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 In this discussion I will attempt to present an (already outdated) overview of the use of Ewens' sampling formula in Bayesian nonparametrics, specifically focusing on the many creative ways the community has built more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].

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 $Y_1, \dots, Y_n$. 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:

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

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 $[n] = \{1, \dots, n\}$ 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

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).

a two-level nested partition of $\{1, \dots, 7\}$. Distributions over nested partitions can be constructed from CRPs in two different ways: by subpartitioning or fragmenting the clusters of a partition recursively in a top-down fashion or by coagulating the clusters of a partition recursively in a bottom-up fashion.

A fragmentation process starts with the trivial partition with just one cluster and recursively fragments clusters using independent CRPs L times to construct an L level nested partition. Such a fragmentation process was called a nested CRP in [5] who explored it as a model for unsupervised learning of topic hierarchies in text analysis. Conversely, a coagulation process can start with another trivial partition with all items in their own clusters, and recursively coagulate clusters in the following way: If π is a partition, let κ be a partition of the *clusters* in π, say, drawn from $\mathrm{CRP}(\pi, \theta)$. The coagulation of π by κ is then $\pi' = \{\bigcup_{c \in \gamma} c : \gamma \in \kappa\}$, where the clusters in π belonging to the same cluster in κ are merged. This coagulation process can be shown to be the dual genealogical process of the hierarchical Dirichlet process [25] (described later), where it is a simple case of the Chinese restaurant franchise.

Fragmentation and coagulation processes are more conveniently represented mathematically as Markov chains on set partitions, with fragmentations being sequences of partition refinements (see Section 3.5 of main article), while coagulations are coarsenings [4]. Viewed in this way, one can also ask for continuous time limits of the Markov chains associated with the nested CRPs and the Chinese restaurant franchise, leading to Dirichlet diffusion trees [20] and Kingman's coalescents (Section 2.4) respectively. It is also possible to construct partition-valued Markov chains with both fragmentations and coagulations in operation. The mathematical properties of such processes were studied in [3], and they were applied to haplotype modeling and genetic imputation in [10, 24].

3. HIERARCHICAL BAYESIAN NONPARAMETRIC MODELS

A common theme across both frequentist and Bayesian statistics is when data are separated into groups and it is important to model groups individually while sharing statistical strength across groups to provide more fine-grained control over model flexibility. In Bayesian statistics this is achieved using hierarchical Bayesian models where each group has an associated random parameter with a common prior distribution across groups parameterized by a random hyperparameter. The randomness of the hyperparameter induces the sharing of statistical strength across groups.

In a Bayesian nonparametric setting, where the random parameter is typically an infinite-dimensional stochastic process, control over model flexibility is arguably even more important than in parametric models. For example, if each group is modeled with a Dirichlet process mixture, with $G_j \sim \mathrm{DP}(\theta, G_0)$ for group j, one can place a hierarchical DP prior on the base distribution, $G_0 \sim \mathrm{DP}(\theta_0, H)$ [25], which induces sharing of the mixture components across groups. Such hierarchical constructions also arise naturally elsewhere in Bayesian nonparametrics, for example, Gaussian processes for regression [26] and beta processes/Indian buffet processes for feature allocations [9, 27].

4. DEPENDENT AND RELATIONAL MODELS

Hierarchical models effectively assume exchangeability among groups and induce relatively simple forms of statistical strength sharing across groups. This can be relaxed to various forms of partial exchangeability. For example, if there are group level covariates, or spatial or temporal structure, then dependent models reflecting this structure may be appropriate. There are two levels at which general dependencies can be induced. At the random measures level, for each covariate value t we introduce a random measure G_t and work with the measure-valued stochastic process (G_t). When each G_t is a DP, such dependent DPs were first explored by MacEachern [18]. At the random partitions level, one instead works with partition-valued stochastic processes, for example, [6, 7]. A significant number of constructions have been provided in the literature and reviewed in [8].

The random set partitions associated with the Ewens sampling formula have also been used in modeling relational data such as social networks and collaborative filtering. These are data where observations (e.g., of links or friendships) are associated with relations between two or more objects, rather than with objects themselves (although there can be object-level covariates). In the infinite relational model [16, 28], objects are partitioned into clusters via the CRP, and observed relations between objects are mediated by the clusters that they belong to. For relational data, de Finetti's theory of exchangeability is generalized to relational exchangeability by Aldous and Hoover [2, 15]; see [21] for an introduction.

5. SUMMARY

The Ewens sampling formula and the associated distributions over partitions, set partitions and probability measures have very many mathematically elegant properties, which have been well explored in the literature and well reviewed in the present paper. With a good understanding of such distributions and in a data-rich world, the Bayesian nonparametrics community is now engaged in the practical uses of Ewens' sampling formula for modeling more complex phenomena. Important approaches have included covariate-dependence, hierarchical Bayesian models, constructions of nested partitions and trees, and applications to non-i.i.d. settings like relational and network data.

ACKNOWLEDGMENTS

Supported in part by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC Grant agreement no. 617071.

REFERENCES

[1] ADAMS, R. P., FOX, E. B., SUDDERTH, E. B. and TEH, Y. W. (2015). Guest editors' introduction to the special issue on Bayesian nonparametrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[2] ALDOUS, D. J. (1985). Exchangeability and related topics. In *École d'été de Probabilités de Saint-Flour, XIII*—1983. *Lecture Notes in Math*. **1117** 1–198. Springer, Berlin. MR0883646

[3] BERESTYCKI, J. (2004). Exchangeable fragmentation-coalescence processes and their equilibrium measures. *Electron. J. Probab.* **9** 770–824 (electronic). MR2110018

[4] BERTOIN, J. (2006). *Random Fragmentation and Coagulation Processes*. *Cambridge Studies in Advanced Mathematics* **102**. Cambridge Univ. Press, Cambridge. MR2253162

[5] BLEI, D. M., GRIFFITHS, T. L. and JORDAN, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM* **57** Art. 7, 30. MR2606082

[6] CARON, F., DAVY, M. and DOUCET, A. (2007). Generalized Polya urn for time-varying Dirichlet process mixtures. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence* **23**.

[7] DUAN, J. A., GUINDANI, M. and GELFAND, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika* **94** 809–825. MR2416794

[8] DUNSON, D. B. (2010). Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics* 223–273. Cambridge Univ. Press, Cambridge. MR2730665

[9] ECK, D., BENGIO, Y. and COURVILLE, A. C. (2009). An infinite factor model hierarchy via a noisy-or mechanism. In *Advances in Neural Information Processing Systems* 405–413.

[10] ELLIOTT, L. and TEH, Y. W. (2012). Scalable imputation of genetic data with a discrete fragmentation–coagulation process. In *Advances in Neural Information Processing Systems*.

[11] GERSHMAN, S. J. and BLEI, D. M. (2012). A tutorial on Bayesian nonparametric models. *J. Math. Psych.* **56** 1–12. MR2903470

[12] GHAHRAMANI, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **371** 20110553, 20. MR3005667

[13] GHOSH, J. K. and RAMAMOORTHI, R. V. (2003). *Bayesian Nonparametrics*. Springer, New York. MR1992245

[14] HJORT, N., HOLMES, C., MÜLLER, P. and WALKER, S., eds. (2010). *Bayesian Nonparametrics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **28**. Cambridge Univ. Press, Cambridge. MR2722987

[15] HOOVER, D. (1979). Relations on probability spaces and arrays of random variables. Technical report, Princeton, NJ.

[16] KEMP, C., TENENBAUM, J. B., GRIFFITHS, T. L., YAMADA, T. and UEDA, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the AAAI Conference on Artificial Intelligence* **21**.

[17] LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519

[18] MACEACHERN, S. (1999). Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*. Amer. Statist. Assoc., Alexandria, VA.

[19] NEAL, R. M. (1992). Bayesian mixture modeling. In *Proceedings of the Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis* **11** 197–211.

[20] NEAL, R. M. (2001). Defining priors for distributions using Dirichlet diffusion trees. Technical Report 0104, Dept. Statistics, Univ. Toronto.

[21] ORBANZ, P. and ROY, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence Special Issue on Bayesian Nonparametrics*.

[22] ORBANZ, P. and TEH, Y. W. (2010). Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer, Berlin.

[23] RASMUSSEN, C. E. (2000). The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems* **12**.

[24] TEH, Y. W., BLUNDELL, C. and ELLIOTT, L. T. (2011). Modelling genetic variations with fragmentation–coagulation processes. In *Advances in Neural Information Processing Systems*.

[25] TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480

[26] TEH, Y. W., SEEGER, M. and JORDAN, M. I. (2005). Semiparametric latent factor models. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics* **10**.

[27] THIBAUX, R. and JORDAN, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics* **11** 564–571.

[28] XU, Z., TRESP, V., YU, K. and KRIEGEL, H.-P. (2006). Infinite hidden relational models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence* **22**.