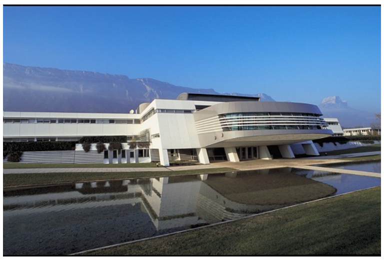
Bayesian nonparametric discovery probabilities

Julyan Arbel, Chargé de recherche, Mistis team
Inria Grenoble Rhône-Alpes
http://www.julyanarbel.com/
julyan.arbel@inria.fr
06 43 28 75 03



Stipend: 500 Euros per month.

Skills required: Minimal prerequisites are a mastering of Probability theory and Statistics. The thesis could be followed by a PhD in statistics.

Context & Objectives

The longstanding problem of discovery probabilities dates back to World War II with no less than Alan Turing codebreaking the Axis forces Enigma coding machine at Bletchley Park. The problem can be simply sketched as follows: an experimenter samples units, say animals, from a population and records their types, say species. Questions in the *discovery probability problem* include:

What is the probability that the next sampled animal coincides with a species already observed a given number of times? or that it is a newly discovered species?

Applications are not limited to ecology but span bioinformatics, genetics, machine learning, multi-armed bandits, and so on...

Classical and highly popular estimators for discovery probabilities were proposed by Good and Turing (GT) (Good, 1953), however they suffer from some inconsistencies. Known results about confidence intervals for GT include McAllester and Schapire (2000); McAllester and Ortiz (2003); Berend and Kontorovich (2013). Bayesian nonparametric (BNP) estimators were first investigated by Lijoi et al. (2007). Interestingly, these BNP estimators take on the form of a convex combination of the GT estimators and some a priori smoothing quantity. Arbel et al. (2016) recently derived the exact posterior distribution of discovery probabilities, leading to credible intervals for BNP estimators.

The objective of the thesis is to *compare performances of BNP and of GT estimators*, both on simulated data and on real data. Specifically, we will study (classical) confidence intervals of GT estimators and (Bayesian) credible intervals of BNP estimators, both for finite sample size and for sample size growing to infinity.

References

Arbel, J., Favaro, S., Nipoti, B., and Teh, Y. W. (2016). Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica, in press.*

Berend, D. and Kontorovich, A. (2013). On the concentration of the missing mass. *Electronic Communications in Probability*, 18(3):1–7.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786.

McAllester, D. and Ortiz, L. (2003). Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4(Oct):895–911.

McAllester, D. A. and Schapire, R. E. (2000). On the Convergence Rate of Good-Turing Estimators. In *COLT*, pages 1–6.