\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{Poi}})$	0.0001	0.0000	0.0000	0.0000	0.0000
	$\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{PD}})$	0.0053	0.0046	0.0038	0.0043	0.0047
	$\check{\mathscr{D}}_n(l;\mathscr{S}_{\mathrm{SGT}})$	0	0.0194	0.0195	0	0
MSE($\hat{\mathscr{D}}_n$)		0.0006	0.0016	0.0007	0.0007	0.0006
MSE($\check{\mathscr{D}}_n$)		0.3475	0.3773	0.3460	0.3575	0.3530
MSE($\check{\mathscr{D}}_n(\mathscr{S}_{\mathrm{Poi}})$)		0.2657	0.2723	0.2765	0.2769	0.2745
MSE($\check{\mathscr{D}}_n(\mathscr{S}_{\mathrm{PD}})$)		0.1748	0.1748	0.1753	0.1746	0.1747
MSE($\check{\mathscr{D}}_n(\mathscr{S}_{\mathrm{SGT}})$)		0.0007	0.0018	0.0014	0.0008	0.0007