MR4632144 62G05 62C20

Favaro, Stefano (I-TRIN-ECS); **Naulet, Zacharie** (F-UPS7-LM)

Near-optimal estimation of the unseen under regularly varying tail populations. (English. English summary)

Bernoulli **29** (2023), *no.* 4, 3423–3442.

Estimating the number of unseen species is a familiar problem in ecological studies. Suppose that for $n, m \geq 1$, $(n+m)$ samples are randomly drawn from an unknown multinomial distribution p, the mass points of which signify a set of species. Also suppose that only the first n elements of those $(n+m)$ samples $\mathbf{X}_{n+m} = (X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+m})$ can be observed. Then, the unseen-species problem boils down to estimating the number $U = U_{n,m} = U_{n,m}(\mathbf{X}_{n+m})$ of hitherto unseen species that will be observed if the remaining m elements are additionally collected. In this article, the unseen-species problem for a sufficiently large $\lambda = m/n > 0$ is studied.

The unseen-species problem of this type requires a model assumption on the underlying p. Based on an analogy of the unseen-species problem to estimating probabilities of rare events in extreme value theory, the authors assume that the tail part of p obeys a power-law distribution. Then they propose to estimate the tail index $\alpha \in (0,1)$ and the number of unseen species $U = U_{n,\lambda n}$ sequentially.

Statistical properties of the maximum likelihood estimator for the tail index $\widehat{\alpha}_n$ and an estimator of the number of unseen species $\widehat{U}_{n,\lambda n} = \widehat{U}_{n,\lambda n}(\widehat{\alpha}_n)$ are explored. The tail index estimator $\widehat{\alpha}_n$ is consistent for α with the rate $n^{-\alpha/2}\sqrt{\log n}$, and $\widehat{U}_{n,\lambda n}(\widehat{\alpha}_n)$ is also consistent for $U_{n,\lambda n}$ all the way up to $\log \lambda \asymp n^{-\alpha/2}/\sqrt{\log n}$. Moreover, these estimators are shown to be minimax near optimal up to a power of $\log n$ factor. It is also demonstrated that estimating $U_{n,\lambda n}$ is harder than α for a sufficiently large λ in the sense that the range $\log \lambda \asymp n^{-\alpha/2}/\sqrt{\log n}$ is the best possible for consistent estimation of $U_{n,\lambda n}$. *Masayuki Hirukawa*

[References]

1. Anevski, D., Gill, R.D. and Zohren, S. (2017). Estimating a probability mass function with unknown labels. *Ann. Statist.* **45** 2708–2735. MR3737907 https://doi.org/10.1214/17-AOS1542 MR3737907
2. Ayed, F., Battiston, M., Camerlenghi, F. and Favaro, S. (2021). On consistent and rate optimal estimation of the missing mass. *Ann. Inst. Henri Poincaré Probab. Stat.* **57** 1476–1494. MR4291456 https://doi.org/10.1214/20-aihp1126 MR4291456
3. Balabdaoui, F. and Kulagina, Y. (2020). Completely monotone distributions: Mixing, approximation and estimation of number of species. *Comput. Statist. Data Anal.* **150** 107014. MR4101998 https://doi.org/10.1016/j.csda.2020.107014 MR4101998
4. Balocchi, C., Favaro, S. and Naulet, Z. (2021). Bayesian nonparametric inference for "species-sampling problems". Preprint. Available at arXiv:2203.06076.
5. Barabási, A.L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* **435** 227.
6. Ben-Hamou, A., Boucheron, S. and Ohannessian, M.I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli* **23** 249–287. MR3556773 https://doi.org/10.3150/15-BEJ743 MR3556773
7. Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1989). *Regular Variation. Encyclopedia of Mathematics and Its Applications* **27**. Cambridge: Cambridge Univ. Press. MR1015093 MR0898871
8. Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: A review. *J. Amer. Statist. Assoc.* **88** 364–373.

9. Camerlenghi, F., Favaro, S., Naulet, Z. and Panero, F. (2021). Optimal disclosure risk assessment. *Ann. Statist.* **49** 723–744. MR4255105 https://doi.org/10.1214/20-aos1975 MR4255105
10. Cancho, R.F. and Solé, R.V. (2003). Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci. USA* **100** 788–791. MR1951090 https://doi.org/10.1073/pnas.0335980100 MR1951090
11. Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.* **87** 210–217. MR1158639 MR1158639
12. Chee, C.-S. and Wang, Y. (2016). Nonparametric estimation of species richness using discrete k-monotone distributions. *Comput. Statist. Data Anal.* **93** 107–118. MR3406199 https://doi.org/10.1016/j.csda.2014.10.021 MR3406199
13. Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Rev.* **51** 661–703. MR2563829 https://doi.org/10.1137/070710111 MR2563829
14. Daley, T. and Smith, A.D. (2013). Predicting the molecular complexity of sequencing libraries. *Nat. Methods* **10** 325–327. https://doi.org/10.1038/nmeth.2375
15. de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction. Springer Series in Operations Research and Financial Engineering*. New York: Springer. MR2234156 https://doi.org/10.1007/0-387-34471-3 MR2234156
16. Drees, H. (1998). Optimal rates of convergence for estimates of the extreme value index. *Ann. Statist.* **26** 434–448. MR1608148 https://doi.org/10.1214/aos/1030563992 MR1608148
17. Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.
18. Favaro, S. and Naulet, Z. (2023). Supplement to "Near-optimal estimation of the unseen under regularly varying tail populations." https://doi.org/10.3150/23-BEJ1589SUPP
19. Feller, W. (1971). *An Introduction to Probability Theory and Its Applications. Vol. II*, 2nd ed. New York: Wiley. MR0270403 MR0270403
20. Fisher, R.A., Corbet, A.S. and Williams, C.B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12** 42–58.
21. Formentin, M., Lovison, A., Maritan, A. and Zanzotto, G. (2014). Hidden scaling patterns and universality in written communication. *Phys. Rev. E* **90** 012817.
22. Gao, Z., Tseng, C.H., Pei, Z. and Blaser, M.J. (2007). Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl. Acad. Sci. USA* **104** 2927–2932.
23. Giguelay, J. and Huet, S. (2018). Testing k-monotonicity of a discrete distribution. Application to the estimation of the number of classes in a population. *Comput. Statist. Data Anal.* **127** 96–115. MR3820312 https://doi.org/10.1016/j.csda.2018.02.006 MR3820312
24. Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probab. Surv.* **4** 146–171. MR2318403 https://doi.org/10.1214/07-PS092 MR2318403
25. Good, I.J. and Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039 https://doi.org/10.1093/biomet/43.1-2.45 MR0077039
26. Hall, P. and Welsh, A.H. (1984). Best attainable rates of convergence for estimates of parameters of regular variation. *Ann. Statist.* **12** 1079–1084. MR0751294 https://doi.org/10.1214/aos/1176346723 MR0751294
27. Hall, P. and Welsh, A.H. (1985). Adaptive estimates of parameters of regular variation. *Ann. Statist.* **13** 331–341. MR0773171 https://doi.org/10.1214/aos/

1176346596 MR0773171

28. Hao, Y. and Li, P. (2020). Optimal prediction of the number of unseen species with multiplicity. In *Advances in Neural Information Processing Systems*.
29. Harald, B.R. (2001). *Word Frequency Distributions*. Berlin: Springer.
30. Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174. MR0378204 MR0378204
31. Huberman, B.A. and Adamic, L.A. (1999). Internet: Growth dynamics of the World-Wide Web. *Nature* **401** 131.
32. Ionita-Laza, I., Lange, C. and Laird, N.M. (2009). Estimating the number of unseen variants in the human genome. *Proc. Natl. Acad. Sci. USA* **106** 5008–5013. MR2496537 https://doi.org/10.1073/pnas.0807815106 MR2496537
33. Jana, S., Polyanskiy, Y. and Wu, Y. (2020). Extrapolating the profile of a finite population. In *Conference on Learning Theory*.
34. Jiao, J., Venkat, K., Han, Y. and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **61** 2835–2885. MR3342309 https://doi.org/10.1109/TIT.2015.2412945 MR3342309
35. Kingman, J.F.C. (1993). *Poisson Processes. Oxford Studies in Probability* **3**. Oxford University Press, New York: The Clarendon Press. MR1207584 MR1207584
36. Kingman, J.F.C., Taylor, S.J., Hawkes, A.G., Walker, A.M., Cox, D.R., Smith, A.F.M., Hill, B.M., Burville, P.J. and Leonard, T. (1975). Random discrete distribution. *J. Roy. Statist. Soc. Ser. B* **37** 1–22. MR0368264 MR0368264
37. Kroes, I., Lepp, P.W. and Relman, D.A. (1999). Bacterial diversity within the human subgingival crevice. *Proc. Natl. Acad. Sci. USA* **96** 14547–14552.
38. Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (N.L. Hjort, C.C. Holmes, P. Müller and S.G. Walker, eds.). *Camb. Ser. Stat. Probab. Math.* **28** 80–136. Cambridge: Cambridge Univ. Press. MR2730661 MR2730661
39. Monechi, B., Ruiz-Serrano, Ã., Tria, F. and Loreto, V. (2017). Waves of novelties in the expansion into the adjacent possible. *PLoS ONE* **12** e0179303. https://doi.org/10.1371/journal.pone.0179303
40. Mossel, E. and Ohannessian, M.I. (2019). On the impossibility of learning the missing mass. *Entropy* **21**. https://doi.org/10.3390/e21010028
41. Muchnik, L., Pei, S., Parra, L.C., Reis, S.D.S., Andrade, J.S., Havlin, S. and Makse, H.A. (2013). Origins of power-law degree distribution in the heterogeneity of human activity in social networks. *Nature Scientific Reports* **3** 1783.
42. Ohannessian, M.I. and Dahleh, M.A. (2012). Rare probability estimation under regularly varying heavy tails. *Journal of Machine Learning Reseach* **23** 1–24.
43. Orlitsky, A., Suresh, A.T. and Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113** 13283–13288. MR3582444 https://doi.org/10.1073/pnas.1607774113 MR3582444
44. Pitman, J. (2003). Poisson-Kingman partitions. In *Statistics and Science: A Festschrift for Terry Speed* (D.R. Goldstein, ed.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **40** 1–34. Beachwood, OH: IMS. MR2004330 https://doi.org/10.1214/lnms/1215091133 MR2004330
45. Pitman, J. (2006). *Combinatorial Stochastic Processes. Lecture Notes in Math.* **1875**. Berlin: Springer. MR2245368 MR2245368
46. Polyanskiy, Y. and We, Y. (2020). Dualizing Le Cam's method for functional estimation, with applications to estimating the unseens. Preprint. Available at arXiv:1902.05616.
47. Rybski, D., Buldyrev, S.V., Havlin, S., Liljeros, F. and Makse, H.A. (2016). Scaling laws of human interaction activity. *Proc. Natl. Acad. Sci. USA* **106** 12640.

48. Tria, F., Loreto, V., Servedio, V.D.P. and Strogatz, S.H. (2014). The dynamics of correlated novelties. *Nature Scientific Reports* **4** 5890.
49. Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794 MR2724359
50. Valiant, G. and Valiant, P. (2013). Estimating the unseen: Improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems* **27** 2157–2165. MR2932019
51. Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory* **62** 3702–3720. MR3506758 https://doi.org/10.1109/TIT.2016.2548468 MR3506758
52. Wu, Y. and Yang, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.* **47** 857–883. MR3909953 https://doi.org/10.1214/17-AOS1665 MR3909953
53. Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley.

Note: This list reflects references listed in the original paper as accurately as possible with no attempt to correct errors.