Statistical Science
2016, Vol. 31, No. 1, 34–36
DOI: 10.1214/15-STS540
Main article DOI: 10.1214/15-STS529

Bayesian Nonparametric Modeling and the Ubiquitous Ewens Sampling Formula

Yee Whye Teh

I would like to thank Harry Crane for a most enlightening review of the many ways and guises in which the Ewens sampling formula pops up throughout statistics and mathematics. Given the simplicity and the almost inevitability of Ewens' sampling formula when working with distributions over partitions, one could say that it plays a similar role for random partitions as the normal distribution plays for random real-valued variables. And just as the normal distribution plays an important role as a core building block for more complex models, for example, hierarchical Bayesian models or graphical models, Ewens' sampling formula and the associated Chinese restaurant process distribution over set partitions and Dirichlet process distribution over probability measures play increasingly important roles as building blocks of more complex Bayesian nonparametric models. Crane has noted, and I agree, that this is "one of the most active areas of statistical research," whose "overwhelming activity forbids any possibility of a satisfactory survey of the topic and promises to quickly outdate the contents of the present section." In this discussion I will attempt to present an (already outdated) overview of the use of Ewens' sampling formula in Bayesian nonparametrics, specifically focusing on the many creative ways the community has built more complex models out of these simpler building blocks. Much of the work is motivated by recent trends toward using the analysis of "Big Data" sets to derive scientific understanding and drive technological progress. Such modern data sets are often not just tall, they are also wide, and not just tall and wide, but also complex and structured, and it is important to model the nontrivial dependencies hidden behind the data.

Good introductions to Bayesian nonparametrics can be found in the collection edited by Hjort et al. [14] and the book by Ghosh and Ramamoorthi [13], while more recent works can be found in the IEEE TPAMI special issue [1] and a number of other forthcoming special issues. Finally, shorter introductions and tutorials for less mathematically inclined readers can be found in [11, 12, 22].

Yee Whye Teh is Professor of Statistical Machine Learning, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, United Kingdom (e-mail: y.w.teh@stats.ox.ac.uk).

1. NONPARAMETRIC MIXTURE MODELS AND CLUSTERING

One of the most popular uses of Ewens' sampling formula in Bayesian nonparametrics is via the Chinese restaurant process (CRP), a distribution over set partitions described in Section 4, for mixture modeling and clustering. Consider a data set of size n modeled as observations of exchangeable random variables $Y_1, \ldots, Y_n$. Assuming that these come from a number of heterogenous sources or clusters, we can model the assignment of the observations to different sources using a partition Π of the index set. If the number of sources is unknown and taking a Bayesian formalism, a sensible prior should then place positive mass over all possible partitions. A simple example of such a prior is given by the Chinese restaurant process $\mathrm{CRP}([n], \theta)$, leading to the following model:

$$\Pi \sim \mathrm{CRP}\big([n], \theta\big), \quad Y_i | \Pi \overset{\text{ind.}}{\sim} F\big(X_c^*\big), \quad X_c^* \overset{\text{i.i.d.}}{\sim} H,$$

where $i \in c \in \Pi$, X_c^* is the unknown parameters describing cluster c in Π, H is its prior, and $\mathrm{CRP}([n], \theta)$ denotes the CRP distribution over partitions of the set $[n] = \{1, \ldots, n\}$ with parameter θ. Such a model was first proposed by Lo [17] for density estimation problems, and rediscovered for clustering in machine learning [19, 23]. It is now commonly known as the Dirichlet process mixture model, so named as the de Finetti measure underlying the CRP mixture is the Dirichlet process $\mathrm{DP}(\theta, H)$.

2. NESTED PARTITIONS AND TREES

In certain applications, for example, phylogenetics and unsupervised categorization learning, it is of interest to model data as arising from a nested collection of clusters. For example, a beagle is a dog is an animal is a living organism. These can be modeled as nested partitions, for example, $\{\{\{1, 4\}, \{5\}\}, \{\{2, 6\}, \{3\}\}, \{\{7\}\}\}$ is