Nonparametric priors for finite unknown cardinalities of sampling spaces

Philipp Benner
Max Planck Institute
for Mathematics in the Sciences
Inselstrasse 22
04103 Leipzig, Germany

Pierre-Yves Bourguignon
Max Planck Institute
for Mathematics in the Sciences
Inselstrasse 22
04103 Leipzig, Germany

Stephan Poppe
Max Planck Institute
for Mathematics in the Sciences
Inselstrasse 22
04103 Leipzig, Germany

Abstract

The Dirichlet process and its popular generalization, the Pitman-Yor process, are often considered as priors in the context of multinomial sampling. They permit inferences on discrete sampling spaces of infinite cardinality. In fact, they a priori assume that there are infinitely many di erent outcomes to be observed, but rule out that there are only finitely many things. This, for instance, limits their usage in species sampling problems, where among other things we have to infer an unknown, but finite cardinality. Following first principles of inductive inference, such as exchangeability, we characterize a new class of nonparametric priors that extends the Pitman-Yor process, and permits the elicitation of a posterior distribution for the cardinality of the sample space, as well as the derivation of non-degenerate probabilities for any number of novel outcomes to appear given a finite sample.

Over the course of the last decade, the Dirichlet process, and even more its generalization into the Pitman-Yor process, have met an indisputable success in the field of discrete Bayesian nonparametrics. One obvious explanation of this success lies in the analytical tractability of the inference procedures they support in settings where there is an infinite sample space to be considered. This benefit is certainly best exemplified by the multiple uses of these processes as priors in clustering problems where the number of clusters is unknown, although applications where these processes drive the generation of the observations themselves have also gained a wide popularity.

While meant to tackle very di erent problems, these applications share one common feature: only properties pertaining to finite samples are queried. In the context of clustering, one is indeed typically interested in the posterior distribution of the random partition of the observations into clusters; in the species sampling problem, the focus is rather on predicting the next outcome in a manner consistent with the possibility that an outcome unobserved so far might appear. Nonetheless, the inference of universal properties of the sample space, such as *how many clusters would one encounter if one were to keep sampling indefinitely many times*, or *how many species are there in total*, is usually avoided. This bias does not really follow from a lack of scientific interest in the answers to these questions, but rather from the full support provided by those processes to the universal hypothesis that infinitely many di erent outcomes would ultimately occur. In other words, any hypothesis set-

ting an upper bound on the cardinality lies outside the support of the Pitman-Yor process as a prior distribution, and therefore cannot receive a positive support even conditional on large (yet finite) datasets.

Another feature of the Pitman-Yor process relates to the prediction rules it gives rise to: the probability that the next outcome will coincide with an already seen one, given past observations, does not depend on the number of outcomes these observations revealed. The recognition of this fact is often the basis for criticism of this prior (Lijoi and Pruenster [2010]), and the relationship between this property and the rate of discovery of new outcomes as data come in has been the matter of in-depth investigations. The above mentioned limitation arguably also relates to this property.

While the Pitman-Yor process yields a predictive probability of a new outcome to appear that depends linearly on the number of unique outcomes seen so far, whereas the Dirichlet process predicts the very same event in a manner insensitive to this piece of information, we argue here that another leap forward is needed in order to enable the inference of the cardinality of an observed sequence of observations -or the number of clusters in a clustering problem.

Inspired by the inductive characterization of the Pitman-Yor Process by Zabell [2005], we propose here a framework that borrows much from previous results by Hintikka, Niiniluoto and Kuipers (see Hintikka and Niiniluoto [1980] and Kuipers [1978]), who made use of a weaker version of the su cientness postulate for extending Carnap's continuum (Carnap [1952]), which characterizes Dirichlet priors. The key relaxation here is to allow predictive probabilities to depend also on the number of di erent outcomes seen so far. Combining their continuum of inductive methods with the powerful results pertaining to partially exchangeable random partitions, see Pitman [1995], we obtain a continuum of inductive methods based on the Hintikka-Niiniluoto-Kuipers' system. It is noteworthy that this is merely an implementation of ideas already present in Kuipers' program. This new continuum comes along with an appropriate set of prior parameters. The continuum's de Finetti's representation involves an adequate multinomial model that appears in Pitman [1995]. A particular choice of an inductive system from this continuum is further investigated, yielding a tractable inference calculus for which closed-form expressions and a straightforward stochastic representation are given.

References

R. Carnap. *The continuum of inductive methods*. University of Chicago Press, 1952.

J. Hintikka and I. Niiniluoto. An axiomatic foundation for the logic of inductive generalization. 1980.

Theo A.F. Kuipers. *Studies in inductive probability and rational expectation*. D. Reidel Publishing Company, 1978.

Antonio Lijoi and Igor Pruenster. Models beyond the dirichlet process. In Nils Lid Hjort, Chris Holmes, Peter Mueller, and Stephen G. Walker, editors, *Bayesian Nonparametrics*, Cambridge Series in Statistical and Probability Mathematics, pages 80–136. Cambridge University Press, 2010.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158, 1995.

S.L. Zabell. *Symmetry and its discontents. Essays on the history of inductive probability*. Cambridge Univ. Press, 2005.