Estimating the number of unseen species under heavy tails

Sulla stima del numero di nuove specie nell'ipotesi di code pesanti

Marco Battiston, Federico Camerlenghi, Emanuele Dolera and Stefano Favaro

Abstract Species sampling is a popular subject in several scientific disciplines. Assuming to be provided with an initial sample of size n, a crucial issue is the estimation of the number of new species that will be observed in an additional sample of size λn, being $\lambda > 0$. The case $\lambda < 1$ has been successfully tackled in [6] and [7], but the most interesting situation $\lambda \geq 1$ has been addressed only recently in [11]. We will show that the solution of [11] is unsatisfying when the species' proportions have regularly varying heavy tails. Under this assumption, we provide another estimator for the number of new species and we empirically show its performance.

Abstract *Il campionamento di specie è di particolare interesse in molti contesti. Avendo a disposizione un campione di ampiezza n, un problema particolarmente rilevante consiste nella stima del numero di nuove specie che verranno osservate nelle prossime λn osservazioni, con $\lambda > 0$. Per il caso $\lambda < 1$ il problema fu risolto in [6, 7], mentre il caso $\lambda \geq 1$ è stato affrontato solo di recente in [11]. Mostreremo che la soluzione proposta in [11] non è soddisfacente quando le porzioni delle varie specie hanno code pesanti. Sotto opportune ipotesi sulle code delle porzioni, proporremo uno stimatore per il numero di nuove specie e mostreremo le sue proprietà attraverso alcune simulazioni.*

Marco Battiston
Department of Statistics, University of Oxford, 24-29 St Giles', OX1 3LB Oxford, UKe-mail: marco.battiston@stats.ox.ac.uk

Federico Camerlenghi
Department of Economics, Management and Statistics, University of Milano–Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy e-mail: federico.camerlenghi@unimib.it

Emanuele Dolera
Department of Mathematics, University of Pavia, via Ferrata 5, 27100 Pavia, Italy e-mail: emanuele.dolera@unipv.it

Stefano Favaro
Department of Economics and Statistics, University of Torino, Corso Unione Sovietica 218/bis, 10134 Torino, Italy e-mail: stefano.favaro@unito.it

Key words: Good-Toulmin type estimators, regular variation, species estimation, two parameter Poisson-Dirichlet prior

1 Introduction

Consider a generic population of individuals $(X_i)_{i\geq 1}$ belonging to different species $(X_i^*)_{i\geq 1}$ with unknown proportions $(p_i)_{i\geq 1}$. Given an initial sample of size n, say $(X_1,\ldots,X_n)$, from the population of interest, a crucial problem is the estimation of hitherto unseen species that will be observed in an additional sample of size λn, being $\lambda > 0$. More precisely, denoted by $N_{n,i}$ the frequency of the i-th species in the sample, one is typically interested to estimate the quantity

$$U_{\lambda n} := \sum_{i\geq 1} \mathbb{1}_{\{N_{n,i}=0\}} \mathbb{1}_{\{N_{\lambda n,i}>0\}},$$

i.e. the number of new species that will be observed in an additional sample $(X_{n+1},\ldots,X_{\lambda n-n})$ of size λn.

The first solution of this problem have been suggested in the seminal contributions of [6] and [7]. To fix the notation, we denote by $M_{n,r}$ the number of species with frequency r in $(X_1,\ldots,X_n)$, for any $1 \leq r \leq n$, and by $m_{n,r}$ the corresponding value in the observed sample. Besides K_n represents the number of distinct species in $(X_1,\ldots,X_n)$, and k_n the observed value. The following estimator for $U_{\lambda n}$

$$\hat{U}_{\lambda n} := -\sum_{j\geq 1} (-\lambda)^j m_{n,j} \tag{1}$$

has been provided in [7]. Such an estimator works very well whenever $\lambda < 1$, but it is useless for $\lambda \geq 1$, due to the exponential growth of the coefficients $(-\lambda)^j$. In order to predict $U_{\lambda n}$ for $\lambda > 1$, [1, 7] suggested to use the so–called Euler's transformation, which converts the series in (1) into another one having the same sum, but featuring a faster convergence of its partial sums. However, no theoretical guarantees for the resulting estimator have been established until the work by [11]. They have been able to define a general estimator $U_{\lambda n}$ for the case $\lambda > 1$, which amounts to be

$$\hat{U}^L_{\lambda n} := -\sum_{j\geq 1} (-\lambda)^j \mathbb{P}[L \geq j] m_{n,j}, \tag{2}$$

where L is a random variable whose tail probability compensates for the growth of $(-\lambda)^j$. If L is the Binomial random variable with parameter $(k,(1+\lambda)^{-1})$ then (2) coincides with the Euler-smoothed estimator of [1], with k being the truncation level of (2).

In order to illustrate the performance of (2), we consider a population of 10^6 species whose proportions p_i's are masses of the Zipf distribution, i.e. $p_i \propto i^{-s}$ for some $s > 0$. The parameter s controls the tail of the distribution, to large values of s corresponds heavy tails distributions. Figure 1 shows the estimator (2) for different

choices of s, i.e. from left to right and top to bottom $s = 0.6, 0.8, 1.0, 1.2, 1.4, 1.6$. All experiments are averaged over 100 iterations. The true value is shown in black, and estimated values are colored according to the three choices of the distribution of L considered in Table 1 of [11]: i) a Poisson distribution with parameter $(2\lambda)^{-1}\log_e(n(\lambda+1)^2/(\lambda-1))$; ii) a Binomial distribution with parameter $(2^{-1}\log_2(n\lambda^2/(\lambda-1)),(\lambda+1)^{-1})$; iii) a Binomial distribution with parameter $(2^{-1}\log_3(n\lambda^2/(\lambda-1)),2(\lambda+2)^{-1})$. Shaded bands correspond to one standard deviation. Figure 1 highlights how the tail behavior of the p_i's affects the experimental performance of the estimator $\hat{U}^L_{\lambda n}$: the heavier the tail of $(p_i)_{i\geq 1}$, or rather the lower the species discovery rate, the worse the performance of $\hat{U}^L_{\lambda n}$. The underestimation phenomenon thus suggests that the methods proposed by [1] and then by [11] are not useful for heavy-tailed p_i's. Indeed those methods rely on analytic considerations aimed at improving the rate of convergence of the estimator (1), without acting on the species compositions $(p_i)_{i\geq 1}$. Heavy-tailed species proportions is a common setting in several application areas (see, e.g., [14, 15]), hence the definition of an estimator for $U_{\lambda n}$ under the assumption of heavy-tailed proportions p_i's is a problem of paramount importance.

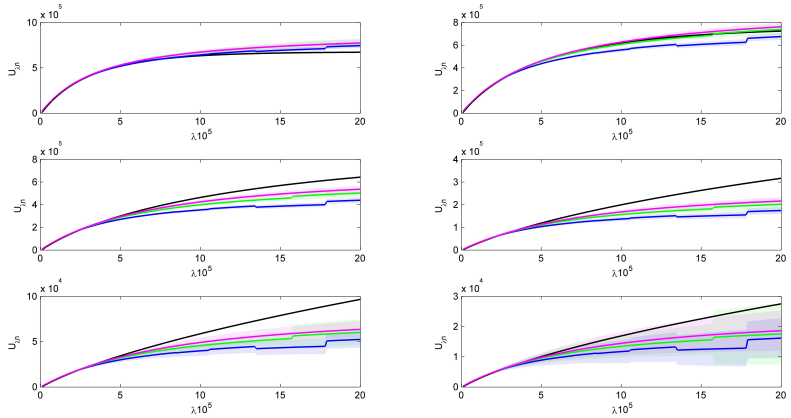

Fig. 1 Estimator of $U_{\lambda n}$ in six Zipf scenarios. The true value is drawn in black, the estimated values are colored in blue (L being the Poisson distribution), green (L being the binomial distribution with success probability $1/(\lambda+1)$) and magenta (L being the binomial distribution with success probability $2/(\lambda+2)$). The shaded bands correspond to one standard deviation.

In the present paper, we introduce the estimator of $U_{\lambda n}$ under heavy–tailed proportions p_i's, showing that it has an opposite behaviour with respect to that highlighted in Figure 1, namely the estimations improve as the parameter s of the Zipf distribution increases. In Section 3 we briefly discuss how to choose the best estimator of $U_{\lambda n}$ among those presented here in relation to the problem at the hand.

Finally we hint possible connections with the Bayesian nonparametric approach, which merit further investigation.

2 Good-Toulmin estimators under regular variation

In the previous section we have seen that the higher the tail of $(p_i)_{i\geq 1}$ (i.e. the higher the parameter s of the Zipf law), the worse the underestimation of $\hat{U}^L_{\lambda n}$. In order to define a suitable estimator for large values of s, we impose a specific assumption on the tails of $(p_i)_{i\geq 1}$, more precisely we resort to the theory of regular variation [8]. In the sequel we will use the notation $f \sim g$ to mean $f/g \to 1$, besides define $\nu(\mathrm{d}x) := \sum_{i\geq 1} \delta_{p_i}(\mathrm{d}x)$ and the measure $\overline{\nu}(x) := \nu[x,1]$. We will say that $(p_i)_{i\geq 1}$ is regularly varying with regular variation index $\alpha \in (0,1)$ if $\overline{\nu}(x) \sim x^{-\alpha}\ell(1/x)$ as $x \downarrow 0$, where $\ell(t)$ is a slowly varying function, that is $\ell(ct)/\ell(t) \to 1$ as $t \to +\infty$ for all $c > 0$. Karlin [8] has proven that in such a context

i) $K_n \overset{\text{a.s.}}{\sim} \mathbb{E}[K_n] \sim \Gamma(1-\alpha)n^{\alpha}\ell(n)$,
ii) $M_{n,r} \overset{\text{a.s.}}{\sim} \mathbb{E}[M_{n,r}] \sim \frac{\alpha\Gamma(r-\alpha)}{r!}n^{\alpha}\ell(n)$

where $\Gamma(\cdot)$ represents the Gamma function. The regular variation index is not known *a priori* and needs to be estimated from the data. This issue can be easily addressed taking the ratio of the number of species with frequency one and the total number of species, namely $\hat{\alpha} := \frac{M_{n,1}}{K_n}$ is a (strongly) consistent estimator of α. For additional details on regular variation refer to [8] and [5].

We now define an estimator for $U_{\lambda n}$, when $\lambda > 1$ and the sequence $(p_i)_{i\geq 1}$ has regularly varying heavy tails. In order to to this, we consider (2) when L is a Binomial random variable with parameters $(2^{-1}\log_2(n\lambda^2/(\lambda-1)), (\lambda+1)^{-1})$, and we tune this estimator under the hypothesis of regular variation, thus obtaining

$$\hat{U}^L_{\lambda n}(\alpha) := k_n \sum_{z=1}^{k_L} \binom{z+\alpha-1}{z} \left(\frac{\lambda}{\lambda+1}\right)^z, \tag{3}$$

where k_L is the truncation level, i.e. $k_L := 2^{-1}\log_2(n\lambda^2/(\lambda-1))$. Note that $k_L \to +\infty$ as $n \to +\infty$.
Finally it is worth noticing that in (3) the regular variation index α is unknown, hence, in order to use the estimator $\hat{U}^L_{\lambda n}(\alpha)$, one should replace α with the corresponding consistent estimator $\hat{\alpha} = \frac{M_{n,1}}{K_n}$.

We now consider the same Zipf's scenarios presented in Section 1 to illustrate the performance of (3). Figure 2 shows $\hat{U}^L_{\lambda n}(\hat{\alpha})$ for different choices for the parameter s of the Zipf distribution, i.e. from left to right and top to bottom $s = 0.6, 0.8, 1.0, 1.2, 1.4, 1.6$. All experiments are averaged over 100 iterations. The true value is shown in black, the estimated value in red, and the shaded band corresponds to one standard deviation. By a comparison between Figure 1 and Figure 2, one immediately realizes that $\hat{U}^L_{\lambda n}(\hat{\alpha})$ has an opposite behaviour with respect to

$\hat{U}^L_{\lambda n}$. That is, the heavier the tail of $(p_i)_{i\geq 1}$, or rather the lower the species discovery rate, the better the performance of $\hat{U}^L_{\lambda n}(\hat{\alpha})$.

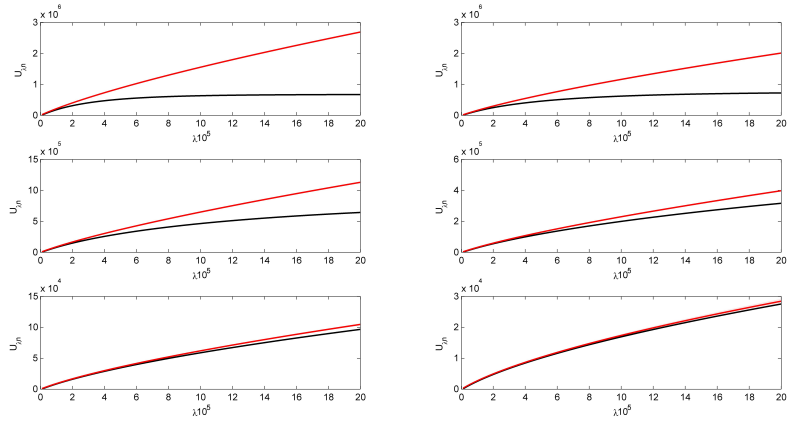

Fig. 2 Estimator of $U_{\lambda n}$ in six Zipf scenarios. The true value is drawn in black, the estimated value in red. The shaded bands correspond to one standard deviation.

3 Discussion

In this paper we focused on the estimation of the number of unseen species that will be observed in a future sample of size λn. The performance of the estimators presented here has been assessed empirically for the ubiquitous Zipf distribution with parameter s. We have shown that the estimator (2) proposed by [11] is useful when $s \leq 1$, but it radically worsens when $s > 1$. For this reason, in Section 2, we have tuned such an estimator under the assumption of regularly varying heavy tails p_i's. In Figure 2, we have empirically shown that $\hat{U}^L_{\lambda n}(\hat{\alpha})$ performs very well when $s > 1$, but not when $s \leq 1$, featuring an opposite behaviour with respect to $\hat{U}^L_{\lambda n}$.
In real applications the parameter s is not given and one has to decide whether to employ either $\hat{U}^L_{\lambda n}$ or $\hat{U}^L_{\lambda n}(\hat{\alpha})$. In order to face this issue one can find an estimate of the parameter s by means of linear regression as suggested in [10], thus using $\hat{U}^L_{\lambda n}$ if the resulting estimator of s is less than 1, $\hat{U}^L_{\lambda n}(\hat{\alpha})$ otherwise.

An interesting open problem which merits further investigation is the connection between the estimator of the number of unseen species $\hat{U}^L_{\lambda n}(\alpha)$ presented here and the Bayesian nonparametric estimator derived in [2] and [3]. The Bayesian viewpoint needs the specification of a prior distribution for the species proportions $(p_i)_{i\geq 1}$, namely one needs to choose a prior distribution for the random probability

measure $\tilde{p} = \sum_{i \geq 1} p_i \delta_{X_i^*}$. In such a context $(X_1, \ldots, X_n)$ is a sample coming from an exchangeable sequence of observations driven by $\tilde{p}$, i.e.

$$\begin{aligned} X_i \,|\, \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \qquad i = 1, \ldots, n, \\ \tilde{p} &\sim \mathscr{P}, \end{aligned} \tag{4}$$

where $\mathscr{P}$ is the distribution of $\tilde{p}$. A common choice for $\mathscr{P}$ is the law of the two parameter Poisson-Dirichlet process, which was introduced in [12] and further investigated in [13]. The sequence $(p_i)_{i \geq 1}$ is such that $p_1 = v_1$ and $p_i = v_i \prod_{1 \leq j \leq i-1}(1 - v_j)$, for any $i \geq 2$, where the v_j's are independent random variables, each v_j is distributed according to a Beta distribution with parameter $(1 - \alpha, \theta + j\alpha)$, for $\alpha \in (0,1)$ and $\theta > -\alpha$. In [4], the authors have proven that the celebrated Good-Turing estimator of the discovery probability is asymptotically equivalent, for a large sample size, to the Bayesian nonparametric one under the assumption of a two parameter Poisson-Dirichlet prior. Analogously, we would like to asses whether $U_{\lambda n}$ is asymptotically equivalent to the regularly varying nonparametric estimator $\hat{U}^L_{\lambda n}(\alpha)$ for specific choices of $\mathscr{P}$, as the sample size n increases.

References

1. Efron, B., Thisted, R.: Estimating the number of unseen species: How many words did Shakespeare know? Biometrika **63**, 435–447 (1976)
2. Favaro, S., Lijoi, A., Mena, R.H., Prünster, I.: Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. J. Roy. Statist. Soc. Ser. B **71**, 993–1008 (2009)
3. Favaro, S., Lijoi, A., Prünster, I.: A new estimator of the discovery probability. Biometrics **68**, 1188–1196 (2012)
4. Favaro, S., Nipoti, B., Teh, Y.W.: Rediscovery of GoodTuring estimators via Bayesian nonparametrics. Biometrics **72**, 136–145 (2016)
5. Gnedin, A., Hansen, B., Pitman, J.: Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. Probab. Surv. **4**, 146–171 (2007)
6. Good, I.J.: The population frequencies of species and the estimation of population parameters. Biometrika **40**, 237–264 (1953)
7. Good, I.J., Toulmin, G.H.: The number of new species, and the increase in population coverage, when a sample is increased. Biometrika **43**, 45–63 (1956)
8. Karlin, S.: Central limit theorems for certain infinite urn schemes. J. Math. Mech. **17**, 373–401 (1967)
9. Lijoi, A., Mena, R.H., Prünster, I.: Bayesian nonparametric estimation of the probability of discovering new species. Biometrika **94**, 769–786 (2007)
10. Motwani, S., Vassilvitskii, S.: Distinct value estimators for power law distributions. In Proceedings of the Workshop on Analytic Algorithms and Combinatorics (2006)
11. Orlitsky, A., Suresh, A.T., Wu, Y.: Optimal prediction of the number of unseen species. Proc. Natl. Acad. Sci. USA **47**, 13283–13288 (2016)
12. Perman, M., Pitman, J., Yor, M.: Size-biased sampling of Poisson point processes and excursions. Probab. Theory Related Fields **92**, 21–39 (1992)
13. Pitman, J., Yor, M.: The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab. **25**, 855–900 (1997)
14. Sampson, G.: Empirical linguistic. Bloomsbury Academic (2002)
15. Thompson, W.K.: Sampling rare or elusive species. Island Press (2004)