Hierarchical Models, Nested Models and Completely Random Measures

Michael I. Jordan
University of California, Berkeley
Berkeley, CA 94720

May 1, 2013

1 Introduction

Statistics has both optimistic and pessimistic faces, with the Bayesian perspective often associated with the former and the frequentist perspective with the latter, but with foundational thinkers such as Jim Berger reminding us that statistics is fundamentally a Janus-like creature with two faces. In creating one field out of two perspectives, one of the unifying ideas emphasized by Berger and others is the Bayesian hierarchy, a modeling framework that simultaneously allows complex models to be created and tames their behavior.

Another general tool for creating complex models while controlling their complexity is by nesting simplified models inside of more complex models, an appeal to the principle of "divide-and-conquer." An example is the classical finite mixture model, where each data point is modeled as arising from a single mixture component. Note that this appeal to divide-and-conquer is quite different from the recursive principle underlying hierarchical modeling—the latter strategy provides a way to share statistical strength among components while the former strategy tends to isolate components. Of course, many complex models involve a blend of these strategies.

If the need to exploit hierarchical and nested structures is compelling in parametric models, it is still more compelling in Bayesian nonparametrics, where the growth in numbers of degrees of freedom creates significant challenges in controlling model complexity. The basic idea of Bayesian nonparametrics is to replace classical finite-dimensional prior distributions with general stochastic processes, thereby allowing an open-ended number of degrees of freedom in a model. The framework expresses an essential optimism—only an optimist could hope to fit a model involving an infinite number of degrees of freedom based on finite data. But it also expresses the pessimism that simplified parametric models may be inadequate to

capture many real-world phenomena, particularly in the setting of large data sets in which increasingly subtle aspects of those phenomena may be revealed. From either perspective, care needs to be taken to exploit and manage the large number of degrees of freedom available within a nonparametric model.

In this article we discuss hierarchical and nested modeling concepts within the framework of Bayesian nonparametrics. To keep the discussion focused, we restrict ourselves to a special class of stochastic processes known as "completely random measures." These random measures have the simplifying property that they assign independent random mass to disjoint regions of a probability space. This property turns out to imply that these measures are discrete (up to a deterministic component that is of limited value for Bayesian modeling). While the discreteness is limiting for some applications, it also has some significant virtues. In particular it provides a natural tool for focusing on structural aspects of models, where the effects of hierarchy and nesting have relatively simple interpretations.

2 Completely Random Measures

Letting Ω denote a measurable space endowed with a sigma algebra $\mathcal{A}$, a *random measure* G is a stochastic process whose index set is $\mathcal{A}$. That is, $G(A)$ is a random variable for each set A in the sigma algebra. A *completely random measure* G is defined by the additional requirement that whenever A_1 and A_2 are disjoint sets in $\mathcal{A}$, the corresponding random variables $G(A_1)$ and $G(A_2)$ are independent. This idea generalizes the notion of "independent increments processes" that is familiar in the special case in which Ω is the real line.

Kingman (1967) presented a way to construct completely random measures based on the nonhomogeneous Poisson process. This construction has significant consequences for Bayesian modeling and computation; in particular, it allows connections to be made to the exponential family and to conjugacy. The construction is as follows (see Figure 1 for a graphical depiction). Consider the product space $\Omega \otimes \Re$, and place a sigma-finite product measure η on this space. Treating η as the rate measure for a nonhomogeneous Poisson process, draw a sample $\{(\omega_i, p_i)\}$ from this Poisson process. From this sample, form a measure on Ω in the following way:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}.$$

We refer to $\{\omega_i\}$ as the *atoms* of the measure G and $\{p_i\}$ as the *weights*.

Clearly the random measure defined in Eq. (2) is completely random because the Poisson process assigns independent mass to disjoint sets. The interesting fact is that all completely random processes can be obtained this way (up to a deterministic component and a fixed set of atoms).

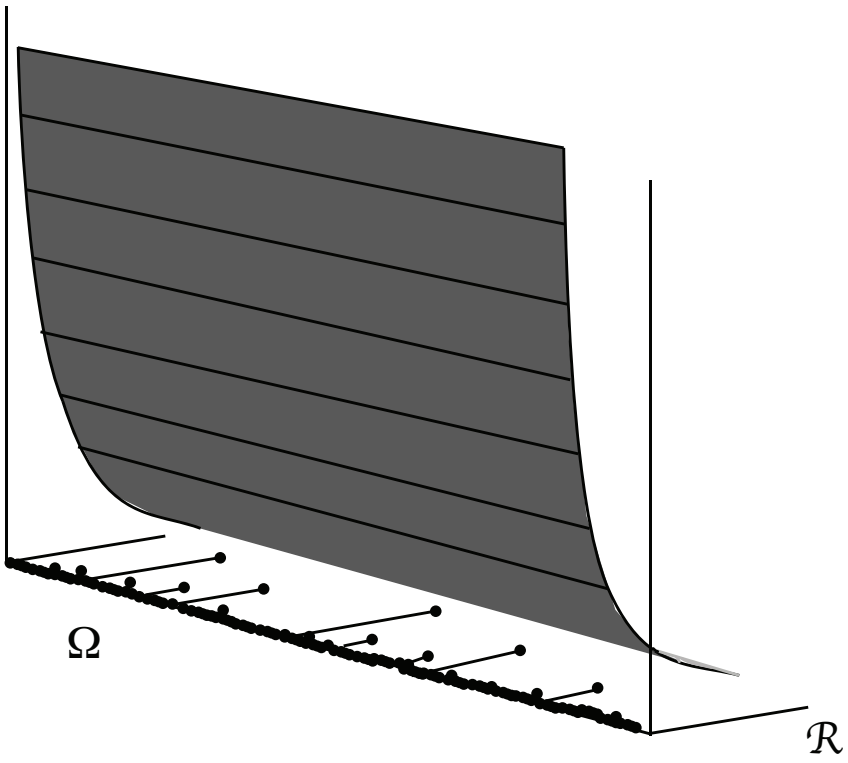

Figure 1: The construction of a completely random measure on Ω from a nonhomogeneous Poisson process on $\Omega \otimes \Re$.

As an example of a completely random measure, define the rate measure η as a product of an arbitrary sigma-finite measure B_0 on Ω and an "improper" beta distribution on $(0, 1)$:

$$\eta(d\omega, dp) = cp^{-1}(1-p)^{c-1} dp\, B_0(d\omega),$$

where $c > 0$. Note that the expression $cp^{-1}(1-p)^{c-1}$ integrates to infinity; this has the consequence that a countably infinite number of points are obtained from the Poisson process. The resulting completely random measure is known as the *beta process*.[1] We denote a draw from the beta process as follows:

$$B \sim \mathrm{BP}(c, B_0),$$

where $c > 0$ is referred to as a *concentration parameter* and where B_0 is the *base measure*. Note that for the beta process the weights $\{p_i\}$ lie in the interval $(0, 1)$. Their sum is finite

[1]For further details on this derivation of the beta process, see Thibaux and Jordan (2007). For an alternative derivation that does not make use of the framework of completely random measures, see Hjort (1990).

(a consequence of Campbell's theorem), with the magnitude of the sum controlled by the concentration parameter c and by $B_0(\Omega)$. The locations of the atoms are determined by B_0.

As a second example, let the rate measure be a product of a base measure G_0 and an improper gamma distribution:

$$\eta(d\omega, dp) = cp^{-1}e^{-cp}dp\, G_0(d\omega).$$

Again the density on p integrates to infinity, yielding a countably infinite number of atoms. The resulting completely random measure is known as the *gamma process*. We write:

$$G \sim \mathrm{GP}(c, G_0)$$

to denote a draw from the gamma proess. Note that the weights $\{p_i\}$ lie in $(0, \infty)$ and their sum is again finite.

It is also of interest to consider random measures that are obtained from completely random measures by normalization. For example, returning to the rate measure defining the gamma process in Eq. (2), let $\{(\omega_i, p_i)\}$ denote the points obtained from the corresponding Poisson process. Form a random probability measure as follows:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i},$$

where $\pi_i = p_i / \sum_{j=1}^{\infty} p_j$. This is the famous *Dirichlet process (DP)* (Ferguson, 1973). We denote a draw from the DP as $G \sim \mathrm{DP}(\alpha_0, H_0)$, where $\alpha_0 = G_0(\Omega)$ and $H_0 = G_0/\alpha_0$. Note that the DP is *not* a completely random measure—for disjoint sets A_1 and A_2, the random variables $G(A_1)$ and $G(A_2)$ are negatively correlated due to the normalization.

3 Marginal Probabilities

At this point it is useful to recall De Finetti's theorem, which states that infinitely exchangeable sequences of random variables are obtained by drawing a random element G and then drawing the elements of the sequence independently conditional on G. Given the ubiquity of this conditional independence motif in Bayesian modeling, it is of interest to ask what kinds of exchangeable sequences are obtained if G is one of the random measures discussed in the previous section.

In the case of the DP, the answer is classical—one obtains the *Pólya urn model* (Blackwell and MacQueen, 1973). This model can be described in terms of the closely related *Chinese restaurant process (CRP)* (Aldous, 1985). Consider a restaurant with an infinite number of tables, listed in some (arbitrary) order. Customers enter the restaurant sequentially. The first customer sits at the first table. Subsequent customers choose a table with probability

proportional to the number of customers already sitting at that table. With probability proportional to a parameter α_0 they start a new table. This defines the CRP as a distribution on partitions of the customers. To define the Pólya urn we augment the model to place a parameter ϕ_k at the kth table, where the $\{\phi_k\}$ are drawn independently from some distribution G_0. All customers sitting at the kth table are assigned the parameter ϕ_k. Letting θ_i denote the parameter assigned to the ith customer, this defines an exchangeable sequence $(\theta_1, \theta_2, \ldots)$.

This connection between the DP and the CRP and Pólya urn model can be understood by noting that the representation of the DP in Eq. (2) is essentially a mixture model with a countably infinite number of components. We can view the CRP as a draw from this mixture model. In particular, let us associate an integer-valued variable W_i to the ith customer as follows:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha_0, G_0) \\ p(W_i = k \,|\, G) &= \pi_k, \quad i = 1, \ldots, n. \end{aligned}$$

In the language of the CRP, the event $\{W_i = k\}$ means that the ith customer sits at the kth table. In essence, the DP defines an infinite collection of random probabilities that, when integrated out according to the De Finetti construction, yield the CRP. The specific rule underlying the CRP—that customers sit at a table proportional to the number of customers already at that table—reflects classical Dirichlet-multinomial conjugacy.

Similar connections can be obtained for priors obtained from completely random measures. Consider in particular the beta process. Here the weights $\{p_i\}$ lie in $(0, 1)$, and thus it is natural to view the beta process as yielding an infinite collection of coin-tossing probabilities. Moreover, given the definition of a completely random measure, we are motivated to toss these coins *independently*. This defines the following hierarchy:

$$\begin{aligned} B &\sim \mathrm{BP}(c, B_0) \\ Z \,|\, B &\sim \mathrm{BeP}(B), \end{aligned}$$

where $Z = \sum_{k=1}^{\infty} z_k \delta_{\omega_k}$ is a completely random measure known as the *Bernoulli process*. The atoms $\{\omega_k\}$ are the same atoms as in B and the weights $\{z_k\}$ are binary values that are equal to one with probability p_k and equal to zero otherwise.

Returning to the De Finetti conditional independence motif, we can draw repeatedly from the Bernoulli process given an underlying draw from the beta process:

$$\begin{aligned} B &\sim \mathrm{BP}(c, B_0) \\ Z_i &\sim \mathrm{BeP}(B), \quad i = 1, \ldots, n. \end{aligned} \tag{1}$$

This defines a binary-valued matrix with n rows and an infinite number of columns. Campbell's theorem can again be invoked to show that we obtain a sparse matrix in which the

number of ones in each row is finite with probability one. Viewing the columns as "latent traits" or "features," this matrix can be viewed as a sparse featural representation of objects, or alternatively as a model in which each object is assigned to a subset of classes. Note that this differs from the Dirichlet process, where each object is assigned to a single class.

It is also possible to define this probability law directly via a sequential process that is known as the *Indian buffet process (IBP)* (Griffiths and Ghahramani, 2006). In the IBP, customers enter a restaurant sequentially and select dishes in the buffet line. Dishes that have been chosen previously by other customers are selected with probability proportional to the number of times they have been selected by the previous customers. Each customer also selects a random number of new dishes according to a Poisson distribution (with decreasing rate). As shown by Thibaux and Jordan (2007), this probability law can be obtained by marginalizing out the beta process in the hierarchy in Eq. (1). Their argument is the analog of the derivation of the CRP from the DP. In particular, as alluded to above, the CRP can be derived as a simple consequence of the fact that a posterior DP is itself a DP (Ferguson, 1973). A similar conjugacy relationship holds for the beta process—a posterior BP is itself a BP (Kim, 1999). The posterior BP contains atoms in its base measure, and these are necessarily present in any subsequent draw from the BP. Indeed, these act like independent coins relative to the rest of the random measure. Posterior updating of the probabilities associated with these coins is classical beta-Bernoulli updating, which is the rule underlying the IBP.

4 Hierarchical Models

Bayesian nonparametric models often incorporate classical finite-dimensional parameters—e.g., location parameters, scale parameters, regression parameters and correlation parameters—and it is common to build hierarchies on these parameters. In this section, however, we wish to consider a more thoroughgoing form of Bayesian hierarchy in which the infinite-dimensional parameters of nonparametric models are also linked via hierarchies. Specifically, we discuss conditional independence hierarchies in which a set of completely random measures, $\{G_1, G_2, \ldots, G_M\}$, are conditionally independent given a base measure G_0, and where G_0 is itself a completely random measure.

To see the value of this kind of construction, let us consider the case of the Dirichlet process, and consider M groups of data, $\{x_{1i}\}$, $\{x_{2i}\}$, and $\{x_{Mi}\}$, where each group is to be modeled as a DP mixture. If we have reason to believe that these groups are related, then we may wish to couple the underlying random probability measures via the following

hierarchical Dirichlet process (HDP) construction (Teh et al., 2006):

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ G_m \,|\, G_0 &\sim \mathrm{DP}(\alpha_0, G_0), \quad m = 1, \ldots, M \\ \theta_{mi} \,|\, G_m &\sim G_m, \quad m = 1, \ldots, M,\ i = 1, \ldots, N_m \\ x_{mi} \,|\, \theta_{mi} &\sim F_{\theta_{mi}}, \quad m = 1, \ldots, M,\ i = 1, \ldots, N_m, \end{aligned} \tag{2}$$

where $\{F_\theta\}$ is a parametric family of distributions. The nature of the coupling that is induced by this hierarchy is easily understood by considering a Chinese restaurant representation. Each random measure G_m corresponds to a Chinese restaurant where the atoms forming that random measure correspond to the tables in the restaurant. Some of these tables tend to have large occupancy—these correspond to the atoms with particularly large weights. All of the customers sitting around a single table can be viewed as belonging to a cluster; this is reflected in the fact that the corresponding parameters θ_{mi} are equal to each other. Now if we have reason to believe that the M groups are related, we might expect that a cluster discovered in group m will be useful in modeling the data in the other groups. To achieve this we need to share atoms not only within groups but also between groups. This is achieved by the specification in Eq. (2): the fact that G_0 is drawn from a DP means that it is atomic, and each G_m re-draws from among these atoms. Thus, atoms are shared among the $\{G_m\}$.

Note that it is also possible to couple the random measures $\{G_m\}$ via a classic parametric hierarchy, but this would not generally achieve the goal of sharing clusters among the groups. For example, suppose that G_0 were a parametric distribution depending on a location parameter μ. Bayesian inference for μ would share statistical strength among the groups by centering their base measure at a common location, but, due to the absolutely continuous nature of G_0, the atoms in G_m would be distinct from those in $G_{m'}$ for $m \neq m'$. That is, none of the θ_{mi} would be equal to $\theta_{m'i}$; there would be no sharing of clusters among groups.

The HDP has been used in a wide variety of applications, including social network analysis (Airoldi et al., 2009), genetic analysis (Xing et al., 2007), computational vision (Sudderth, 2006; Kivinen et al., 2007), natural language parsing (Liang et al., 2009), information retrieval (Cowans, 2004) and music segmentation (Ren et al., 2008).

The sharing of atoms achieved via the hierarchical nature of the HDP is also useful for the beta process and other members of the family of completely random measures. Recall that the beta process can be viewed as providing a featural description of a collection of objects, where the weights $\{p_i\}$ are probabilities of the objects possessing or not possessing a given feature. In the setting of multiple collections of objects, it may be useful to transfer the features discovered in one collection to other collections. As an example of this *hierarchical beta process* construction, Thibaux and Jordan (2007) presented an application to document modeling, where the groups correspond to document corpora. Each document is represented as a binary vector indexed by the vocabulary items, and this probability vector is modeled

as a draw from a Bernoulli process. The sparseness of word occurrences in documents means that it is useful to transfer probabilities between corpora.

Similarly, it is useful to consider hierarchies based on the gamma process, where the data might consist (for example) of counts of items in some open-ended vocabulary and the gamma process encodes the Poisson rates underlying these counts.

5 Nested Models

Hierarchical models provide a way to share atoms among multiple random measures, a useful idea when these measures are viewed as related. It can also be worthwhile, however, to consider the opposite end of the spectrum, where atoms are separated into different, non-interacting groups. From a modeling point of view, this allows complex models to be built of simpler components. There is also a computational argument. When atoms are shared, the inferential problem of computing the posterior distribution becomes a highly-coupled problem in which each data point has an inferential impact on each atom. Such highly-coupled problems can be difficult to solve numerically; in particular, in the MCMC setting the correlations introduced by the coupling can increase the mixing time of MCMC algorithms. The coupling also has an effect on the difficulty of implementing inference algorithms; in particular, it makes it difficult to use divide-and-conquer strategies.

Nesting is a general strategy for building complex models out of simpler components. To illustrate, let us consider a nested version of the Chinese restaurant process (Blei et al., 2010). In the *nested CRP*, the restaurant metaphor is extended to a set of restaurants organized according to a branching structure, where individuals partake of a sequence of dinners as they proceed down a path in the tree. All individuals enter a fixed restaurant at the root node of the tree. They select a table according to the usual CRP rule (i.e., they sit at a table with probability proportional to the number of customers who have previously selected the table). The table also has a card on it giving the address of a restaurant where the customers will eat the following night. This construction recurses, yielding an infinitely-branching tree where each customer follows a particular path down the tree. After n customers have entered the tree, there will be up to n paths selected in the tree, with some paths selected by multiple customers. The depth of the tree can be fixed and finite, or it can be infinite.

The nested CRP defines a sequence of distributions on partitions, one for each level of the tree. To turn this into a random measure, we introduce atoms drawn from a base measure G_0 at the tables in the restaurants (one atom per table). One possible option is to consider a tree of fixed depth and to place atoms only at the tables in the restaurants at the leaves of the tree, but it can also be useful to place atoms throughout the tree. In either case, the construction separates atoms into distinct groups. In particular, having selected a branch at the root node, only the atoms in the clade below that branch are available.

This nested construction can also be expressed using random measures directly. In partic-

ular, consider the following specification of a two-level *nested Dirichlet process (nDP)* (Rodríguez et al., 2008):

$$\begin{aligned} G &\sim \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*} \\ G_k^* &= \sum_{j=1}^{\infty} \pi_{kj} \delta_{\theta_{kj}}, \end{aligned} \tag{3}$$

where the weights $\{\pi_{kj}\}$ and $\{\pi_k^*\}$ are obtained as in Eq. (2). We see that G is a draw from a DP that selects among an infinite set of components $\{G_k^*\}$, where each component is itself a DP. Note that the atoms associated with the lower-level DPs are distinct (assuming a continuous base measure). From the point of view of the nCRP formalism the specification in Eq. (3) corresponds to a two-level tree in which atoms are associated only with the tables at the leaves of the tree. The top-level restaurant implements the choice among the G_k and each G_k corresponds to a restaurant at the second level of the tree. More generally, the nDP can be extended to an arbitrary number of levels, and a K-level nCRP is obtained by integrating out the Dirichlet processes in a K-level nDP.

Rodríguez et al. (2008) discussed an application of the two-level nDP to a problem in health care where the goal is to model an outcome variable for the hospitals in the fifty US states. The intra-state distribution is generally multi-modal and thus it is natural to use DP mixtures for each state. Moreover, there are inter-state similarities as well, and one approach to capturing these similarities is to cluster the states. This is done at the higher level of the nDP by allowing similar states to select the same low-level DP. Note the difference between this approach and an HDP-based approach, where all atoms would be shared among all of the groups; here, atoms are shared only when states fall in the same cluster.

There are also natural applications of infinite-level nested models. Blei et al. (2010) presented an application of the nCRP to the problem of discovering topics in document collections. A topic is defined to be a probability distribution across the words in some vocabulary. The nCRP is augmented to place a topic at each table at every restaurant in the tree. The generation of the words in a document is modeled as follows. The first step is to select a path down the (infinite) tree according to the nCRP. Fixing that path, we repeatedly pick a level in the tree using the GEM ("stick-breaking") distribution and pick a word from the topic distribution at that level on the selected path.[2] Given that nodes at the higher levels in the tree tend to be shared across multiple documents (e.g., the root node is shared across all documents), there is statistical pressure to force topics at higher levels in the tree to concentrate on words that are useful across many documents. Topics at lower levels can focus on more specialized words. Thus the model yields an *abstraction hierarchy*.

[2]The GEM distribution is closely related to the Dirichlet process; the GEM probabilities can be obtained by randomly permuting the weights $\{\pi_k\}$ in the Dirichlet process according to a size-biased permutation.

It is also of interest to consider nesting for random measures other than the DP. In particular, we can define a *nested beta process* in the following way:

$$\begin{aligned} B &\sim \text{BeP}\left(\sum_{k=1}^{\infty} p_k^* \delta_{B_k^*}\right) \\ B_k^* &= \sum_{j=1}^{\infty} p_{kj} \delta_{\theta_{kj}}. \end{aligned}$$

This defines a random measure B that is a collection of atoms, each of which is a beta process. Instead of picking a single path down a tree as in the nDP, this definition makes it possible to pick multiple paths down a tree. This construction is quite natural in applications; indeed, in the setting of document modeling it is a natural generalization of *latent Dirichlet allocation (LDA)* (Blei et al., 2003). LDA is a topic model that allow documents to range over arbitrary collections of topics. In the nCRP model, topics are restricted to lie along a single path of the tree, differing only in level of abstraction but not in thematic content. A model based on the nested BP would allow a document to range over both thematic content and level of abstraction.

Similarly, it is of interest to consider a *nested gamma process* construction, which could be used to select multiple branches at each level of a tree, each of which is associated with a count or a rate.

6 Discussion

We have reviewed some recent developments in Bayesian nonparametrics. Our discussion has focused on completely random measures, a broad class of random measures that have simplifying representational and computational properties. Additional random measures can be obtained by normalization of such measures; in particular, the Dirichlet process can be obtained in this way.

The proliferation of parameters (i.e., atoms) in models based on completely random measures calls for organizational principles to control these models in statistically and computationally sound ways. We have focused on two such principles—hierarchy and nesting. While familiar in the parametric setting, these principles have only recently begun to be exploited fully and explicitly in the nonparametric setting. We anticipate many further developments in this vein. We also note that the theory of Bayesian nonparametrics is in an early stage of development, and in particular we have not yet seen theoretical results in which hierarchy and nesting play an explicit role. Given the important role these concepts play in parametric theory, we expect to see analogous theory emerging in nonparametrics. Finally, we note that hierarchy and nesting are applicable to a wide range of Bayesian nonparametric

models that lie outside of the class of completely random measures; indeed, the Pólya tree is one instance of nesting that is well known in Bayesian nonparametrics.

Acknowledgments

I would like to thank Percy Liang, Kurt Miller, Erik Sudderth and Yee Whye Teh for helpful comments.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2009). Mixed-membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

Aldous, D. (1985). Exchangeability and related topics. In *Ecole d'Eté de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1:353–355.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian inference of topic hierarchies. *Journal of the ACM*, 57.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cowans, P. (2004). Information retrieval using hierarchical Dirichlet processes. In *Proceedings of the Annual International Conference on Research and Development in Information Retrieval*, volume 27, pages 564–565.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Griffiths, T. L. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA. MIT Press.

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294.

Kim, Y. (1999). Nonparametric Bayesian estimators for counting processes. *The Annals of Statistics*, 27:562–588.

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

Kivinen, J., Sudderth, E., and Jordan, M. I. (2007). Learning multiscale representations of natural scenes using Dirichlet processes. In *IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil.

Liang, P., Jordan, M. I., and Klein, D. (2009). Probabilistic grammars and hierarchical Dirichlet processes. In *The Handbook of Applied Bayesian Analysis*, Oxford, UK. Oxford University Press.

Ren, L., Dunson, D., and Carin, L. (2008). The dynamic hierarchical Dirichlet process. In *Proceedings of the International Conference on Machine Learning*, Helsinki, Finland.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103:1131–1154.

Sudderth, E. (2006). *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, Department of Electrical Engineering and Computer Science, MIT.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

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*, volume 11.

Xing, E. P., Jordan, M. I., and Sharan, R. (2007). Bayesian haplotype inference via the Dirichlet process. *Journal of Computational Biology*, 14:267–284.