Journal of Mathematical Psychology 56 (2012) 1–12

Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Tutorial

A tutorial on Bayesian nonparametric models

Samuel J. Gershman [a,*], David M. Blei [b]

[a] Department of Psychology and Princeton Neuroscience Institute, Princeton University, Princeton NJ 08540, USA

[b] Department of Computer Science, Princeton University, Princeton NJ 08540, USA

ARTICLE INFO

Article history:
Received 25 January 2011
Received in revised form
4 August 2011
Available online 1 September 2011

Keywords:
Bayesian methods
Chinese restaurant process
Indian buffet process

ABSTRACT

A key problem in statistical modeling is model selection, that is, how to choose a model at an appropriate level of complexity. This problem appears in many settings, most prominently in choosing the number of clusters in mixture models or the number of factors in factor analysis. In this tutorial, we describe Bayesian nonparametric methods, a class of methods that side-steps this issue by allowing the data to determine the complexity of the model. This tutorial is a high-level introduction to Bayesian nonparametric methods and contains several examples of their application.

Contents

1. Introduction

How many classes should I use in my mixture model? How many factors should I use in factor analysis? These questions regularly exercise scientists as they explore their data. Most scientists address them by first fitting several models, with different numbers of clusters or factors, and then selecting one using model comparison metrics (Claeskens & Hjort, 2008). Model selection metrics usually include two terms. The first term measures how well the model fits the data. The second term, a complexity penalty, favors simpler models (i.e., ones with fewer components or factors).

Bayesian nonparametric (BNP) models provide a different approach to this problem (Hjort, Holmes, Müller, & Walker, 2010). Rather than comparing models that vary in complexity, the BNP approach is to fit a single model that can adapt its complexity to the data. Furthermore, BNP models allow the complexity to grow as more data are observed, such as when using a model to perform prediction. For example, consider the problem of clustering data.

* Correspondence to: Department of Psychology, Princeton University, Princeton NJ 08540, USA.
E-mail addresses: sjgershm@princeton.edu (S.J. Gershman), blei@cs.princeton.edu (D.M. Blei).

0022-2496/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2011.08.004

The traditional mixture modeling approach to clustering requires the number of clusters to be specified in advance of analyzing the data. The Bayesian nonparametric approach estimates how many clusters are needed to model the observed data and allows future data to exhibit previously unseen clusters.[1]

Using BNP models to analyze data follows the blueprint for Bayesian data analysis in general (Gelman, Carlin, Stern, & Rubin, 2004). Each model expresses a *generative process* of the data that includes hidden variables. This process articulates the statistical assumptions that the model makes, and also specifies the joint probability distribution of the hidden and observed random variables. Given an observed data set, data analysis is performed by *posterior inference*, computing the conditional distribution of the hidden variables given the observed data. Loosely, posterior inference is akin to "reversing" the generative process to find the distribution of the hidden structure that likely generated the observed data. What distinguishes Bayesian nonparametric models from other Bayesian models is that the hidden structure is assumed to grow with the data. Its complexity, e.g., the number of mixture components or the number of factors, is part of the posterior distribution. Rather than needing to be specified in advance, it is determined as part of analyzing the data.

In this tutorial, we survey Bayesian nonparametric methods. We focus on Bayesian nonparametric extensions of two common models, mixture models and latent factor models. As we mentioned above, traditional mixture models group data into a prespecified number of latent clusters. The Bayesian nonparametric mixture model, which is called a Chinese restaurant process mixture (or a Dirichlet process mixture), infers the number of clusters from the data and allows the number of clusters to grow as new data points are observed.

Latent factor models decompose observed data into a linear combination of latent factors. Different assumptions about the distribution of factors lead to variants such as factor analysis, principal component analysis, independent component analysis, and others. As for mixtures, a limitation of latent factor models is that the number of factors must be specified in advance. The Indian buffet process latent factor model (or Beta process latent factor model) infers the number of factors from the data and allows the number of factors to grow as new data points are observed.

We focus on these two types of models because they have served as the basis for a flexible suite of BNP models. Models that are built on BNP mixtures or latent factor models include those tailored for sequential data (Beal, Ghahramani, & Rasmussen, 2002; Fox, Sudderth, Jordan, & Willsky, 2008, 2009; Paisley & Carin, 2009), grouped data (Navarro, Griffiths, Steyvers, & Lee, 2006; Teh, Jordan, Beal, & Blei, 2006), data in a tree (Johnson, Griffiths, & Goldwater, 2007; Liang, Petrov, Jordan, & Klein, 2007), relational data (Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006; Miller, Griffiths, & Jordan, 2009; Navarro & Griffiths, 2008) and spatial data (Duan, Guindani, & Gelfand, 2007; Gelfand, Kottas, & MacEachern, 2005; Sudderth & Jordan, 2009).

This tutorial is organized as follows. In Sections 2 and 3, we describe mixture and latent factor models in more detail, starting from finite-capacity versions and then extending these to their infinite-capacity counterparts. In Section 4, we summarize the standard algorithms for inference in mixture and latent factor models. Finally, in Section 5, we describe several limitations and extensions of these models. In Appendix A, we detail some of the mathematical and statistical foundations of BNP models.

We hope to demonstrate how Bayesian nonparametric data analysis provides a flexible alternative to traditional Bayesian (and non-Bayesian) modeling. We give examples of BNP analysis of published psychological studies, and we point the reader to the available software for performing her own analyses.

2. Mixture models and clustering

In a mixture model, each observed data point is assumed to belong to a cluster. In posterior inference, we infer a grouping or clustering of the data under these assumptions—this amounts to inferring both the identities of the clusters and the assignments of the data to them. Mixture models are used for understanding the group structure of a data set and for flexibly estimating the distribution of a population.

For concreteness, consider the problem of modeling response time (RT) distributions. Psychologists believe that several cognitive processes contribute to producing behavioral responses (Luce, 1986), and therefore it is a scientifically relevant question how to decompose observed RTs into their underlying components. The generative model we describe below expresses one possible process by which latent causes (e.g., cognitive processes) might give rise to observed data (e.g., RTs).[2] Using Bayes' rule, we can invert the generative model to recover a distribution over the possible set of latent causes of our observations. The inferred latent causes are commonly known as "clusters".

2.1. Finite mixture modeling

One approach to this problem is finite mixture modeling. A finite mixture model assumes that there are K clusters, each associated with a parameter θ_k. Each observation y_n is assumed to be generated by first choosing a cluster c_n according to $P(c_n)$ and then generating the observation from its corresponding observation distribution parameterized by θ_{c_n}. In the RT modeling problem, each observation is a scalar RT and each cluster specifies a hypothetical distribution $F(y_n|\theta_{c_n})$ over the observed RT.[3]

Finite mixtures can accommodate many kinds of data by changing the data generating distribution. For example, in a Gaussian mixture model the data – conditioned on knowing their cluster assignments – are assumed to be drawn from a Gaussian distribution. The cluster parameters θ_k are the means of the components (assuming known variances). Fig. 1 illustrates data drawn from a Gaussian mixture with four clusters.

Bayesian mixture models further contain a prior over the mixing distribution $P(c)$, and a prior over the cluster parameters: $\theta \sim G_0$. (We denote the prior over cluster parameters G_0 to later

[1] The origins of these methods are in the distribution of random measures called the *Dirichlet process* (Antoniak, 1974; Ferguson, 1973), which was developed mainly for mathematical interest. These models were dubbed "Bayesian nonparametric" because they place a prior on the infinite-dimensional space of random measures. With the maturity of Markov chain Monte Carlo sampling methods, nearly twenty years later, Dirichlet processes became a practical statistical tool (Escobar & West, 1995). Bayesian nonparametric modeling is enjoying a renaissance in statistics and machine learning; we focus here on their application to latent component models, which is one of their central applications. We describe their formal mathematical foundations in Appendix A.

[2] A number of papers in the psychology literature have adopted a mixture model approach to modeling RTs (e.g., Ratcliff & Tuerlinckx, 2002; Wagenmakers, van der Maas, Dolan, & Grasman, 2008). It is worth noting that the decomposition of RTs into constituent cognitive processes performed by the mixture model is fundamentally different from the diffusion model analysis (Ratcliff & Rouder, 1998), which has become the gold standard in psychology and neuroscience. In the diffusion model, behavioral effects are explained in terms of variations in the underlying parameters of the model, whereas the mixture model attempts to explain these effects in terms of different latent causes governing each response.

[3] The interpretation of a cluster as a psychological process must be made with caution. In our example, the hypothesis is that some number of cognitive processes produces the RT data, and the mixture model provides a characterization of the cognitive process under that hypothesis. Further scientific experimentation is required to validate the existence of these processes and their causal relationship to behavior.

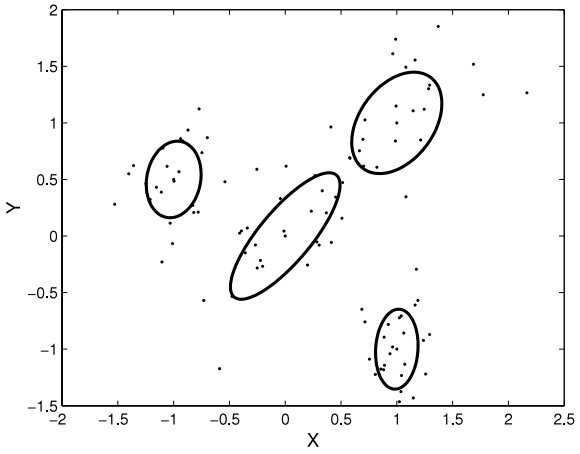

Fig. 1. *Draws from a Gaussian mixture model.* Ellipses show the standard deviation contour for each mixture component.

make a connection to BNP mixture models.) In a Gaussian mixture, for example, it is computationally convenient to choose the cluster parameter prior to be Gaussian. A convenient choice for the distribution on the mixing distribution is a Dirichlet. We will build on fully Bayesian mixture modeling when we discuss Bayesian nonparametric mixture models.

This generative process defines a joint distribution over the observations, cluster assignments, and cluster parameters,

$$P(\mathbf{y}, \mathbf{c}, \theta) = \prod_{k=1}^{K} G_0(\theta_k) \prod_{n=1}^{N} F(y_n|\theta_{c_n}) P(c_n), \tag{1}$$

where the observations are $\mathbf{y} = \{y_1, \ldots, y_N\}$, the cluster assignments are $\mathbf{c} = \{c_1, \ldots, c_N\}$, and the cluster parameters are $\theta = \{\theta_1, \ldots, \theta_K\}$. The product over n follows from assuming that each observation is conditionally independent given its latent cluster assignment and the cluster parameters. Returning to the RT example, the RTs are assumed to be independent of each other once we know which cluster generated each RT and the parameters of the latent clusters.

Given a data set, we are usually interested in the cluster assignments, i.e., a grouping of the data.[4] We can use Bayes' rule to calculate the posterior probability of assignments given the data:

$$P(\mathbf{c}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{c})P(\mathbf{c})}{\sum_{\mathbf{c}} P(\mathbf{y}|\mathbf{c})P(\mathbf{c})}, \tag{2}$$

where the likelihood is obtained by marginalizing over settings of θ:

$$P(\mathbf{y}|\mathbf{c}) = \int_{\theta} \left[\prod_{n=1}^{N} F(\mathbf{y}|\theta_{c_n}) \prod_{k=1}^{K} G_0(\theta_k) \right] \mathrm{d}\theta. \tag{3}$$

A G_0 that is conjugate to F allows this integral to be calculated analytically. For example, the Gaussian is the conjugate prior to a Gaussian with fixed variance, and this is why it is computationally convenient to select G_0 to be Gaussian in a mixture of Gaussians model.

The posterior over assignments is intractable because computing the denominator (marginal likelihood) requires summing over every possible partition of the data into K groups. (This problem becomes more salient in the next section, where we consider the limiting case $K \to \infty$.) We can use approximate methods, such as Markov chain Monte Carlo (McLachlan & Peel, 2000) or variational inference (Attias, 2000); these methods are discussed further in Section 4.

2.2. The Chinese restaurant process

When we analyze data with the finite mixture of Eq. (1), we must specify the number of latent clusters (e.g., hypothetical cognitive processes) in advance. In many data analysis settings, however, we do not know this number and would like to learn it from the data. BNP clustering addresses this problem by assuming that there is an infinite number of latent clusters, but that a finite number of them is used to generate the observed data. Under these assumptions, the posterior provides a distribution over the number of clusters, the assignment of data to clusters, and the parameters associated with each cluster. Furthermore, the predictive distribution, i.e., the distribution of the next data point, allows for new data to be assigned to a previously unseen cluster.

The BNP approach finesses the problem of choosing the number of clusters by assuming that it is infinite, while specifying the prior over infinite groupings $P(\mathbf{c})$ in such a way that it favors assigning data to a small number of groups. The prior over groupings is called the *Chinese restaurant process* (CRP; Aldous, 1985; Pitman, 2002), a distribution over infinite partitions of the integers; this distribution was independently discovered by Anderson (1991) in the context of his rational model of categorization (see Section 6.1 for more discussion of psychological implications). The CRP derives its name from the following metaphor. Imagine a restaurant with an infinite number of tables,[5] and imagine a sequence of customers entering the restaurant and sitting down. The first customer enters and sits at the first table. The second customer enters and sits at the first table with probability $\frac{1}{1+\alpha}$, and the second table with probability $\frac{\alpha}{1+\alpha}$, where α is a positive real. When the nth customer enters the restaurant, she sits at each of the occupied tables with probability proportional to the number of previous customers sitting there, and at the next unoccupied table with probability proportional to α. At any point in this process, the assignment of customers to tables defines a random partition. A schematic of this process is shown in Fig. 2.

More formally, let c_n be the table assignment of the nth customer. A draw from this distribution can be generated by sequentially assigning observations to classes with probability

$$P(c_n = k|\mathbf{c}_{1:n-1}) = \begin{cases} \dfrac{m_k}{n-1+\alpha} & \text{if } k \leq K_+ \\ \quad \text{(i.e., } k \text{ is a previously occupied table)} \\ \dfrac{\alpha}{n-1+\alpha} & \text{otherwise} \\ \quad \text{(i.e., } k \text{ is the next unoccupied table)} \end{cases} \tag{4}$$

where m_k is the number of customers sitting at table k, and K_+ is the number of tables for which $m_k > 0$. The parameter α is called the *concentration parameter*. Intuitively, a larger value of α will produce more occupied tables (and fewer customers per table).

The CRP exhibits an important invariance property: the cluster assignments under this distribution are *exchangeable*. This means that $p(\mathbf{c})$ is unchanged if the order of customers is shuffled (up to label changes). This may seem counter-intuitive at first, since the process in Eq. (4) is described sequentially.

[4] Under the Dirichlet prior, the assignment vector $\mathbf{c} = [1, 2, 2]$ has the same probability as $\mathbf{c} = [2, 1, 1]$. That is, these vectors are equivalent up to a "label switch". Generally, we do not care about what particular labels are associated with each class; rather, we care about *partitions*—equivalence classes of assignment vectors that preserve the same groupings but ignore labels.

[5] The Chinese restaurant metaphor is due to Pitman and Dubins, who were inspired by the seemingly infinite seating capacity of Chinese restaurants in San Francisco.

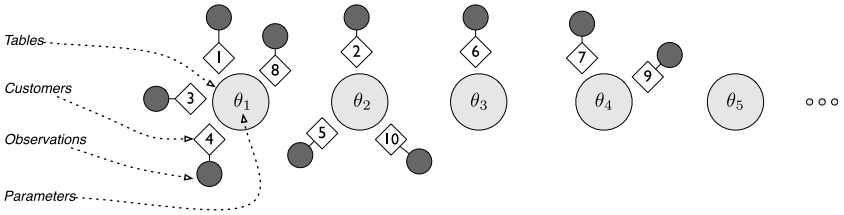

Fig. 2. *The Chinese restaurant process*. The generative process of the CRP, where numbered diamonds represent customers, attached to their corresponding observations (shaded circles). The large circles represent tables (clusters) in the CRP and their associated parameters (θ). Note that technically the parameter values $\{\theta\}$ are not part of the CRP *per se*, but rather belong to the full mixture model.

Consider the joint distribution of a set of customer assignments $c_{1:N}$. It decomposes according to the chain rule,

$$p(c_1, c_2, \ldots, c_N) = p(c_1)p(c_2 \mid c_1)p(c_3 \mid c_1, c_2) \cdots p(c_N \mid c_1, c_2, \ldots, c_{N-1}), \quad (5)$$

where each terms comes from Eq. (4). To show that this distribution is exchangeable, we will introduce some new notation. Let $K(c_{1:N})$ be the number of groups in which these assignments place the customers, which is a number between 1 and N. (Below, we will suppress its dependence on $c_{1:N}$.) Let I_k be the set of indices of customers assigned to the kth group, and let N_k be the number of customers assigned to that group (i.e., the cardinality of I_k).

Now, examine the product of terms in Eq. (5) that correspond to the customers in group k. This product is

$$\frac{\alpha \cdot 1 \cdot 2 \cdots (N_k - 1)}{(I_{k,1} - 1 + \alpha)(I_{k,2} - 1 + \alpha) \cdots (I_{k,N} - 1 + \alpha)}. \quad (6)$$

To see this, notice that the first customer in group k contributes probability $\frac{\alpha}{I_{k,1}-1+\alpha}$ because he is starting a new table, the second customer contributes probability $\frac{1}{I_{k,2}-1+\alpha}$ because he is sitting a table with one customer at it, the third customer contributes probability $\frac{2}{I_{k,3}-1+\alpha}$, and so on. The numerator of Eq. (6) can be more succinctly written as $\alpha(N_k - 1)!$

With this expression, we now rewrite the joint distribution in Eq. (5) as a product over per-group terms,

$$p(c_{1:N}) = \prod_{k=1}^{K} \frac{\alpha(N_k - 1)!}{(I_{k,1} - 1 + \alpha)(I_{k,2} - 1 + \alpha) \cdots (I_{k,N_k} - 1 + \alpha)}. \quad (7)$$

Finally, notice that the union of I_k across all groups k identifies each index once, because each customer is assigned to exactly one group. This simplifies the denominator and lets us write the joint as

$$p(c_{1:N}) = \frac{\alpha^K \prod_{k=1}^{K}(N_k - 1)!}{\prod_{i=1}^{N}(i - 1 + \alpha)}. \quad (8)$$

Eq. (8) reveals that Eq. (5) is exchangeable. It only depends on the number of groups K and the size of each group N_k. The probability of a particular seating configuration $c_{1:N}$ does not depend on the order in which the customers arrived.

2.3. Chinese restaurant process mixture models

The BNP clustering model uses the CRP in an infinite-capacity mixture model (Anderson, 1991; Antoniak, 1974; Escobar & West, 1995; Rasmussen, 2000). Each table k is associated with a cluster and with a cluster parameter θ_k, drawn from a prior G_0. We emphasize that there are an infinite number of clusters, though a finite data set only exhibits a finite number of active clusters. Each data point is a "customer", who sits at a table c_n and then draws its observed value from the distribution $F(y_n|\theta_{c_n})$. The concentration parameter α controls the prior expected number of clusters (i.e., occupied tables) K_+. In particular, this number grows logarithmically with the number of customers N: $\mathbb{E}[K_+] = \alpha \log N$ (for $\alpha < N/\log N$). If α is treated as unknown, one can put a hyperprior over it and use the same Bayesian machinery discussed in Section 4 to infer its value.

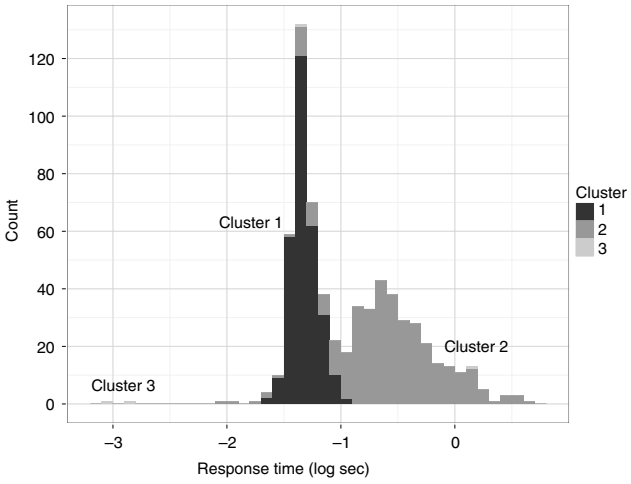

Fig. 3. *Response time modeling with the CRP mixture model*. An example distribution of response times from a two-alternative forced-choice decision making experiment (Simen et al., 2009). Colors denote clusters inferred by 100 iterations of Gibbs sampling.

Returning to the RT example, the CRP allows us to place a prior over partitions of RTs into the hypothetical cognitive processes that generated them, without committing in advance to a particular number of processes. As in the finite setting, each process k is associated with a set of parameters θ_k specifying the distribution over RTs (e.g., the mean of a Gaussian for log-transformed RTs). Fig. 3 shows the clustering of RTs obtained by approximating the posterior of the CRP mixture model using Gibbs sampling (see Section 4); in this figure, the cluster assignments from a single sample are shown. These data were collected in an experiment on two-alternative forced-choice decision making (Simen et al., 2009). Notice that the model captures the two primary modes of the data, as well as a small number of left-skewed outliers.

By examining the posterior over partitions, we can infer the assignment of RTs to hypothetical cognitive processes and the number of hypothetical processes. In addition, the (approximate) posterior provides a measure of confidence in any particular clustering, without committing to a single cluster assignment. Notice that the number of clusters can grow as more data are observed. This is a natural regime for many scientific applications, and it makes the CRP mixture robust to new data that is far away from the original observations.

When we analyze data with a CRP, we form an approximation of the joint posterior over all latent variables and parameters. In practice, there are two uses for this posterior. One is to

examine the likely partitioning of the data. This gives us a sense of how data are grouped, and how many groups the CRP model chose to use. The second use is to form predictions with the posterior predictive distribution. With a CRP mixture, the posterior predictive distribution is

$$P(y_{n+1}|\mathbf{y}_{1:n}) = \sum_{\mathbf{c}_{1:n+1}} \int_{\theta} P(y_{n+1}|c_{n+1},\theta) \times P(c_{n+1}|\mathbf{c}_{1:n}) P(\mathbf{c}_{1:n},\theta|\mathbf{y}_{1:n}) \mathrm{d}\theta. \quad (9)$$

Since the CRP prior, $P(c_{n+1}|\mathbf{c}_{1:n})$, appears in the predictive distribution, the CRP mixture allows new data to possibly exhibit a previously unseen cluster.

3. Latent factor models and dimensionality reduction

Mixture models assume that each observation is assigned to one of K components. Latent factor models weaken this assumption: each observation is influenced by each of K components in a different way (see Comrey & Lee, 1992, for an overview). These models have a long history in psychology and psychometrics (Pearson, 1901; Thurstone, 1931), and one of their first applications was to modeling human intelligence (Spearman, 1904). We will return to this application shortly.

Latent factor models provide dimensionality reduction in the (usual) case when the number of components is smaller than the dimension of the data. Each observation is associated with a vector of component activations (latent factors) that describes how much each component contributes to it, and this vector can be seen as a lower dimensional representation of the observation itself. When fit to data, the components parsimoniously capture the primary modes of variation in the observations.

The most popular of these models—factor analysis (FA), principal component analysis (PCA) and independent component analysis (ICA)—all assume that the number of factors (K) is known. The Bayesian nonparametric variant of latent factor models we describe below, allows the number of factors to grow as more data are observed. As with the BNP mixture model, the posterior distribution provides both the properties of the latent factors and how many are exhibited in the data.[6]

In classical factor analysis, the observed data is a collection of N vectors, $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, each of which are M-dimensional. Thus, $\mathbf{Y}$ is a matrix where rows correspond to observations and columns correspond to observed dimensions. The data (e.g., intelligence test scores) are assumed to be generated by a noisy weighted combination of latent factors (e.g., underlying intelligence faculties):

$$\mathbf{y}_n = \mathbf{G}\mathbf{x}_n + \boldsymbol{\epsilon}_n, \quad (10)$$

where $\mathbf{G}$ is a $M \times K$ factor loading matrix expressing how latent factor k influences observation dimension m, $\mathbf{x}_n$ is a K-dimensional vector expressing the activity of each latent factor, and $\boldsymbol{\epsilon}_n$ is a vector of independent Gaussian noise terms.[7] We can extend this to a sparse model by decomposing the factor loading into the product of two components: $G_{mk} = z_{mk} w_{mk}$, where z_{mk} is a binary "mask" variable that indicates whether factor k is "on" ($z_{mk} = 1$) or "off" ($z_{mk} = 0$) for dimension m, and w_{mk} is a continuous weight variable. This is sometimes called a "spike and slab" model (Ishwaran & Rao, 2005; Mitchell & Beauchamp, 1988) because the marginal distribution over x_{mk} is a mixture of a (typically Gaussian) "slab" $P(w_{mk})$ over the space of latent factors and a "spike" at zero, $P(z_{mk} = 0)$.

We take a Bayesian approach to inferring the latent factors, mask variables, and weights. We place priors over them and use Bayes' rule to compute the posterior $P(\mathbf{G}, \mathbf{Z}, \mathbf{W}|\mathbf{Y})$. In contrast, classical techniques like ICA, FA and PCA fit point estimates of the parameters (typically maximum likelihood estimates).

As mentioned above, a classic application of factor analysis in psychology is to the modeling of human intelligence (Spearman, 1904). Spearman (1904) argued that there exists a general intelligence factor (the so-called g-factor) that can be extracted by applying classical factor analysis methods to intelligence test data. Spearman's hypothesis was motivated by the observation that scores on different tests tend to be correlated: participants who score highly on one test are likely to score highly on another. However, several researchers have disputed the notion that this pattern arises from a unitary intelligence construct, arguing that intelligence consists of a multiplicity of components (Gould, 1981). Although we do not aspire to resolve this controversy, the question of how many factors underlie human intelligence is a convenient testbed for the BNP factor analysis model described below.

Since in reality the number of latent intelligence factors is unknown, we would like to avoid specifying K and instead allow the data to determine the number of factors. Following the model proposed by Knowles and Ghahramani (2011), $\mathbf{Z}$ is a binary matrix with a finite number of rows (each corresponding to an intelligence measure) and an infinite number of columns (each corresponding to a latent factor).

Like the CRP, the infinite-capacity distribution over $\mathbf{Z}$ has been furnished with a similarly colorful culinary metaphor, dubbed the *Indian buffet process* (IBP; Griffiths & Ghahramani, 2005, 2011). A customer (dimension) enters a buffet with an infinite number of dishes (factors) arranged in a line. The probability that a customer m samples dish k (i.e., sets $z_{mk} = 1$) is proportional to its popularity h_k (the number of prior customers who have sampled the dish). When the customer has considered all the previously sampled dishes (i.e., those for which $h_k > 0$), she samples an additional Poisson(α/N) dishes that have never been sampled before. When all M customers have navigated the buffet, the resulting binary matrix $\mathbf{Z}$ (encoding which customers sampled which dishes) is a draw from the IBP.

The IBP plays the same role for latent factor models that the CRP plays for mixture models: it functions as an infinite-capacity prior over the space of latent variables, allowing an unbounded number of latent factors (Knowles & Ghahramani, 2011). Whereas in the CRP, each observation is associated with only one latent component, in the IBP each observation (or, in the factor analysis model described above, each dimension) is associated with a theoretically infinite number of latent components.[8] A schematic of the IBP is shown in Fig. 4. Comparing to Fig. 2, the key difference between the CRP and the IBP is that in the CRP, each customer sits at a single table, whereas in the IBP, a customer can sample several dishes. This difference is illustrated in Fig. 5, which shows random draws from both models side-by-side.

Returning to the intelligence modeling example, posterior inference in the infinite latent factor model yields a distribution over matrices of latent factors which describe hypothetical intelligence structures:

$$P(\mathbf{X}, \mathbf{W}, \mathbf{Z}|\mathbf{Y}) \propto P(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \mathbf{Z}) P(\mathbf{X}) P(\mathbf{W}) P(\mathbf{Z}). \quad (11)$$

Exact inference is intractable because the normalizing constant requires a sum over all possible binary matrices. However, we can

[6] Historically, psychologists have explored a variety of rotation methods for enforcing sparsity and interpretability in FA solutions, starting with early work summarized by Thurstone (1947). Many recent methods are reviewed by Browne (2001). The Bayesian approach we adopt differs from these methods by specifying a preference for certain kinds of solutions in terms of the prior.

[7] The assumption of Gaussian noise in Eq. (10) is not fundamental to the latent factor model, but is the most common choice of noise distribution.

[8] Most of these latent factors will be "off" because the IBP preserves the sparsity of the finite Beta-Bernoulli prior (Griffiths & Ghahramani, 2005). The degree of sparsity is controlled by α: for larger values, more latent factors will tend to be active.

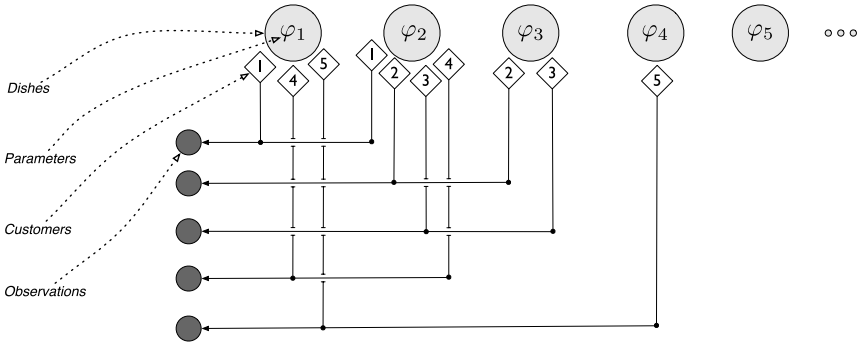

Fig. 4. *The Indian buffet process*. The generative process of the IBP, where numbered diamonds represent customers, attached to their corresponding observations (shaded circles). Large circles represent dishes (factors) in the IBP, along with their associated parameters (φ). Each customer selects several dishes, and each customer's observation (in the latent factor model) is a linear combination of the selected dish's parameters. Note that technically the parameter values $\{\phi\}$ are not part of the IBP *per se*, but rather belong to the full latent factor model.

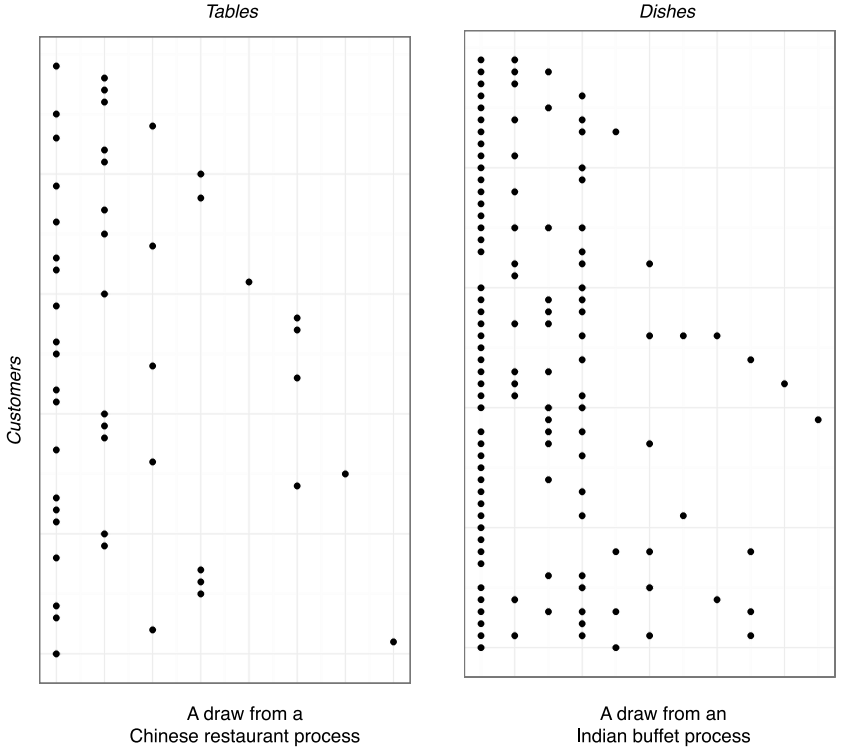

Fig. 5. *Draws from the CRP and IBP*. (*Left*) Random draw from the Chinese restaurant process. (*Right*) Random draw from the Indian buffet process. In the CRP, each customer is assigned to a single component. In the IBP, a customer can be assigned to multiple components.

approximate the posterior using one of the techniques described in the next section (e.g., with a set of samples). Given posterior samples of $\mathbf{Z}$, one typically examines the highest-probability sample (the *maximum a posteriori*, or MAP, estimate) to get a sense of the latent factor structure. As with the CRP, if one is interested in predicting some function of $\mathbf{Z}$, then it is best to average this function over the samples.

Fig. 6 shows the results of the IBP factor analysis applied to data collected by Kane et al. (2004). We consider the 13 reasoning tasks administered to 234 participants. The left panel displays a histogram of the factor counts (the number of times $z_{mk} = 1$ across posterior samples). This plot indicates that the dataset is best described by a combination of around $4 - 7$ factors; although this is obviously not a conclusive argument against the existence of a general intelligence factor, it suggests that additional factors merit further investigation. The right panel displays the first factor loading from the IBP factor analysis plotted against the g-factor, demonstrating that the nonparametric method is able to extract a structure consistent with classical methods.[9]

4. Inference

We have described two classes of BNP models—mixture models based on the CRP and latent factor models based on the IBP. Both types of models posit a generative probabilistic process of a collection of observed (and future) data that includes hidden structure. We analyze data with these models by examining the posterior distribution of the hidden structure given the observations; this gives us a distribution over which latent structure likely generated our data.

[9] It is worth noting that the field of intelligence research has developed its methods far beyond Spearman's g-factor. In particular, *hierarchical* factor analysis is now in common use. See Kane et al. (2004) for an example.

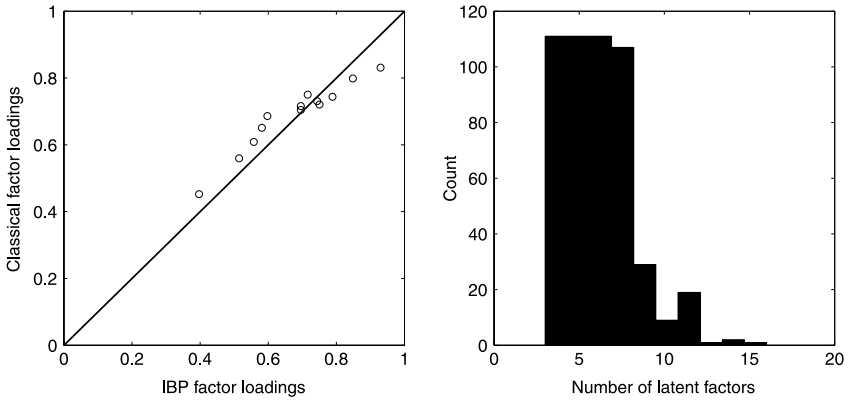

Fig. 6. *IBP factor analysis of human performance on reasoning tasks.* (*Left*) Histogram of the number of latent factors inferred by Gibbs sampling applied to reasoning task data from Kane et al. (2004). 1000 samples were generated, and the first 500 were discarded as burn-in. (*Right*) Relationship between the loading of the first factor inferred by IBP factor analysis and Spearman's *g* (i.e., the loading of the first factor inferred by classical factor analysis; Spearman, 1904).

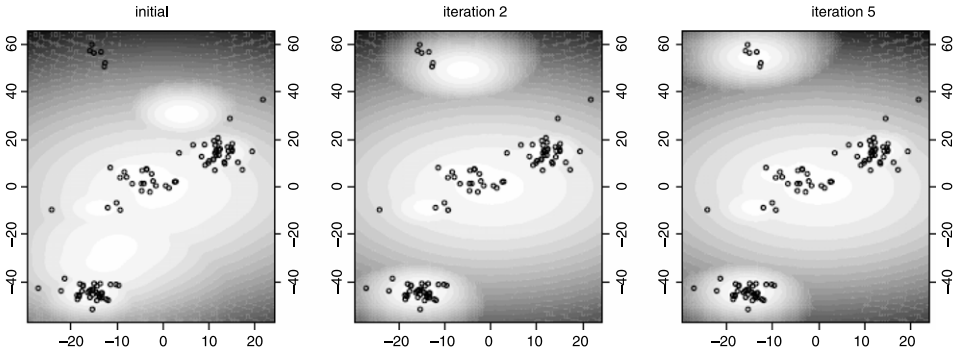

Fig. 7. *Inference in a Chinese restaurant process mixture model.* The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.
Source: Figure reproduced with permission from Blei and Jordan (2006).

Thus, the basic computational problem in BNP modeling (as in most of Bayesian statistics) is computing the posterior. For many interesting models, including those discussed here, the posterior is not available in closed form. There are several ways to approximate it. While a comprehensive treatment of inference methods in BNP models is beyond the scope of this tutorial, we will describe some of the most widely-used algorithms. In Appendix B, we provide links to software packages implementing these algorithms.

The most widely used posterior inference methods in Bayesian nonparametric models are Markov Chain Monte Carlo (MCMC) methods. The idea MCMC methods is to define a Markov chain on the hidden variables that has the posterior as its equilibrium distribution (Andrieu, De Freitas, Doucet, & Jordan, 2003). By drawing samples from this Markov chain, one eventually obtains samples from the posterior. A simple form of MCMC sampling is Gibbs sampling, where the Markov chain is constructed by considering the conditional distribution of each hidden variable given the others and the observations. Thanks to the exchangeability property described in Section 2.2, CRP mixtures are particularly amenable to Gibbs sampling—in considering the conditional distributions, each observation can be considered to be the "last" one and the distribution of Eq. (4) can be used as one term of the conditional distribution. (The other term is the likelihood of the observations under each partition.) Neal (2000) provides an excellent survey of Gibbs sampling and other MCMC algorithms for inference in CRP mixture models (see also Escobar & West, 1995; Fearnhead, 2004; Ishwaran & James, 2001; Jain & Neal, 2004; Rasmussen, 2000; Wood & Griffiths, 2007). Gibbs sampling for the IBP factor analysis model is described in Knowles and Ghahramani (2011).

MCMC methods, although guaranteed to converge to the posterior with enough samples, have two drawbacks: (1) the samplers must be run for many iterations before convergence and (2) it is difficult to assess convergence. An alternative approach to approximating the posterior is *variational inference* (Jordan, Ghahramani, Jaakkola, & Saul, 1999). This approach is based on the idea of approximating the posterior with a simpler family of distributions and searching for the member of that family that is closest to it.[10] Although variational methods are not guaranteed to recover the true posterior (unless it belongs to the simple family of distributions), they are typically faster than MCMC and convergence assessment is straightforward. These methods have been applied to CRP mixture models (Blei & Jordan, 2006; Kurihara, Welling, & Teh, 2007, see Fig. 7 for an example) and IBP latent factor models (Doshi-Velez, Miller, Berkeley, Van Gael, & Teh, 2009; Paisley, Zaas, Woods, Ginsburg, & Carin, 2010). We note that variational inference usually operates on a random measure representation of CRP mixtures and IBP factor models, which are described in Appendix A. Gibbs samplers that operate on this representation are also available (Ishwaran & James, 2001).

As we mentioned in the introduction, BNP methods provide an alternative to model selection over a parameterized family of models.[11] In effect, both MCMC and variational strategies for posterior inference provide a data-directed mechanism for

[10] Distance between probability distributions in this setting is measured by the Kullback–Leibler divergence (relative entropy).

[11] The Journal of Mathematical Psychology has published two special issues (Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006) on model selection which review a broad array of model selection techniques (both Bayesian and non-Bayesian).

simultaneously searching the space of models and finding optimal parameters. This is convenient in settings like mixture modeling or factor analysis because we avoid needing to fit models for each candidate number of components. It is essential in more complex settings, where the algorithm searches over a space that is difficult to efficiently enumerate and explore.

5. Limitations and extensions

We have described the most widely used BNP models, but this is only the tip of the iceberg. In this section we highlight some key limitations of the models described above, and the extensions that have been developed to address these limitations. It is worth mentioning here that we cannot do full justice to the variety of BNP models that have been developed over the past 40 years; we have omitted many exciting and widely-used ideas, such as Pitman–Yor processes, gamma processes, Dirichlet diffusion trees and Kingman's coalescent. To learn more about these ideas, see the recent volume edited by Hjort et al. (2010).

5.1. Hierarchical structure

The first limitation concerns *grouped* data: how can we capture both commonalities and idiosyncrasies across individuals within a group? For example, members of an animal species will tend to be similar to each other, but also unique in certain ways. The standard Bayesian approach to this problem is based on *hierarchical models* (Gelman et al., 2004), in which individuals are coupled by virtue of being drawn from the same group-level distribution.[12] The parameters of this distribution govern both the characteristics of the group and the degree of coupling. In the nonparametric setting, hierarchical extensions of the Dirichlet process (Teh et al., 2006) and beta process (Thibaux & Jordan, 2007) have been developed, allowing an infinite number of latent components to be shared by multiple individuals. For example, hierarchical Dirichlet processes can be applied to modeling text documents, where each document is represented by an infinite mixture of word distributions ("topics") that are shared across documents.

Returning to the RT example from Section 2, imagine measuring RTs for several subjects. The goal again is to infer which underlying cognitive process generated each response time. Suppose we assume that the same cognitive processes are shared across subjects, but they may occur in different proportions. This is precisely the kind of structure the HDP can capture.

5.2. Time series models

The second limitation concerns *sequential* data: how can we capture dependencies between observations arriving in a sequence? One of the most well-known models for capturing such dependencies is the hidden Markov model (see, e.g., Bishop, 2006), in which the latent class for observation n depends on the latent class for observation $n-1$. The infinite hidden Markov model (HMM; Beal et al., 2002; Paisley & Carin, 2009; Teh et al., 2006) posits the same sequential structure, but employs an infinite number of latent classes. Teh et al. (2006) showed that the infinite HMM is a special case of the hierarchical Dirichlet process.

As an alternative to the HMM (which assumes a discrete latent state), a linear dynamical system (also known as an autoregressive moving average model) assumes that the latent state is continuous and evolves over time according to a linear-Gaussian Markov process. In a *switching* linear dynamical system, the system can have a number of dynamical modes; this allows the marginal transition distribution to be non-linear. Fox et al. (2008) have explored nonparametric variants of switching linear dynamical systems, where the number of dynamical modes is inferred from the data using an HDP prior.

5.3. Spatial models

Another type of dependency arising in many datasets is *spatial*. For example, one expects that if a disease occurs in one location, it is also likely to occur in a nearby location. One way to capture such dependencies in a BNP model is to make the base distribution G_0 of the DP dependent on a location variable (Duan et al., 2007; Gelfand et al., 2005). In the field of computer vision, Sudderth and Jordan (2009) have applied a spatially-coupled generalization of the DP to the task of image segmentation, allowing them to encode a prior bias that nearby pixels belong to the same segment.

We note in passing a burgeoning area of research attempting to devise more general specifications of dependencies in BNP models, particularly for DPs (Blei & Frazier, 2010; Griffin & Steel, 2006; MacEachern, 1999). These dependencies could be arbitrary functions defined over a set of covariates (e.g., age, income, weight). For example, people with similar age and weight will tend to have similar risks for certain diseases.

More recently, several authors have attempted to apply these ideas to the IBP and latent factor models (Doshi-Velez & Ghahramani, 2009; Miller, Griffiths, & Jordan, 2008; Williamson, Orbanz, & Ghahramani, 2010).

5.4. Supervised learning

We have restricted ourselves to a discussion of *unsupervised* learning problems, where the goal is to discover hidden structure in data. In *supervised* learning, the goal is to predict some output variable given a set of input variables (covariates). When the output variable is continuous, this corresponds to *regression*; when the output variable is discrete, this corresponds to *classification*.

For many supervised learning problems, the outputs are non-linear functions of the inputs. The BNP approach to this problem is to place a prior distribution (known as a *Gaussian process*) directly over the space of non-linear functions, rather than specifying a parametric family of non-linear functions and placing priors over their parameters. Supervised learning proceeds by posterior inference over functions using the Gaussian process prior. The output of inference is itself a Gaussian process, characterized by a mean function and a covariance function (analogous to a mean vector and covariance matrix in parametric Gaussian models). Given a new set of inputs, the posterior Gaussian process induces a predictive distribution over outputs. Although we do not discuss this approach further, Rasmussen and Williams (2006) is an excellent textbook on this topic.

Recently, another nonparametric approach to supervised learning has been developed, based on the CRP mixture model (Hannah, Blei, & Powell, 2010; Shahbaba & Neal, 2009). The idea is to place a DP mixture prior over the inputs and then model the mean function of the outputs as conditionally linear within each mixture component (see also Meeds & Osindero, 2006; Rasmussen & Ghahramani, 2002, for related approaches). The result is a marginally non-linear model of the outputs with linear sub-structure. Intuitively, each mixture component isolates a region of the input space and models the mean output linearly within that region. This is an example of a *generative* approach to supervised learning, where the joint distribution over both the inputs and outputs is modeled. In contrast, the Gaussian process approach described above is a *discriminative* approach, modeling only the conditional distribution of the outputs given the inputs.

[12] See also the recent issue of Journal of Mathematical Psychology (Volume 55, Issue 1) devoted to hierarchical Bayesian models. Lee (2010) provides an overview for cognitive psychologists.

6. Conclusions

BNP models are an emerging set of statistical tools for building flexible models whose structure grows and adapts to data. In this tutorial, we have reviewed the basics of BNP modeling and illustrated their potential in scientific problems.

It is worth noting here that while BNP models address the problem of choosing the number of mixture components or latent factors, they are not a general solution to the model selection problem which has received extensive attention within mathematical psychology and other disciplines (see Claeskens & Hjort, 2008, for a comprehensive treatment). In some cases, it may be preferable to place a prior over finite-capacity models and then compare Bayes factors (Kass & Raftery, 1995; Vanpaemel, 2010), or to use selection criteria motivated by other theoretical frameworks, such as information theory (Grünwald, 2007).

6.1. Bayesian nonparametric models of cognition

We have treated BNP models purely as a data analysis tool. However, there is a flourishing tradition of work in cognitive psychology on using BNP models as theories of cognition. The earliest example dates back to Anderson (1991), who argued that a version of the CRP mixture model could explain human categorization behavior. The idea in this model is that humans adaptively learn the number of categories from their observations. A number of recent authors have extended this work (Griffiths, Canini, Sanborn, & Navarro, 2007; Heller, Sanborn, & Chater, 2009; Sanborn, Griffiths, & Navarro, 2010) and applied it to other domains, such as classical conditioning (Gershman, Blei, & Niv, 2010).

The IBP has also been applied to human cognition. In particular, Austerweil and Griffiths (2009a) argued that humans decompose visual stimuli into latent features in a manner consistent with the IBP. When the parts that compose objects strongly covary across objects, humans treat whole objects as features, whereas individual parts are treated as features if the covariance is weak. This finding is consistent with the idea that the number of inferred features changes flexibly with the data.

BNP models have been fruitfully applied in several other domains, including word segmentation (Goldwater, Griffiths, & Johnson, 2009), relational theory acquisition (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010) and function learning (Austerweil & Griffiths, 2009b).

6.2. Suggestions for further reading

A recent edited volume by Hjort et al. (2010) is a useful resource on applied Bayesian nonparametrics. For a more general introduction to statistical machine learning with probabilistic models, see Bishop (2006). For a review of applied Bayesian statistics, see Gelman et al. (2004).

Acknowledgments

We are grateful to Andrew Conway, Katherine Heller, Irvin Hwang, Ed Vul, James Pooley and Ken Norman who offered comments on an earlier version of the manuscript. The manuscript was greatly improved by suggestions from the reviewers. We are also grateful to Patrick Simen, Andrew Conway and Michael Kane for sharing their data and offering helpful suggestions. This work was supported by a graduate research fellowship from the NSF to SJG. DMB is supported by ONR 175-6343, NSF CAREER 0745520, AFOSR 09NL202, the Alfred P. Sloan foundation, and a grant from Google.

Appendix A. Foundations

We have developed BNP methods via the CRP and IBP, both of which are priors over combinatorial structures (infinite partitions and infinite binary matrices). These are the easiest first ways to understand this class of models, but their mathematical foundations are found in constructions of random distributions. In this section, we review this perspective of the CRP mixture and IBP factor model.

The Dirichlet process

The Dirichlet process (DP) is a distribution over distributions. It is parameterized by a concentration parameter $\alpha > 0$ and a base distribution G_0, which is a distribution over a space Θ. A random variable drawn from a DP is itself a distribution over Θ. A random distribution G drawn from a DP is denoted $G \sim \mathrm{DP}(\alpha, G_0)$.

The DP was first developed in Ferguson (1973), who showed its existence by appealing to its finite dimensional distributions. Consider a measurable partition of Θ, $\{T_1, \ldots, T_K\}$.[13] If $G \sim \mathrm{DP}(\alpha, G_0)$ then every measurable partition of Θ is Dirichlet-distributed,

$$(G(T_1), \ldots, G(T_K)) \sim \mathrm{Dir}(\alpha G_0(T_1), \ldots, \alpha G_0(T_K)). \tag{12}$$

This means that if we draw a random distribution from the DP and add up the probability mass in a region $T \in \Theta$, then there will on average be $G_0(T)$ mass in that region. The concentration parameter plays the role of an inverse variance; for higher values of α, the random probability mass $G(T)$ will concentrate more tightly around $G_0(T)$.

Ferguson (1973) proved two properties of the Dirichlet process. The first property is that random distributions drawn from the Dirichlet process are discrete. They place their probability mass on a countably infinite collection of points, called "atoms",

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}. \tag{13}$$

In this equation, π_k is the probability assigned to the kth atom and θ_k^* is the location or value of that atom. Further, these atoms are drawn independently from the base distribution G_0.

The second property connects the Dirichlet process to the Chinese restaurant process. Consider a random distribution drawn from a DP followed by repeated draws from that random distribution,

$$G \sim \mathrm{DP}(\alpha, G_0) \tag{14}$$

$$\theta_i \sim G \quad i \in \{1, \ldots, n\}. \tag{15}$$

Ferguson (1973) examined the joint distribution of $\theta_{1:n}$, which is obtained by marginalizing out the random distribution G,

$$p(\theta_1, \ldots, \theta_n \mid \alpha, G_0) = \int \left(\prod_{i=1}^{n} p(\theta_i \mid G) \right) \mathrm{d}P(G \mid \alpha, G_0). \tag{16}$$

He showed that, under this joint distribution, the θ_i will exhibit a *clustering property*—they will share repeated values with positive probability. (Note that, for example, repeated draws from a Gaussian do not exhibit this property.) The structure of shared values defines a partition of the integers from 1 to n, and the distribution of this partition is a Chinese restaurant process with parameter α. Finally, he showed that the unique values of θ_i shared among the variables are independent draws from G_0.

[13] A partition of Θ defines a collection of subsets whose union is Θ. A partition is measurable if it is closed under complementation and countable union.

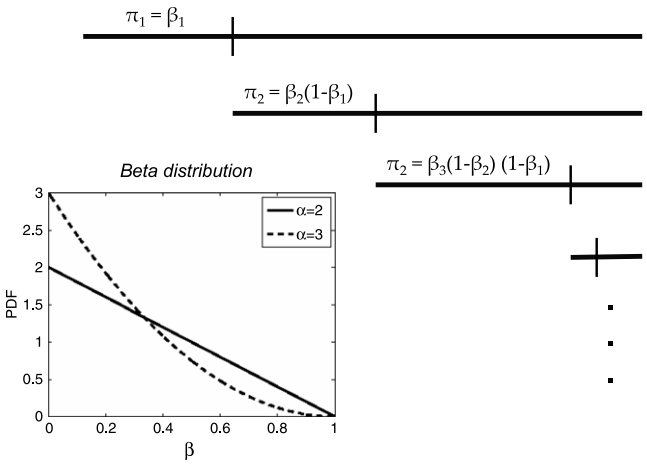

Fig. 8. *Stick-breaking construction.* Procedure for generating π by breaking a stick of length 1 into segments. Inset shows the beta distribution from which β_k is drawn, for different values of α.

Note that this is another way to confirm that the DP assumes exchangeability of $\theta_{1:n}$. In the foundations of Bayesian statistics, De Finetti's representation theorem (De Finetti, 1931) says that an exchangeable collection of random variables can be represented as a conditionally independent collection: first, draw a data generating distribution from a prior over distributions; then draw random variables independently from that data generating distribution. This reasoning in Eq. (16) shows that $\theta_{1:n}$ are exchangeable. (For a detailed discussion of De Finetti's representation theorems, see Bernardo and Smith (1994).)

Dirichlet process mixtures

A DP mixture adds a third step to the model above Antoniak (1974),

$$G \sim \mathrm{DP}(\alpha, G_0) \tag{17}$$

$$\theta_i \sim G \tag{18}$$

$$x_i \sim p(\cdot \mid \theta_i). \tag{19}$$

Marginalizing out G reveals that the DP mixture is equivalent to a CRP mixture. Good Gibbs sampling algorithms for DP mixtures are based on this representation (Escobar & West, 1995; Neal, 2000).

The stick-breaking construction

Ferguson (1973) proved that the DP exists via its finite dimensional distributions. Sethuraman (1994) provided a more constructive definition based on the *stick-breaking representation* (Fig. 8).

Consider a stick with unit length. We divide the stick into an infinite number of segments π_k by the following process. First, choose a beta random variable $\beta_1 \sim \mathrm{beta}(1, \alpha)$ and break of β_1 of the stick. For each remaining segment, choose another beta distributed random variable, and break off that proportion of the remainder of the stick. This gives us an infinite collection of weights π_k,

$$\beta_k \sim \mathrm{Beta}(1, \alpha) \tag{20}$$

$$\pi_k = \beta_k \prod_{j=1}^{k-1}(1-\beta_j) \quad k = 1, 2, 3, \ldots. \tag{21}$$

Finally, we construct a random distribution using Eq. (13), where we take an infinite number of draws from a base distribution G_0 and draw the weights as in Eq. (21). Sethuraman (1994) showed that the distribution of this random distribution is a $\mathrm{DP}(\alpha, G_0)$.

This representation of the Dirichlet process, and its corresponding use in a Dirichlet process mixture, allows us to compute a variety of functions of posterior DPs (Gelfand & Kottas, 2002) and is the basis for the variational approach to approximate inference (Blei & Jordan, 2006).

The beta process and Bernoulli process

Latent factor models admit a similar analysis (Thibaux & Jordan, 2007). We define the random measure B as a set of weighted atoms:

$$B = \sum_{k=1}^{K} w_k \delta_{\theta_k}, \tag{22}$$

where $w_k \in (0, 1)$ and the atoms $\{\theta_k\}$ are drawn from a base measure B_0 on Θ. Note that in this case (in contrast to the DP), the sum of the weights does not sum to 1 (almost surely), which means that B is not a probability measure. Analogously to the DP, we can define a "distribution on distributions" for random measures with weights between 0 and 1—namely the beta process, which we denote by $B \sim \mathrm{BP}(\alpha, B_0)$. Unlike the DP (which we could define in terms of Dirichlet-distributed marginals), the beta process cannot be defined in terms of beta-distributed marginals. A formal definition requires an excursion into the theory of completely random measures, which would take us beyond the scope of this Appendix (see Thibaux & Jordan, 2007).

To build a latent factor model from the beta process, we define a new random measure

$$X_n = \sum_{k=1}^{K} z_{nk} \delta_{\phi_k}, \tag{23}$$

where $z_{nk} \sim \mathrm{Bernoulli}(w_k)$. The random measure X_n is then said to be distributed according to a *Bernoulli process* with base measure B, written as $X_n \sim \mathrm{BeP}(B)$. A draw from a Bernoulli process places unit mass on atoms for which $z_{nk} = 1$; this defines which latent factors are "on" for the nth observation. N draws from the Bernoulli process yield an IBP-distributed binary matrix $\mathbf{Z}$, as shown by Thibaux and Jordan (2007).

In the context of factor analysis, the factor loading matrix $\mathbf{G}$ is generated from this process by first drawing the atoms and their weights from the beta process, and then constructing each $\mathbf{G}$ by turning on a subset of these atoms according to a draw from the Bernoulli process. Finally, observation $\mathbf{y}_n$ is generated according to Eq. (10).

Stick breaking construction of the beta process

A "double-use" of the same breakpoints β leads to a stick-breaking construction of the beta process (Teh, Görür, & Ghahramani, 2007); see also Paisley et al. (2010). In this case, the weights correspond to the length of the remaining stick, rather than the length of the segment that was just broken off: $\pi_k = \prod_{j=1}^{k}(1-\beta_j)$.

The infinite limit of finite models

In this section, we show BNP models can be derived by taking the infinite limit of a corresponding finite-capacity model. For mixture models, we assume that the class assignments $\mathbf{z}$ were drawn from a multinomial distribution with parameters $\pi = \{\pi_1, \ldots, \pi_K\}$, and place a symmetric Dirichlet distribution with concentration parameter α on π. The finite mixture model can be summarized as follows:

$$\begin{aligned} &\pi|\alpha \sim \mathrm{Dir}(\alpha), \qquad z_n|\pi \sim \pi \\ &\theta_k|G_0 \sim G_0, \qquad \mathbf{y}_n|z_n, \theta \sim F(\theta_{z_n}). \end{aligned} \tag{24}$$

When $K \to \infty$, this mixture converges to a Dirichlet process mixture model (Ishwaran & Zarepour, 2002; Neal, 1992; Rasmussen, 2000).

To construct a finite latent factor model, we assume that each mask variable is drawn from the following two-stage generative process:

$$w_k|\alpha \sim \mathrm{Beta}(\alpha/K, 1) \tag{25}$$

$$z_{nk}|w_k \sim \mathrm{Bernoulli}(w_k). \tag{26}$$

Table B.1
Software packages implementing various Bayesian nonparametric models.

Model	Algorithm	Language	Author	Link
CRP mixture model	MCMC	Matlab	Jacob Eisenstein	http://people.csail.mit.edu/jacobe/software.html
CRP mixture model	MCMC	R	Matthew Shotwell	http://cran.r-project.org/web/packages/profdpm/index.html
CRP mixture model	Variational	Matlab	Kenichi Kurihara	http://sites.google.com/site/kenichikurihara/academic-software
IBP latent factor model	MCMC	Matlab	David Knowles	http://mlg.eng.cam.ac.uk/dave
IBP latent factor model	Variational	Matlab	Finale Doshi-Velez	http://people.csail.mit.edu/finale/new-wiki/doku.php?id=publications_posters_presentations_code

Intuitively, this generative process corresponds to creating a bent coin with bias w_k, and then flipping it N times to determine whether to activate factors $\{z_{1k}, \ldots, z_{Nk}\}$. Griffiths and Ghahramani (2005) showed that taking the limit of this model as $K \to \infty$ yields the IBP latent factor model.

Appendix B. Software packages

We present a table (Table B.1) of several available software packages implementing the models presented in the main text.

References

Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour XIII* (pp. 1–198). Berlin: Springer.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, *2*, 1152–1174.

Attias, H. (2000). A variational Bayesian framework for graphical models. *Advances in Neural Information Processing Systems*, *12*, 209–215.

Austerweil, J., & Griffiths, T. (2009a). Analyzing human feature learning as nonparametric Bayesian inference. *Advances in Neural Information Processing Systems*, *21*, 386–492.

Austerweil, J., & Griffiths, T. L. (2009b). Analyzing human feature learning as nonparametric Bayesian inference. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems*: *Vol. 21* (pp. 97–104).

Beal, M., Ghahramani, Z., & Rasmussen, C. (2002). The infinite hidden Markov model. *Advances in Neural Information Processing Systems*, *14*, 577–584.

Bernardo, J., & Smith, A. (1994). *Bayesian theory*. Chichester: Wiley.

Bishop, C. (2006). *Pattern recognition and machine learning*. New York: Springer.

Blei, D., & Frazier, P. (2010). Distance dependent Chinese restaurant processes. In *International conference on machine learning*.

Blei, D., & Jordan, M. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, *1*(1), 121–144.

Browne, M. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Claeskens, G., & Hjort, N. (2008). *Model selection and model averaging*. Cambridge University Press.

Comrey, A., & Lee, H. (1992). *A first course in factor analysis*. Lawrence Erlbaum.

De Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale (4)*.

Doshi-Velez, F., & Ghahramani, Z. (2009). Correlated non-parametric latent feature models. In *International conference on uncertainty in artificial intelligence*.

Doshi-Velez, F., Miller, K., Berkeley, C., Van Gael, J., & Teh, Y. (2009). Variational inference for the Indian buffet process. In *Proceedings of the international conference on artificial intelligence and statistics* (pp. 137–144).

Duan, J., Guindani, M., & Gelfand, A. (2007). Generalized spatial Dirichlet process models. *Biometrika*, *94*(4), 809.

Escobar, M., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430).

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, *14*(1), 11–21.

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

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2008). Nonparametric Bayesian learning of switching linear dynamical systems. *Advances in Neural Information Processing Systems*, *21*.

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2009). Sharing features among dynamical systems with beta processes. In *Advances in Neural Information Processing Systems, 22*.

Gelfand, A., & Kottas, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *11*, 289–305.

Gelfand, A., Kottas, A., & MacEachern, S. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, *100*(471), 1021–1035.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*.

Gershman, S., Blei, D., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197.

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A Bayesian framework for word segmentation: exploring the effects of context. *Cognition*, *112*(1), 21–54.

Gould, S. (1981). *The mismeasure of man*. New York: Norton.

Griffin, J., & Steel, M. (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association*, *101*(473), 179–194.

Griffiths, T., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the 29th annual conference of the cognitive science society* (pp. 323–328).

Griffiths, T., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. *Advances in Neural Information Processing Systems*, *18*, 475.

Griffiths, T., & Ghahramani, Z. (2011). The Indian buffet process: an introduction and review. *Journal of Machine Learning Research*, *12*, 1185–1224.

Grünwald, P. (2007). *The minimum description length principle*. The MIT Press.

Hannah, L., Blei, D., & Powell, W. (2010). Dirichlet process mixtures of generalized linear models. In *Proceedings of the international conference on artificial intelligence and statistics. Vol. 13*.

Heller, K., Sanborn, A., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems*: *Vol. 22* (pp. 727–735).

Hjort, N., Holmes, C., Müller, P., & Walker, S. G. (Eds.). (2010). *Bayesian nonparametrics: principles and practice*. Cambridge University Press.

Ishwaran, H., & James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

Ishwaran, H., & Rao, J. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, 730–773.

Ishwaran, H., & Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, *30*(2), 269–283.

Jain, S., & Neal, R. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, *13*(1), 158–182.

Johnson, M., Griffiths, T., & Goldwater, S. (2007). Adaptor grammars: a framework for specifying compositional nonparametric Bayesian models. *Advances in Neural Information Processing Systems*, *19*.

Jordan, M., Ghahramani, Z., Jaakkola, T., & Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233.

Kane, M., Hambrick, D., Tuholski, S., Wilhelm, O., Payne, T., & Engle, R. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189.

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kemp, C., Tenenbaum, J., Griffiths, T., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the national conference on artificial intelligence*: *Vol. 21* (p. 381). Menlo Park, CA, Cambridge, MA, London: AAAI Press, MIT Press, 1999.

Kemp, C., Tenenbaum, J., Niyogi, S., & Griffiths, T. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165–196.

Knowles, D., & Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modelling. *Annals of Applied Statistics*, *5*, 1534–1552.

Kurihara, K., Welling, M., & Teh, Y. (2007). Collapsed variational Dirichlet process mixture models. In *Proceedings of the international joint conference on artificial intelligence. Vol. 20* (p. 19).

Lee, M. (2010). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*(1), 1–7.

Liang, P., Petrov, S., Jordan, M., & Klein, D. (2007). The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 688–697).

Luce, R. (1986). *Response times: their role in inferring elementary mental organization*. Oxford University Press.

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science* (pp. 50–55).

McLachlan, G., & Peel, D. (2000). *Finite mixture models*: *Vol. 299*. Wiley-Interscience.

Meeds, E., & Osindero, S. (2006). An alternative infinite mixture of Gaussian process experts. *Advances in Neural Information Processing Systems*, *18*, 883.
Miller, K., Griffiths, T., & Jordan, M. (2008). The phylogenetic indian buffet process: a non-exchangeable nonparametric prior for latent features. In *International conference on uncertainty in artificial intelligence*.
Miller, K., Griffiths, T., & Jordan, M. (2009). Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems*, (pp. 1276–1284).
Mitchell, T., & Beauchamp, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.
Myung, I., Forster, M., & Browne, M. (2000). Special issue on model selection. *Journal of Mathematical Psychology*, *44*(1–2).
Navarro, D., & Griffiths, T. (2008). Latent features in similarity judgments: a nonparametric Bayesian approach. *Neural Computation*, *20*(11), 2597–2628.
Navarro, D., Griffiths, T., Steyvers, M., & Lee, M. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122.
Neal, R. (1992). Bayesian mixture modeling. In *Maximum entropy and Bayesian methods: proceedings of the 11th international workshop on maximum entropy and Bayesian methods of statistical analysis* (pp. 197–211).
Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 249–265.
Paisley, J., & Carin, L. (2009). Hidden Markov models with stick-breaking priors. *IEEE Transactions on Signal Processing*, *57*(10), 3905–3917.
Paisley, J., Zaas, A., Woods, C., Ginsburg, G., & Carin, L. (2010). Nonparametric factor analysis with beta process priors. In *Proceedings of the international conference on machine learning*.
Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series (6)*, *2*(11), 559–572.
Pitman, J. (2002). Combinatorial stochastic processes. *Techincal report 621*. Notes for Saint Flour Summer School. Dept. Statistics. UC, Berkeley.
Rasmussen, C. (2000). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, *12*, 554–560.
Rasmussen, C., & Ghahramani, Z. (2002). Infinite mixtures of Gaussian process experts. In *Advances in neural information processing systems: Vol. 14* (pp. 881–888). MIT Press.
Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning*. New York: Springer.
Ratcliff, R., & Rouder, J. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347.
Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481.
Sanborn, A., Griffiths, T., & Navarro, D. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144.
Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*(2), 639–650.
Shahbaba, B., & Neal, R. (2009). Nonlinear models using Dirichlet process mixtures. *The Journal of Machine Learning Research*, *10*, 1829–1850.
Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. (2009). Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1865.
Spearman, C. (1904). General intelligence, objectively determined and measured. *The American Journal of Psychology*, 201–292.
Sudderth, E., & Jordan, M. (2009). Shared segmentation of natural scenes using dependent Pitman–Yor processes. *Advances in Neural Information Processing Systems*, *21*, 585–1592.
Teh, Y., Görür, D., & Ghahramani, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Proceedings of the international conference on artificial intelligence and statistics. Vol. 11*.
Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.
Thibaux, R., & Jordan, M. (2007). Hierarchical beta processes and the indian buffet process. In *Proceedings of the international conference on artificial intelligence and statistics*.
Thurstone, L. (1931). Multiple factor analysis. *Psychological Review*, *38*(5), 406.
Thurstone, L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.
Vanpaemel, W. (2010). Prior sensitivity in theory testing: an apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.
Wagenmakers, E., van der Maas, H., Dolan, C., & Grasman, R. (2008). Ez does it! extensions of the ez-diffusion model. *Psychonomic Bulletin & Review*, *15*(6), 1229–1235.
Wagenmakers, E., & Waldorp, L. (2006). Editor's introduction. *Journal of Mathematical Psychology*, *50*(2), 99–100.
Williamson, S., Orbanz, P., & Ghahramani, Z. (2010). Dependent Indian buffet processes. In *International conference on artificial intelligence and statistics*.
Wood, F., & Griffiths, T. (2007). Particle filtering for nonparametric Bayesian matrix factorization. *Advances in Neural Information Processing Systems*, *19*, 1513.