A Brief Introduction to Bayesian Nonparametric Methods for Clustering and Time Series Analysis

Scott Niekum[1]

Abstract

Many descriptions of Bayesian nonparametric methods assume advanced mathematical and statistical proficiency. The goal of this tutorial is to provide a conceptual introduction to Bayesian nonparametrics that assumes only basic knowledge of standard Bayesian statistics, while also containing a few key derivations that provide mathematical insight into the presented methods. We begin by reviewing the motivation for Bayesian nonparametric methods, including DeFinetti's theorem. The Dirichlet process and the Chinese restaurant process (and their hierarchical counterparts) are then introduced in a clustering scenario that provides a mathematical and conceptual foundation for understanding more complex models. After reviewing the basics of Hidden Markov Models, these ideas are extended to time series analysis and augmented with priors that enable partial sharing of structure across multiple time series—the Beta process and the Indian buffet process. Finally, we close with a brief discussion of inference via Markov Chain Monte Carlo sampling methods.

1 Bayesian Nonparametrics

Graphical models and Bayesian inference have seen great success in artificial intelligence and machine learning applications spanning many fields including natural language processing [5], computer vision [17], social science [15], genetics [3], and medicine [13]. The Bayesian paradigm provides principled mechanisms to allow the specification of prior beliefs, model dependencies between variables, and perform efficient inference. However, one persistent difficulty in Bayesian inference is that of *model selection*. Often, it is not known *a priori* how complex a model must be to capture all the important structure of a dataset without overfitting, making it difficult to even provide a reasonable prior over such models. A classic example of this is choosing the polynomial degree in polynomial regression; too low of a degree will miss important characteristics of the data, while too high of a degree will begin to fit noise. This problem is also commonly found when trying to determine the appropriate number of components in a mixture model or the number of hidden states in a Hidden Markov Model.

Many techniques have been developed to select amongst models with fixed, finite parameterizations, including cross validation, Bayesian information criterion, and maximization of model evidence. However, many of these methods rely on heuristics, approximations, or the specification of a prior over model complexity, which may not be practical to specify for complex data. By contrast,

[1] Scott Niekum is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA sniekum@cmu.edu

Bayesian nonparametric methods sidestep having to explicitly perform model comparison by using an infinite parameterization that can determine an appropriate model complexity directly from data in a fully Bayesian manner. The next section will discuss how such parameterizations are possible. These infinite, nonparametric representations can be used to flexibly discover structure in data, such as appropriate numbers of clusters or dynamical systems that best describe observations.

1.1 De Finetti's Theorem

A common simplifying assumption in statistics is that of *independence*, in which the joint probability of data can be expressed as the product of the probabilities of each individual data point:

$$p(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} p(x_i). \tag{1}$$

A somewhat weaker assumption that nonparametric Bayesian methods leverage is that of *exchangeability.* A finite sequence of random variables is considered to be finitely exchangeable if every possible permutation of the random variables has an identical joint distribution, making the order in which data arrives irrelevant. An infinite sequence is considered infinitely exchangeable if every finite subset is finitely exchangeable. It can be seen from this that independence implies exchangeability, but that exchangeability does not necessarily imply independence.

De Finetti's theorem [6] states that a sequence $x_1, x_2, \dots$ of binary random variables is infinitely exchangeable if and only if there exists a random variable θ with cumulative distribution function Q on $[0, 1]$, such that for all n:

$$p(x_1, x_2, \dots, x_n) = \int_0^1 \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} dQ(\theta), \tag{2}$$

or equivalently:

$$p(x_1, x_2, \dots, x_n) = \int_0^1 p(x_1, x_2, \dots, x_n|\theta) dQ(\theta), \tag{3}$$

where $Q(\theta)$ is typically (but not required to be) well-behaved, such that $dQ(\theta) = Q'(\theta)d\theta = p(\theta)d\theta$. De Finetti only proved this for the case of infinite sequences of binary variables, but more general formulations exist, such as that of Hewitt and Savage [12], that extend this result to arbitrarily distributed real-valued random variables.

This formulation provides important insight into the nature of exchangeability. Specifically, it reveals the surprising fact that given the parameter θ, the data $x_1, x_2, \dots, x_n$ is conditionally independent and identically distributed (IID). In other words, the joint probability distribution $p(x_1, x_2, \dots, x_n)$ of an exchangeable sequence can always be represented as a mixture of IID sequences of random variables with mixing distribution $p(\theta)$. The key thing to notice here is that the parameterization of θ is not restricted in complexity; in fact, the parameterization may need to be infinite-dimensional, providing motivation for the search for nonparametric Bayesian methods.

Thus, we have seen that exchangeable data can be viewed as being conditionally IID from a mixture, which will make efficient Bayesian inference possible for this type of data. However, such mixtures may be arbitrarily complex, and may even require an infinite-dimensional parameterization to describe. Next, we will examine how to formalize and perform inference over such distributions.

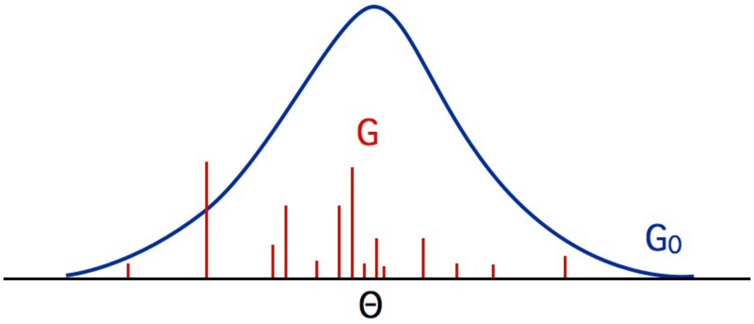

Figure 1: A discrete draw G from a Dirichlet process parameterized by G_0 over parameter space $\boldsymbol{\theta}$.

1.2 The Chinese Restaurant Process and the Dirichlet Process

The Chinese restaurant process (CRP) [1] is a discrete-time stochastic process that produces exchangeable data and is often used to illustrate a generative model for cluster data in Bayesian nonparametrics. Informally, it can be imagined as a Chinese restaurant that contains an infinite number of empty tables that each have infinite capacity; at each time step, a customer enters the restaurant and either joins a previously occupied table, or sits at a new empty table. The i^{th} customer will choose to sit at an empty table with probability $\frac{\alpha}{i-1+\alpha}$ (all empty tables are identical; this is a *total* probability for sitting at any of them), where α is a "concentration" parameter that determines the degree of dispersion of customers to different tables. The probability of the customer instead choosing a particular occupied table z is proportional to the number of people s_z already sitting at it, causing a clustering or "rich-get-richer" effect, and is defined as $\frac{s_z}{i-1+\alpha}$. Additionally, the first person to sit at each table chooses a unique dish for everyone at the table to share from an infinite menu, where the dish corresponds to a vector of parameters θ_i that parameterize some distribution $F(\cdot)$. In other words, in terms of a clustering problem, each customer is a data point, each table is a cluster or mixture component in a mixture model, and the table's dish represents the parameters of that mixture component (for example, the mean and variance of a Gaussian).

The sequence of customer assignments generated by the CRP is not IID, but it is exchangeable, so according to de Finetti's theorem, there exists a representation of the sequence that is conditionally IID with respect to some parameter, which in this case is θ. The mixing distribution G over the various possible settings of θ_i is simply defined by the number of customers sitting at the each of the corresponding tables. Thus, since each θ_i is a parameterization of a distribution, or a mixture component, G can be viewed as the mixing distribution of a mixture model.

However, there are an infinite number of tables, so G must be an infinite dimensional categorical distribution (a special case of a multinomial distribution where the number of trials is equal to 1). To take a Bayesian perspective on inference in such a model, we must specify a prior on G. If G were a fixed-dimensional, finite categorical distribution, then the appropriate conjugate prior would be a Dirichlet distribution. In this case, the appropriate prior is an infinite dimensional extension of the Dirichlet distribution, the Dirichlet Process.

Let us consider this model from a generative Bayesian perspective. The generative model for observations generated from a CRP partitioning is called a Dirichlet Process Mixture Model (DPMM) and can be specified as follows. A Dirichlet process (DP), parameterized by a base distribution

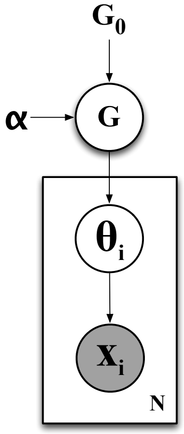

Figure 2: A Dirichlet Process Mixture Model (DPMM), where the rectangle is *plate notation*, denoting N copies of the outlined structure.

G_0 over a parameter space $\boldsymbol{\theta}$, and a concentration parameter α, is used as a prior over the distribution G of mixture components. For data points X, mixture component parameters θ, and a parameterized distribution $F(\cdot)$, the DPMM can be written as [16]:

$$\begin{aligned}
&G|\alpha, G_0 \sim DP(\alpha, G_0)\\
&\theta_i|G \sim G\\
&x_i|\theta_i \sim F(\theta_i),
\end{aligned}$$

where "$\sim$" can be read as "drawn from" or "distributed as".

A draw G from a Dirichlet process is discrete with probability 1, as shown in the example in Figure 1. This draw provides the set of *atoms*, or mixture components, to which data points are assigned according to the distribution G, corresponding to the assignment of customers to tables in the CRP. Finally, each data point x_i can be generated by drawing from the distribution parameterized by the assigned mixture component θ_i. In the CRP analogy, this means that a customer x_i walks in and sits at the table with dish θ_i (in the generative view, the tables/dishes can be seen as generating the customers). Note that G must integrate to 1. From this, it can be seen that the Dirichlet process can be interpreted as a distribution over probability distributions. Figure 2 shows the corresponding graphical model for the DPMM.

To show the relationship between the Dirichlet process and the CRP, let us first examine the simpler case of a finite mixture, defining it in such a way that the assignment of data to components is more explicit. Define $c_i \in \{1 \ldots k\}$ as the index of the mixture component assigned to observation x_i , where k is the number of mixture components, and π_j as the mixing weight on the j^{th} component. Thus, the generative model for this finite mixture is:

$$\begin{aligned}
&\theta_i|G_0 \sim G_0\\
&\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha/k, \ldots, \alpha/k)\\
&c_i|\boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi})\\
&x_i|\boldsymbol{\theta}, c_i \sim F(\theta_{c_i}).
\end{aligned}$$

Thus, the conditional probability of a mixture assignment[1] c_i, given all the other assignments $\mathbf{c}_{-i}$, can be found by integrating out the weights $\boldsymbol{\pi}$:

$$p(c_i = z \mid \mathbf{c}_{-i}, \alpha) = \int p(c_i = z \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \mathbf{c}_{-i}, \alpha) d\boldsymbol{\pi}. \tag{4}$$

The first term in the integral, is simply equal to the z^{th} component of the weight vector:

$$p(c_i = z \mid \boldsymbol{\pi}) = \pi_z. \tag{5}$$

The second term is the posterior probability of the weight vector, given all but the i^{th} mixture assignment, and can be written as:

$$p(\boldsymbol{\pi} \mid \mathbf{c}_{-i}, \alpha) \propto p(\mathbf{c}_{-i} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \alpha). \tag{6}$$

Since the first term has a categorical distribution and the second term is Dirichlet distributed, by conjugacy, the posterior $p(\boldsymbol{\pi} \mid \mathbf{c}_{-i}, \alpha)$ is also Dirichlet distributed. Define the normalization function for the Dirichlet distribution as:

$$\begin{aligned} \mathcal{Z}(\boldsymbol{\beta}) = & \int \prod_{j=1}^{k} \pi_j^{\beta_j - 1} d\pi & (7) \\ = & \frac{\prod_{j=1}^{k} \Gamma(\beta_j)}{\Gamma(\sum_{j=1}^{k} \beta_j)}, & (8) \end{aligned}$$

where $\Gamma(\cdot)$ is the standard Gamma function and β_j is a concentration parameters that can be conceptualized as a pseudocount of the number of times that the j^{th} event has previously been observed. Define the vector $\mathbf{s} = (s_1, \ldots, s_k)$, where s_j is the total number of assignment variables in $\mathbf{c}_{-i}$ that indicate mixture component j. Using this, the posterior probability of the weight vector can be written in terms of the previous number of counts, $\mathbf{s}$:

$$p(\boldsymbol{\pi} \mid \mathbf{c}_{-i}, \alpha) = \frac{1}{\mathcal{Z}(\mathbf{s} + \frac{\alpha}{k})} \prod_{j=1}^{k} \pi_j^{s_j + \frac{\alpha}{k} - 1}. \tag{9}$$

Finally, combining equations 5 and 9, and using the fact that $\Gamma(x+1) = x\Gamma(x)$, we can rewrite the

[1] Note that due to the exchangeable nature of the data, we can always assume that the observation c_i in question is observed last given all the others, such that i=N, the total number of observations.

posterior for the assignment variables (equation 4) as:

$$\begin{aligned}
& p(c_i = z \mid \mathbf{c}_{-i}, \alpha) && (10) \\
& \propto \frac{1}{\mathcal{Z}(\mathbf{s} + \frac{\alpha}{k})} \int \pi_z \prod_{j=1}^{k} \pi_j^{s_j + \frac{\alpha}{k} - 1} d\boldsymbol{\pi} && (11) \\
& = \frac{\mathcal{Z}(\mathbf{s} + \frac{\alpha}{k} + \mathbf{1}^{(z)})}{\mathcal{Z}(\mathbf{s} + \frac{\alpha}{k})} && (12) \\
& = \frac{\prod_{j=1}^{k} \Gamma\left(s_j + \frac{\alpha}{k} + \mathbf{1}_j^{(z)}\right)}{\prod_{j=1}^{k} \Gamma\left(s_j + \frac{\alpha}{k}\right)} \frac{\Gamma\left(\sum_{j=1}^{k} s_j + \frac{\alpha}{k}\right)}{\Gamma\left(\sum_{j=1}^{k} s_j + \frac{\alpha}{k} + \mathbf{1}_j^{(z)}\right)} && (13) \\
& = \frac{\left[\prod_{j \neq z} \Gamma\left(s_j + \frac{\alpha}{k}\right)\right] \Gamma\left(s_z + \frac{\alpha}{k} + 1\right)}{\prod_{j=1}^{k} \Gamma\left(s_j + \frac{\alpha}{k}\right)} \frac{\Gamma\left(\sum_{j=1}^{k} s_j + \frac{\alpha}{k}\right)}{\left(\sum_{j=1}^{k} s_j + \frac{\alpha}{k}\right) \Gamma\left(\sum_{j=1}^{k} s_j + \frac{\alpha}{k}\right)} && (14) \\
& = \frac{\left[\prod_{j=1}^{k} \Gamma\left(s_j + \frac{\alpha}{k}\right)\right] \left(s_z + \frac{\alpha}{k}\right)}{\prod_{j=1}^{k} \Gamma\left(s_j + \frac{\alpha}{k}\right)} \frac{1}{\sum_{j=1}^{k} s_i + \frac{\alpha}{k}} && (15) \\
& = \frac{s_z + \frac{\alpha}{k}}{N - 1 + \alpha}, && (16)
\end{aligned}$$

where N is the total number of observations (including the current one whose mixture assignment is in question), and $\mathbf{1}^{(z)}$ is a vector of length k with a 1 in the z^{th} position, and zeros elsewhere.

Now, the behavior of this posterior probability can be examined as the number of mixture components k goes to infinity. Since only a finite number of samples, $N - 1$, have been observed so far, then only a finite number of the counts $s_1, \ldots, s_\infty$ are non-zero. This divides the mixture components into two sets: a set $\mathcal{Q}$ that contains components with non-zero counts and a set $\hat{\mathcal{Q}}$ that contains components with zero counts. First, consider the probability that a new observation gets assigned to one particular mixture component z that already has a non-zero count s_z:

$$\begin{aligned}
p(c_i = z \in \mathcal{Q} \mid \mathbf{c}_{-i}, \alpha) &= \lim_{k \to \infty} \left(\frac{s_z + \frac{\alpha}{k}}{N - 1 + \alpha} \right) && (17) \\
&= \frac{s_z}{N - 1 + \alpha}. && (18)
\end{aligned}$$

Next, consider the total probability that a new observation gets assigned to any component that does not already have an associated observation (i.e. all components z such that the corresponding count s_z is equal to zero):

$$
\begin{aligned}
p(c_i \in \hat{\mathcal{Q}} \mid \mathbf{c}_{-i}, \alpha) = & \quad \lim_{k\to\infty} \left(\sum_{z \in \hat{\mathcal{Q}}} \frac{s_z + \frac{\alpha}{k}}{N-1+\alpha} \right) && (19) \\
= & \quad \lim_{k\to\infty} \left(\sum_{z \in \hat{\mathcal{Q}}} \frac{\frac{\alpha}{k}}{N-1+\alpha} \right) && (20) \\
= & \quad \frac{\alpha}{N-1+\alpha} \lim_{k\to\infty} \left(\sum_{z \in \hat{\mathcal{Q}}} \frac{\frac{1}{k}}{N-1+\alpha} \right) && (21) \\
= & \quad \frac{\alpha}{N-1+\alpha} \lim_{k\to\infty} \left(\frac{|\hat{\mathcal{Q}}|}{k(N-1+\alpha)} \right) && (22) \\
= & \quad \frac{\alpha}{N-1+\alpha}, && (23)
\end{aligned}
$$

since $|\hat{\mathcal{Q}}| \to \infty$ as $k \to \infty$.

It can be seen that these probabilities are identical to those in the CRP for a customer joining an occupied table and an unoccupied table, respectively, showing the correspondence between the CRP and a Dirichlet process prior. Thus, the CRP is an illustrative description of a Bayesian prior over an infinite dimensional categorical distribution—in other words, the Dirichlet process. Equations 18 and 23 also illustrate that most of the data will have a tendency to cluster into a small number of components, since the likelihood of a new component forming becomes very small as N gets large. In fact, it can be shown that the number of non-zero components increases roughly logarithmically with respect to N [15].

Given a data set, inference can be performed on a DPMM to find an appropriate number of mixture components and their associated parameters that best explain the data without overfitting (of course, subject to the clustering strength assumptions made by setting the concentration parameter or its hyperparameters), in a fully Bayesian manner. This sidesteps the difficult problem of model selection and provides a principled framework for representing distributions of arbitrary complexity with mixtures of simpler distributions, such as Gaussians. This technique can be used for tasks such as unsupervised clustering of data and density estimation of complex distributions from samples. For a more complete discussion of how to perform inference in such a model, or how the full DPMM can be derived as the limit of finite mixture models, see [18].

1.3 The Chinese Restaurant Franchise and the Hierarchical Dirichlet Process

Now that the connection between the CRP and a Dirichlet process prior has been established, other similar metaphors can be explored that describe new classes of Bayesian nonparametric priors and their properties. One possible shortcoming of the standard Dirichlet process mixture model is that all component parameters θ_i are drawn from the same distribution G—in other words, that all data is drawn from the same underlying mixture distribution. However, in many data sets, data may be come from several distinct, but related, groups or distributions. By explicitly modeling these groups, the underlying distribution of the data can often be estimated more accurately and efficiently.

For example, imagine the distribution over words from a physics textbook compared to that of a chemistry textbook. Each component in a mixture model describing either book could represent a topic that generates words; since both books are science books, many such topics would appear in both books. Thus, to model the joint distribution of words across both books, it would be desirable to have a model that allowed some parameters, or *atoms*, to be shared across books, while still being able to model each book as having unique topics and distributions of topics. The Hierarchical Dirichlet Process, and a related metaphor, the Chinese restaurant franchise (CRF) [19], allow such sharing.

The CRF can be described similarly to the CRP, but can share mixture components amongst groups (of data) that may have different, but related, characteristics. The primary difference is that the CRF assigns each group (or set of customers) to a separate restaurant. Each restaurant still has an infinite number of tables, but instead of each table's dish being globally unique, a dish can be chosen at additional tables at other restaurants in the franchise. Again, the first person to sit at each table chooses that table's dish from the menu, but in the CRF, the probability of choosing a particular dish is proportional to the number of tables that have already chosen that dish franchise-wide. Thus, dishes (i.e. mixture components such as topics) can be shared across restaurants, but each restaurant can have a unique distribution of dishes.

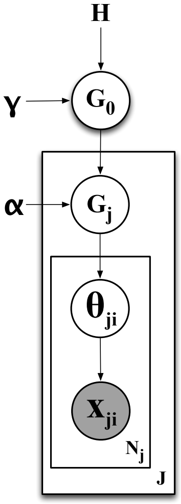

Figure 3: A Hierarchical Dirichlet Process Mixture Model (HDPMM).

Just as the CRP describes a Dirichlet process prior, the CRF corresponds to a hierarchical Dirichlet process prior (HDP). The generative model for data sets produced by an HDP mixture model is shown in Figure 3 and can be written as [19]:

$$\begin{aligned} &G_0|\gamma, H \sim DP(\gamma, H) \\ &G_j|\alpha, G_0 \sim DP(\alpha, G_0) \\ &\theta_{ji}|G_j \sim G_j \\ &x_{ji}|\theta_{ji} \sim F(\theta_{ji}), \end{aligned}$$

where H is a base distribution, γ is a concentration parameter, j is the index of a data group (i.e. an index that corresponds to all data generated from the atoms of a particular G_j), and both

G_0 and G_j are discrete distributions. Here, the Dirichlet process that serves as a prior over G_0 defines a prior over the entire parameter space, whereas G_0 and G_j represent the franchise-wide and restaurant-specific menus, respectively. This model is said to be hierarchical since there are two levels (i.e. draws that take place in the generative model) at which atoms are selected that are then later used at lower levels; the atoms from G_0 can be shared amongst the various G_j, since G_0 is used as a base distribution to parameterize a second Dirichlet process that acts as a prior over G_j, encouraging sharing amongst the groups. Returning to our earlier example, it can be seen that an HDP mixture can be used to jointly model the word distribution from several documents that may share some topics, while also retaining unique topics and distributions over topics as well.

2 Time Series Analysis

Time-series data present unique challenges for analysis, since observations are temporally correlated. Clearly, time-series data are not independent and identically distributed (nor are they exchangeable), but weaker assumptions can often be made about the data to make inference tractable.

Define a sequence to be a *first-order Markov chain* if:

$$p(x_n|x_1, x_2, \ldots, x_{n-1}) = p(x_n|x_{n-1}). \tag{24}$$

In other words, given the previous observation x_{n-1}, an observation x_n is conditionally independent of all other previous observations. To capture longer-range interactions, this concept can be extended to higher orders, such that an observation is dependent on the previous M observations:

$$p(x_n|x_1, x_2, \ldots, x_{n-1}) = p(x_n|x_{n-1}, \ldots, x_{n-M}). \tag{25}$$

One way to tractably model time-series data is through the use of a *state space model*, in which each observation x_i has a corresponding latent variable or *hidden state* z_i associated with it. The latent variables $z_1, \ldots, z_n$ form a Markov chain and *emit* the observations $x_1, \ldots, x_n$ based on conditional distributions of the form $p(x|z)$. Figure 4 shows the graphical representation of a state space model.

When the latent variables $\mathbf{z}$ in a state space model are discrete[2], we obtain the standard Hidden Markov Model (HMM). The standard HMM is defined by the number of states K that the latent variables can take on, a $K \times K$ transition probability matrix $\boldsymbol{\pi}$ (with rows $\boldsymbol{\pi}_k$) that describes the probabilities $p(z_i|z_{i-1})$, and a parameterized distribution[3] $F(\cdot)$ that describes the conditional probabilities $p(x_i|z_i)$. The generative model for an HMM can be written as:

$$\begin{aligned} z_i &\sim \boldsymbol{\pi}_{z_{i-1}} \\ x_i &\sim F(\theta_{z_i}), \end{aligned}$$

where θ_{z_i} is a parameter vector associated with state z_i. In other words, the HMM can describe time series data with a mixture model in which the latent mixture component indices have a temporal relationship as a first-order Markov chain.

[2] The term "Hidden Markov Model" is typically used to refer to a model with discrete latent variables. When latent variables are continuous, a discrete transition matrix can no longer be used to describe transitions, instead requiring formulations like linear-Gaussian systems. These models are sometimes referred to by different names in the literature, rather than as an HMM.

[3] Technically, each hidden state z can have its own unique parameterized distribution $F_z(\cdot)$, but in practice all hidden state emission distributions usually share a common form, such as a Gaussian distribution. However, each hidden state still has a unique set of parameters θ_z for this distribution.

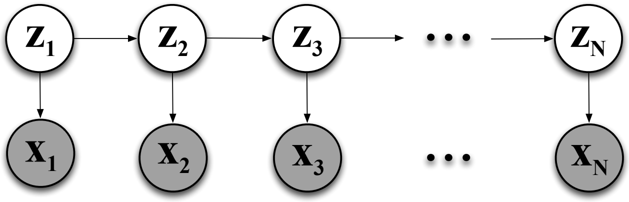

Figure 4: A state space model, or standard HMM when the latent variables $\mathbf{z}$ are discrete.

One drawback of the standard HMM is that the observation x_i is conditionally independent of any other observation x_j, given the generating hidden state z_i. This independence assumption is clearly not well founded for much time-series data, such as robot demonstrations, in which the observations, as well as the hidden states, have temporal dependencies. The autoregressive HMM (AR-HMM) addresses this by adding links between successive observations, forming a Markov chain, as shown in Figure 5. This can be extended to a M^{th} order AR-HMM, as shown in Figure 6.

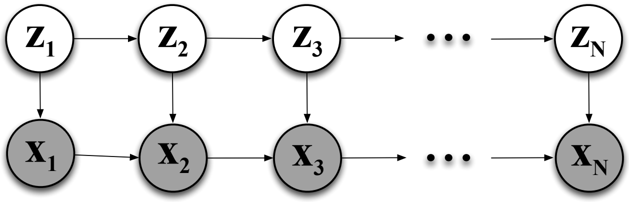

Figure 5: A first-order autoregressive HMM.

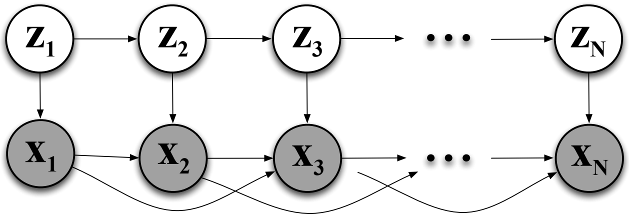

Figure 6: Additional links can be added to make an M^{th}-order autoregressive HMM. This example shows a second-order AR-HMM.

AR-HMMs face a problem in cases where the observations are not discrete—a simple transition matrix cannot be used to describe the probabilities $p(x_i|z_i, x_{i-1}, \ldots, x_{i-M})$. For example, demonstration data is often comprised of continuously valued state-action pairs representing robot joint poses and actuations. In this case the conditional probability density function over x_i must be able to be written in terms of a continuous function of its predecessors. For example, a linear dynamical system governing the conditional distribution can be used, such that:

$$p(x_i|z_i, x_{i-1}, \ldots, x_{i-M}) = \sum_{j=1}^{M} A_{j,z_i} x_{i-j} + \mathbf{e}_i(z_i),$$

for transition matrices A_{j,z_i} and covariance terms $\mathbf{e}_i(z_i)$. In this way, time-series data can be flexibly modeled as a mixture of linear dynamical systems.

However, an issue facing all of these methods is that of model selection—guessing or inferring the optimal model size (the number of hidden states). In the next section, we will discuss how Bayesian nonparametric methods can deal with this issue naturally, similar to how the Dirichlet process mixture model is able to virtually use an infinite number of clusters.

2.1 The Indian Buffet Process and the Beta Process

HDP priors can be useful for modeling time-series data by acting as a prior over the transition matrix in a Hidden Markov model, forming an HDP-HMM [2]. In this case, the groups being modeled are the rows of the transition matrix; in other words, state-specific transition distributions. The HDP prior allows each transition distribution to be unique, while sharing global characteristics such as the overall popularity of particular states. Since the nonparametric prior allows this transition matrix to be infinite, an appropriate number of states (and their corresponding emission distributions) can be found that best explain the data without overfitting.

The HDP-HMM is an effective model for analyzing a single time series. However, it is often desirable to be able to jointly analyze multiple time series sequences, each of which may have unique transition dynamics—for instance, demonstrations of several different robotic tasks. Furthermore, each sequence may only exhibit some subset of the larger set of states that are observed across all the sequences. It is relatively straightforward to extend the HDP-HMM so that it jointly models the transition and emission parameters of all the sequences. However, such a model assumes that all the sequences exhibit the same set of states and transition between them in an identical manner, precluding the desired flexibility. Instead, a more flexible *featural* model is required, in which each sequence can exhibit some subset of a larger library of states and transition between them in a unique manner.

The culinary metaphor that describes a statistical model with the properties that we desire is the Indian Buffet Process (IBP) [10]. In the IBP, the first customer enters the restaurant and select Poisson(α) dishes from an infinite buffet. After that, the i^{th} customer can select multiple dishes from the buffet; each existing dish z is selected with probability $\frac{s_z}{i}$, where s_z is the number of times dish z has previously been selected. The customer then also selects Poisson(α/i) new dishes. The IBP describes a nonparametric Bayesian prior known as the Beta process.

Recall that a draw from a Dirichlet process is a probability distribution—a draw from a Dirichlet process results in an infinite number of weights on point masses that sum to 1. A draw from a Beta process is also an infinite collection of point masses, but the weight of each point is drawn from a Beta distribution parameterized by the value of the base distribution at that point. Thus, the weights from a Beta process draw can each be interpreted as a $[0, 1]$ probability and do not sum to one. This draw can be seen as an infinite vector of probabilities corresponding to the chance of "heads" on an infinite set of unfair coins. The featural property of Beta processes stems from this view—in the generative view, when generating any given piece of data, a particular feature manifests with probability proportional to its weight in the draw from the Beta process.

A Beta process prior can be used over the transition matrices of hidden Markov models for multiple time series sequences, much like the HDP, but elicits a featural representation, in which hidden states can be shared across sequences. One such model is the Beta Process Autoregressive Hidden Markov Model (BP-AR-HMM) [8], shown in Figure 7, in which hidden states correspond to dynamic modes defined by linear dynamical systems. The BP-AR-HMM is able to jointly model a library of dynamical modes over many time series sequences, while allowing each time series to exhibit

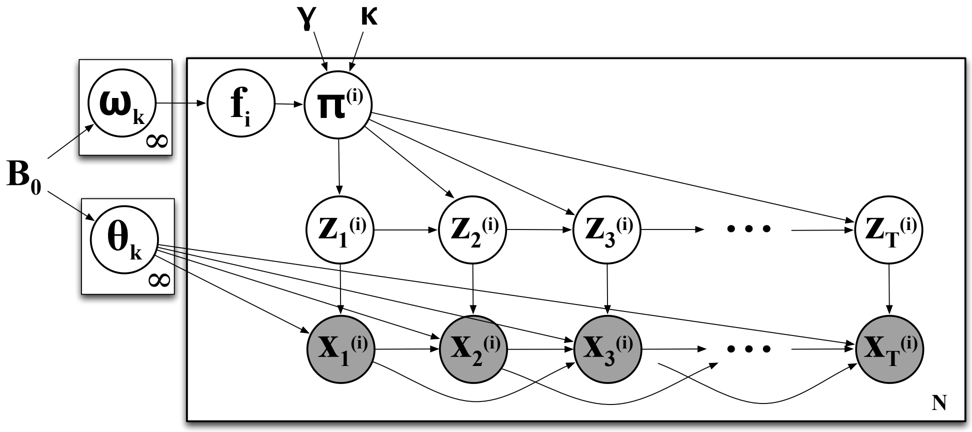

Figure 7: The Beta Process Autoregressive Hidden Markov Model (BP-AR-HMM). Here, the Beta Process draw B has been separated into masses $\boldsymbol{\omega}$ and parameters $\boldsymbol{\theta}$. Each parameter θ_k consists of r transition matrices $\mathbf{A}$ and a covariance term $\mathbf{e}$.

some subset of those modes and switch between them in a unique manner. Thus, a potentially infinite library of modes can be constructed in a fully Bayesian way, in which modes are flexibly shared between time series, and an appropriate number of modes is inferred directly from the data, without the need for model selection. In other words, each time series corresponds to a customer in the IBP and each dynamical mode corresponds to a dish.

The generative model for the BP-AR-HMM can be summarized as follows [8]:

$$
\begin{aligned}
&B|B_0 \sim \mathrm{BP}(1, B_0)\\
&X_i|B \sim \mathrm{BeP}(B)\\
&\pi_j^{(i)}|\boldsymbol{f_i}, \gamma, \kappa \sim \mathrm{Dir}([\gamma, ..., \gamma+\kappa, \gamma, ...] \otimes \boldsymbol{f_i})\\
&z_t^{(i)} \sim \pi^{(i)}_{z_{t-1}^{(i)}}\\
&\mathbf{y}_t^{(i)} = \sum_{j=1}^{r} A_{j, z_t^{(i)}} \mathbf{y}_{t-j}^{(i)} + \mathbf{e}_t^{(i)}(z_t^{(i)})
\end{aligned}
$$

First, a draw B from a Beta Process (BP) provides a set of global weights for the potentially infinite number of states. Then, for each time series, an X_i is drawn from a Bernoulli Process (BeP) parameterized by B. Each X_i can be used to construct a binary vector $\boldsymbol{f_i}$ indicating which of the global features, or states, are present in the i^{th} time series. Thus, B encourages sharing of features amongst multiple time series, while the X_i leave room for variability. Next, given the features that are present in each time series, for all states j, the transition probability vector $\pi_j^{(i)}$ is drawn from a Dirichlet distribution with self transition bias κ. A state $z_t^{(i)}$ is then drawn for each time step t from the transition distribution of the state at the previous time step. Finally, given the state at each time step and the *order* of the model, r, the observation is computed as a sum of state-dependent linear transformations of the previous r observations, plus mode-dependent noise.

3 Inference and Markov Chain Monte Carlo Methods

Several nonparametric Bayesian methods have been discussed that can model various types of data of unknown complexity and sidestep the problem of model selection in a principled Bayesian manner. Thus far, these models have been explored from a generative point of view, with little discussion of how inference can be performed. In general, efficient Bayesian inference is model-specific and tailored to the unique statistical properties of the model. However, there are several core techniques and principles that are generally useful in performing interference in nonparametric Bayesian models, which we will now introduce.

First, how is it possible to perform inference on a model that has an infinite number of parameters? While it is not possible to optimize an infinite number of parameters, for problems like clustering with the DPMM, where only the assignments of data points to mixture components matter, it is possible to integrate the parameters out and look for high-likelihood configurations of the auxiliary variables that assign each observed data point to a mixture component. Other models like the BP-AR-HMM use incremental techniques like birth and death proposals [8] so that there are only a finite number of parameters to work with at any given time; the number of parameters can incrementally grow and shrink unboundedly, but is always practically limited by the size of the data.

Algorithm 1 Metropolis-Hastings Sampling

```
Given: Distributions $\widetilde{p}(\mathbf{z}), q(\mathbf{z}|\mathbf{z}')$
Input: Starting configuration $\mathbf{z}^{(0)}$

  1. Initialize $\mathbf{z}^{(0)}$
  2. For $\tau = 0, \ldots, T$:
      (a) Sample $\mathbf{z}^* \sim q(\mathbf{z}|\mathbf{z}^{(\tau)})$
      (b) Calculate acceptance probability:
          $$A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(\mathbf{z}^*)q(\mathbf{z}^{(\tau)}|\mathbf{z}^*)}{\widetilde{p}(\mathbf{z}^{(\tau)})q(\mathbf{z}|\mathbf{z}^{(\tau)})}\right)$$
      (c) Sample $u \sim \text{uniform}(0, 1)$
      (d) Assign $\mathbf{z}^{(\tau+1)}$:
            if $A(\mathbf{z}^*, \mathbf{z}^{(\tau)}) > u$ then
              $\mathbf{z}^{(\tau+1)} = \mathbf{z}^*$
            else
              $\mathbf{z}^{(\tau+1)} = \mathbf{z}^{(\tau)}$
            end if
  3. Return $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T+1)}$
```

Inferring the MAP distribution of the hidden variables in nonparametric models is often difficult, especially in the non-conjugate case. However, in applications like clustering, only unbiased samples from the posterior are required, rather than the full distribution. Leveraging this, Markov chain Monte Carlo (MCMC) sampling methods can be used to draw samples from the posterior when

Algorithm 2 Gibbs Sampling

```
Given: Conditionals $p(z_1|\mathbf{z}_{-1}), \ldots, p(z_M|\mathbf{z}_{-M})$
Input: Starting configuration $\mathbf{z}^{(0)}$

  1. Initialize $\mathbf{z}^{(0)}$
  2. For $\tau = 0, \ldots, T$:
       Sample $z_1^{(\tau+1)} \sim p(z_1|z_{-1}^{(\tau)})$
       $\vdots$
       Sample $z_M^{(\tau+1)} \sim p(z_M|z_{-M}^{(\tau)})$
  3. Return $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T+1)}$
```

inferring the full posterior is impossible or intractable due to problems like high-dimensionality, difficulty of integration, and non-conjugacy.

Define z_i as the i^{th} variable in a model and $\mathbf{z}_{-i}$ as the set of all other variables in $\mathbf{z}$ except z_i. In cases where the conditionals $p(z_i|\mathbf{z}_{-i})$ cannot be written as a known distribution that can easily be sampled from, Metropolis-Hastings sampling [14, 11] can be used to obtain samples of the joint distribution $p(\mathbf{z})$. Algorithm 1 describes Metropolis-Hastings sampling, given a starting configuration of the variables $\mathbf{z}^{(0)}$, a *proposal distribution* $q(\mathbf{z}|\mathbf{z}')$ that is easy to sample from, and an unnormalized distribution $\widetilde{p}(\mathbf{z}) = \frac{1}{\mathcal{Z}}p(\mathbf{z})$, where $\mathcal{Z}$ may be unknown. It is assumed that while $p(\mathbf{z})$ cannot easily be sampled from, $\widetilde{p}(\mathbf{z})$ can be easily evaluated at a single point.

Under mild conditions, the Metropolis-Hastings algorithm creates a Markov chain whose stationary distribution approximates $p(\mathbf{z})$. It does this through a careful choice of the acceptance probability function that determines whether a step in the Markov chain is accepted or rejected. This function is designed so that the distribution being sampled at each time step is invariant and equal to the correct distribution $p(\mathbf{z})$. For more details on the derivation, see [4].

The Metropolis-Hastings algorithm has several drawbacks, most notably that it requires a “burn-in period”—a number of early samples that are thrown out, because of the bias introduced by the starting configuration. Additionally, successive samples are correlated, so if independent samples are desired, some number of samples must be ignored between each independent sample. An appropriate number can often be found by looking at the autocorrelation of samples.

When the conditional distributions $p(z_i|\mathbf{z}_{-i})$ can be written in a standard form for which the CDF can easily be calculated, a special case of Metropolis-Hastings sampling called Gibbs sampling [9] can be used. Algorithm 2 outlines the Gibbs sampling procedure. Gibbs sampling uses the conditionals $p(z_i|\mathbf{z}_{-i})$ as proposal distributions, cycling through and sampling from them one at a time. Instead of calculating an acceptance probability, it can be shown that under this choice of proposal distribution, the new draw should always be accepted. Thus, when the conditionals are available in a standard form, Gibbs sampling can show substantial efficiency gains over more general Metropolis-Hastings sampling.

The full details of inference for the models described earlier are outside the scope of this document, but there exist many good references on this topic [7, 15, 16, 18].

References

[1] D. Aldous, I. Ibragimov, and J. Jacob. *Exchangability and Related Topics*. Lecture notes in mathematics. Springer-Verlag, 1983.

[2] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden markov model. In *Machine Learning*, pages 29–245. MIT Press, 2002.

[3] M. Beaumont and B. Rannala. The bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–61, 2004.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] B. de Finetti. *Funzione Caratteristica Di un Fenomeno Aleatorio*, pages 251–299. 6. Memorie. Academia Nazionale del Linceo, 1931.

[7] E. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Ph.D. thesis, MIT, Cambridge, MA, 2009.

[8] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. Joint modeling of multiple related time series via the beta process. *arXiv:1111.4226*, 2011.

[9] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[10] T. L. Griffiths and Z. Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

[11] W. Hastings. Monte carlo samping methods using markov chains and their applications. *Biometrika*, pages 97–109, 1970.

[12] E. Hewitt and L. J. Savage. Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955.

[13] P. J. F. Lucas, L. C. van der Gaag, and A. Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial intelligence in medicine*, 30(3):201–14, 2004.

[14] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[15] D. J. Navarro, T. L. Griffiths, M. Steyvers, and M. D. Lee. Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, 50(2):101–122, 2006.

[16] R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

[17] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *In NIPS*, pages 1097–1104. MIT Press, 2004.

[18] C. E. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000.

[19] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.